Chapter 8

THEORIES OF PERSISTENT INEQUALITY AND INTERGENERATIONAL MOBILITY

THOMAS PIKETTY

CNRS-CEPREMAP, Paris, E-mail: thomas.piketty@cepremap.cnrs.fr

Contents

1. Introduction 430
   1.1. The dimensions of conflict about intergenerational mobility 430
   1.2. Organization of this chapter 435
2. Persistent inequality and the family transmission of wealth 436
   2.1. The contribution of inheritance to the persistence of inequality 436
   2.2. The long-run dynamics of wealth inequalities with exogenous savings 438
   2.3. The long-run dynamics of wealth inequalities with dynastic utility functions 440
   2.4. "Active" vs. "passive" inheritance: the costs of redistribution 442
3. Persistent inequality and the family transmission of ability 446
   3.1. The transmission of productive abilities 446
   3.2. Efficient inequality and the costs of redistribution 448
   3.3. Taste-based persistent inequality 451
4. Persistent inequality and the imperfect capital market 453
   4.1. Credit constraints vs. first-best credit 453
   4.2. What do we know about the importance of credit constraints for mobility? 456
   4.3. Poverty traps vs. low-mobility traps 459
5. Persistent inequality and local segregation 462
   5.1. Models of inefficient segregation into unequal neighborhoods 462
   5.2. The policy implications of local segregation 464
   5.3. Other levels of segregation 467
6. Persistent inequality and self-fulfilling beliefs 468
   6.1. The theory of discrimination 468
   6.2. Sociologists' theories of self-fulfilling inequality 471
References 473

Handbook of Income Distribution, Volume 1. Edited by A. B. Atkinson and F. Bourguignon

Keywords: Income distribution, intergenerational mobility

JEL codes. D30, D31, D63

1. Introduction

This chapter aims to survey existing theories of persistent inequality across generations. That is, unlike other theory- oriented chapters in this Handbook, we are concerned with total economic inequality, both in wealth and in earnings, and we concentrate upon the intergenerational mobility dimension of total inequality. The questions we ask in this chapter are the following: what determines the degree of transmission and persistence of inequality across generations? What are the policy implications of the various existing theories?

Although the scope of this chapter is primarily theoretical, we will also offer a nonexhaustive, non-technical survey of existing empirical work about intergenerational mobility and persistent inequality between dynasties. Instead of presenting this body of empirical evidence in a separate section, we will refer to empirical studies when needed in order to confirm, contradict or illustrate the different theoretical models. Although existing evidence is scarce, we believe that such a straightforward confrontation between theories and empirical evidence is particularly needed in this field. The question of intergenerational mobility has always been one of the most controversial issues indeed, both in actual political conflicts and in academic writings by social scientists, and conflicting theories in this area have very often been motivated by conflicting qualitative perceptions of the extent of mobility (and conversely....). Before we describe the organization of the chapter and the main theoretical models of intergenerational mobility, it is useful to briefly recall some basic background about the controversies which characterize the history of this field.

1.1. The dimensions of conflict about intergenerational mobility

As a first approximation, one can say that controversies about intergenerational mobility have been dominated during most of the nineteenth and twentieth centuries by a violent conflict between what Erikson and Goldthorpe (1992) call the "liberal theory" of industrialization on the one hand, and the Marxist theory (and various socialist theories) on the other hand.[1] According to the "liberal theory", the industrial society is characterized by an irreversible commitment to technical and economic rationality, and therefore by high and rising rates of social mobility and equality of opportunity, as procedures of social selection become more and more rational. The Marxist theory basically says the

[1] See Erikson and Goldthorpe (1992: Chap. 1) and subsequent references. Erikson and Goldthorpe concentrate on the post-World War II, academic section of this intellectual and political conflict, but similar controversies did already exist long before (at least since the industrial revolution).

opposite: capitalist societies are characterized by class reproduction, whereby a small number of capitalist dynasties reproduce themselves from generation to generation and a large and growing number of working-class dynasties is being exploited by capitalist dynasties from generation to generation.[2]

What is striking about these two conflicting viewpoints is that they combine conflicting empirical claims about mobility (is actual mobility low or high in industrial societies?) with conflicting theoretical claims about the working of the market system: are market economies characterized by rationality, efficiency and openness, or do they just perpetuate initial inequalities? Note also that the basic premise of both theories is that mobility should be high. In particular, the liberal theory implicitly assumes that allocative efficiency requires a high level of social mobility, presumably because the intergenerational correlation of ability and other efficiency-relevant individual characteristics is assumed to be low. Marxist and socialist theories obviously make this assumption as well, but they claim that the market system is unable to allocate individual talents as they should be and to achieve this high and efficient mobility level.

In its most extreme form, this conflict between liberal and Marxist theories of intergenerational mobility is by now well behind us. On the one hand, following the spectacular improvement of living standards in capitalist countries and the tragic failure of communist systems, nobody seems to support any longer the Marxist theory of mass proletarianization and class reproduction under capitalism. On the other hand, the optimist view of high and perpetually increasing mobility rates in market societies has proven to be excessively naive. During the past decades, sociologists in many countries have collected a large body of survey evidence about occupations and social status of parents and children, allowing them to compute mobility matrices and various other mobility measures. This type of data does not generally allow for easy and reliable comparisons of mobility measures over time and across countries, given the substantial variability of occupational categories and social status scales. It is remarkable however that all comparative empirical studies of social mobility rates, based upon different data sets collected at different points in time, have found very similar mobility matrices across industrial nations, and in particular no significant difference between Europe and the US.[3] This comparison between the US and various European countries has always played a central role in controversies about social mobility. At least since the time of Tocqueville, the "liberal theory" would seem to predict that a more open and market-oriented society such as the US should lead to significantly higher mobility rates. These empirical studies by sociologists also seem to show that there has been no significant

[2] In its most extreme dogmatic form, another version of the Marxist response is to dismiss the "bourgeois" question of mobility altogether (see Erickson and Goldthorpe, 1992: pp. 9–10).

[3] See, e.g., Lipset and Bendix (1959, 1966), Erikson and Goldthorpe (1985, 1992) and the references therein. See also the historical study by Kaelble (1985), who compares social mobility rates in various western cities over the 1840–1920 period, and finds no significant difference across western countries.

change in mobility rates over time, at least since World War II.[4] Comparative studies of educational mobility also suggest a high level of commonality and inertia of mobility rates, both over time and across countries.[5]

More recently, following the development of large panel data sets with economic variables spanning across several generations, economists have started to measure intergenerational mobility. These economic measures of intergenerational mobility should in principle offer more reliable cross-country and time-series comparisons. Preliminary results seem to confirm the sociologists finding about the absence of any distinctive US pattern: intergenerational correlation coefficients for both total income and labor earnings seem to be very similar across developed countries (see the recent survey of Bjorklund and Jantti (1998)). Overall, the relative consensus at the end of the twentieth century seems to be that commonality and inertia are the main characteristics of intergenerational mobility: mobility rates just do not seem to vary very much.

This relative consensus obviously does not imply that the issue of intergenerational mobility is no longer controversial. First, there are still some disagreements about whether the extent of mobility is that similar across countries. For example, Bjorklund and Jantti (1998) note that when discussing with their US colleagues, they "were struck by the strong belief that the US is a more open society with higher intergenerational mobility than Western European ones". Although there does not seem to exist any strong scientific evidence to confirm this "US exceptionalism" thesis, it is fair to say that there is sufficient uncertainty about these cross-country comparisons to explain how such disagreements can persist. Careful cross-country comparisons of mobility patterns are still in their infancy. Although we can be relatively confident that mobility rates do not differ enormously across comparable countries, it is by no means impossible that, as better data sets become available and more detailed comparative studies develop, we become able to identify interesting cross-country variations. For instance, a recent comparative study has found higher intergenerational educational and occupational mobility in the US than in Italy, which can be viewed as consistent with the liberal theory of mobility in industrial societies.[6]

Next, and most importantly, a relative consensus about the level of mobility in industrial societies obviously does not provide us with a consensus about a theory of intergenerational mobility. Many different theoretical models are consistent with a given level of mobility, and the kind of empirical evidence that would be needed in order

[4] See Erikson and Goldthorpe (1992), who offer the most complete comparative study of occupational mobility rates to date. Whether there was a significant increase in mobility rates before World War II is unclear: Lipset and Bendix (1959) conjectured that all countries reach a high mobility threshold as they industrialize; the historical study by Kaelble (1985) suggest that mobility rates did increase during the shift from family firms to large corporations, due to the emergence of a large class of nonowner business executives and associated upwardly-mobile careers (see Section 4.2 below).

[5] See Shavit and Blossfeld (1993).

[6] See Ichino et al. (1997). The authors conclude that higher mobility rates in the US could result from the higher mobility incentives implied by higher earnings inequality in the US, and that in any case that the Italian public school system seems to fail to deliver higher mobility.

to discriminate between these different models is even more uncertain and scarce than evidence about mobility levels. In particular, a relative consensus about actual mobility rates would not tell us very much about whether actual mobility is "high" or "low" and whether we should (and could) do something about it. This chapter will try to demonstrate that the issue of intergenerational mobility is still very controversial, but that disagreements between various existing theories span over many different dimensions, as opposed to the simple, one-dimensional conflict between liberal and Marxist theories referred to above. In the extreme form of the "liberal vs. marxist" conflict, things were indeed very simple: everybody agreed that mobility should and could be high, but the "right-wing" (i.e., pro-laissez-faire) view claimed that a mixture of free market and laissez-faire policies was sufficient to generate such an outcome, while the "left-wing" (i.e., pro-interventionist) view claimed that markets were so grossly imperfect that only a radical destruction of the free market system could make it happen. In practice, things can be more complicated.

First, there is no reason to believe that the socially-optimal level of intergenerational mobility should be high. If one believes that low intergenerational mobility is due to the high heritability of ability, and that the distortionary costs of welfare redistribution are very high, then it is perfectly reasonable to argue that public intervention should not try to interfere too much with the efficient functioning of the private choices and contractual arrangements made by families and markets, even though this laissez-faire process leads to little intergenerational mobility. Historically, this "conservative" type of right-wing view has been at least as widespread as the "liberal" type referred to above. Conservative right-wing views about mobility have been very influential not only in traditional societies, but also in advanced liberal societies such as the US, where there is long tradition of academic writing about the social efficiency of an "hereditary meritocracy" and the evils of egalitarian beliefs about individual abilities (see Herrnstein and Murray (1994) for the latest episode of this tradition). Although economists rarely use these terms to describe their theories, it is interesting to note that both types of laissez-faire theories are also present in the very important writings of Chicago economists about intergenerational mobility. On the one hand, Becker and Tomes (1986) interpret the high level of mobility that they observe in the US primarily in the liberal right-wing way (ability is moderately heritable and markets are highly efficient). On the other hand, Mulligan (1997) interprets the low level of mobility that he observes in the US primarily in the conservative right-wing way (persistent inequality derives from efficient parental and market choices, and there is not much one can do about it).[7]

Left-wing views are in a sense more homogenous: unlike right-wing views, they all share the basic premise that intergenerational mobility in the ideal society should be high. However they strongly disagree about what should be done in order to achieve this high and efficient mobility. Left-wing theories traditionally emphasize market imperfections and their inefficient, negative impact on intergenerational mobility. But there are

[7] See Sections 2-4 below, and especially Section 4.2.

different ways to analyze market imperfections: one can believe that markets have some imperfections that make inequality more persistent than it ought to be, without inferring from this claim that the only possible remedy is the abolition of private property and the market system altogether. At the very least, one needs to distinguish between "radical" left-wing views, of which Marxist and socialist theories of social mobility are the primary example, and "liberal" left-wing views, according to which market imperfections need to be corrected in a market-friendly manner.[8] In fact, left-wing, pro-interventionist theories of intergenerational mobility do not necessarily rely on any market imperfection at all. It is logically consistent to believe that observed mobility is the outcome of a market process that is basically efficient (in the Pareto sense), but that the distortionary costs of pure redistribution are relatively low, and that opportunities for consumption and welfare should be equalized between dynasties to a substantial extent.[9]

The very fact of locating the various views on a one-dimensional left vs. right scale can be in itself very misleading. For instance, it is not obvious how one would locate on such a one-dimensional axis the theory of social mobility developed in Plato's Republic.[10] On the one hand, Plato obviously does not believe that decentralized choices and the price system can set social priorities in the appropriate way. He recommends for instance that smart kids be taken away from their lower-class families, because the latter may not know how to raise them properly. This very activist view of mobility-enhancing policies would first seem to be very close to radical left-wing views, who have often advocated the need to socialize the education of children in order to counteract the family transmission and reproduction process. But on the other hand, Plato insists that bright lower-class kids are the exception rather than the rule, and that the ideal society should merely be characterized by a high degree of hereditary reproduction of rulers, warriors and producers. This makes Plato much closer to the conservative right-wing view of the "hereditary meritocracy" than to most left-wing views.

What this Plato example shows is not only that our modern concepts of right vs. left may not be very appropriate to classify the theories of the past. It also shows that there are deep reasons why the radical left and the conservative right are often much closer than what a one-dimensional classification would suggest. If we push it to the extreme (as historical experiments often did ...), the radical left's strong emphasis on market imperfections requires strong beliefs about the inequality of abilities between individuals: without the help of some enlightened elite, disadvantaged individuals are unable to interact in society and can easily be exploited, so that social justice and high mobility may require a very authoritarian hierarchical structure. Conversely, the conservative right's strong emphasis on the inequality of ability between dynasties can easily lead to question the capability of low-ability individuals to interact in society and on the

[8] See Sections 4–6 below.

[9] See Sections 2–3 below, and especially Sections 2.4 and 3.2.

[10] See, e.g., Merllié and Prévot (1991: p. 15) for an introduction to Plato's theory.

market place, which explains why the conservative right often advocates authoritarian, anti-market policies in some domains.[11]

1.2. Organization of this chapter

In order to distinguish as clearly as possible between the different dimensions of conflicts about persistent inequality and intergenerational mobility, the rest of this chapter will be organized as follows.

We will first deal with theoretical models of intergenerational mobility based upon Pareto-efficient markets (Sections 2 and 3). Section 2 concentrates on the process of (nonhuman) wealth transmission from parents to children, while Section 3 concentrates on the process of ability transmission. The assumption of efficient markets imply that policy intervention in these theoretical models is motivated solely by distributive justice considerations. That is, the only policy question is whether we should have a large redistributive tax on inheritance and/or labor earnings, so as to make consumption and welfare inequality less persistent than it would otherwise be. We will see that different theoretical models of the family transmission process have different implications regarding the distortionary costs of such redistributive policies, and that existing evidence does not allow us to discriminate very sharply between them.

We will then review the main existing theories of persistent inequality based upon market inefficiencies (Sections 4, 5 and 6). Section 4 deals with the intergenerational mobility consequences of imperfect credit markets. Section 5 discusses theories of persistent inequality based upon local segregation into unequal communities. Section 6 reviews theories of persistent inequality based upon self-fulfilling beliefs, and in particular the theory of discrimination. All of these theories imply that inequality is more persistent than what the simple family transmission of wealth and ability would imply if markets were perfect. Moreover, the extra persistence is inefficient, in the sense that appropriate corrective policies can raise intergenerational mobility and output at the same time. This attractive possibility obviously depends on the empirical relevance of these transmission mechanisms: if they do not account for a large fraction of persistent inequality, then we are back to the inequality/efficiency trade-off. As we will see, more empirical evidence is needed before we can give a precise estimate of how much these mechanisms contribute to the intergenerational transmission of inequality.

Finally, note that many other mechanisms of "inefficient inequality" have been explored by economists, although they are not covered in this chapter. For instance, the theory of employer monopsony implies that firms will pay wages below marginal products, even though this reduces labor supply, so that minimum-wage redistribution would be efficiency-improving.[12] More generally, the existence of mobility costs or firm-specific human capital can lead to hold-up problems and allow employers to pay wages below

[11] See especially Section 3.2 below.

[12] See, e.g., Card and Krueger (1995) for recent empirical research on local monopsony and the efficiency effects of minimum wages.

marginal products (or employees to charge wages above marginal products ...), in which case salary scales and centralized constraints on wages can have positive distributive and efficiency effects at the same time.[13] Another important example is the keynesian theory, one popular version of which claims that redistributing purchasing power towards wage-earners can generate both a fairer distribution of income and positive expansionary effects for everybody.[14] All these theories play an important role in the way many people think about inequality and redistribution (rightly or wrongly), but they will be neglected in this chapter, because they do not deal explicitly with the issue of intergenerational mobility and persistent inequality across generations. In particular, we will assume throughout the chapter that wages are equal to marginal products, just as in the textbook model of competitive labor markets, so that fiscal redistribution is the only form of redistribution that can possibly be justified.

2. Persistent inequality and the family transmission of wealth

The most obvious channel explaining why inequality can persist across generations is the transmission of wealth from parents to children through inheritance. We first describe how inheritance contributes to raise inequality and to make it more persistent across generations (Section 2.1). We then show that most theoretical models of inheritance and inequality dynamics predict that wealth inequality and its effects on intergenerational mobility should indeed persist in the long-run (Sections 2.2 and 2.3). Finally, we use these theoretical models to analyze the prospects for raising welfare mobility through progressive inheritance taxation (Section 2.4).

2.1. The contribution of inheritance to the persistence of inequality

Consider a simple infinite-horizon model where each dynasty i lives during one period and has exactly one offspring.[15] Total income of dynasty i at period t can be written as the sum of two terms:

$$y_{it} = v_t a_{it} + r_t w_{it}. \tag{2.1}$$

The first term, $v_t a_{it}$, is the labor income of dynasty i at period t: it is the product of the wage rate v_t and of its productive ability parameter a_{it} (measured in efficiency labor units). The second term, $r_t w_{it}$, is the capital income of dynasty i at period t: it is the product of the interest rate r_t and of the wealth w_{it} transmitted by dynasty i from

[13] Thurow's (1975) theory of income distribution is largely based on the idea that there exist direct policy interventions on the labor market that would be both redistributive and efficiency-improving.

[14] See Murphy et al. (1989) for a modern modeling of how income distribution can affect demand composition and the level of economic activity.

[15] For a discussion of differential fertility behaviour, see Section 2.2 below. For a discussion of marriage, assortative mating and their effects on the persistence of inequality, see Section 5.3 below.

generation $t-1$ to generation t. We note $G_t(w)$ the distribution of wealth inherited by generation t. This section concentrates on the process of (non-human) wealth transmission. The process of ability transmission, and in particular the possible impact of wealth inequalities on ability transmission (e.g., because of imperfect credit), will be analyzed in Sections 3–6 below. At this stage, we take as given some exogenous law of motion for abilities. Although most results of Section 2 can easily be generalized, for simplicity we will mainly consider the following cases: uniform labor earnings ($\forall i, t, a_{it} = 1$); random labor earnings with zero intergenerational transmission ($\forall i, t, a_{it} = 1 + \epsilon_{it}$, where ϵ_{it}, is an error term with zero mean, variance σ_ϵ^2 and zero serial correlation); random labor earnings with first-order serial correlation ($\forall i, t, a_{it} = 1 - \rho + \rho a_{it-1} + \epsilon_{it}$, where ρ is the intergenerational correlation of ability).

The first obvious implication of Eq. (2.1) is that as long as a_{it} and w_{it} are not negatively correlated, the inequality of total income will tend to be larger than the inequality of labor earnings. The standard deviation of total income is simply equal to the sum of the standard deviation of labor earnings and the standard deviation of capital income in case ability and wealth are uncorrelated, and it is even larger if the correlation is positive.[16] In practice, one does indeed observe that total income inequality is always larger than the inequality of labor earnings.[17]

If one further assumes the inheritance w_{it+1} left by dynasty i to generation $t+1$ to be an increasing function $S(y_{it})$ of income y_{it}, then one obtains the following transition equation for total income:

$$y_{it+1} = v_{it+1} + r_{t+1} S(y_{it}). \tag{2.2}$$

Equation (2.2) shows that the second obvious implication of inheritance is that it tends to perpetuate the inequality of living standards across generations. For instance, Eq. (2.2) implies that even if the intergenerational correlation of labor earnings is assumed to be zero, the intergenerational correlation of total income is positive. More generally, Eq. (2.2) implies that the intergenerational income correlation will always be larger than the intergenerational earnings correlation, as long as the ability-wealth correlation is not negative. This second implication is also confirmed by recent empirical evidence. Mulligan (1997) uses the PSID to estimate these intergenerational correlations, and he finds that the correlation coefficients for consumption and total income fall in the 0.7–0.8 range, while the intergenerational correlation of earnings is about 0.5. These estimates are probably the most reliable estimates to date (see Section 4.2 below for a discussion of downward biases in previous estimates). Note that this is a very large

[16] Through this chapter, we will mostly refer to rudimentary measures of inequality and mobility such as standard deviations, coefficients of variation and intergenerational correlations, simply because they are very convenient in loglinear models. See, e.g., Cowell's chapter 2 in this Handbook for a survey of existing inequality measures.

[17] See, e.g., Davies and Shorrocks' chapter 11 in this Handbook. This simple fact shows that the main purpose of wealth accumulation is not to smooth life-time or integenerational earnings shocks (in which case income inequality should be lower than earnings inequality).

difference. For instance, an intergenerational correlation of 0.7 means that if parents of children i are five times richer (in total income) than parents of children j, then children i will be on average about 3.1 times richer (in total income) than children j. A correlation of 0.5 means that children of parents who are five times richer (in earnings) will be "only" about 2.2 times richer (in earnings).[18] This shows that inheritance is a very powerful mechanism to transmit inequality across generations, and this explains why the inheritance channel of inequality transmission has attracted so much attention.

2.2. *The long-run dynamics of wealth inequalities with exogenous savings*

From a theoretical viewpoint, should we expect these two properties (inheritance raises inequality and makes it more persistent across generations) to hold in the long-run? If the inheritance function $S(y)$ is concave and if there is no inequality of labor earnings ($\forall i a_{it} = 1$), then one can easily show that the answer is negative. As Stiglitz (1969) pointed out, the concavity of inheritance and the equalizing effect of labor earnings imply that wealth inequality will decline slowly over time and that each dynasty will eventually own the same steady-state wealth. To see this, assume that gross output is given by a standard, concave production function $f(k_t)$, where $k_t = w_t$ is the capital stock per labor unit, i.e., the average of w_{it} across all dynasties. Wealth depreciates at rate $\delta > 0$ (i.e., net output is equal to $f(k) - \delta k$). Dynastic and aggregate transition equations are given by:

$$w_{it+1} = S(v_t + r_t w_{it}) + (1 - \delta) w_{it}, \tag{2.3}$$

$$w_{t+1} = S(f(w_t)) + (1 - \delta) w_t. \tag{2.4}$$

Equation (2.4), together with the concavity of $S(y)$, imply that aggregate wealth w_t will converge to a unique long-run wealth level w_∞. In the special case where savings are linear ($S(y) = sy$), w_∞ is simply given by $sf(w_\infty) = \delta w_\infty$. The fact that the capital stock per labor unit converges to w_∞ implies that the interest rate r_t converges to $r_\infty = f'(w_\infty)$, while the wage rate v_t converges to $v_\infty = f(w_\infty) - r_\infty w_\infty$. Equation (2.3) then implies that all dynasties will converge to the same long-run wealth level $w_\infty = sv_\infty/(\delta - sr_\infty)$, irrespective of the initial wealth distribution $G_0(w)$. That is, initial wealth inequalities do not persist in the long-run.

However, this conclusion ceases to hold if any of the assumptions is relaxed. For instance, if inheritance behavior is better approximated by a convex savings function $S(y)$, i.e., if the savings rate of the poor is smaller than the savings rate of the rich, such as in the Kaldorian class savings model, then wealth inequalities will persist in the long-run (see Bourguignon (1981) for such an extension of the Stiglitz model). That is, the

[18] That is, $5^{0.7} = 3.1$, while $5^{0.5} = 2.2$. All the intergenerational correlation estimates referred to in this chapter are obtained by regressing the log of children's income (or consumption, or earnings) on the log of parental income (see Mulligan, 1997: Chaps. 6 and 7).

long-run distribution of wealth $G_\infty(w)$ will depend on the initial distribution $G_0(w)$. In general, there will exist multiple long-run wealth levels $w_{1\infty}, w_{2\infty}, \ldots, w_{n\infty}$, and the long-run wealth level of each dynasty can be expressed as a function of their initial wealth w_{i0}. In steady-state, wealthy dynasties have income and consumption levels that are permanently higher than those of poorer dynasties, although all dynasties have the same labor income.

Another reason why wealth inequalities might not decline over time is differential fertility behavior. If one assumes that dynasty i has $1 + n_i$ children, then Eq. (2.3) becomes:

$$w_{it+1} = S(v_t + r_t w_{it})/(1 + n_i) + (1 - n_i - \delta)w_{it}. \tag{2.4}$$

It is obvious from Eq. (2.4) that differential fertility behavior can have the same effects as convex savings functions: if poor dynasties tend to have more kids than wealthy dynasties, then wealth inequalities can persist in the long-run even if all dynasties have the same savings rate. This kind of analysis of how different savings behavior, family structure and inheritance patterns generate more or less persistent inequality has a long tradition in economics.[19]

Even in the absence of convex inheritance functions or differential fertility behavior, wealth inequalities persist in the long-run if we assume that labor earnings are unequally distributed. For instance, if abilities are perfectly transmitted across generations ($\forall t$, $a_{it} = a_i$), then the long-run wealth distribution amplifies the inequality of labor earnings: with linear savings, w_{it} converges toward $w_{i\infty} = sa_i v_\infty/(\delta - sr_\infty)$ (see Stiglitz, 1969: p. 394). The long-run standard deviation of total income is larger than that of labor earnings, and the multiplicity factor is an increasing function of the savings rate s. The intergenerational correlations of income, earnings and consumption are all equal to 1. If we assume abilities to be drawn at random at each generation ($\forall i, t, a_{it} = 1 + \epsilon_{it}$), then the transition equation $w_{it+1} = (1 - \delta)w_{it} + s(r_t w_i + v_t a_{it})$ implies that the wealth distribution $G_t(w)$ converges to a long-run distribution $G_\infty(w)$ with mean w_∞ (such as $sf(w_\infty) = \delta w_\infty$) and variance σ_w^2 given by:

$$\sigma_w^2 = s^2 r_\infty^2 \sigma_\epsilon^2/(1 - (1 - \delta + sr_\infty)^2). \tag{2.5}$$

Equation (2.5) shows that the long-run standard deviation of wealth is an increasing function of the savings rate and of the variance of shocks. In this model, the long-run standard deviation of total income is again larger than that of earnings, and the long-run integenerational correlation of total income is positive, although the intergenerational correlation of earnings is permanently equal to zero. More generally, if one assumes

[19] The concern about how the poor's high fertility might lead to persistent poverty dates back at least to Malthus and Ricardo. James Meade has also written extensively about the interplay between savings behavior, family patterns and inequality dynamics (see Atkinson (1980) and subsequent references). See Chu (1991) for a recent analysis of the effect of primogeniture on long-run inequality and mobility.

some positive heritability of abilities ($\forall i, t, a_{it} = 1 - \rho + \rho a_{it-1} + \epsilon_{it}$), then one can easily show that the long-run correlation between wealth and ability is positive,[20] so that the standard deviation and intergenerational correlation of total income are larger than the standard deviation and intergenerational correlation of labor earnings. That is, the two key properties pointed out in Section 2.1 hold in the long-run.

2.3. *The long-run dynamics of wealth inequalities with dynastic utility functions*

How would this analysis differ if one explicitly models inheritance behavior instead of taking as given some exogenous savings function $S(y)$? In general, there are different private motives that can contribute to explain the existence of inheritance. First, bequests might just be the unintended side-product of precautionary savings in a world of imperfect insurance. That is, each generation saves during its lifetime in order to self-insure against negative shocks to its earnings potential, and imperfections on the annuity market imply that accidental bequests are passed on to the next generation at the time of death. The exact form of the inheritance function $S(y)$ that one can derive from such a model depends on the specific structure of lifetime earnings shocks, risk aversion, the degree of insurance market imperfections, etc.[21] There does not seem to be any general presumption as to whether the resulting $S(y)$ function should be concave, linear or convex.

Next, bequests can be motivated by intergenerational altruism. There exists two different ways of modeling bequests and intergenerational altruism. Becker and Tomes (1979) and Atkinson (1980) are two often cited papers that explicitly incorporate intergenerational altruism in general-equilibrium, Stiglitz-type models. One can either assume that the bequest enters directly into the utility function of the parents, or that parents care about their children's utility per se. The first formulation depends entirely on the specific form of the parental utility function $U(c_{it}, b_{it+1})$. For instance, if the utility function over parental consumption and bequest has a Cobb–Douglas form ($U(c, b) = c^{1-s}b^s$), then the inheritance function is linear ($S(y) = sy$). The second formulation also depends on the specific way one assumes parents to care about future generations' utility levels. The following form of Beckerian dynastic utility function has become very popular among economic theorists:

$$U_{it} = \sum_{S \geq t} U(c_{is})/(1 + \theta_i)^s, \tag{2.6}$$

$\theta_i \geq 0$ is the rate of time preference: a low θ_i means that dynasty i is very altruistic towards its children, and conversely. Assume that each dynasty can perfectly forecast

[20] Simple computations give the following formula for the long-run covariance between wealth and ability: $\text{cov}(w_i, a_i) = sv_\infty\sigma_a^2/(1 - \delta + sr_\infty)$, with $\sigma_a^2 = \sigma_\epsilon^2/(1 - \rho)^2$.

[21] Note that in general the resulting $S(.)$ function might depend on wealth on w and not only on income y. See Davies and Shorrocks' chapter 11 in this Handbook for more on savings and inheritance behavior.

the ability parameters a_{it} of its future generations, or at least that each dynasty can purchase complete insurance contracts against such risks.[22] Under the assumption of perfect capital markets, utility maximization implies that the consumption level of future generations will not depend on their ability shock. For any dynamic process from which abilities are drawn, the trade-off between parental consumption and children's consumption leads to the following first-order condition:

$$U'(c_{it})/U'(c_{it+1}) = (1 + r_{t+1})/(1 + \theta_i). \tag{2.7}$$

This first-order condition has very strong impications for the dynamics of the wealth distribution. First, Eq. (2.7) implies that if some dynasties have a permanently higher θ_i than some other dynasties, then the consumption level of more altruistic dynasties will grow at a higher rate than the consumption level of less altruistic dynasties. In the long run, the relative consumption share of less altruistic dynasties goes to zero, and the most altruistic dynasties own all the wealth (Mayshar and Benninga (1996)). We will come back later to this extreme form of taste-based persistent inequality (see Section 3.2 below).

Next, in the case where all dynasties have the same rate of time preference ($\forall i \theta_i = \theta$), Eq. (2.7) implies that a necessary condition for the economy to be in a steady-state is $r_\infty = \theta$. With a concave, net-of-depreciation production function $f(k)$, this implies that the steady-state average wealth w_∞ per efficiency labor unit must be such that $f'(w_\infty) = r_\infty = \theta$. Conversely, any consumption distribution $G_\infty(c)$ that is consistent with an average wealth equal to w_∞ can be a steady-state, where "consistent" simply means that average consumption c_∞ is equal to long-run average output $f(w_\infty)$. In the special case with uniform labor earnings ($\forall i\, a_{it} = 1$), any wealth distribution $G_\infty(w)$ such that the average wealth is equal to w_∞, can be a steady-state. Dynasty i with long-run wealth $w_{i\infty}$ consumes $c_{i\infty} = v_\infty + r_\infty w_{i\infty}$ at each period. In the general case where productive abilities are drawn from some arbitrary dynamic process, dynastic long-run wealth may vary with the specific ability shock of each generation, but the important point is that each dynasty will converge towards a fixed consumption level. That is, irrespective of what the intergenerational correlation of labor earnings might be, the theoretical prediction of the dynastic utility model is that the long-run intergenerational correlation of consumption should be equal to 1. This theoretical prediction can be viewed as an extreme form of the more general prediction according to which the intergenerational correlation of consumption and total income should be higher than that of labor earnings.[23]

Mulligan (1997) has recently pointed out that this very strong theoretical prediction has strong implications regarding how we should model intergenerational altruism. Mulligan argues that since we do observe regression to the mean in consumption across

[22] That is, risks about the future ability parameters of its future generations.

[23] These steady-state results can be generalized to models with balanced growth, such as those surveyed by Bertola's chapter 9 in this Handbook.

generations (the observed intergenerational consumption correlation is less than 1; see Section 2.1 above), it must be the case that altruism is not randomly distributed across dynasties and that poor dynasties are on average more altruistic than wealthy dynasties.[24] Mulligan then develops a theoretical model of endogenous altruism where the poor turn out to be more altruistic than the rich, so that the predicted intergenerational correlation of consumption is less than 1. The basic idea of Mulligan's model is that the amount of time spent per kid increases altruism: since rearing costs include time costs, high wage rate dynasties will spend less time with their children and will love them less. Note that this theory differs from the Becker–Barro (1988) theory of fertility and quality/quantity trade-offs, according to which fixed monetary rearing costs induce wealthy parents to choose to have more kids of lower average quality (i.e., less altruism per kid). This allows Becker and Barro to predict regression to the mean in consumption in the dynastic utility model, but Mulligan argues that the predicted positive relationship between income and fertility is counterfactual. In contrast, Mulligan's model predicts that wealthy dynasties have both less kids and less altruism per kid. Mulligan (1997) concludes that his theoretical model is the only model that can simultaneously account for all the observed facts.

Mulligan's reasoning is not entirely convincing, however. First, the dynastic utility model predicts a unitary intergenerational correlation of consumption only if we assume perfect insurance markets. In practice, one can very well imagine why even very altruistic parents cannot guarantee with absolute certainty that their children will enjoy some fixed consumption level, irrespective of their labor earnings. Obvious moral hazard reasons can easily explain why there must be some degree of regression to the mean in consumption across generations in the dynastic utility model, with no need for a theory of endogenous altruism. Next, regardless of this imperfect insurance issue, one must bear in mind that the dynastic utility model described by Eqs. (2.6) is primarily a convenient theoretical construction, rather than a well-documented explanation of how people actually behave. Models with exogenous savings $S(y)$, which can be rationalized by models of inheritance based upon precautionary savings or direct utility for bequests, can easily explain why the intergenerational correlation of consumption is both larger than the intergenerational earnings correlation and smaller than 1. More empirical evidence seems to be needed before we take too seriously the implications of Eq. (2.7) for the theory of intergenerational altruism (see below).

2.4. *"Active" vs. "passive" inheritance: the costs of redistribution*

Most theories of justice would argue that it is unfair that two individuals with exactly the same behavior and characteristics enjoy vastly unequal consumption and welfare levels, simply because one individual received a large inheritance and the other did not. For

[24] If wealthy dynasties were more altruistic, then we would observe no regression to the mean and the wealth distribution would diverge (just as in the case where some dynasties have a rate of time preference that is permanently higher than that of other dynasties).

instance, according to Rawls' difference principle, we should try to improve as much as possible the prospects of the children who receive no inheritance.[25]

The obvious way to correct for the unfair persistence of inequality implied by the family transmission of wealth would be to tax inheritance and to redistribute the tax revenues to all individuals. If wealth inequalities tend naturally to decline over time, such as in the model with concave savings and uniform labor earnings, then the redistributive taxation of inheritance does not only redistribute income and welfare today: it also increases the rate at which wealth is equalized (Stiglitz, 1969: p. 392). More generally, in models where full equality of wealth is a steady-state, i.e., in models with uniform labor earnings (either with exogenous savings or with dynastic preferences), it is sufficient to redistribute wealth at a 100% rate at $t = 0$ in order to reach a permanent steady-state with no wealth inequality. However, in more realistic models with unequal labor earnings, the economy always returns to a steady-state regime of persistent wealth inequalities (see Sections 2.2 and 2.3 above). In these more realistic models, redistributive inheritance taxation needs to be permanent in order to reduce permanently the intergenerational transmission of inequality trough inheritance.

Such a permanent taxation of inheritance is likely to have some adverse effects on the level of bequests. The magnitude of these adverse effects depends crucially on how one models inheritance behavior. If inheritance is the unintended side-product of precautionary savings and life-cycle wealth accumulation, then inheritance taxation has obviously no effect on the level of pre-tax bequests. That is, the distortionary costs of redistributive inheritance taxation are negligible if inheritance is primarily a "passive" phenomenon. On the other hand, if inheritance is primarily motivated by intergenerational altruism and is the outcome an "active" choice process, then the distortionary costs are potentially large. Several empirical studies have shown that intergenerational transfers are at least partly motivated by intergenerational altruism: for instance, households do not seem to annuitize their wealth as much as they could.[26] However, economists vastly disagree about what part of total wealth accumulation and transfers can be explained by intergenerational altruism and what part can be explained by life-cycle accumulation and precautionary savings, i.e., about the relative importance of "active" and "passive" inheritance.[27]

Moreover, intergenerational altruism per se does not necessarily imply that the effect of taxation on pre-tax bequests is negative. If bequests enter directly into the utility function of the parents ($U_i = U(c_{it}, b_{it+1})$), then the effect of taxation on pre-tax bequests can be positive or negative, depending on whether the elasticity of substitution between parental consumption and bequest is smaller or larger than 1 (see Atkinson, 1980: p. 178 and subsequent references). In the special case of a Cobb–Douglas utility

[25] See, e.g., Sen's chapter 1 in this handbook for a survey of distributive justice theories.

[26] See, e.g., Bernheim (1991).

[27] See, e.g., Kessler and Masson (1989), Kotlikoff (1988) and Modigliani (1988) for conflicting empirical viewpoints.

function ($U(c, b) = c^{1-s}b^s$), the elasticity of substitution is equal to 1, and pre-tax bequests do not depend on the level of inheritance taxation.

However, if intergenerational altruism is better described by dynastic utility functions given by Eq. (2.6), then redistributive inheritance taxation has unambiguously negative effects on capital accumulation.[28] This is because when parents care about their children's consumption, inheritance taxation acts as a capital income tax, and capital income taxes are well-known to have negative accumulation effects in models with infinite-horizon, dynastic preferences. For instance, if all dynasties have the same rate of time preference θ and can perfectly insure against all future ability shocks, Eq. (1.7) implies that if inheritance is taxed at rate τ, then the long-run, pre-tax interest rate r_∞ will be shown that $(1 - \tau)r_\infty = \theta$. That is, the long-run capital stock per capita k_∞ will decline until the point where the after-tax rate of return is again equal to θ, i.e., the new long-run k_∞ will be such that $(1 - \tau)f'(k_\infty) = \theta$. It follows that long-run income depends negatively on the rate of redistributive inheritance taxation.

Some authors have used this simple result in order to conclude that the socially-optimal rate of all forms of capital taxation, and in particular of inheritance taxation, should be equal to zero (see, e.g., Lucas, 1990). This very strong conclusion seems excessive. First, as was already pointed out, the infinite-horizon, dynastic utility model is not the only available theoretical model, and the question of its empirical relevance usually receives far less attention than the careful derivation of its theoretical implications. Next and most importantly, even if higher tax rates on inheritance do imply lower long-run average wealth, which seems like the most likely case, this obviously does not imply that the socially-optimal tax rate should be equal to zero. In order to make a proper welfare analysis, one needs to compare the distortionary costs of inheritance taxation, as measured by the long-run fall in average income, with the redistributive gains. In the standard dynastic utility model, one can show that in the long-run, even zero-wealth individuals will loose more from the distortionary costs of the tax than they will gain from its redistributive impact.[29] But in a world of permanent growth in living standards, the interpretation of such a result is somewhat complicated: the low-wealth individuals who benefit from redistributive inheritance taxation in the short-run enjoy lower welfare levels than those who are affected by the distortionary effects of taxation in the long-run, and it is not obvious how one should balance the two effects. In other words, even if we knew with certainty that inheritance taxation, as it has been applied in the US during the

[28] Note that we have little direct empirical evidence as to whether intergenerational altruism is better described by utility for bequests, by dynastic utility functions, or by other mathematical representations. From a theoretical perspective, Bernheim and Bagwell (1988) and Abel and Bernheim (1991) have argued that if the dynastic model leads to a number of implausible implications if we take it too seriously, and therefore that we should be extremely cautious when we use it for policy purposes.

[29] If bequests are taxed at rate t and the tax revenues are used to finance a lump-sum transfer, then the long-run net income of a prolaterian dynasty with zero wealth is equal to $v_\infty + T_\infty$, i.e., the sum of the wave rate $v_\infty = f(k_\infty) - r_\infty k_\infty$ and the lump-sum transfer $T_\infty = tr_\infty k_\infty$. That is, $v_\infty + T_\infty = f(k_\infty) - (1 - t)r_\infty k_\infty = f(k_\infty) - \theta k_\infty$. It follows that the long-run net income of zero-wealth dynasties is maximized if $f'(k_\infty) = \theta$, i.e., if $t = 0$ (see Judd, 1985).

twentieth century, has caused an average income loss of 10% by 1998 (which we do not know), this would not automatically mean that total social welfare during the twentieth century would have been higher in the absence of all inheritance tax revenues. From a practical policy perspective, the only interesting question is the magnitude of the adverse effects of redistributive inheritance taxation and the speed at which these negative effects are produced, as compared to the size and timing of positive distributive effects.

Under special assumptions, one can show that the tax-induced decline in absolute wealth dispersion can be smaller than the fall in average wealth, so that redistributive inheritance taxation can actually lead to a long-run rise of relative wealth inequality. This paradoxical result (redistribution increases long-run inequality) has been given high prominence by Becker and Tomes (1979: pp. 1175–1178).[30] To see how it works, consider the model with linear savings and i.i.d. ability shocks (see Section 2.2 above). Equation (2.5) shows that the long-run standard deviation of wealth is an increasing function of the savings rate s. If we assume that s is a decreasing function of the inheritance tax rate t (for instance because the elasticity of substitution between parental consumption and bequests is larger than 1), then it follows that inheritance taxation leads to decline in the long-run standard deviation of wealth. However, long-run average wealth also declines ($w_\infty = sv_\infty/(\delta - sr_\infty)$). One way to measure long-run, relative wealth inequality is to compute the coefficient of variation of the long-run distribution of wealth:

$$CV(s) = \sigma_w^2/w_\infty^2 = (\delta - sr_\infty)\sigma_\epsilon^2/(2 - \delta + sr_\infty). \tag{2.8}$$

Equation (2.8) shows average wealth falls more rapidly than the standard deviation of s when s declines, so that $CV(s)$ is a decreasing function of s. Therefore inheritance taxation and lower savings rate can lead to a long-term rise of relative inequality. However, as Atkinson (1980: p. 178) has pointed out, this is again a theoretical result, and one can easily construct other theoretical models with different specifications of savings behavior where the standard deviation of wealth would decline more than the average wealth.

Overall, we just seem to have very little practical knowledge about the socially-optimal rate of redistributive inheritance taxation. After a quick review of how cross-country and time-series variations of tax progressivity might have affected observed intergenerational mobility, Mulligan (1997: p. 218) is led to the obvious conclusion: "much more research (...) are necessary to arrive at a strong conclusion regarding the unimportance of progressive taxes for intergenerational mobility".

[30] See also Stiglitz (1978).

3. Persistent inequality and the family transmission of ability

Intergenerational wealth transfers make consumption and welfare more persistent across generations than labor earnings. According to the best available estimates, the intergenerational correlation goes up from about 0.5 for earnings to about 0.7 for consumption and total income (see Section 2.1 above). However, although wealth transfers are a very powerful transmission mechanism, these figures show that the main component (at least 70%) of the intergenerational correlation of welfare is due to the persistent inequality of labor earnings, and any useful theory of intergenerational mobility must address this fact. Some theories attribute a large fraction of the intergenerational earnings correlation to market inefficiencies, and in particular to wealth transfers themselves (see Sections 4–6 below). In this section, we focus on theories based upon efficient markets, according to which persistent earnings inequality can be explained either by a combination of direct family transmission of productive abilities and efficient human capital investments (Sections 3.1 and 3.2), or by the family transmission of ambition and other tastes that are conducive to high productive ability (Section 3.3).

3.1. The transmission of productive abilities

In Section 2, we considered a simple model of ability transmission, where productive abilities were measured in labor efficiency units and were given by the following transition equation:

$$a_{it} = 1 - \rho + \rho a_{it-1} + \epsilon_{it}. \tag{3.1}$$

In order to introduce human capital investments and to distinguish between pure ability endowments and human capital investments, Eq. (3.1) can be broken down into two separate equations (see, e.g., Becker and Tomes, 1986):

$$e_{it} = 1 - \rho + \rho e_{it-1} + \epsilon_{it}, \tag{3.2}$$

$$a_{it} = A(e_{it}, h_{it}). \tag{3.3}$$

Equation (3.2) relates the pure ability endowment of generation t to that of the previous generation, where ρ measures the intergenerational correlation of ability endowments. Becker and Tomes (1979, 1986) emphasize that pure ability endowments should be interpreted in a broad sense. That is, Eq. (3.2) measures not only the genetic transmission of innate abilities, but also the cultural transmission of family characteristics through childhood learning and family interaction. The relative importance of genetic vs. cultural transmission has always been a very controversial issue. In fact, even Herrnstein and Murray (1994), who have often been accused of overestimating the importance of genetic transmission, recognize that from the few reliable adoption studies that we have, childhood family environment seems to be more important than

genetic factors per se.[31] In any case, the relevant question from a policy perspective is whether one can do something about these early childhood environmental factors. If the inequality of ability endowments is primarily determined by childhood learning through interaction with the parents at a very early age, and if this nurturing process is associated with the personality and behavior of the parents rather than with material wealth per se, then there is not much one can do about persistent inequality of abilities, aside from mass adoption programs. In other words, if "culture" means nurture at the family level, then the nature vs. culture debate is almost irrelevant (see Becker and Tomes, 1986).[32]

The other key component of Eq. (3.1) is Eq. (3.3), which simply says that ability endowments e_{it} and human capital investments h_{it} translate into productive ability parameters a_{it} (measured in efficiency labor units). Becker and Tomes (1986) argue that ability endowments and human capital investments are likely to be complementary (i.e., $\partial^2 A/\partial e \partial h > 0$), so that allocative efficiency requires that high endowed ability kids benefit from higher human capital investments. Whatever the exact pattern of efficient investments might be, these efficient levels of human capital investments will be undertaken if one assumes credit and education markets to be first-best efficient. As Becker and Tomes (1986: p. S10) put it: "access to capital markets to finance investments in children separates the transmission of earnings from the generosity of resources of parents". That is, bright kids will always find sufficient credit on the market to finance their human capital investment as long as their investment is profitable, irrespective of their parental wealth. Becker and Tomes (1986) also introduce credit constraints into their framework, so that h_{it} can also depend on parental wealth w_{it} per se, but their conclusion is that credit constraints must be unimportant in the real world (see Section 4.2. below for an evaluation of their empirical argument).

This theory of efficient ability transmission has strong policy implications. First, it implies that public intervention should not try to interfere directly with the process of ability formation. If markets are efficient, then it is useless to finance public subsidies to human capital investments or to attempt to equalize opportunities in education, since all efficient investments were already made in the first place. Compensatory responses of parents would tend to undo their potential positive impact, so that such policies would have purely distortionary effects (Becker and Tomes, 1986: pp. S16–S17). In particular, such policies will not lead to higher mobility.[33] As Mulligan (1997: pp. 247–248) puts it,

[31] See Herrnstein and Murray (1994: pp. 410–413) and subsequent references, and especially the well-known French adoptions studies of Schiff et al. (1982) and Schiff and Lewontin (1986).

[32] The point is obviously that "culture" might also include socially-inefficient processes of inequality transmission, such as local segregation (see Sections 4–6 below).

[33] See also Conlisk (1974) for an early model showing under what conditions attempts to equalize educational opportunities can have negative effects, in the form of a decline of mobility rates. Conlisk's model is not based upon compensatory responses of parents, however: no choice process is formalized in the Conlisk model, which belongs to the class of what Goldberger (1989) refers to as the "mechanical" models of intergenerational mobility. Conlisk's result is based on the interpretation of equal opportunity policies as a reduction of the variance of random ability shocks (so that equalizing opportunities can reduce the probability of social ascent of bright poor kids).

"rather than reducing inequality, government subsidization of schooling may only have the effect of transferring resources from taxpayers to educators and richer families who are more likely to choose many years of schooling for their children". That is, the first implication of the theory of efficient ability transmission is that there is not much to do about the persistent inequality of abilities and labor earnings.

3.2. *Efficient inequality and the costs of redistribution*

However, the fact that we should not interfere with the efficient process of ability transmission does not imply that there should be no redistribution at all. If children are not responsible for the ability that they inherit from their parents, then even though we cannot redistribute productive abilities, it would seem to be fair to redistribute consumption and welfare, just as in the case of nonhuman wealth transmission (see Section 2.4 above). In the same way as in the case of redistributive inheritance taxation, the key question is that of the magnitude of the distortionary costs of a redistributive tax on labor earnings. Although a great deal of effort has been devoted to the empirical evaluation of these distortionary costs, economists vastly disagree about their magnitude.[34] In order to illustrate what these disagreements involve, Piketty (1995) developed a simple intergenerational mobility model where agents try to learn about the magnitude of the incentive costs of redistribution. Assume that labor income y_{it} of dynasty i at period t can take one of two positive values y_0 and y_1, with $y_1 > y_0 > 0$. The probability of obtaining a high income y_1 is given by the following equations:

$$\text{Proba}(y_{it} = y_1 | y_{it-1} = y_0, e_{it} = e) = \pi + \theta e, \tag{3.4}$$

$$\text{Proba}(y_{it} = y_1 | y_{it-1} = y_1, e_{it} = e) = \pi + \Delta\pi + \theta e, \tag{3.5}$$

$\theta > 0$ measures the extent to which individual achievement is responsive to individual effort e_{it}. Effort should be interpreted in a broad sense: it includes all actions that are within one's control and that can have an impact on achievement. $\Delta\pi > 0$ measures ex ante inequality between lower-class and upper-class children. For instance, if abilities are highly heritable, then $\Delta\pi$ should be large. Piketty (1995) assumes no market imperfection, so that the only redistributive policy that can possibly be justified is a redistributive tax on labor incomes y_0 and y_1. Effort is assumed to be private information, so that redistribution entails distortionary costs in the form of lower effort. One can easily show that distortionary costs are an increasing function of the income responsiveness of effort θ. It follows that the socially optimal rate of redistribution τ is low if economic success depends mostly on individual effort (θ high and $\Delta\pi$ low), and conversely that τ is high if economic success depends mostly on ex ante inequality (θ low and $\Delta\pi$ high).[35]

[34] See, e.g., Feldstein (1995) and Slemrod (1995) for some of the latest developments of this long-standing controversy.

[35] Piketty (1995) assumes a Rawlsian social welfare function (maximization of the expected utility of lower-class children), but the same qualitative property would hold with any utilitarian welfare function.

Piketty (1995) then assumes that dynasties use their own dynastic mobility experience to rationally update their probability beliefs μ_{it} about θ and $\Delta\pi$. One can show that this rational learning process will generally not result into complete learning of the true parameters (unless dynasties are sufficiently patient, so that they are ready to experiment during several generations effort levels which they believe to be inefficient in the short-run). In the long-run, "left-wing" dynasties believing that ex ante inequality is large and that the incentive costs of redistribution are low coexist with "right-wing" dynasties believing the opposite. Since they have stronger beliefs in individual effort, right-wing dynasties put in more effort and tend to be richer (whatever the true parameters might be). This implies that even though all dynasties have the same distributive objective, high-income individuals favor less redistribution than low-income individuals. This provides an example of a model where all agents agree about the aggregate mobility level, but disagree about how much incentives and mobility would be altered by redistribution, and therefore disagree about the socially-optimal level of redistribution. Just like economists, agents in this model would need large-scale social experiments in order to solve their disagreements. Unfortunately, reliable natural experiments are very difficult to design in the social sciences.

Some important ingredients are missing in the conflict over the socially-efficient level of redistribution described in the Piketty (1995) model. First, all right-wing dynasties in the model belong to the "liberal right-wing" type (see Section 1.1): they believe that the heritability of ability is low ($\Delta\pi$ low) and that market processes of social selection are highly responsive to individual effort (θ high).[36] This is because the model assumes that the incentive costs of redistribution are determined solely by the income responsiveness θ of children's effort input. That is, family choices are assumed not to be responsive to redistributive taxation: $\Delta\pi$ simply measures the mechanical transmission of inequality from parents to children, and a high $\Delta\pi$ means that the incentive costs of redistribution are low. However, as Becker's work on intergenerational mobility and the family has repeatedly emphasized, families do choose how much to invest in their children, and government intervention might tend to distort these choices. The theoretical models developed by Becker and his followers do not only describe how family wealth transfers might be adversely affected by government interference and redistributive taxation (see Section 2 above). Chicago economists also stress that families make many other choices, such as how much time they spend with their children, that might affect the labor earnings potential of future generations (and not only their capital income). For instance, Mulligan (1993) estimates that about 20% of the intergenerational transmission of earnings inequality can be attributed to the quality/quantity trade-offs made by parents.[37] A redistributive tax on future earnings might induce parents to spend less time

[36] All left-wing dynasties also belong to the "liberal" left-wing type.

[37] That is, Mulligan estimates (with PSID data) that the intergenerational earnings correlation would be about 20% lower if richer parents did not choose to have fewer kids of higher average quality (this is keeping everything else constant: as we already explained in Section 2.3, richer parents always tend to spend less time with their children than poorer parents in Mulligan's model; but the point is that if they did not choose to have

with their children and therefore to "produce" less productive ability, which might be detrimental to everybody in the long-run, in the same way as in the case of redstributive inheritance taxation (see Section 2.4 above). The potential sensitivity of nurturing and family choices to government policies implies that one can simultaneously believe that $\Delta\pi$ is high and that the socially-optimal, incentive-constrained level of redistribution is low. This corresponds to the "conservative right-wing" view referred to in Section 1.1.

In fact, this strong emphasis on the "active" family and how family choices might be distorted by all forms of government intervention, both by direct interventions on educational markets and by pure welfare redistribution, is the main contribution of Gary Becker and his followers to the study of intergenerational mobility. This is what Becker right responded to Goldberger (1989), who expressed some skeptical view about what Becker's contribution really was (as compared to standard mechanical models of intergenerational transmission). Becker (1989) summarized the main negative results about government interventions derived from his models, and explained to Goldberger that such results could not have been derived in a purely mechanical model. It is fair to say that Chicago economists have spent more energy in deriving the laissez-faire implications of their theoretical models rather than in trying to estimate empirically what the distortionary costs of activist policies really are. But one cannot deny that the introduction of utility maximization and active family behavior into the analysis of intergenerational mobility has important policy implications that purely mechanical models do not have.

The other important limitation of the Piketty (1995) model is that it seems to imply that left-wingers should be happy if the genetic component of inequality transmission was very important. That is, if the mechanical component of $\Delta\pi$ (i.e., the component that is beyond the family's control) is very high, then the incentive costs of redistribution are very low. In the extreme case where earnings inequality results entirely from genetic IQ inequality, then one can equalize consumption across dynasties at no incentive cost. However this theory would also imply that there is no hope of doing anything about the inequality of occupations and labor market status, whereas left-wing theories usually stress that such inequalities are due (at least in part) to market inefficiencies that can be corrected (see Sections 4–6 below). Moreover, a strong emphasis on IQ inequality often leads to questioning the ability of low-IQ segments of the population to make sensible choices. For instance, Herrnstein and Murray (1994) argue that one consequence of modernity is that "it has become much more difficult for a person of low cognitive ability to figure out why marriage is a good thing", and they recommend that we impose tough and simple rules on low-IQ individuals.[38] This illustrates how liberal, pro-market right-wing views about social inequality can easily shift to conservative, authoritarian and

fewer kids than poorer parents, they would spend even less time per child than they actually do, and inequality would be less persistent).

[38] "The old bargain from the man's viewpoint—get married, because that's the only way you're going to be able to sleep with the lady—was the kind of incentive that did not require a lot of intellect to process and had an all-powerful effect on behavior" (Herrnstein and Murray, 1994: p. 544).

anti-laissez-faire right-wing views:[39] If one's basic premise is that individual abilities are so unequally distributed that no policy can do anything about it, then one can easily be led to conclude that low-ability individuals have a limited ability to interact in society (including in markets), and that government policies should try to regulate their behavior, possibly in an authoritarian and anti-market manner. More generally, a strong emphasis on IQ inequality might also lead to question the relevance of Rawlsian and welfarist criteria of distributive justice: if the poor are so stupid, then why we should care about their consumption level? However, such non-welfarist arguments have become less and less popular over time, and incentive-based arguments against redistribution are usually produced as well. For instance, Herrnstein and Murray (1994: chapters 17–19) also argue that everybody (including low-IQ taxpayers) would gain if we chose to reward bright and successful children rather than to subsidize hopeless low-IQ neighborhoods and to encourage welfare dependency.

3.3. Taste-based persistent inequality

Sociologists have also been interested for a long time in the family transmission of productive abilities. The "reference group" theory formulated by Merton (1953) and Boudon (1973, 1974) has been particularly influential. The basic idea of the theory is that individuals tend to compare their social achievements to the "reference group" from which they come. As a consequence, agents with lower-class origins are less motivated to make human capital investments and to acquire high productive abilities, since they have less to prove to the outside world and they can easily maintain their initial social position. Conversely, agents with upper-class origins are more motivated and are able to maintain their initial social position. According to this theory, the intergenerational persistence of labor earnings inequality follows from the intergenerational transmission of ambition and taste for economic success. This theory can be formalized in a model where agents care about their "social prestige" or "social status" (defined as the public beliefs about one's ability), abilities are not directly observable, and earnings and labor market achievements act as a signal of one's ability.[40] In such a model, one can show that the status motive tends to amplify the persistence of inequality across generations.[41]

Although this sociological theory is very different from the Becker–Tomes or Herrnstein–Murray theories, the policy implications are fairly similar. Boudon (1973,

[39] See Section 1.1.

[40] It is interesting to note that economists who emphasize the role of private concern for relative status usually stress the rationale for government intervention arising from the status externality, whereas sociologists are mostly interested in the consequence for the intergenerational persistence of inequality (see Piketty, 1998). This probably reflects the fact that most economists are mainly concerned with the optimal size of monetary transfers and redistributive taxation, whereas sociologists are more interested in persistent occupational inequality per se.

[41] See Piketty (1998). More specifically, if agents with upper-class origins are expected to maintain the irinitial position with a very high probability, then the status motive will tend to magnify the inequality of ambition and effort levels and to make inequality more persistent.

1974) argues forcefully that there is nobody to blame for the low educational and economic performance of lower-class kids and the intergenerational persistence of inequality: this is just the unavoidable consequence of the family transmission of ambition. According to Boudon, the reason why for a given educational score at age 10, lower-class children tend to leave school earlier than upper-class children is not because of credit constraints, disadvantaged neighborhood environment or discrimination (see Sections 4–6), but rather because upper-class parents encourage their children not to leave school and reward educational achievements more than lower-class parents do. Boudon concludes that the only possible way to improve somewhat the educational achievements of lower-class kids would be to limit drastically the influence that parents have on their children, for instance by reducing their participation to school boards and class councils. This is not quite as tough to implement as mass adoption programs (see Section 3.1 above), but this means once again that the only possible way to do something about persistent inequality requires a major conflict between the government and the family, and therefore that we might prefer to be modest and accept the world as it is (as Boudon repeatedly suggests). In contrast, left- wing theories argue that market inefficiencies rather than the family are responsible for persistent inequality, and therefore that we do not need to initiate a fight against the family in order to reduce the persistence of inequality (see Section 4–6 below, and especially Section 6.2 on anti-"reference group" sociological theories).

If one is ready to assume that families can transmit their tastes across generations, then one can also construct other, more extreme taste-based theories of persistent inequality. For instance, if different dynasties are characterized by different rates of time preference in the model with dynastic preferences, then the most patient dynasties will become richer and richer, while the least patient dynasties will become poorer and poorer (see Section 2.3). The concern about how consumption and wealth will be distributed in the long-run between dynasties with heterogeneous preferences has a long tradition in economics, and can be found for instance in the writings of Rae, Ramsey and Irving Fisher.[42] Empirical evidence about time discount rates during one's lifetime seems to show that the poor do indeed discount the future at a substantially higher rate that the rich.[43] Assuming that this dynastic heterogeneity in "tastes" does explain a significant fraction of the intergenerational persistence of inequality, the policy implications are far from clear, however. The key question is where the heterogeneity of tastes comes from and whether it can be altered. If heterogeneous behavior and attitudes are due to some "culture of poverty", which is itself the consequence of neighborhood segregation or other socially-inefficient market processes, then activist redistributive policies are called for (see Sections 4–6 below). But if heterogeneous tastes come from direct family transmission and can be altered only at a very high cost, as in the "reference group" theory, then the only thing one can do is to redistribute consumption, the extent of which can

[42] See Mayshar and Benninga (1996) and subsequent references.

[43] See, e.g., Green et al. (1996) and Lawrence (1991) for recent evidence.

be severely limited by incentive considerations. If heterogeneous behavior comes from a dynastic learning process with limited experimentation, then the policy conclusions can be even more anti-redistribution. For instance, if one believes that persistent poverty is due to the fact that poor dynasties underestimate the returns to individual effort, then one may want to implement even less redistribution than would otherwise be the case (or even negative redistribution, from the poor to the rich), so as to induce the poor to experiment with high effort levels and to learn about the true returns to effort.[44] In other words, if poor dynasties are somehow responsible for their wrong behavior, then very little redistribution is called for.

4. Persistent inequality and the imperfect capital market

The simplest market failure theory of persistent inequality is the theory of imperfect credit: if credit markets are imperfect, then dynasties with little initial wealth face limited investment opportunities, and they remain poor. Credit constraints imply that the consequence of intergenerational wealth transfers is not only to make welfare and consumption inequality more persistent than earnings inequality (see Section 2): wealth transfers can also contribute to make earnings differentials more persistent across generations than they would otherwise be. We first briefly review why credit market imperfections might arise and describe the basic implications for the theory of intergenerational mobility (Section 4.1). We then ask the following question: what evidence do we have about the likely importance of credit constraints for intergenerational mobility? (Section 4.2). Finally, a number of theoretical contributions have recently explored some new implications of credit constraints for the dynamics of occupational structure, wealth inequality and intergenerational mobility, and we summarize the main ideas of these theories in Section 4.3.

4.1. Credit constraints vs. first-best credit

Credit or wealth constraints are said to arise whenever the opportunity to invest depends not only on the "technological" viability of the investment (rate of return, risk, ability of the entrepreneur, ...), but also on the initial wealth (or collateral) of the would-be entrepreneur per se. The idea of credit constraints is probably as old as capitalist economies. Although Marx and other nineteenth century socialist theorists do not refer explicitly to the concept of credit constraints, the belief that such constraints are pervasive in capitalist economies implicitly plays a central role in their analysis of the capitalist system. Their basic premise is that initial wealth and capital ownership per se are the key determinants of class reproduction and persistent inequalities on the workplace. This could not happen in a world with first-best credit, where initial wealth per se should be

[44] See Piketty (1995: p. 563, footnote 31).

irrelevant from the viewpoint of productive efficiency and should have no consequence on the distribution of earnings.

It is only recently however that formal theories describing precisely the microeconomic origin of credit constraints have been developed. It is by now well understood that the source of credit constraints is the commitment power of initial wealth: without a sufficient personal stake in the investment project, the would-be entrepreneur has no way no commit that he will reveal the truth to the lender (adverse-selection), nor that he will take the right actions to ensure that the lender will be paid back (moral-hazard).[45] Depending on the exact technological and informational parameters, this will result in equilibrium into some specific credit-rationing curve $k(w, r)$: $k(w, r) \geq w$ is the maximal capital investment a would-be entrepreneur with initial wealth w can undertake when the market interest rate is r, i.e., $k(w, r) - w$ is the maximal credit that lenders accept to offer. In contrast, with first-best credit $k(w, r)$ does not depend on w and is uniquely determined by technological opportunities alone. Note that credit constraints are likely to be particularly severe regarding children's human capital investments, since parents have a limited ability to commit on behalf of their children.

The first consequence of credit constraints is that unequal wealth may prevent some profitable investment from being undertaken. In other words, the inherited distribution of wealth $F(w)$ may not be output-maximizing. By allowing a larger number of able children and entrepreneurs to educate and invest, wealth redistribution can reduce inequality, raise the intergenerational mobility of earnings and increase output at the same time. This was first pointed out by Loury (1981), who introduced credit constraints into a Becker–Tomes (1979)-type model of intergenerational mobility. Limited borrowing ability thus provides the basic justification for redistributive public funding of education.[46] More generally, capital market imperfections imply that the usual results about the long-run efficiency costs of capital income taxation (see Section 2.4) are no longer valid: one needs to compare these efficiency costs not only with the distributive gains, but also with the efficiency gains resulting from previously unfinanced investments.[47] Credit constraints and the commitment value of initial wealth also imply that occu-

[45] See, e.g., Jaffee and Stiglitz (1990) and Bardhan and Bowles' chapter 10 in this Handbook for a survey.

[46] Public intervention in educational markets can obviously be justified by simpler considerations, such as the idea that young children and their ill-informed parents are unable to choose the education they need (for instance, illiterate parents may not be able to fully appreciate what literacy would bring to their children). Although modern economists usually dislike such "paternalistic" concerns and favor market-friendly policy interventions (see below), such concerns do play an important role in the way many people think about intergenerational mobility (see Section 1.1).

[47] Chamley (1996) shows that the efficient, long-run capital income tax rate can be positive in a model with imperfect capital markets. Note that this general result can be the consequence not only of credit market imperfections but also of insurance market imperfections: Aiyagari (1994) shows that imperfect insurance markets imply excessive precautionary savings, in the sense that a lump-sum transfer financed by capital income taxation can be welfare-improving. See Benabou (1997) for a recent attempt to quantify the efficiency gains of redistribution resulting from previously unfinanced profitable educational investments (he concludes that they are roughly comparable to the distortionary costs, and therefore the socially-optimal trade-off leads to "reasonable", interior solutions).

pational choice, i.e., who becomes a wage-earner, who becomes self-employed, etc., is partly determined by the distribution of wealth, even if the latter is unrelated to the distribution of productive abilities (Newman, 1991). Banerjee and Newman (1994) stress that these consequences of limited commitment power are the key economic implications of poverty: poor people have little to loose, and therefore have little credit and career opportunities. This implies that the contractual relationships governing the organization of production that emerge in equilibrium have no reason in general to be output-maximizing.[48] Again, the general implication is that appropriate corrective policies can have both positive distributive effects and positive efficiency effects.

But the explicit microeconomic modeling of credit constraints does not only allow modern economists to rationalize what older generations already knew. It also allows for a more balanced welfare analysis of capital market imperfections. First, the fact that wealth redistribution can be output-improving in the presence of credit constraints does not necessarily imply that wealth redistribution can be Pareto-improving. In general, market equilibria with credit constraints are second-best Pareto-efficient. For instance, it is well-known that sharecropping contracts are privately efficient: no policy can simultaneously raise the productivity of the tenant and the income of the landlord. The only way to raise productivity and output is to redistribute property rights away from the landlord. One cannot simply redistribute the higher output level so as to make everybody better off "after" the efficiency gains have been realized, since this would cancel the positive incentive effects of wealth redistribution, and private contracting could have done the same thing if that was incentive-compatible. This illustrates a more general lesson that can be drawn from microeconomic theory: incentive constraints apply both to private contracting and to activist policies. That is, the same informational and incentive reasons that imply the existence of credit constraints also imply that governments should be cautious before they try to make the credit market more efficient. This simple fact has been dramatically overlooked by radical "remedies" to credit imperfections, such as the abolition of private property, collective ownership or the centralization of credit. In contrast, modern theories of credit market imperfections suggest market-friendly corrective policies, such as a transparent system of educational subsidies or wealth transfers, with limited interference with how actual investments are being made by private individuals. If private individuals are short of cash rather than short of rationality, then governments should try to provide them with the former rather than with the latter.

[48] For instance, Legros and Newman (1995) consider a model where production can be organized either in partnerships, whereby agents with moderate wealth share the investment costs, or in "hierarchical" firms where one rich agent makes the investment and monitors low-wealth wage-earners. They show that hierarchical firms will tend to dominate in equilibrium even though partnerships lead to higher output (since there is no labor wasted in monitoring), simply because wealthy agents use hierarchical firms to extract a larger share of a smaller pie.

4.2. *What do we know about the importance of credit constraints for mobility?*

What empirical evidence do we have about the extent to which credit constraints contribute to make inequality more persistent across generations? First, we do have extensive evidence showing that credit constraints do exist at the micro level. For instance, many empirical studies in developing countries have shown that redistributing the property of the land, or more generally securing the tenure of the land, can raise the incentives and productivity of poor farmers.[49] In developed countries, there is also extensive evidence that for given investment opportunities, firms' investment behavior depends heavily on their cash flows and retained earnings, although first-best credit would predict the opposite.[50] However, although these different pieces of empirical evidence of the micro level are suggestive, they obviously do not allow us to give a precise estimate of how much credit constraints are likely to affect aggregate intergenerational mobility at the macro level.

It is equally difficult to draw strong conclusions from traditional sociological studies about educational achievements and occupations across generations. For instance, the fact that, for given standardized test scores at age 10, lower-class children tend to leave school earlier than upper-class children does not necessarily imply that wealth constraints are binding. It is also consistent with the "reference group" theory of intergenerational mobility (see Section 3.2), or with the existence of some mismeasured endowed ability differential. Sociologists have also shown that for given educational achievements, upper-class children tend to reach higher-status and better-paid occupations than lower-class children.[51] This could be due to the fact that wealth constraints make it more difficult for low-wealth children to translate educational achievements into occupational outcomes. But this is also consistent with a post-school "reference group" theory, or with the fact that educational achievements are very difficult to measure and that the error term is correlated with parental status.

One empirical argument that has been put forward by Gary Becker is that since observed earnings mobility is so high, it must be the case that credit constraints are not very important. Until recently, the few existing studies by economists of the intergenerational correlation of earnings in the US usually found some very low estimates. For instance, Behrman and Taubman (1985: p. 147) estimate an intergenerational correlation of at most 0.2 and conclude: "the members of this sample come from a highly mobile society". Becker and Tomes (1986: p. 269) refer to a couple of similarly low estimates and reach the following conclusion: "The evidence suggests that neither the inheritability of (ability) endowments by sons nor the propensity to invest in children's human capi-

[49] See, e.g., Banerjee and Gathak (1996) for a recent empirical analysis of the productivity effects of land reform in West Bengal.

[50] See, e.g., Gilchrist and Himmelberg (1995) and Lamont (1997).

[51] See Goux and Maurin (1996) for recent evidence. In particular, Goux and Maurin show that, for given educational achievements, the effect of parental status on children is very strong all along one's occupational career (even more so that at the entry level).

tal because of capital constraints is large".[52] Becker's 1988 presidential address to the American Economic Association similarly concluded: "In every country with data that I have seen (...) low earnings as well as high earnings are not strongly transmitted from fathers to sons. (...) Evidently, abilities and other endowments that generate earnings are only weakly transmitted from parents to children" (Becker, 1988: p. 10).

However, these very low estimates of the intergenerational earnings correlation have been rejected by the more recent and reliable literature. Solon (1992) and Zimmerman (1992) have convincingly argued that previous estimates have been biased downwards by unrepresentative samples and measurement errors. The most important source of downward bias in previous studies derives from the use of single-year or short-run measures of earnings. The existence of large, short-run variations in earnings makes it impossible to estimate properly the true intergenerational correlation of life-time earnings based on such short-term measures. Solon and Zimmerman use better data sets than previous studies, correct for measurement errors by using multi-year income averages, and both estimate intergenerational earnings correlation coefficients in the 0.4–0.5 range. Mulligan (1997) further refines the Solon-Zimmerman approach to measurement errors and concludes that the correct estimate is likely to be at least equal to 0.5. Dearden et al. (1997) use a similar methodology with British data and also find an intergenerational earnings correlation in the 0.5–0.6 range. One must bear in mind that whether the intergenerational correlation is 0.2 or 0.5 has enormous consequences for actual mobility rates. If the intergeneration earnings correlation was equal to 0.2, as argued by Gary Becker and pre-Solon–Zimmerman estimates, this would mean that if parents are five times richer, then children will be on average less than 40% richer, and grand-children less than 7% richer. But if the correlation is equal to 0.5, as more recent and reliable studies seem to suggest, this means that if parents are five times richer, children will be more than 2.2 times richer, and the grand-children 50% richer.[53] To put it another way, a son whose father's status is in the fifth percentile has a 37% chance to rise above the median if the intergenerational correlation is 0.2, and a 17% chance to rise above the median if the intergenerational correlation is 0.5.[54]

In fact, some authors have pointed out a long time ago that simple raw estimates of intergenerational earnings correlation suffered from serious downward biases (see, e.g., Bowles (1972)). In the early 1980s, Atkinson (1981) and Atkinson et al. (1983) had already tried to correct for measurement errors and had found an intergenerational earnings correlation of 0.45 with British data. Becker's (1988) faith in very low estimates probably reflects when Bjorklund and Jantti (1998) describe about their US colleagues' faith in US exceptionalism (see Section 1.1 above).

Needless to say, one cannot conclude from the fact that intergenerational earnings correlation is pretty high that credit constraints are important. Low mobility might just

[52] The page number refers to the Becker (1991) reprint version of the Becker and Tomes (1986) article.

[53] $5^{0.2} = 1.38$ and $1.38^{0.2} = 1.066$, while $5^{0.5} = 2.23$ and $2.23^{0.5} = 1.495$ (see Section 2.1).

[54] See Solon (1992: p. 404). Note that this is assuming bivariate normally, which may overestimate the mobility chances of dynasties at the bottom and at the top of the distribution.

result from an efficient process of ability transmission in families and on human capital markets (see Section 3). For instance, Mulligan (1997) finds that the intergenerational correlation of earnings is at least equal to 0.5, but he agrees with Gary Becker about the fact that credit constraints must be unimportant in the real world. Mulligan's empirical argument is more sophisticated than that of Becker, however. Using PSID data, Mulligan (1997: Chap. 8) compares the intergenerational correlation of earnings and consumption of children who have received financial transfers from their parents at age 30 with that of children who did not receive such transfers, and finds that correlation coefficients are not significantly different between the two groups.[55] To the extent that the second group is more likely than the first group to suffer from wealth constraints, this can be taken as evidence that wealth constraints are not very important for mobility. Mulligan further concludes that since credit constraints are unimportant, they cannot possibly explain why consumption regresses to the mean across generations, and therefore that his model of endogenous altruism is the only model of intergenerational mobility that can simultaneously explain all the observed facts (see Section 2.3 above). Given that the available information used by Mulligan to identify credit-constrained dynasties can hardly have been viewed as satisfactory, such a strong negative conclusion about the importance of credit constraints seems premature. But Mulligan's empirical strategy is promising and clearly illustrates what the empirical work of the future should look like: the extensive use of richer and richer panel data sets should allow us to make progress on such issues.

Finally, note that it is by no mean impossible that the importance of credit constraints for intergenerational mobility does vary enormously over time and across countries. For instance, the historical study by Kaelble (1986) argues that the major change that occurred in the history of social mobility since the industrial revolution is the shift from the middle-size family firm to the large corporation. According to Kaelble, the consequence of the transition to "corporate capitalism" was that capital became less and less a precondition for the business career, which led to a slow decline of the business family and the emergence of a large class of non-owner business executives and of associated upwardly-mobile careers. Kaelble stresses the fact that this transition was very slow: he finds that the proportion of fathers of the business elite who were themselves businessmen was high and rising in all industrialized countries until the interwar years. Kaelble concludes that the initial effect of the industrial revolution on social mobility was probably a negative one, because of the transmission of property and land and crucial role of access to capital in the new world, and that the history of social mobility since the industrial revolution should be seen as a crisis and a subsequent response, rather than as self-sustained growth of mobility rates. From a completely different perspective, Herrnstein and Murray (1994: Chap. 1) also argue that capital barriers have become less and less important over time: they display some graphical evidence showing that IQ has

[55] Mulligan also uses information about the expectation of receiving such transfers.

progressively become more important than social origins per se in order to be admitted in top universities in the US over the course of the twentieth century.

4.3. Poverty traps vs. low-mobility traps

The simplest theoretical implication of credit constraints is the existence of poverty traps: dynasties with little initial wealth can remain poor forever. The following model, which is a slightly simplified version of the model of Galor and Zeira (1993), illustrates how it works.[56] Assume linear savings and a very extreme form of moral-hazard-induced credit-rationing: borrowers can always "take the money and run" at no cost, so that in effect the credit market completely collapses ($k(w, r) = w$). Further assume that each generation can either earn a subsistence income y or make a fixed investment I that yields a net return RI, with $RI > y$. Galor and Zeira (1993) choose to interpret the fixed investment I as a human capital investment, but this is obviously inessential. Credit constraints imply that at each period t, all agents whose initial wealth w_t is smaller than I earn y, while agents with $w_t \geq I$ earn RI, so that transitional equations can be written:

$$\text{If } w_{it} < I,\ w_{it+1} = (1-\delta)w_{it} + sy, \tag{4.1}$$

$$\text{If } w_{it} > I,\ w_{it+1} = (1-\delta)w_{it} + sRI. \tag{4.2}$$

If we assume the savings rate s to be small enough so that $sy + (1-\delta I) < I$ and the rate of return R to be high enough so that $sRI + (1-\delta)I > I$, then we have a poverty trap: poor dynasties starting with $w_0 < I$ earn a low income y and remain poor forever ($w_t \to w^0 = sy/\delta < I$), while rich dynasties starting with w_0 earn a high-income RI and remain rich ($w_t \to w^1 = sRI/\delta > I$). That is, if the initial distribution of wealth $F_0(w)$ is characterized by a mass $F_0(I)$ of poor dynasties and a mass $1 - F_0(I)$ of rich dynasties, then so will be the long-run distribution $F_\infty(w)$: initial wealth inequality persists in the long-run. This persistence would immediately disappear with first-best credit: everybody would invest I irrespective of one's initial wealth, and all dynasties would converge to the same wealth level, for any initial wealth distribution. This shows that the assumption of fixed costs or increasing returns is not sufficient to make the initial distribution relevant if it is not supplemented with the assumption of credit constraints. Conversely, poverty traps rely on a threshold effect and a technological nonconvexity and would not arise with credit constraints alone. Without the assumption of a fixed-size investment, poor dynasties could slowly accumulate by starting with small investment levels and eventually catch up with the rich. It is the combination of nonconvex technologies and credit constraints that produce nonconvexities in transition equations and the possibility of poverty traps. In effect, this combination gives rise to a dynamic model that is very similar to the Bourguignon (1981) model with a nonconvex savings function $S(y)$ described in Section 2.2. above.

[56] Freeman (1996) also offers a model of persistent inequality based upon borrowing constraints and a poverty trap.

The recent literature has explored more sophisticated dynamic implications of credit constraints. One important finding of the recent literature is that with credit constraints we actually do not need nonconvexities and threshold effects to conclude that credit constraints can have important long-run effects. Consider a model where agents can invest at any level according to a concave production function $f(k)$, but where moral-hazard in entrepreneurial effort leads to a credit-rationing curve $k(w, r)$ (see Section 4.1 above). Under natural assumptions, one can show that credit constraints become more and more binding as the market interest rate r goes up $(dk(w, r)/dr < 0)$.[57] Risks from investment are imperfectly insurable (because of moral-hazard), so that individual transitions $w_{it+1}(w_{it})$ are stochastic. With suitable concavity assumptions, one can ensure that individual transitions $w_{it+1}(w_{it})$ exhibit no threshold effect, i.e., that all dynasties can switch between any two wealth levels in a finite time with positive probability. If we assume that the market interest rate r is exogenously fixed, then this ergodicity property is sufficient to ensure global convergence, i.e., the fact that the long-run distribution $F_{\propto r}(w)$ does not depend on the initial distribution $F_0(w)$.

However things are different when the interest rate is endogenously determined by the supply and demand of capital. Note first that with credit constraints the equilibrium interest rate is not simply given by "the" marginal product of capital, since the latter varies across production units. In other words the equilibrium interest rate r_t now depends on the entire wealth distribution $F_t(w)$ at period t. One can then show that depending on the exact initial distribution $F_0(w)$ there will exist different possible long-run distributions $F_{\propto 1}(w)$, $F_{\propto 2}(w)$, associated with different long-run interest rates $r_{\propto 1}$, $r_{\propto 2}$, ...(see Piketty, 1997). The intuition is the following: initial distributions with a large population of low-wealth agents lead to a high demand for capital and to high interest rates, which in turn imply that it takes a long time for low-wealth agents to accumulate and rebuild their collateral, so that the initially large mass of poor agents is self-reproducing. Conversely, low initial interest rates lead to high wealth mobility, high accumulation and low equilibrium interest rates. Such a multiplicity will arise whenever an interest rate rise strengthens credit constraints more than it strengthens the accumulation of the rich, i.e., whenever $|dk(w, r)/dr|$ is large enough. The steady-states with higher interest rates have at the same time less wealth mobility and a lower aggregate output and capital stock. One key difference between this type of "low-mobility trap" and the poverty trap described earlier is that the latter can be eliminated once and for all by pushing all poor agents above the threshold, whereas the former is more perverse and requires continuous downward pressures on the interest rate (through fiscal or credit policy) in order to shift the economy to a lower interest rate, higher mobility development path.

This phenomenon of low-mobility traps is actually very general, and it has first been pointed out by Banerjee and Newman (1993) in a context that is slightly different

[57] For an endogenous derivation of such a curve, see Piketty (1997), whose moral-hazard credit model is an extension of that of Aghion and Bolton (1997).

from the Piketty (1997) model that we just described. Banerjee and Newman consider a dynamic accumulation/distribution model with a fixed exogenous interest rate r, but with an endogenous wage rate v_t playing a role that is similar to the endogenous interest rate in the previous discussion. In their model, the wage rate is the equilibrium market price of monitored labor. They consider a world where moral-hazard-induced credit constraints prevent poor agents from investing in large projects but where rich agents can use a technology to monitor poor agents working as wage earners. That is, unlike in the previous model where everybody was an entrepreneur, there are three possible occupations in their model: wage earners (who are too poor to make any investment on their own), self-employed (who finance and run their own investment) and entrepreneurs (who finance large investments and monitor wage earners). The equilibrium wage rate v_t is determined by the equality between the number of agents "choosing" to become wage-earners and the number of wage-earners required by entrepreneurs, and thus depends on the entire wealth distribution $F_t(w)$. One can easily see how this can generate long-run effects of the initial wealth distribution: an initially large mass of poor agents with no other option than becoming a wage-earner leads to a low wage rate and little upward mobility for wage earners, while an initially small mass of poor agents leads to high wage rates and high mobility between wage-earners and self-employed, which reproduces the forces leading to high wage rates. Depending on the initial distribution $F_0(w)$, the economy will therefore converge to different possible long-run distributions $F_{\alpha 1}(w)$, $F_{\alpha 2}(w)$, ... associated with different long-run wage rates $v_{\alpha 1}$, $v_{\alpha 2}$, ... Although the original Banerjee and Newman (1993) did assume a fixed cost technology (so as to simplify transitional dynamics), the Piketty (1997) model described above clearly shows that their central result would also hold with a standard, concave technology. If both models were combined, i.e., if both the interest rate and the wage rate depend on the wealth distribution, then the general conclusion would be that both long-run factor prices can depend on the initial wealth distribution. Note that this stands in great contrast with models based upon first-best credit, where equilibrium factor prices do not depend at all on the distribution of wealth.

Of course, whether this two-way interaction between the wealth distribution and equilibrium factor prices can be sufficiently strong in practice to generate such long-term effects depends on the empirical magnitude of credit constraints. Banerjee and Newman (1993) point out that historical evidence seems to suggest that this is plausible. Several historians have argued that the two different initial distributions of land in France and in Britain in the early 1800s in the aftermath of the French Revolution did generate persistently divergent development trajectories: the large population of British landless peasants pushed industrial wages down and fostered early industrial development, while the large population of small French landowners delayed the industrial revolution and had long-run implications for French economic development.[58] Such long-run effects

[58] See Banerjee and Newman (1993: p. 292) and subsequent preferences. Note that in the context of the Banerjee–Newman model, the UK trajectory would appear as a low wage, low output, "industrial trap". This controversial welfare interpretation can easily be modified by introducing learning-by-doing-type externalities

of the initial wealth distribution on mobility and development would be impossible to explain in a world of first-best capital markets, but can be accounted for by a Banerjee–Newman-type model.

This two-way interaction between the distribution of wealth and equilibrium factor prices implied by credit constraints can also generate other interesting and empirically plausible development patterns. For instance, Aghion and Bolton (1997) show that this interaction can generate trajectories characterized by a declining price of capital and an endogenous Kuznets curve. During the initial stage of development, little capital is available, the equilibrium interest rate is high and strong credit constraints imply that only the rich can invest and wealth, mobility is low and income inequalities tend to widen. The capital accumulation of the rich progressively forces the interest rate to drop, so that credit constraints become less binding, mobility rises and inequality begins to decline.

5. Persistent inequality and local segregation

The importance of local segregation into unequal communities for understanding intergenerational mobility has long been emphasized by sociologists.[59] Formal economic models of equilibrium segregation into unequal neighborhoods have been developed more recently, however. These models are important because they show under what conditions local segregation can be socially inefficient, which is the key question from a policy perspective. We first review the main contributions of these recent theoretical models (Section 5.1). We then analyze their empirical and policy implications (Section 5.2). Finally, we discuss the role of other levels of local segregation (Section 5.3).

5.1. Models of inefficient segregation into unequal neighborhoods

Consider first the following model due to Benabou (1993). Agents must choose to live in one of two spatially distinct neighborhoods and whether to obtain a low education (cost C_L) or a high education (cost C_H). These costs $C_L(x)$ and $C_H(x)$ depend negatively upon the fraction x of one's neighbors choosing to obtain a high education, reflecting the positive external effects of education on one's neighbors (in the classroom, as a role model, ...). Whether or not it is socially optimal to get all agents choosing a high education to live in the same neighborhood depends on the slope of the total educational cost function $C(x)$ given by the following equation:

$$C(x) = xC_H(x) + (1 - x)C_L(x). \tag{5.1}$$

in the large-scale, industrial sector, so that productivity and wages are eventually higher in the industrial development path.

[59] The scientific study of ghettos and residential segregation in the US sociological tradition dates back to the Chicago school of sociology in the interwar period, up to the more recent works of William Julius Wilson (see, e.g., Wilson, 1987).

If $C(x)$ is convex, then for any given optimal number of high-education agents, it is less costly to divide them equally between the 2 neighborhoods. Conversely, segregation is optimal if $C(x)$ is concave.

The key point is that whether segregation or integration will prevail in laissez-faire equilibrium depends on a different condition. Benabou shows that the condition for equilibrium segregation is the following:

$$C'_H(x) < C'_L(x). \tag{5.2}$$

If condition (5.2) holds, i.e., if the marginal private benefits of having more educated neighbors are higher if one chooses high education, then integration is inherently unstable and market forces push towards stable segregation. The intuition is that if the two neighborhoods have initially a marginally different composition, this condition implies that high-education agents are ready to pay a marginally higher rent to live in the better neighborhood, which leads to more segregation, and so on. The reason why the conditions for social optimality and decentralized equilibrium are different is that high-education agents only take into account their marginal private benefits of moving to a better neighborhood and do not internalize the marginal costs they impose on their initial neighborhood by diminishing the fraction of high-education agents. The failure of the price system is that the housing market does not charge the true social costs of moving: market rents are the same for everybody, whereas a socially-optimal price system should charge a higher rent to high-education movers to a high-education area (or, alternatively, a lower rent to high-education movers to a low-education area). These two conditions also highlight which parameter configurations typically lead to inefficient segregation: if $C'_H(x) < C'_L(x)$ but both slopes are very close, then $C(x)$ will be convex if $C_H(x)$ and $C_L(x)$ are convex, i.e., if the benefits of living with educated people exhibit decreasing returns. Conversely if these returns are increasing segregation will be socially optimal, and so will be the decentralized equilibrium.[60] Note that unlike in the case of credit constraints (see Section 4.1 above), local externalities can make market equilibria inefficient in the Pareto sense: corrective policies cannot only raise total output but also raise everybody's welfare.

In the original Benabou (1993) model, all agents are ex ante equally endowed in human capital, which allows us to identify in a very transparent way the conditions for inefficient segregation. In a dynamic world however, human capital inequality and segregation reinforce each other over time, and segregation leads to lower intergenerational mobility than would otherwise be the case. The model and the conditions for inefficient segregation described above can easily be extended to such a setting. Moreover, Benabou (1996a) has also shown that even if it is less costly in the short-run to have segregated neighborhoods in order to produce more human capital ($C(x)$ concave), segregation may

[60] Assume for instance that $C_L(x) = (1-a)C_H(x) - c$, with a sufficiently close to 0. Then $C''(x) = C''_H(x) - ad^2((1-x)C_H(x))/dx^2$, which is arbitrarily close to $C''_H(x)$.

not be efficient in the long-run because it tends to amplify future human capital inequality, which can be harmful for total output. Under these conditions there can be a trade-off between minimizing the short-run costs of existing inequality through segregation and minimizing the long-run costs of inequality through mixing and integration. Again, the theoretical model allows us to identify the exact conditions that need to be empirically estimated: the values of the complementarity parameters of the local interaction process and of the global production function determine whether segregation is inefficient from the viewpoint of long-run growth.

In the Benabou model, the forces pushing towards segregation or integration are the pure forces of local externalities (peer effects) and the housing market. This framework can be extended in several directions. First, segregation could be supported by other institutions than a competitive housing market, such as the possibility for local communities to enact zoning regulations, so as to restrict access to their neighborhood to agents meeting specific criteria (income, age, landowner/tenant status, ...), which will in general exacerbate segregation (Durlauf, 1996; Fernandez and Rogerson, 1996). The effects on efficiency are unclear however, since such institutions might also allow communities to internalize the relevant local externalities.

Next, individual motives for segregation can be more complex than direct peer effects. If each community decides how much fiscal revenue to allocate to schools and the quality of schooling depends on the level of educational spendings, this creates an incentive to locate in a wealthy neighborhood, even if there is no direct peer effect at the neighborhood level. The local external effects $C_L(x)$ and $C_H(x)$ of the Benabou model can be interpreted as a reduced form of this "fiscal channel". Explicit models of this "fiscal channel" for local segregation can be found in Fernandez and Rogerson (1994) and Benabou (1996b). Whether segregation will take place and whether it is efficient then depend on the shape of the marginal benefits of having better-funded schools, just as in the Benabou model. For the same reasons as in the Benabou model, there is no reason to suspect that housing prices will lead to an efficient level of segregation. For instance, Fernandez and Rogerson (1994) estimate that the positive output efficiency effects of a switch from local educational finance to federal, redistributive educational finance would be substantial in the US.

Note also that the forces behind inefficient segregation always tend to be magnified by imperfect capital markets. For instance, if one adds to the Benabou model that agents are initially unequal and face credit constraints, then poor agents might be unable to move to a better neighborhood even if $C'_H(x) > C'_L(x)$, i.e., even if their marginal benefits of moving are higher (see Benabou, 1996b).

5.2. *The policy implications of local segregation*

The fact that intergenerational mobility depends not only on parental characteristics but also on the composition of the local neighborhood is well documented (see, e.g., Borjas (1992, 1995) on the effect of ethnic residential segregation). However the fact that segre-

gation matters does not necessarily imply that segregation is socially inefficient. If $C(x)$ was concave (see Eq. (5.1) above), then it would be socially inefficient to try to raise intergenerational mobility by forcing unequal dynasties to live together in homogeneous neighborhoods. If $C(x)$ was concave, it would be more cost-effective to spend available resources to help educate credit-constrained children (Section 4), to fight discrimination (Section 6), or simply to redistribute consumption and welfare if one believes that credit constraints and discrimination are unimportant and markets are efficient (Sections 2 and 3).

It is very difficult however to measure empirically whether the conditions for inefficient segregation identified by the theoretical models are met in practice. One of the main difficulties is the fact that measured neighborhood effects may reflect a spurious correlation, induced by the possibility that the same factors which lead to particular location choices also lead to particular socioeconomic outcomes. Cutler and Glaeser (1997) have developed an ingenious empirical methodology in order to correct for these biases in the formation of racial ghettos, and they find that a one standard deviation decrease in segregation would eliminate one-third of the black-white differential in schooling and employment outcomes.[61] Although they do not compare these costs of segregation with the benefits of segregation enjoyed by well-off neighborhoods, their findings are suggestive.

A very skeptical empirical argument about local segregation has recently been developed by Kremer (1997). Kremer estimates with PSID data that a child's educational attainment can be expressed as 0.39 times the educational attainment of the child's parents, plus 0.15 times the average educational attainment in the census tract in which the child grew up, plus an error term with a standard deviation of 1.79 years of schooling.[62] Kremer concludes that moving from no educational segregation to complete educational segregation would increase the steady-state standard deviation of education by only 9% (from 1.95 years to 2.13 years),[63] and therefore that the magnitudes of the effects are just too small to justify much public concern about residential segregation. However, whether 9% of the steady-state standard deviation should be viewed as small or large is a matter of perspective: Kremer's estimate also imply if the parental persistence parameter was equal to 0 (instead of 0.39), then the steady-state standard deviation of education would decline by only about 8% (from 1.95 to 1.79), although a persistence parameter of 0.39 does mean substantial persistence of inequality across generations (see Section 4.2 above). To put it another way, Kremer's findings also indicate that moving from no segregation to complete segregation would increase the intergenerational persistent

[61] Cutler and Glaeser exploit variations across US cities in a number of exogenous factors that are likely to have an impact on the probability of having ghettos, such as the number of rivers and naturally divided neighborhoods.

[62] Mulligan (1993) estimates (also with PSID) data that about 10% of the observed intergenerations correlation of earnings can be attributed to residential segregation. Note however that he uses county-level averages to identify neighborhood effects, which may severely underestimate the true effects of finer local neighborhoods.

[63] $1.95 = 1.79/(1 - 0.39^2)^{1/2}$, and $2.13 = 1.79/(1 - 0.54^2)^{1/2}$. See Kremer (1997: p. 116).

parameter by about 40% (from 0.39 to 0.54), which most observers would view as a substantial effect. Moreover, linear estimates tend to underestimate the effects on mobility at the bottom and at the top of the distribution. Cooper et al. (1994) also use PSID data, but their methodology allows them to find that neighborhoods effects on mobility are highly nonlinear, and that for a given parental income group, intergenerational income correlations can vary by a factor of two depending on the average income of the parents' neighborhood. Finally, note that the key question raised by the theoretical models is not whether neighborhood effects on mobility are small or large, but whether they are larger for more disadvantaged children than for less disadvantaged children, i.e., whether $C(x)$ is convex or concave. The point is that if the conditions for inefficient segregation apply, then it is possible to raise output and intergenerational mobility at the same time, which would look like an interesting thing to do, even if the orders of magnitude were not enormous. From that viewpoint, the Cooper et al. (1994) findings about the strong nonlinearity of neighborhood effects would seem to indicate that Benabou's conditions for inefficient segregation are likely to be satisfied. If marginal changes in neighborhood composition produce large effects on the mobility prospects of kids at the very bottom of the distribution and moderate effects on the mobility prospects of middle-class kids, then integrated neighborhoods might well be socially efficient.

The theoretical models also raise the question of whether the peer effect channel or the fiscal channel of local interaction is most important. If residential segregation matters mostly because of its effect on local funding for education, then one does not need to force neighborhoods to be socially integrated in order to correct the negative effects of segregation: it is sufficient to redistribute educational resources across neighborhoods, for instance through a uniform national system of educational finance. However if local externalities operate mostly through the peer effect channel, i.e., via direct interaction between children who are in the same school, role models, etc.,[64] rather than educational finance per se, then more radical policies are necessary: one needs to intervene directly on the housing market, e.g., by subsidizing low-rent housing in wealthy neighborhoods, and/or to force children coming from unequal neighborhoods to go to the same school, e.g., via busing policies.[65] Such radical policies are very difficult to implement, because of their strong interference with what most parents consider as their purely private choices. This explains why low estimates of the effects of redistributing educational

[64] Roemer and Wets (1994) have recently proposed endogenizing peer effects in informational terms. They assume that agents are uncertain about the shape of the convex relationship between the level of human capital investment and the resulting market income, and that they learn about this relationship through linear extrapolation of the (human capital investment, market income) vector that they observe in their social neighborhood. They show that this can generate persistent inequality in human capital investment and market income between otherwise identical neighborhoods, which provides yet another example of a "self-fulfilling" theory of inequality (see Section 6).

[65] This opposition between educational finance and housing/busing policies should not be overestimated, however. In practice, some degree of local responsibility over educational finance can be beneficial for other reasons, so that even if local interaction operates only through the fiscal channel, it might be socially efficient to force neighborhoods to be somewhat integrated.

finance across neighborhoods have usually been interpreted as negative results from a policy perspective.[66] For instance, the conclusion of the Coleman (1966) report, who argued that financial transfers to the schools of disadvantaged neighborhoods had little effect on educational performance, was that there is not much to do about local segregation and persistent inequality.[67] However such results could also be interpreted as the proof that more radical housing and busing policies are necessary.

5.3. Other levels of segregation

Residential segregation is not the only level of segregation that can have tremendous consequences for intergenerational mobility. Other potentially important levels of local interaction include the family and the firm.

At the level of the family, it is obvious why positive assortative mating can contribute to make inequality more persistent across generations: if children's abilities depend on the characteristics of both parents, then the fact that men and women with similar characteristics tend to mate together makes intergenerational mobility lower than it would be under random matching. Kremer (1995) argues that a cumulative mechanism might exist along similar lines: if higher human capital inequality increases the incentives to marry with someone of similar human capital level, then higher human capital inequality between parents leads to higher human capital inequality between children, and so on. Kremer illustrates the relevance of these cumulative dynamics by contrasting the US case with that of Brazil.[68] However the key difference between residential segregation and assortative mating is that the housing price system is unable to internalize all relevant external effects (see the Benabou model in Section 5.1 above), whereas potential partners should in principle be able to internalize the effects of assortative mating on their children. Gary Becker has repeatedly argued that positive assortative mating is likely to be an efficient market outcome (see especially Becker, 1991: Chap. 4). Things can be different in models where marriage patterns result from private concern about relative status and some self-enforced social norm.[69]

If human capital acquisition is influenced by one's coworkers (just as by one's neighbors and one's parents), then skill segregation at the firm level can also contribute to make inequality more persistent. Kremer and Maskin (1995) have argued that higher human capital inequality can increase the incentives of high-skill workers to break away

[66] This is again another example of the "do nothing or hit the family" dilemma (see Section 3.2).

[67] Coleman's empirical results have however been challenged by more recent estimates using post-school wages rather than standardized tests (see, e.g., Card and Krueger, 1992).

[68] Whether marital sorting has recently increased in the US is controversial. Kremer (1997) finds the correlation between spouses' education has declined somewhat. However Meyer (1995) estimates that almost half of the total increase of US household income inequality during the past 20 years can be accounted for by the rise in the correlation between spouses' earnings (she also finds that almost 40% of the rise of this correlation can be attributed to the rise of divorce: marital sorting is on average higher for second marriages than for first marriages, probably because more information about potential partners' permanent attributes is available at the age of second marriages).

[69] See Cole et al. (1992) for such a model.

from low-skill workers and to work together. For instance, assume that there two possible human capital levels in the population, h_1 and h_2, with $h_1 > h_2$. If production requires two workers (one "manager" and one "assistant") and output is given by $Y = h_A h_M^2$ (where h_A is the assistant's human capital and h_M is the manager's human capital), then one can show that it will be efficient for high-human-capital agents to work together if and only if the human capital ratio h_1/h_2 is larger than some threshold $\lambda > 1$. Similar intuitions can be obtained with more general production functions (see Kremer and Maskin, 1995). Kremer and Maskin show that this process might be relevant to account for the recent evolution of wage inequality in western countries: they find that in almost every production sector the variance of the distribution of firm-level mean wages has increased much more rapidly than the mean variance of the firm-level distribution of wages. In the Kremer–Maskin model, equilibrium skill segregation between firms is efficient: forcing firms to be more integrated would diminish total output, and it would again be more efficient to have a purely redistributive tax on earnings and to let the market do its job. However, this need not be the case in general. In case the initial productivity of lower-skill workers in high-skill firms is very low, then wealth constraints might prevent lower-skill workers from joining such firms, even if the long-run productivity effects of interacting with high-skill workers were higher for them than for higher-skill workers. This illustrates once again the importance of distinguishing between local segregation as a general channel of inequality transmission and local segregation as a source of inefficient persistent inequality.

6. Persistent inequality and self-fulfilling beliefs

Can self-fulfilling beliefs alone generate persistent inequality across generations? One answer is given by the well-known model of statistical discrimination (Phelps, 1968; Arrow, 1973). We first review the basic mechanism and policy implications of the theory of discrimination (Section 6.1). A number of sociologists have also been interested in phenomena of self-fulfilling inequality, and we will briefly review these theories in Section 6.2.

6.1. The theory of discrimination

Assume that two social groups (say, the blacks and the whites) have the same distribution $G(c)$ of private costs c to become a qualified worker. These costs can measure the heterogeneity of tastes with respect to human capital investment, as well as the heterogeneity of investment potentials and abilities. If employers could perfectly observe whether a given would-be employee has made the investment or not, then there would exist a unique threshold c^* below which individuals would choose to invest. There would be no systematic inequality between the two social groups. Assume however that employers only observe a noisy signal θ of workers' qualification. Then under appropriate assumptions

there exists a discriminatory hiring policy (θ_B, θ_W) which is self-fulfilling. This can be an equilibrium because if employers are expected to promote to qualified tasks black workers with a $\theta \geq \theta_B$ and white workers with a $\theta \geq \theta_W < \theta_B$, then this discourages black workers and induces them to become qualified less often than the whites (the threshold cost c_B is lower than c_W). This in turn validates employers' discriminatory priors. That is, persistent intergenerational inequality between two social groups with homogenous characteristics has been generated solely out of self-fulfilling beliefs. More generally, self-fulfilling discriminatory beliefs can make the inequality between social groups with initially unequal characteristics more persistent than it would otherwise be.

Although the development of this theory of statistical discrimination was primarily inspired by racial discrimination in the US, one can apply this same logic to other observationally distinguishable groups than blacks and whites. For instance, Acemoglu (1995) shows in a model where employers imperfectly observe whether unemployed have paid the cost to recover their skills that an equilibrium where unemployed do not incur this cost and are discriminated by employers can be supported by self-fulfilling beliefs. He then shows how this can justify policies of positive discrimination towards long-term unemployed, although such policies would seem inefficient in a model where the latter are simply less productive. More generally, persistent inequality through self-fulfilling beliefs can be generated between men and women, upper-caste and lower-caste dynasties, high-wealth and low-wealth dynasties, etc. Wilson (1987) has argued that residential segregation tends to increase labor market discrimination. The idea is that if employers can more easily associate a particular set of would-be employees to a specific, disadvantaged neighborhood, then self-fulfilling discriminatory beliefs can more easily develop. Since credit constraints also tend to make residential segregation more likely (see Section 5.1), this means that credit imperfections, local segregation and discrimination can operate together and lead to a cumulative process of socially-inefficient persistent inequality.

Assume that persistent intergenerational inequality is due to discrimination, at least in part. What would socially-efficient redistribution look like? First, note that statistical discrimination is grossly inefficient: first-best efficiency would require all individuals with similar characteristics to make the same investments. That is, corrective policies could simultaneously raise output and make inequality less persistent, just as in the case of credit constraints and residential segregation (see Sections 4 and 5). The ideal corrective policy would be to force employers to use the same testing requirements $\theta_B = \theta_W$ for every social group. This would immediately put an end to self-fulfilling discriminatory beliefs. The problem is that it might be difficult to observe the threshold θ which is applied by employers. This issue of optimal second-best anti-discrimination policies has recently been addressed by Coate and Loury (1993). Coate and Loury argue that direct anti-discrimination policies are in general not enforceable, and that in practice affirmative action policies look much more like quotas: employers must end up with the same distribution of black and white workers in their qualified and unqualified tasks. Coate and Loury then distinguish between two cases. First, they show

that if initial discrimination is not complete, in the sense that a positive fraction of black workers ends up in qualified tasks, then quota-type policies are generally dominated by a policy of state-financed income subsidy to black workers promoted to qualified tasks, which would gradually eliminate discrimination. The intuition is that quotas can lead to "patronizing" hiring policies whereby employers reduce their standards θ_B so much in order to meet the quota that black workers have even less incentives to become qualified than in the previous situation. However, if we start from a situation where θ_B is so high that no black worker is allocated to qualified tasks (complete discrimination), then quota-type affirmative action is the only way to make progress. In any case, note that the optimal policy tool (race-specific income subsidies to promotion, or quotas) would be difficult to justify without a model of persistent inequality based upon self-fulfilling beliefs. Coate and Loury's analysis of corrective policies in the discrimination model also illustrates what may be the most important contribution of formal economic modeling to our understanding of inequality and redistribution. In the same way as in the case of credit constraints (see Section 4.1), formal economic modeling of discrimination makes transparent the fact that the informational imperfections that generate market failures and inefficient inequalities also apply to government and public policies, thereby making government intervention more difficult than it would otherwise be.

Whether unconventional policy tools such as quotas and other affirmative-action policies are appropriate depends however on whether one believes that statistical discrimination is an important source of inequality. For instance, Friedman (1962) argues that to the extent that discrimination does exist, i.e., to the extent that persistent earnings inequality does not simply derive from the "normal" family transmission of unequal abilities, discrimination is due primarily to consumers' tastes rather than to employers' discriminatory behavior. Friedman concludes that there is not much to do about discrimination per se: redistribution should rather take the form of a transparent redistributive tax on labor earnings, and the rates of such a negative income tax should be relatively low, so as to minimize the incentive costs of redistribution. Herrnstein and Murray (1994: Chaps. 19–20) argue that persistent racial inequality can easily be accounted for by the transmission of unequal cognitive abilities across generations, with no need for a theory of discrimination. They also argue that the average cognitive ability of blacks in top universities and top occupations is now lower than that of their white counterparts, which shows how inefficient affirmative action policies in higher education and on the workplace have been. Most labor economists seem to have a more balanced view of the empirical relevance of employers' discrimination. According to Freeman (1973, 1981), a large part of the narrowing of the black/white earnings gap since 1964 can be attributed to the Civil Rights movement and the development of affirmative action policies. Neal and Johnson (1996) also find evidence of current labor market discrimination, although they argue that the most important part of racial inequality in the US can be explained by a skill gap that can be measured by test scores at age 16, i.e., before labor market discrimination. It is true however that it is extremely difficult to distinguish empirically the

effects of discriminatory beliefs from the many other channels through which unequal social and cultural attitudes can be transmitted across generations.[70]

6.2. Sociologists' theories of self-fulfilling inequality

Unlike economists, whose formal models of "self-fulfilling inequality" are to a large extent unconventional and do not reflect the way mainstream economists usually think about labor markets and earnings inequality, mainstream sociologists' theories have always emphasized the role of "beliefs" and related cultural attitudes in the generation of persistent inequality between dynasties. In particular, the very influential works of French sociologist Pierre Bourdieu all describe the various social processes through which individuals from lower-class backgrounds are being discouraged by the "dominant discourse" from making adequate mobility-enhancing investments, especially within the schooling system (see, e.g., Bourdieu and Passeron, 1964, 1970).[71] For instance, lower-class children can be discouraged by school teachers who tell them they have no chance of going to a good college and they should opt for a more "reasonable" school orientation. Lower class individuals can also be discouraged by bosses who tell them that they will never be sufficiently able to be a manager and that should better accept their life as a factory worker. More generally, lower class individuals can be discouraged by the general "dominant discourse" produced by politicians, opinion leaders and the capitalist system as a whole, according to which persistent inequality is basically efficient and lower-class dynasties should better accept their inferior role. According to Bourdieu, they can be discouraged to such an extent that they "internalize" entirely the low probabilities of social ascent enjoyed by their peers within the current structure of social inequality and adopt behaviors that validate the "dominant discourse".

There exists an obvious similarity between Bourdieu's theory and the theory of self-fulfilling discriminatory beliefs described above. In both cases, inequality is persistent simply because the elite expects inequality to be persistent, so that the poor are discouraged and validate the elite's expectation. One important difference is that the theory of statistical discrimination puts the blame exclusively on employers, which leads to a number of specific policy recommendations (see Section 6.1 above). In contrast, Bourdieu's theory tends to put the blame on the society as a whole, so that the set of appropriate policy tools is potentially enormous. For instance, Bourdieu implicitly suggests that one just needs to change the "dominant discourse", e.g., by writing books.

Note that the emphasis by sociologists on beliefs and non-economic sources of persistent inequality is not confined to "left-wing" theories. Indeed, most sociologists agree

[70] See Ichino and Ichino (1997) for a recent attempt to distinguish between persistent inequality due to the intergenerational transmission of cultural attitudes and persistent inequality due to discrimination, in the context of the inequality between southern and northern Italian workers (they use labor market data about north-south migrants).

[71] Other radical analysis of how the conservative ideology of the school system might contribute to make inequality more persistent than it would otherwise be have also been produced in the US (see, e.g., Bowles and Gintis, 1975).

that intergenerational inequality persists above and beyond the pure transmission of ability differentials, but they disagree about the extent to which this economically useless inequality can be altered by policy. For instance, one of the main counter-arguments to Bourdieu's theories is based upon the "reference group" theory, according to which inequality is persistent simply because lower-class families transmit less ambition and taste for economic success than upper-class families (see Section 3.3 above). It is interesting to note that the conflict between Bourdieu's theory and the reference group theory is at the origin of the major, long-standing controversy within French sociology, after the reference group theory had been advocated by Boudon (1973, 1974). To a large extent, this conflict is still the major dividing line in the French community of academic sociologists, where scholars need to affiliate themselves either as pro-Bourdieu or pro-Boudon. The conflict about the existence of efficient corrective policies is obvious: both theories describe persistent inequality as a self-fulfilling phenomena, whereby poor dynasties remain poor because of their lack of ambition, but Bourdieu puts the blame on society and suggests that radical activism can easily raise social mobility, whereas Boudon insists that there is nobody to blame for it, unless one is ready to destroy the family institution.[72] As we have seen in the case of statistical discrimination, the parameters involved in these conflicting theories are extremely difficult (if not impossible) to measure empirically, which may explain why these political conflicts are so persistent.

The claims made by radical sociologists should however not be dismissed on purely a priori grounds. Needless to say, the idea that discourse and beliefs alone can have a significant impact on "real" economic inequality sounds inherently suspicious to most economists. Economists legitimately consider that redistributive taxation can at least alleviate with certainty the "real" inequality of living standards, as compared to the uncertain outcomes of "beliefs politics". On the other hand, it is probably true for instance that the dramatic improvement of the relative economic position of women during the twentieth century, which probably constitutes the most spectacular redistribution that has ever happened, has not happened through fiscal redistribution and economic policies but rather through the evolution of beliefs and mental attitudes towards women. In the same way, it is not impossible that the discourses of Gandhi have done more to modify social attitudes toward lower-caste Indians and to improve their economic prospects than any straight economic redistribution could ever have. It is also possible that all those phenomena that noneconomists attribute to "discourse" or changing mental attitudes towards others can actually be explained by some underlying "technological" evolution of demand and supply, so that discourse is just a veil, but this needs to be proven rather than assumed. The challenge for economists is to be able to recognize and measure the relative importance of such channels of inequality transmission and at the same time maintain the rigor and the intellectual standards of the discipline.

[72] See Section 3.3 and Piketty (1998). Note that the Piketty (1995) model of dynastic learning about the income responsiveness of effort (see Section 3.2 above) offers another theory of self-fulfilling inequality where there is nobody to blame (different dynasties just happen to have different experimentation trajectories).

References

Abel, A.B. and D.B. Bernheim (1991), Fiscal policy with impure intergenerational altruism, Journal of Political Economy 59: 1687–1711.

Acemoglu, D. (1995), Public policy in a model of long-term unemployment, Economica 62: 161–178.

Aghion, P. and P. Bolton (1997), A trickle-down theory of growth and development with debt-overhang, Review of Economic Studies 64: 151–172.

Aiyagari, S. (1994), Optimal Capital Income Taxation with Incomplete Markets, Borrowing Constraints and Constant Discounting, mimeo, Minneapolis.

Arrow, K. (1972), The theory of Discrimination, in O. Ashenfelter and A. Rees, eds., Discrimination on Labor Markets (Princeton University Press).

Atkinson, A.B. (1980), Inheritance and the redistribution of wealth, in G.A. Hughes and G.M. Heal, eds., Public Policy and the Tax System: Essays in Honour of James Meade (Allen and Unwin, London).

Atkinson, A.B. (1981), On intergenerational income mobility in Britain, Journal of Post-Keynesian Economics 3: 194–218.

Atkinson, A.B., A.K. Maynard and C.G. Trinder (1983), Parents and Children: Incomes in Two Generations (Heinemann, London).

Banerjee, A. and M. Ghatak (1996), Empowerment and efficiency: the economics of tenancy reform, mimeo, MIT.

Banerjee, A. and A. Newman (1993), Occupational choice and the process of development, Journal of Political Economy 101: 274–299.

Banerjee, A. and A. Newman (1994), Poverty, incentives and development, American Economic Review 84: 211–216.

Becker, G. (1988), Family economics and macro behavior, American Economic Review 78: 1–13.

Becker, G. (1989), On the economics of the family: reply to a skeptic, American Economic Review 79: 514–518.

Becker, G. (1991), A Treatise on the Family (Harvard University Press).

Becker, G. and N. Tomes (1979), An equilibrium theory of the distribution of income and intergenerational mobility, Journal of Political Economy 87: 1153–1189. (reprinted as chapter 7 of Becker (1991)).

Becker, G. and N. Tomes (1986), Human capital and the rise and fall of families, Journal of Labor Economics 4: S1–S39. (reprinted as supplement to chapter 7 of Becker (1991)).

Becker, G. and R. Barro (1988), A reformulation of the economic theory of fertility, Quarterly Journal of Economics 103: 1-25. (reprinted as supplement to chapter 5 of Becker (1991)).

Behrman, J.R. and P. Taubman (1985), Intergenerational earnings mobility in the United States: some estimates and a test of Becker's intergenerational endowments model, Review of Economics and Statistics 67: 144–151.

Benabou, R. (1993), Workings of a city: location, education, production, Quarterly Journal of Economics 108: 619–653.

Benabou, R. (1996a), Heterogeneity, stratification and growth: macroeconomic implications of community structure and school finance, American Economic Review 86: 584–609.

Benabou, R. (1996b), Equity and efficiency in human capital investment: the local connection, Review of Economic Studies 63: 237–264.

Benabou, R. (1997), What Level of Redistribution Maximizes Long-Run Output? mimeo, New York University.

Bernheim, D.B. (1991), How strong are bequest motives? evidence based on estimates of the demand for life insurance and annuities, Journal of Political Economy 99: 899–927.

Bernheim, D.B. and K. Bagwell (1988), Is everything neutral? Journal of Political Economy 96: 308–338.

Bjorklund, A. and M. Jantti (1998), Intergenerational mobility of economic status: is the United States different? Working Paper (Swedish Institute for Social Research, Stockholm).

Borjas, G.J. (1992), Ethnic capital and intergenerational mobility, Quarterly Journal of Economics 107: 123–150.

Borjas, G.J. (1995), Ethnicity, neighborhoods and human capital externalities, American Economic Review 85: 365–390.

Bourdieu, P. and J.C. Passeron (1964), Les héritiers (Les éditions de Minuit, Paris).
Bourdieu, P. and J.C. Passeron (1970), La reproduction (Les éditions de Minuit, Paris).
Boudon, R. (1973), L'inégalité des chances (Armand Colin, Paris).
Boudon, R. (1974), Education, Opportunity and Social Inequality (Wiley, New York).
Bourguignon, F.(1981), Pareto-superiority of unegalitarian equilibria in Stiglitz' model of wealth distribution with convex savings function, Econometrica 49: 1469–1475.
Bowles, S. (1972), Schooling and inequality from generation to generation, Journal of Political Economy 80: S219–S251.
Bowles, S. and H. Gintis (1976), Schooling in Capitalist America (Basic Books, New York).
Card, D. and A. Krueger (1992), Does school quality matter? Journal of Political Economy 100: 1–40.
Card, D. and A. Krueger (1995), Myth and Measurement: the New Economics of the Minimum Wage (Princeton University Press).
Chamley, C. (1996), Capital Income Taxation, Income Distribution and Borrowing constraints, mimeo (DELTA, Paris).
Chu, C.Y. (1991), Primogeniture, Journal of Political Economy 99: 78–99.
Coate, S. and G. Loury (1993), Will affirmative-action policies eliminate negative stereotypes? American Economic Review 83: 1220–1240.
Cole, H., G. Mailath and A. Postlewaite (1992), Social norms, savings behavior and growth, Journal of Political Economy 100: 1092–1125.
Coleman, J. (1966), Equality of Educational Opportunity, Report to the US Department of Health, Education and Welfare.
Conlisk, J. (1974), Can equalization of opportunity reduce social mobility? American Economic Review 64: 80–90.
Cooper, S., S. Durlauf and P. Johnson (1994), On the evolution of economic status across generations, American Statistical Association, Business and Economics Section, Papers and Proceedings 50–58.
Cutler, D.M. and E.L. Glaeser (1997), Are ghettos good or bad? Quarterly Journal of Economics 112: 827–872.
Dearden, L., S. Machin and H. Reed (1997), Intergenerational mobility in Britain, Economic Journal 107: 47–66.
Durlauf, S. (1996), A theory of persistent income inequality, Journal of Economic Growth 1: 75–93.
Erikson, R. and J.H. Goldthorpe (1985), Are American rates of social mobility exceptionally high? European Sociological Review 1: 1–22.
Erikson, R. and J.H. Goldthorpe (1992), The Constant Flux: A Study of Class Mobility in Industrial Societies (Clarendon Press, Oxford).
Feldstein, M. (1995), The effects of marginal tax rates on taxable income: a panel study of the 1986 tax reform act, Journal of Political Economy 103: 551–572.
Fernandez, R. and R. Rogerson (1994), Public Education and Income Distribution: A Quantitative Evaluation of School Finance Reform, NBER Working Paper 4883.
Fernandez, R. and R. Rogerson (1996), Income distribution, communities and the quality of public education, Quarterly Journal of Economics 111: 135–164.
Freeman, R. (1973), Changes in the labor market status of black Americans, 1948–1972, Brookings Papers on Economic Activity 1: 67–120.
Freeman, R. (1981), Black economic progress after 1964: who has gained and why? in S. Rosen, ed., Studies in Labor Markets (Chicago University Press).
Freeman, S. (1996), Equilibrium income inequality among identical agents, Journal of Political Economy 104: 1047–1064.
Friedman, M. (1962), Capitalism and Freedom (The University of Chicago Press).
Galor, O. and J. Zeira (1993), Income distribution and macroeconomics, Review of Economic Studies 60: 35–52.
Gilchrist, S. and C.P. Himmelberg (1995), Evidence on the role of cash flow for investment, Journal of Monetary Economics 36: 541–572.
Goldberger, A. (1989), Economic and mechanical models of intergenerational transmission, American Economic Review 79: 504–513.

Goux, D. and E. Maurin (1996), Meritocracy and social heredity in France: some aspects and trends, European Sociological Review.

Green, L., J. Myerson, D. Lichtman, S. Rosen and A. Fry (1996), Temporal discounting in choices between delayed rewards: the role of age and income, Psychology and Aging 11: 79–84.

Herrnstein, R. and C. Murray (1994), The Bell Curve: Intelligence and Class Structure in American Life (The Free Press).

Ichino, A., D. Checchi and A. Rustichini (1997), More Equal but Less Mobile? Education Financing and Intergenerational Mobility in Italy and the U.S., mimeo (Universita Bocconi and IGIER).

Ichino, A. and P. Ichino (1997), Culture, Discrimination and Individual Productivity: Regional Evidence from Personnel Data in a Large Italian Firm, CEPR Discussion Paper No. 1709.

Jaffee, D. and J. Stiglitz (1990), Credit rationing, in B. Friedman and F. Hahn, eds., Handbook of Monetary Economics, Vol. 2 (Elsevier Science Publishers).

Judd, K. (1985), Redistributive taxation in a simple perfect foresight model, Journal of Public Economics 28: 59–83.

Kaelble, H. (1986), Social Mobility in the Nineteenth and Twentieth Centuries: Europe and America in Comparative Perspective (St. Martin's Press, New York).

Kessler, D. and A. Masson (1989), Bequests and wealth accumulation: are some pieces of the puzzle missing? Journal of Economic Perspectives 3: 141–152.

Kotlikoff, L.J. (1988), Intergenerational transfers and savings, Journal of Economic Perspectives 2: 41–58.

Kremer, M. (1994), The Dynamics of Inequality: US vs Brazil, mimeo, MIT.

Kremer, M. (1997), How much does sorting increase inequality? Quarterly Journal of Economics 112: 115–140.

Kremer, M. and E. Maskin (1995), Segregation by Skill and the Rise in Inequality, NBER Working Paper 5718.

Lamont, O. (1997), Cash flow and investment: evidence from internal capital markets, Journal of Finance 52: 83–109.

Lawrence, E.C. (1991), Poverty and the rate of time-preference: evidence from panel data, Journal of Political Economy 99: 54–77.

Legros, P. and A. Newman (1996), Wealth effects and the theory of organisation, Journal of Economic Theory 70: 312–341.

Lipset, S.M. and R. Bendix (1959), Social Mobility in Industrial Society (University of California Press, Berkeley).

Lipset, S.M. and R. Bendix (1966), Class, Status, and Power: Social Stratification in Comparative Perspective, 2nd. edn. (The Free Press, New York).

Loury, G. (1981), Intergenerational transfers and the distribution of earnings, Econometrica 49: 843–867.

Lucas, R. (1990), Supply-side economics: an analytical review, Oxford Economic Papers 42: 34–45.

Mayshar, J. and S. Benninga (1996), Heterogeneity of Intertemporal Tastes and Efficient Capital Markets as Sources for Inequality: Extending a Theme by Rae and Ramsey with some Policy Implications, mimeo, Hebrew University.

Merllié, D. and J. Prévot (1991), La mobilité sociale (La Découverte, Paris).

Merton, R. (1953), Reference group theory and social mobility, in R. Bendix and S.M. Lipset, eds., Class, Status and Power (The Free Press, New York).

Meyer, C.(1995), Income Distribution and Family Structure, PhD dissertation (Department of Economics, MIT).

Modigliani, F. (1988), The role of intergenerational transfers and life cycle saving in the accumulation of wealth, Journal of Economic Perspectives 2: 15–40.

Mulligan, C. (1993), On Intergenerational Altruism, Fertility and the Persistence of Economic Status, PhD dissertation (Department of Economics, University of Chicago).

Mulligan, C. (1997), Parental Priorities and Economic Inequality (The University of Chicago Press).

Murphy, K.M., A. Shleifer and R.W. Vishny (1989), Income distribution, market size and industrialization, Quarterly Journal of Economics 104: 537–564.

Neal, D.A. and W.R. Johnson (1996), The role of pre-market factors in black-white wage differences, Journal of Political Economy 104: 869–895.

Newman, A. (1991), The Capital Market, Inequality and the Employment Relation, mimeo, Columbia University.
Phelps, E. (1968), The statistical theory of racism and sexism', American Economic Review **62**, 659–661.
Piketty, T.(1995), Social mobility and redistributive politics, Quarterly Journal of Economics 110: 551–584.
Piketty, T.(1997), The dynamics of the wealth distribution and the interest rate with credit-rationing, Review of Economic Studies 64: 173–189.
Piketty, T. (1998), Self-fulfilling beliefs about social status, Journal of Public Economics 70:1 115–132.
Roemer, J. and R. Wets (1994), Neighborhood Effects on Belief Formation and the Distribution of Education and Income, mimeo, UC Davis.
Schiff, M., M. Duyme, A. Dumaret and S. Tomkiewicz (1982), How much could we boost scholastic achievement and IQ scores? A direct answer from a French adoption study, Cognition 12: 165–196.
Schiff, M. and R. Lewontin (1986), Education and Class: the Irrelevance of IQ Genetic Studies (Clarendon Press, Oxford).
Shavit, Y. and H.P. Blossfeld (1993), Persistent Inequality: Changing Educational Attainment in Thirteen Countries (Westview Press).
Slemrod, J. (1995), Income creation or income shifting? Behavioral responses to the Tax Reform Act of 1986, American Economic Review 85: 175–180.
Solon, G. (1992), Intergenerational income mobility in the United States, American Economic Review 82: 393–408.
Stiglitz, J. (1969), Distribution of income and wealth among individuals, Econometrica 37: 382–397.
Stiglitz, J. (1978), Equality, taxation and inheritance, in W. Krell and A.F. Shorrocks, eds, Personal Income Distribution (North-Holland, Amsterdam).
Thurow, L. (1975), Generating Inequality: Mechanisms of Distribution in the US Economy (Basic Books, New York).
Wilson, W.J. (1987), The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy (University of Chicago Press).
Zimmerman, D.J. (1992), Regression toward mediocrity in economic stature, American Economic Review 82: 409–429.