*Economic Quarterly—Volume 97, Number 3—Third Quarter 2011—Pages 189–193*

# Introduction to the Special Issue on Modern Macroeconomic Theory

Andreas Hornstein

The great recession of 2007–2009 has generated significant external criticism of the way economists study and try to understand aggregate economic outcomes. Modern macroeconomic theory, in particular, has been criticized for its representation of the economy through highly stylized environments that abstract from distributional issues, ignore or minimize linkages between the financial and nonfinancial sectors of the economy, and, in general, rely too much on highly aggregative frameworks. This issue collects four articles that describe how modern macroeconomic research has dealt with some of these issues as part of a research program that has been ongoing for more than a decade.

The first article by Nobuhiro Kiyotaki provides a short history of modern business cycle theory and how it has evolved to potentially address the role of the financial sector in the aggregate economy. Kiyotaki starts with the neoclassical growth model as a reference point for most of modern business cycle theory. This modelling framework, originally known as "real business cycle" theory, starts with the stark abstraction of one representative household and one representative producer in a competitive environment without any frictions on the interactions of consumers and producers. From the perspective of this model, business cycles are driven by exogenous shocks, and the dynamics of the cycle essentially reflect the dynamics of the shocks. In other words, there is only a weak model-internal mechanism that propagates shocks. Kiyotaki then studies a sequence of well-defined deviations from this reference point and asks what deviations are more likely to affect the baseline interpretation of business cycles. Kiyotaki first shows how heterogeneity in consumption and production can be easily accommodated in this

The views expressed do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: andreas.hornstein@rich.frb.org.

framework given the assumption of complete markets. In a second step, Kiyotaki shows how non-competitive markets, either because of market power or limitations on the interactions of agents, can be introduced into the baseline model. Neither of these modifications affect the interpretation of business cycles as being driven by shocks. Finally, Kiyotaki argues that restrictions on the set of available financial contracts significantly affect the way exogenous shocks are propagated in the model economy.

The second article by Vincenzo Quadrini elaborates on the role of financial frictions for production decisions. Quadrini illustrates these financial frictions in a simple example where entrepreneurs have to acquire capital to operate an intertemporal production technology. Again, financial frictions are introduced relative to the baseline complete markets framework. Quadrini discusses the two most popular models of market incompleteness—the costly state verification (CSV) model and the collateral constraint (CC) model. Both frameworks limit entrepreneurs to the use of two financial instruments: contingent debt (equity or net worth) and non-contingent debt. In the CSV model, non-contingent debt is the optimal response to a limited information problem, and an entrepreneur's net worth limits his ability to issue debt and finance investment projects. In the CC model, posting collateral allows the entrepreneur to obtain credit despite his inability to credibly commit to the repayment of debt. The main question then becomes how these financial frictions can amplify the effects of shocks to the economy or be themselves a source of shocks to the economy. Quadrini illustrates the basic mechanism for amplification and propagation in the simple model, and surveys the results from more "realistic" models.

The third article by Fatih Guvenen surveys recent research on household heterogeneity in the absence of complete markets. We might be interested in household heterogeneity for two reasons. First, even though we assume in the baseline "real business cycle" model that aggregate consumption and labor supply decisions can be modelled through a representative household construct, we might worry that "distributions" of ability, income, or wealth do matter for the behavior of these aggregate outcomes. Second, observed inequality of income and wealth often gives rise to attempts to redistribute resources. In order to address the costs and benefits of such a policy, one first needs a theory that accounts for the currently observed inequality across households. If we care about inequality because of implied differences in "well-being," then we should care about inequality in consumption and leisure, and we should care about income inequality only to the extent that it gives rise to consumption inequality. Much of the research surveyed by Guvenen studies how, in the absence of complete markets, income inequality gets translated into consumption and wealth inequality. If the level of income and its distribution are exogenous, the redistribution problem is simplified since any attempt to influence consumption and wealth inequality does not feed back into either the

level or the distribution of income. But economists are always worried about the labor supply effects of tax policies, that is that at least part of income levels and inequality are endogenous. In standard models, these labor supply effects show up as variations in hours worked or labor market participation decisions. In his survey, Guvenen emphasizes a different labor supply decision, namely the accumulation of human capital. Overall, Guvenen shows that accounting for heterogeneity of households in environments with incomplete markets is feasible, but it also requires the application of advanced computational tools. In the absence of controlled experiments, researchers are essentially compelled to construct artificial worlds with a population of heterogeneous households. Once the consumption and labor supply decisions of the households in the model mirror the observed behavior of households, we can ask how changes in the artificial environment will affect outcomes.

The fourth article by Diego Restuccia deviates somewhat from the immediate concerns of the U.S. economy and studies the issues of output determination in a global framework. During the "Great Recession," U.S. real gross domestic product (GDP) declined by 5 percent from 2007 to 2009, and, as of 2011, real GDP is now arguably 10 percent below its long-run trend growth path. While these changes of real output are large, they pale in comparison to observed cross-country income differences: In 2005, the average per capita income in the richest countries was about 65 times that of the poorest countries. Restuccia first surveys the evidence on cross-country differences in per capita income. He shows that, although it appears that cross-country per capita income inequality has been increasing over the last 30 years, for individual countries there are success stories and then there are failures. The recent, most prominent examples for countries that have been catching up with the leading world economy—the United States—are China and India. However, there are countries such as Zimbabwe and Venezuela that have been falling behind the United States more and more. Restuccia then argues that the process of structural transformation, that is, the transition from a predominantly agricultural economy to an industrialized economy, and then to a service-oriented economy, can account for some of these differences. In particular, he points to the relatively low levels of agricultural productivity in poor countries as a major source of income differences. Essentially, Restuccia argues that cross-country differences in aggregate productivity and per capita income can be attributed to differences in sectoral productivities resulting in differences in resource allocation. Restuccia then surveys theories that attribute differences in sectoral productivity to distortions that lead to the inefficient allocation of resources across production establishments. Restuccia's survey reflects how the baseline neoclassical model of production can be modified to account for heterogeneity in production, first at the industry level, then at the establishment level. These modifications are matched to observations, and we can see how much they contribute to differences in aggregate output.

The four articles in this issue represent part of a research program in macroeconomics that takes the basic stochastic growth model with complete markets as its point of departure. Work in this research program then adds various sources of frictions and heterogeneity on the consumption and production side, including restrictions on the set of available markets, and the ability of market participants to pledge to repay debts. This procedure allows macroeconomists to evaluate the contributions of the various features that allow model economies to capture more dimensions of available empirical evidence relative to a common benchmark model. Another line of research that is part of this program, but is not addressed by these articles, departs from the baseline growth model by introducing nominal price rigidities in order to address monetary non-neutrality.[1] In fact, until the Great Recession, research on the role of nominal price rigidities and monetary policy institutions in particular, received more attention in macroeconomics in general than did research on financial market frictions. This ranking of different lines of research simply reflected the historical experience with the U.S. economy and other advanced economies: Apparent inflation-output tradeoffs were considered to be much more important than financial-market instability. For example, in the U.S. economy the stock market crash of 1987 had no appreciable impact on the aggregate economy, and the boom in equity prices in the 1990s, with a subsequent crash in 2001, was followed by one of the shallowest recessions in post-WWII history. For many macroeconomists, the Great Recession changed the perception on how important financial markets might be for the economy. Consequently, attention among economists has shifted more toward the lines of research that emphasize financial market frictions and that are described in this special issue. The fact that economists continue to discuss the causes and consequences of the Great Depression should, however, give one pause to expect any time soon a coherent and generally accepted narrative of the Great Recession and how it relates to the preceding collapse of the housing bubble and the ensuing financial crisis.[2]

---

[1] For an introduction, see Galí (2008).

[2] Lo (forthcoming), in a very instructive survey of the literature on the financial crisis, both by academics and journalists, observes that no single narrative has yet emerged from that literature, and that, even for a number of commonly accepted "stylized facts" of the financial crisis, there is no clear cut empirical evidence.

## REFERENCES

Galí, Jordi. 2008. *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework*. Princeton, N.J.: Princeton University Press.

Lo, Andrew W. Forthcoming. "Reading About the Financial Crisis: A 21-Book Review." *Journal of Economic Literature*.

*Economic Quarterly—Volume 97, Number 3—Third Quarter 2011—Pages 195–208*

# A Perspective on Modern Business Cycle Theory

Nobuhiro Kiyotaki

The global financial crisis and recession that started in 2007 with the surge of defaults of U.S. subprime mortgages is having a large impact on recent macroeconomic research. The framework of modern macroeconomics that has replaced traditional Keynesian economics since the 1970s has been widely criticized. Many of the criticisms have focused on the assumptions of the representative agent and its abstraction from firm and household heterogeneity. Critics are also skeptical about the model's ability to explain unemployment and financial crises because it abstracts from market frictions and irrationality. As a result, modern macroeconomics has often been attacked for its futility in providing policy insight in the way that traditional Keynesian economics has done.[1] Some criticisms are constructive and others are misleading. I would like to present my thoughts on what I believe are the contributions and shortcomings of modern macroeconomic theory, in particular the business cycle theory, by responding to some of these criticisms.[2]

## 1. REAL BUSINESS CYCLE THEORY

For the past few decades, real business cycle (RBC) theory has been the focal point of debates in business cycle studies.[3] According to the standard

This is an English translation of my Japanese article "A Perspective on Modern Business Cycle Theory" in *The 75 Years History of Japanese Economic Association*, edited by the Japanese Economic Association (2010). The article is based on my plenary talk at the Japanese Economic Association annual meeting in October 2009. I would like to thank Raoul Minetti, Mako Saito, and Akihisa Shibata for thoughtful comments on the lecture slides and the earlier draft. The opinions expressed in this article do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: kiyotaki@princeton.edu.

[1] For example, see Krugman (2009).

[2] Because of the limitations of space and my expertise, I will not deal with economic growth or the empirical studies of business cycles. I also admit that the references are heavily biased to my own work.

[3] See Kydland and Prescott (1982) and Prescott (1986) for examples.

RBC approach, the competitive equilibrium of the market economy achieves resource allocation that maximizes the representative household's expected utility given the constraints on resources. Although the RBC approach has often been criticized for its abstraction from firm and household heterogeneity, these charges are incorrect. Instead, it would be more accurate to view the RBC framework as one with heterogeneous firms and households all playing a part in the social division of labor under an ideal market mechanism. The real business cycle theory is a business cycle application of the Arrow-Debreu model, which is the standard general equilibrium theory of market economies.

Let us briefly outline the mechanics of an RBC model. Consider an economy with a homogeneous product that can be either consumed or invested. Labor, capital, and land are homogeneous inputs to production and total supply of land is normalized to unity. There are a number of infinitely lived households ($h = 1, 2, ..., H$) and firms ($j = 1, 2, ..., J$). A household's preference is given by the discounted expected utility of consumption and disutility of work:

$$E_0 \left\{ \sum_{t=0}^{\infty} \beta^t \left[ u_h \left( c_{ht} \right) - d_h \left( n_{ht} \right) \right] \right\}. \tag{1}$$

Firm $j$'s maximum output is a function of the factors of production: capital, land, and labor $(k_{jt}, l_{jt}, n_{jt})$ represented by the production function

$$y_{jt} = f(k_{jt}, l_{jt}, n_{jt}; z_{jt}), \tag{2}$$

where productivity of firm $z_{jt}$ follows a Markov process. In the goods market equilibrium, aggregate output equals aggregate consumption and investment:

$$\sum_{j=1}^{J} y_{jt} = \sum_{h=1}^{H} c_{jt} + \sum_{j=1}^{J} k_{jt+1} - (1 - \delta) \sum_{j=1}^{J} k_{jt}, \tag{3}$$

where $\delta$ is the depreciation rate of capital.

Here, we assume that markets are complete, that is, there exists a complete set of Arrow securities so that state-contingent claims to goods and factors of production for every possible future state can be traded at the initial period. We also assume that capital, land, and labor can be allocated freely across firms every period and all markets are perfectly competitive. Under these assumptions, the competitive equilibrium achieves an allocation that maximizes the weighted average of all individual households' expected utilities with constant weights $\lambda_h$ given the resource constraints (Negishi 1960).

We define the representative household's utility function as the weighted average of all household utilities:

$$u(C) = Max_{c_h} \sum_{h=1}^{H} \lambda_h u_h(c_h), \text{ s.t. } \sum_{h=1}^{H} c_h = C,$$
$$d(N) = Min_{d_h} \sum_{h=1}^{H} \lambda_h d_h(n_h), \text{ s.t. } \sum_{h=1}^{H} n_h = N.$$

The aggregate production function is defined as total output given the efficient allocation of factors of production and can be written as

$$Y_t = A_t F(K_t, N_t) = Max_{k_{jt}, l_{jt}, n_{jt}} \sum_{j=1}^{J} f(k_{jt}, l_{jt}, n_{jt}; z_{jt})$$

$$\text{s.t. } \sum_{j=1}^{J} k_{jt} = K_t, \quad \sum_{j=1}^{K} l_{jt} = 1, \quad \sum_{j=1}^{J} n_{jt} = N_t. \tag{4}$$

Here, aggregate productivity $A_t$ is a function of $z_{jt}$ for all $j$. The competitive equilibrium is described by aggregate quantities $(C_t, N_t, Y_t, K_{t+1})$ as a function of the state variables $K_t$ and $\{z_{jt}\}_{j=1,2,\ldots J}$ that maximize the expected utility of the representative household

$$E_0 \sum_{t=0}^{\infty} \beta^t [u(C_t) - d(N_t)], \tag{5}$$

subject to the resource constraint

$$C_t + K_{t+1} - (1-\delta) K_t = A_t F(K_t, N_t).$$

Note that the representative household is not an assumption; it arises as an implication of constant Negishi weights under complete markets as in Negishi (1960). The aggregate production function is also constructed under the assumption that production is efficient in competitive markets without friction. Therefore, the real business cycle theory does not blindly abstract from firm and household heterogeneity. By assuming that markets are functioning "well," we reduce an otherwise general model to one of the representative agent with an aggregate production function and analyze the business cycle phenomenon in this simplified economy.

Now I wish to discuss, in an intuitive manner, how the real business cycle theory explains the fluctuation of aggregate quantities $(C_t, N_t, Y_t, K_{t+1})$ by a shock to aggregate productivity. Suppose that aggregate productivity suddenly increases temporarily. Following this shock, marginal product of labor will increase, leading to a rise in the real wage and therefore the quantity

of labor supplied. The combined effect of higher productivity and increased use of labor will cause output to rise. But since the productivity increase is temporary, future output is expected to increase less than present output, and permanent income and consumption do not increase as much as present output. Thus, from the goods market equilibrium condition (output = consumption + investment), investment and, hence, next period capital stock will increase. This will increase next period marginal productivity of labor, labor input, and output, leading to another cycle of aggregate quantity increases and so on. Hence, we notice that a temporary shock to productivity has precipitated a persistent rise in aggregate quantities.

However, the biggest problem with the propagation mechanism described above is that short-term changes in investment have little impact on capital stock. At the same time, if there is a persistent increase of output, permanent income and consumption will increase almost as much as current income, which leaves little room for investment to rise. Therefore, we conclude that capital plays a limited role in the propagation of a productivity shock. Furthermore, the substitution and wealth effects of a productivity shock on labor supply work to cancel each other out: At a higher real wage, the representative agent is willing to supply more labor, but the higher aggregate productivity also increases the agent's wealth, which in turn reduces the labor supply. Unless the substitution effect is very large, the overall fluctuations of labor will not be very large. As a result, we need large and persistent aggregate productivity shocks in order to explain the business cycle phenomenon. Because RBC models are missing a powerful propagation mechanism whereby small shocks to the economy amplify and produce large fluctuations, they rely on large exogenous shocks. But the question is where do these exogenous shocks come from? It is difficult to identify such shocks even with the recent global recession or the 1930s Great Depression.

## 2. OTHER SHOCKS

While exogenous shocks to productivity were the main source of shock in early RBC analysis, the framework was later extended to include the effects of other potential shocks. For example, in the face of a global downturn, what would be the effects of a decreasing demand for exports? We cannot address this question in a perfectly competitive economy since individual firms are assumed to make their production decisions by taking market prices as given, that is, they cannot perceive changes in demand directly. Therefore, let us assume a monopolistically competitive economy in which each firm $j$ sells a differentiated good. The quantity of aggregate output, which can be used as either a consumption good or investment good, is a function of many differentiated goods as

$$Y_t = \left[ \sum_{j=1}^{J} \left( x_{jt} \right)^{\frac{1}{\theta}} \left( y_{jt} \right)^{\frac{\theta-1}{\theta}} \right]^{\frac{\theta}{\theta-1}},$$

where $\theta$ is the elasticity of substitution between differentiated goods and $\theta > 1$. The parameter $x_{jt}$ is an exogenous idiosyncratic demand shock to firm $j$'s product. If we let $p_{jt}$ be the price of each good, the price index that corresponds to the above aggregate output is

$$P_t = \left[ \sum_{j=1}^{J} x_{jt} \left( p_{jt} \right)^{1-\theta} \right]^{\frac{1}{1-\theta}}.$$

Since households and firms use differentiated goods such that their consumption and investment levels are maximized subject to their budget constraints, for a given level of aggregate output produced, aggregate demand for short, real income of each firm is given by

$$\left( \frac{p_{jt}}{P_t} \right) y_{jt} = x_{jt} \left( Y_t \right)^{\frac{1}{\theta}} \left( y_{jt} \right)^{1-\frac{1}{\theta}}.$$

Export demand shocks, which shift aggregate demand $Y_t$, will affect real income of firms, and will change production, employment, consumption, and investment levels the way productivity shocks did in the previous section. Although a monopolistically competitive economy yields inefficient equilibrium resource allocations, key features of the business cycle are not significantly different from those of a perfectly competitive economy. In fact, a monopolistically competitive equilibrium corresponds to a perfectly competitive market equilibrium with a value-added tax that redistributes the tax revenue lump sum. Therefore, simply adding monopolistic competition to an RBC model cannot account for the business cycle phenomenon, and some other source of friction such as price stickiness or a different type of shock is necessary.

Now, instead of a shock to aggregate demand, let us consider a shock to the quality of capital. Assume that a fraction $\psi_{t+1}$ of differentiated goods becomes obsolete between periods $t$ and $t+1$, and that the capital used as inputs in the production of those goods also becomes obsolete. In other words, the idiosyncratic demand parameter $x_{jt}$ of goods affected by the obsolescence shock becomes zero, and the corresponding amount of demand shifts toward new goods. As a result, the productive capital stock will decrease to

$$K_{t+1} = I_t + (1 - \psi_{t+1})(1 - \delta)K_t. \tag{6}$$

This lower capital stock induces a decrease in output and employment. If there are no other sources of friction in the economy, however, investments will increase, encouraging the expansion of labor supply, and contribute to a quick recovery of output. This is similar to the adjustment process of an economy with initial stock of capital lower than the steady-state equilibrium level in the Neoclassical optimal growth model. Again, incorporating capital obsolescence shocks is insufficient to explain standard cases of recessions in which investment and employment are depressed instead of booming. (See Section 4 for more explanation.)

## 3. LABOR MARKET FRICTION

Real business cycle theory is often criticized for its lack of implications for the cyclical behavior of unemployment. This issue has been partially addressed in the Diamond-Mortensen-Pissarides framework that incorporates matching frictions that exist in the labor market between workers and firms.[4] Matching theory assumes that it is costly and time consuming to find productive matches because workers and jobs are heterogeneous. In order to include this feature into the macroeconomic model, they introduced an aggregate job-matching function written as an increasing function of job vacancies $v_t$ and the number of unemployed workers (difference between the workforce and employment level, $\bar{N}_t - N_t$):

$$N_{t+1} = \mu_t M(v_t, \bar{N}_t - N_t) + (1 - \delta_{nt})N_t.$$

$\mu_t$ represents the efficiency of the job matching and $\delta_{nt}$ is an exogenous parameter that measures the rate at which current job matches are destroyed. We assume firms incur a recruitment cost of $\chi$ units of the output good per vacancy. The goods market equilibrium condition is

$$Y_t = C_t + K_{t+1} - (1 - \delta)K_t + \chi v_t.$$

After firms and workers are matched, wages are determined by Nash bargaining. We assume the Hosios condition (Hosios 1990) (under which the firm's bargaining power is equal to the elasticity of the number of aggregate job matches with respect to the vacancies) is satisfied. Each household consists of many workers, and is therefore able to diversify labor income risk from unemployment. The competitive equilibrium of such an economy with search maximizes the expected utility of the representative household.

[4] See Mortensen and Pissarides (1994), Merz (1995), and Andolfatto (1996) for descriptions of the Diamond-Mortensen-Pissarides framework and its incorporation into RBC models.

In a search and matching model, search is an investment of current resources for future returns, and we expect substantial fluctuations in unemployment only when labor productivity and demand are expected to change persistently. However, according to Shimer (2005), even with persistent labor productivity shocks, fluctuations in unemployment will be small if the marginal product of labor is significantly larger than the marginal cost of labor supply (marginal rate of substitution between labor supply and consumption $d'(\bar{N})/u'(C)$) and if wages are determined by Nash bargaining. Therefore, search models appear to be limited to explaining fluctuations in unemployment of young workers fresh out of school and old workers nearing retirement for whom the difference between the marginal product of labor and marginal cost of labor supply is small.

## 4. HETEROGENEITY AND CREDIT LIMITS

In an Arrow-Debreu economy that underlies RBC theory, credit is considered to be a particular kind of exchange: The borrower receives present goods (or purchasing power to buy goods at present) in exchange for paying the purchasing power at a future date. In this economy, there is an auctioneer who has the authority to enforce all contracts for all the contingencies, thus eliminating any failure of payment in the future. Therefore, an exchange between present goods and future goods in this market is not subject to any frictions and is no different from an exchange between two present goods. If, however, this enforcing auctioneer is absent in a decentralized market economy, then a borrower can default on his payment in the future. Anticipating the possibility of default, the creditor requires collateral for the loans and makes the amount of credit contingent on the value of the collateral. In order to analyze the business cycle in economies with such credit constraints, we assume that it takes one period to transform inputs into output. Instead of production function (2), we use

$$y_{jt+1} = f\left(k_{jt}, l_{jt}, n_{jt}; z_{jt}\right). \tag{7}$$

We assume that the maturity of all outstanding debt is one period and that the debt repayment in the next period $b_{jt}$ cannot exceed a fraction $\phi$ of the expected value of the collateral, which in this model we assume to be land:

$$b_{jt} \leq \phi E_t(q_{t+1} l_{jt}). \tag{8}$$

Here $q_{t+1}$ is the price of the land in period $t+1$ and $l_{jt}$ is the amount of land on collateral. We assume that the repayment amount is independent of the state of the borrower or the economy (i.e., the debt is noncontingent). So even though we rationalize the imposition of the borrowing constraint by

reference to the possibility of default, we assume that there is no default in equilibrium. The entrepreneur's budget constraint is

$$\begin{aligned} & c_{jt} + k_{jt} + q_t l_{jt} + w_t n_{jt} \\ = \; & \left[ y_{jt} + \left(1 - \psi_t\right)(1 - \delta)\, k_{jt-1} + q_t l_{jt-1} - b_{jt-1} \right] + \frac{b_{jt}}{r_t}. \end{aligned} \tag{9}$$

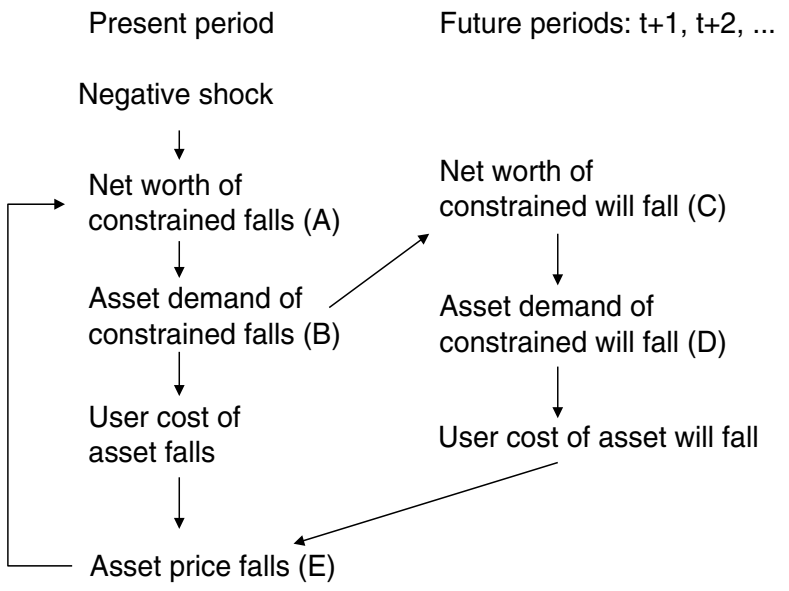
The left-hand side of the equation is the entrepreneur's expenditure on consumption (or dividend) and factors of production—capital, land, and labor. (We assume the entrepreneur must buy capital and land and cannot rent their services.) The right-hand side represents the firm's sources of finance where internal finance is in the square bracket—net worth that equals output plus undepreciated capital and land from the previous period net of repayment of old debt. The last term on the right-hand side is the external finance derived from new debt (calculated as the present value of next period's repayment on loans discounted by the gross real interest rate $r_t$). Each entrepreneur chooses a sequence of consumption, investment, output, and debt in order to maximize the discounted expected utility subject to the constraints of technology, credit, and available funds.

Now let us examine the difference between the RBC model and the economy in which producers are heterogeneous in productivity $z_{jt}$ and are credit constrained.[5] First, if there is limited contract enforcement, then insurance is incomplete. Because the insurance company is aware of the fact that the insurees may not pay in the future, the company, as a precautionary measure, demands premium payments upon entering an insurance contract. Thus, producers and households with low net worth may not purchase insurance with full coverage. When the economy is then hit by various shocks, the net worth of firms and households with partial insurance coverage fluctuates, which in turn requires an adjustment of the Negishi weights. As a result, we can no longer maintain the assumptions of the representative household approach. In addition, when borrowing constraints exist, firms must rely on internal finance—their net worth—as a source of financing inputs. When the borrowing constraint is binding for some firms, the marginal product of capital, land, and labor across firms will no longer be the same. Thus, the assumptions for the existence of a representative firm no longer hold, and an aggregate production function such as given by equation (4) no longer exists. Now, aggregate productivity of the economy will fluctuate endogenously with credit levels and net worth as emphasized in Kiyotaki and Moore (1997a) and Kiyotaki (1998).

When productive firms borrow up to the credit limit and also use their own net worth to finance additional investments that the loans could not cover, the

[5] See Bernanke and Gertler (1989), Kiyotaki and Moore (1997a, 1997b), Kiyotaki (1998), and Bernanke, Gertler, and Gilchrist (1999) for examples of RBC models with credit constraints.

**Figure 1 Credit Cycles**




impact of a small shock to total productivity, investment, and net worth is large. In order to explain the propagation of the effects of the shock, let's assume that net worth of all firms has declined because of the obsolescence of some of their products, and thus the capital used to produce those goods has also become obsolete. Because highly productive firms have outstanding debt from the previous period, the leverage effect of the debt will result in a sharp reduction of net worth (refer to Figure 1, point A). These productive firms will decrease their demand for capital and land because they cannot borrow more (Figure 1, point B) and aggregate productivity will fall as the share of investment of productive firms declines. Because it will take some time for the highly productive firms to recover their preshock level of net worth (Figure 1, point C), their demand for assets (capital and land) and labor will be constrained for a while and therefore aggregate productivity and aggregate demand for assets and investment are also expected to be stagnant for a while (Figure 1, point D). Under these expectations, current period asset prices drop (Figure 1, point E) and the balance sheets of the highly productive firms further deteriorate (Figure 1, point A). As a result, the small aggregate shock causes a persistent decrease of the share of investment by credit-constrained and highly productive producers, which leads to a persistent decline of aggregate productivity. Thus, with borrowing constraints, the fall in asset price is responsible for the magnified drop in output.

Joan Robinson once said, "the essence of Keynesian economics is its recognition of the central role of time in human lives. People live in the present moment which is continuously moving from an unknown future to the irrevocable past."[6] Because the demand for assets by productive firms that face credit constraints depends on each firm's own net worth (which equals accumulated past savings), it takes time for them to recover from a negative shock to net worth (i.e., the effects of shocks are persistent as firms are held back by their past savings). Meanwhile, asset prices are driven down by expectations of a prolonged stagnation in future asset demand (i.e., expectations about the future affect present asset prices). Notice how the asset market serves as a platform on which past savings and expectations about the future interact in present time.

In the real business cycle model with no constraints on borrowing, shocks from the obsolescence of goods and capital trigger higher investment, leading to a quick recovery of capital stock and output (dotted line in Figure 2.) In contrast, in an economy where borrowing constraints exist, the obsolescence shocks significantly reduce net worth and investment of productive firms, further decreasing capital stock and output. Since it takes time for highly productive, yet credit constrained, firms to recover their net worth and investment levels, total output and productivity will both fall persistently (solid line in Figure 2).

Reinhart and Rogoff (2009) claim that, although financial crises are much like forest fires in the sense that it is difficult to predict when and where they will occur, certain conditions set the stage for crises. They present the following as indicators of an emerging crisis for a country: 1) asset prices rise rapidly, and especially the price-rent ratio for real estate increases sharply; 2) the amount of debt expands faster than aggregate output and asset values, leading to a higher leverage (the ratio of total assets to net worth); and 3) the country experiences massive capital inflows. In the presence of borrowing constraints, firm heterogeneity in productivity, and diverse investment opportunities, if an adverse shock arrives when the overall leverage level of the economy is high enough, the powerful propagation mechanism as described in Figure 1 will take effect in asset prices, credit levels, and outputs, forcing the economy into financial crisis.[7]

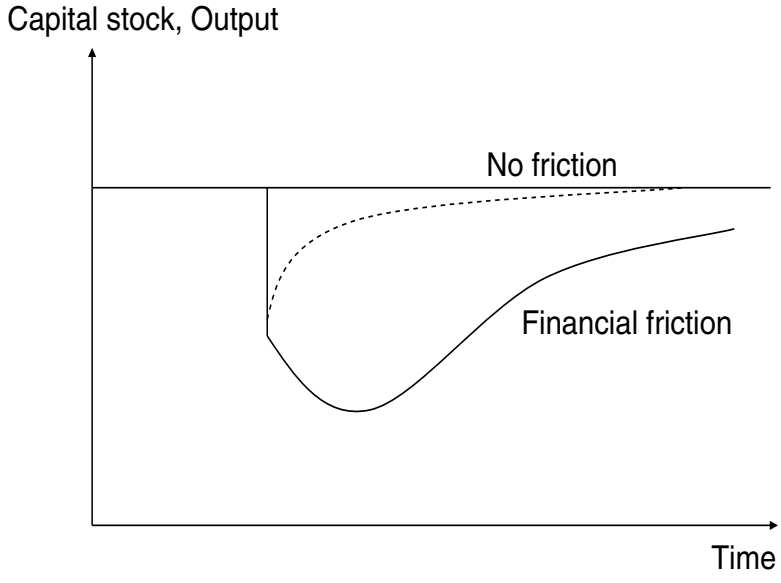
When a financial crisis is accompanied by a banking crisis, I expect that the financial system will cause additional problems that will aggravate the crisis.

---

[6] Introduction of the 1973 Japanese edition of Joan Robinson (1971).

[7] Kiyotaki and Moore (1997a) explain how total output, asset price, and debt fluctuate cyclically when exposed to an exogenous shock, while Matsuyama (2008) suggests that the fluctuation can occur even in the absence of a shock. Aoki, Benigno, and Kiyotaki (2009) distinguish domestic and foreign credit limits, and show that, if the domestic economy has an underdeveloped financial system, it becomes prone to both expansion and contraction after capital account liberalization, as Reinhart and Rogoff (2009) suggest.

**Figure 2 Impulse Response to Capital Quality Shock**




To analyze such crises, we need to look beyond the credit constraints of non-financial borrowers and consider the role of financial intermediaries and their financing constraints. Theories of financial intermediation have developed since Diamond and Dybvig (1983), and others such as Williamson (1987) have extended macroeconomic models to include banks. Although there is active recent research on the source of problems caused by financial intermediaries and their markets (especially "wholesale" or "interbank" financial markets), there is not yet a standard macroeconomic model for analysis of financial intermediation.[8]

In addition, note that it is the leverage effects from debt obligations that induce the net worth and investment of highly productive firms to persistently decline in the presence of borrowing constraints. If firms issue preferred stock or other securities whose returns are contingent on the firm performance instead of taking out loans to finance their investments, the leverage effects will not materialize. Therefore, in order to justify the propagation mechanism,

[8] Kiyotaki and Moore (1997b, 2008) analyze the effects of productivity and liquidity shocks on aggregate production in an economy where firms are involved in both production and financial intermediation. Gertler and Kiyotaki (2011) study the moral hazard problem of financial intermediaries, the relationship between their balance sheets and business cycles, and the effects of broad monetary policies. These articles also provide more references to the literature.

we need to first explain why firms would choose to borrow and not issue contingent securities in procuring their funds. We also need to explain why firms choose not to issue common stocks in order to recover net worth when it is deteriorating.

## 5. CONCLUSION

In this article, I explain that business cycles in an economy of heterogeneous firms and households can be analyzed using the representative agent approach if their interactions take place in an economy without frictions and complete markets. However, in an economy where markets do not function smoothly because of frictions such as credit constraints, the representative household framework may no longer be appropriate and aggregate productivity changes endogenously with the distribution of wealth and productivity of firms. Thus, I argue that the interaction of heterogeneous firms and households in the presence of credit constraints is important for business cycle analysis. Finally, I would like to propose some questions and directions for future research.

While in the presence of borrowing constraints, capital and land does not move between firms so that the marginal products of capital and land are not equalized across firms; the allocation of capital and land will gradually adjust in a similar way that water flows downhill. For example, firms with high marginal products of capital and land do not consume or pay out dividends in excess, and hence accumulate net worth. As a result, they will eventually be less constrained by external finance constraints. Even in an economy where capital and land do not move freely, if labor can move freely between firms, the marginal product of labor will be equalized across firms. Then, the marginal product of capital and land will also become more equal across firms. One suggestion for future research is to study how the distribution of productivity and net worth of firms evolves and how persistent the differences in marginal products of inputs across firms are.

In order to obtain a deeper understanding of the importance of firm heterogeneity, we need to analyze what determines and changes the productivity of individual firms. According to Bernard et al. (2003), firm labor productivity in an industry varies widely between less than one-fourth and more than four times the average labor productivity of the industry. Differences in human and physical capital can account for only a small fraction of these productivity differences among firms. In order to explain the diverse productivity across firms, we need to consider the accumulation process of both tangible as well as intangible assets. As modern growth theory has attempted to extend its models to include endogenous technical progress in addition to the accumulation of the factors of production, perhaps it is about time for modern business cycle theories to look into the source and the propagation of the shocks by

exploring the endogenous evolution of an individual firm's productivity in general equilibrium.

## REFERENCES

Andolfatto, David. 1996. "Business Cycles and Labor-Market Search." *American Economic Review* 86 (March): 112–32.

Aoki, Kosuke, Gianluca Benigno, and Nobuhiro Kiyotaki. 2009. "Capital Flows and Asset Prices." In *NBER International Seminar on Macroeconomics 2007*, edited by Richard Clarida and Francesco Giavazzii. Chicago: University of Chicago Press, 175–216.

Bernanke, Ben S., and Mark Gertler. 1989. "Agency Costs, Net Worth, and Business Fluctuations." *American Economic Review* 79 (March): 14–31.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." In *Handbook of Macroeconomics*, edited by John B. Taylor and Michael Woodford. Amsterdam: North-Holland; 1,341–93.

Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum. 2003. "Plants and Productivity in International Trade." *American Economic Review* 93 (September): 1,268–90.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91 (June): 401–19.

Gertler, Mark, and Nobuhiro Kiyotaki. 2011. "Financial Intermediation and Credit Policy in Business Cycle Analysis." In *Handbook of Monetary Economics*, 3(A), edited by Benjamin M. Friedman and Michael Woodford. Amsterdam: North-Holland, 547–99.

Hosios, Arthur J. 1990. "On the Efficiency of Matching and Related Models of Search and Unemployment." *Review of Economic Studies* 57 (April): 279–98.

Kiyotaki, Nobuhiro. 1998. "Credit and Business Cycles." *Japanese Economic Review* 49 (March): 18–35. Japanese translation in *Survey of Modern Economics 1998*, edited by M. Ohtsuki, K. Ogawa, K. Kamiya, and K. Nishimura. Tokyo: Toyo-Keizai Sjimpo-sha, 29–51.

Kiyotaki, Nobuhiro, and John Moore. 1997a. "Credit Cycles." *Journal of Political Economy* 105 (April): 211–48.

Kiyotaki, Nobuhiro, and John Moore. 1997b. "Credit Chains." Mimeo, London School of Economics.

Kiyotaki, Nobuhiro, and John Moore. 2008. "Liquidity, Business Cycles and Monetary Policy." Mimeo, London School of Economics and Princeton University.

Krugman, Paul. 2009. "How Did Economists Get So Wrong?" *New York Times Magazine*, 6 September, MM36.

Kydland, Finn E., and Edward C. Prescott. 1982. "Time to Build and Aggregate Fluctuations." *Econometrica* 50 (November): 1,345–70.

Matsuyama, Kiminori. 2008. "Aggregate Implications of Credit Market Imperfections." In *NBER Macroeconomic Annual 2007*, edited by Daron Acemoglu, Kenneth Rogoff, and Michael Woodford. Chicago: University of Chicago Press, 1–60.

Merz, Monika. 1995. "Search in the Labor Market and the Real Business Cycle." *Journal of Monetary Economics* 36 (November): 269–300.

Mortensen, Dale T., and Christopher A. Pissarides. 1994. "Job Creation and Job Destruction in the Theory of Unemployment." *Review of Economic Studies* 61 (July): 397–415.

Negishi, Takashi. 1960. "Welfare Economics and Existence of an Equilibrium for a Competitive Economy." *Metroeconomica* 12 (June): 92–7.

Prescott, Edward C. 1986. "Theory Ahead of Business Cycle Measurement." *Carnegie-Rochester Conference Series on Public Policy* 25 (January): 11–44. Reprinted in the Federal Reserve Bank of Minneapolis *Quarterly Review* 10 (Fall): 9–22.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. *This Time is Different*. Princeton, N. J.: Princeton University Press.

Robinson, Joan. 1971. *Economic Heresies*. New York: Basic Books. Translated into Japanese by Hirofumi Uzawa (1973). Tokyo: Nihon Keizai Shin-bunsha.

Shimer, Robert. 2005. "The Cyclical Behavior of Equilibrium Unemployment and Vacancies." *American Economic Review* 95 (March): 25–49.

Williamson, Stephen D. 1987. "Financial Intermediation, Business Failures, and Real Business Cycles." *Journal of Political Economy* 95 (December): 1,196–216.

*Economic Quarterly—Volume 97, Number 3—Third Quarter 2011—Pages 209–254*

# Financial Frictions in Macroeconomic Fluctuations

Vincenzo Quadrini

The financial crisis that developed starting in the summer of 2007 has made it clear that macroeconomic models need to allocate a more prominent role to the financial sector for understanding the dynamics of the business cycle. Contrary to what has been often reported in popular press, there is a long and well-established tradition in macroeconomics of adding financial market frictions in standard macroeconomic models and showing the importance of the financial sector for business cycle fluctuations. Bernanke and Gertler (1989) is one of the earliest studies. Kiyotaki and Moore (1997) provide another possible approach to incorporating financial frictions in a general equilibrium model. These two contributions are now the classic references for most of the work done in this area during the last 25 years.

Although these studies had an impact in the academic field, formal macroeconomic models used in policy circles have mostly developed while ignoring this branch of economic research. Until recently, the dominant structural model used for analyzing monetary policy was based on the New Keynesian paradigm. There are many versions of this model that incorporate several frictions such as sticky prices, sticky wages, adjustment costs in investment, capital utilization, and various types of shocks. However, the majority of these models are based on the assumption that markets are complete and, therefore, there are no financial market frictions. After the financial crisis hit, it became apparent that these models were missing something crucial about the behavior of the macroeconomy. Since then there have been many attempts

I am indebted to Andreas Hornstein and Felipe Schwartzman for very detailed and insightful suggestions that resulted in a much improved version of this article. Of course, I am the only one responsible for possible remaining errors and imprecisions. The opinions expressed in this article do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: quadrini@usc.edu.

to incorporate financial market frictions in otherwise standard macroeconomic models. What I would like to stress here is that the recent approaches are not new in macroeconomics. They are based on ideas already formalized in the macroeconomic field during the last two and a half decades, starting with the work of Bernanke and Gertler (1989). In this article I provide a systematic description of these ideas.

## 1. WHY MODELING FRICTIONS IN FINANCIAL MARKETS?

Before adding complexity to the model, we would like to understand why it is desirable to have meaningful financial markets in macroeconomic models, besides the obvious observation that they seem to have played an important role in the recent crisis. One motivating observation is that the flows of credit are highly pro-cyclical. As shown in the top panel of Figure 1, the change in credit market liabilities moves closely with the cycle. In particular, debt growth drops significantly during recessions. The only exception is perhaps for the household sector in the 2001 recession. However, the growth in debt for the business sector also declined in 2001. Especially sizable is the drop in the most recent recession. The pro-cyclicality of corporate debt is also shown in Covas and Den Haan (2011) using Compustat data.

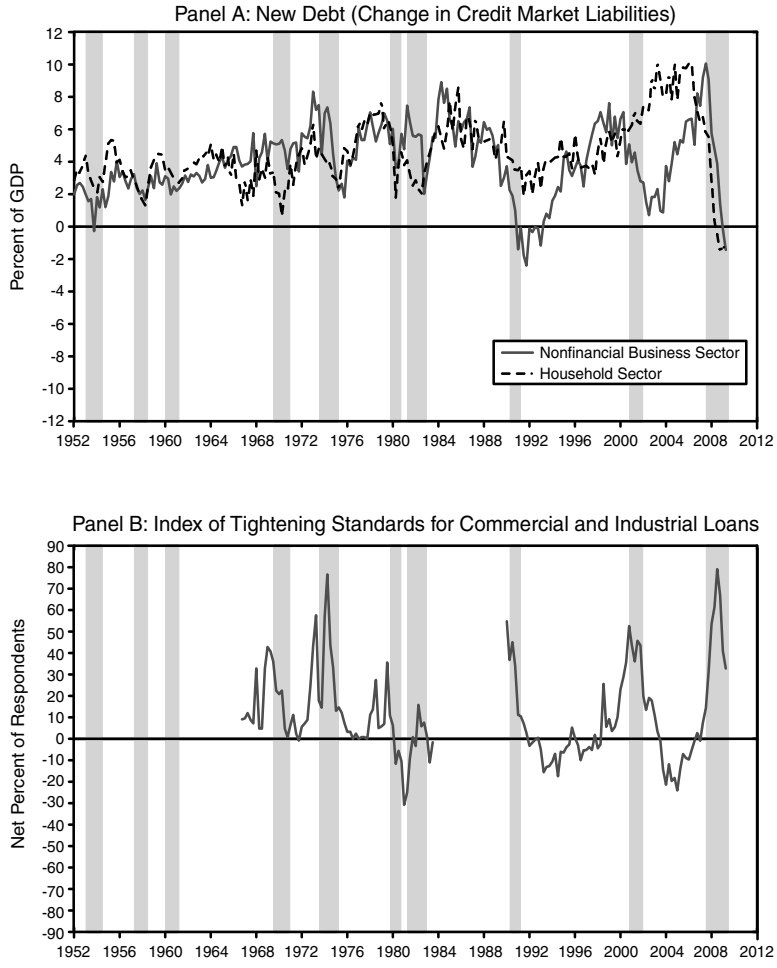
The cyclical properties of financial markets can be seen not only by the aggregate dynamics of credit flows (as shown in the top panel of Figure 1), but also by indicators of tightening credit standards. The bottom panel of Figure 1 plots the net fraction of senior bank managers reporting tightening credit standards for commercial and industrial loans in a survey conducted by the Federal Reserve Board. Clearly, more and more banks tighten their credit standard during recessions. Other indicators of credit tightening such as credit spreads, that is, interest rate differentials between bonds with differing ratings, convey a similar message as shown in Gilchrist, Yankov, and Zakrajsek (2009).

If markets were complete, the financial structure of individual agents, being households, firms, or financial intermediaries, would be indeterminate. We would then be in a Modigliani and Miller (1958) world and there would not be reasons for the financial flows to follow a cyclical pattern. However, the fact that credit flows are highly pro-cyclical and the index of tightening standards is countercyclical suggests that the complete-market paradigm has some limitations. This is especially true for the index of credit tightening.[1]

Of course, Figure 1 does not tell us whether it is the macroeconomic recession that causes the contraction in credit growth or the credit contraction

[1] Although the pro-cyclicality of financial flows does not contradict Modigliani and Miller since the financial structure is indeterminate, when markets are complete there is no reason for lenders to change their "credit standards" over the business cycle. Here I interpret the index of credit standards as reflecting the characteristics of an individual borrower that are required to receive a loan. So it is something additional to the market-clearing risk-free interest rate.

**Figure 1 Debt and Credit Market Conditions**




Notes: Panel A shows change in the volume of credit market instruments in the households and business sector divided by gross domestic product. The data are from the Flow of Funds of the Federal Reserve Board. Panel B shows the index of credit tightening in commercial and industrial loans. The data are from the Senior Loan Officer Opinion Survey on Bank Lending Practices from the Federal Reserve Board. The survey was not conducted from 1984–1990. I thank Egon Zakrajsek for making available the historical data from 1967–1983.

that causes or amplifies the macroeconomic recession. It would then be convenient to distinguish three possible channels linking financial flows to real economic activity.

1. *Real activity causes movements in financial flows.* One hypothesis is that investment and employment respond to changes in real factors such as movements in productivity. In this case, borrowers cut their debt simply because they need less funds to conduct economic transactions. If this was the only linkage between real and financial flows, the explicit modeling of the financial sector would be of limited relevance for understanding movements in real economic activities.

2. *Amplification.* The second hypothesis is that the initial driving force of movements in economic activities are nonfinancial factors such as drops in productivity or monetary policy shocks. However, as investment and employment fall, the credit ability of borrowers deteriorates more than the financing need after the drop in economic activity. This could happen, for instance, if the fall in investment generates a fall in the market value of assets used as collateral. The presence of financial frictions will then generate a larger decline in investment and employment compared to the decline we would observe in absence of financial frictions. Therefore, financial frictions *amplify* the macroeconomic impact of the exogenous changes.

3. *Financial shocks.* A third hypothesis is that the initial disruption arises in the financial sector of the economy. There are no initial changes in the nonfinancial sector. Because of the disruption in financial markets, fewer funds can be channeled from lenders to borrowers. As a result of the credit tightening, borrowers cut on spending and hiring, and this generates a recession. I will refer to these types of exogenous changes as "credit" or "financial" shocks.

Most of the literature in dynamic macrofinance has focused on the second channel, that is, on the "amplification" mechanism generated by financial market frictions. More specifically, the central hypothesis is that financial frictions "exacerbate" a recession but are not the "cause" of the recession. Something wrong (a negative shock) first happens in the nonfinancial sector. This could be caused by "exogenous" changes in productivity, monetary aggregates, interest rates, preferences, etc. These shocks would generate a macroeconomic recession even in absence of financial market frictions. With financial frictions, however, the magnitude of the recession becomes much bigger.

The third channel, that is, the analysis of financial shocks as a "source" of business cycle fluctuations, has received less attention in the literature. More recently, however, a few studies have explored this possibility. In this article I will present the main theoretical ideas about the second and third channels, that is, "amplification" and "financial shocks." I will not focus on the first hypothesis only because, as observed above, if this was the most relevant channel of linkage between real and financial flows, the explicit modeling of the

financial sector would be of limited relevance for understanding movements in real macroeconomic activities.

## 2. MODELING FINANCIAL FRICTIONS

Technically, financial frictions emerge when trade in certain assets cannot take place. In an Arrow-Debreu world with state-contingent trades, markets for some contingencies are missing and, therefore, there is a limit to the feasible range of intertemporal and intratemporal trades. In practical terms this implies that agents are unable to anticipate or postpone spending (for consumption or investment) or insure against uncertain events (to smooth consumption or investment). Of course, this becomes relevant only if agents are heterogeneous. Therefore, any models with financial frictions share the following features:

1. *Missing markets:* Some asset trades are not available or feasible.

2. *Heterogeneity:* Agents are heterogeneous in some important dimension.

I should clarify that these two features are necessary but not sufficient for incomplete markets to play an important role. That we need heterogeneity is obvious. If all agents are homogeneous, there is no reason to trade claims intertemporally or intratemporally. So the fact that some markets are missing becomes irrelevant. Also, if agents could trade any type of contingency, we would have an economy with complete markets. On the other hand, the fact that some markets are missing may be irrelevant if in equilibrium agents choose voluntarily not to trade in these markets. Therefore, market incompleteness and heterogeneity must take specific configurations. In the next two subsections, I will describe first the most common approaches to modeling missing markets and then I will discuss the most common approaches used to generate heterogeneity.

### Missing Markets

The approaches used to model missing markets can be divided into two categories: "exogenous" market incompleteness and "endogenous" market incompleteness.

1. *Exogenous market incompleteness.* The first category includes models that impose exogenously that certain assets cannot be traded. For example, it is common to assume that agents can hold bonds (issue debt if negative) but they cannot hold assets with payoffs contingent on information that becomes available in the future. This approach does not attempt to explain why certain assets cannot be traded but it takes a

more pragmatic approach. Since a large volume of financing observed in the real economy is in the form of standard debt contracts, while the volume of state contingent contracts is limited, it makes sense to assume that debt contracts are the only financial instruments that are available. A further restriction, which is also exogenously imposed, is that the total amount of debt cannot exceed a certain limit (exogenous borrowing constraint). Of course, the goal of this literature is not to explain why markets are incomplete but to understand the consequences of market incompleteness.

2. *Endogenous market incompleteness.* The second category includes models in which the set of feasible contracts are derived from agency problems. The idea is that markets are missing because parties are not willing to engage in certain trades because they are not enforceable or incentive-compatible. What this means is that the borrower is unable to borrow or insure against the risk because, with high liabilities and full insurance, he or she would act against the interests of the lender. Typically, endogenous market incompleteness is derived from two agency problems:

   (a) *Limited enforcement.* The idea of limited enforcement is that the lender is fully capable of observing whether or not the borrower is fulfilling his or her contractual obligations. However, there are no tools the lender can use to enforce the contractual obligations. For example, even if the lender knows that the borrower is not exerting effort or is diverting funds, it may be difficult to prove it in court. There could also be legal limits to what the lender can enforce. For example the law does not allow the lender to force the borrower to work in order to repay the debt (no slavery).

   (b) *Information asymmetry.* Information asymmetries also limit the ability of lenders to force the borrowers to fulfil their obligations. In this case, the limit derives from the inability to observe the borrower's action. For example, if the repayment depends on the performance of the business and the performance depends on unobservable effort, the borrower may have an incentive to choose low effort.

   From a technical point of view, models with limited enforcement are typically easier to analyze than models with information asymmetries. Both models, however, share a common property: higher is the net worth of borrowers and higher is the (incentive-compatible) financing that can be raised externally—a recurrent factor in the theoretical analysis that will be conducted in the remaining sections of this article.

**Heterogeneity**

There are many approaches used in the literature to generate heterogeneity. One popular approach is that agents are ex-ante identical but they are subject to idiosyncratic shocks. Therefore, the heterogeneity derives from the assumption that at any point in time each agent receives a different shock. For example, in the Bewley (1986) economy agents receive stochastic endowments. Because at any point in time there are agents with low endowments while others have high endowments, it will be optimal to sign state-contingent contracts that insure the endowment risks and allow for consumption smoothing. With these contracts, agents receive payments when their endowments are low and make payments when their endowments are high.

If markets were complete, the analysis of this model would be simple. With incomplete markets, however, the model generates high dimensional heterogeneity. Even if agents are initially or ex-ante homogeneous, in the long run there will be a continuum of asset holdings. Because the state dimensionality makes the characterization of the equilibrium challenging, the majority of applications of Bewley-type economies have abstracted from aggregate uncertainty and business cycle fluctuations. An exception is Krusell and Smith (1998). Other exceptions include Cooley, Marimon, and Quadrini (2004), where the heterogeneity is on the production side and, more recently, Guerrieri and Lorenzoni (2010) and Khan and Thomas (2011). In general, however, the majority of studies investigating the importance of financial frictions for macroeconomic fluctuations have tried alternative approaches that keep the degree of heterogeneity small.

A common approach is to assume that there are only two types of agents with permanent differences in preferences and/or technology. In equilibrium one agent ends up being the *borrower* and the other the *lender*. Alternatively, there could be a continuum of heterogeneous agents but their aggregate behavior can be characterized by a single representative agent thanks to linear aggregation. This is the case, for example, in Carlstrom and Fuerst (1997); Bernanke, Gertler, and Gilchrist (1999); and Miao and Wang (2010). Although entrepreneurs face uninsurable idiosyncratic risks and there is a distribution of entrepreneurs over net worth, the linearity of technology and preferences allows for the derivation of aggregate policies that are independent of the distribution. So, effectively, the reduced form in these models is also characterized by only two representative agents: households/workers and entrepreneurs.

Still, the fact that firms are owned by entrepreneurs and external financing is limited is not enough for financial frictions to play an important role. Even if entrepreneurs (firms) are temporarily financially constrained, that is, they would like to borrow more than they are allowed, over time they could save enough resources to make the financial constraints nonbinding. Therefore, further assumptions need to be made in order for the borrowing constraints to also be relevant in the long run. This is achieved in different ways.

1. *Finite life span.* A common modeling approach is based on the assumption that borrowers have a finite life span. For example, in overlapping generations models, it is commonly assumed that newborn agents have no initial assets and, therefore, they are financially constrained in the first stage of their lives. Over time agents accumulate assets and become unconstrained. However, since there is a continuous entrance of newborns, at any point in time there are always some agents who face binding financial constraints. A similar idea is applied in industry dynamics models where exiting firms are replaced by new entrant firms.

2. *Different discounting.* Another common approach is to assume that borrowers are infinitely lived but they discount the future more heavily than lenders. What this implies is that the cost of external financing is lower than the cost of internal funds. As a result, debt is preferred to internal funds. This insures that borrowers do not save enough to make the borrowing constraint irrelevant. Then, unanticipated shocks could lead to a larger spending response of borrowers because of the binding constraint.

3. *Tax benefits.* A similar approach to the differential discounting is the assumption that there are tax benefits of debt. For example, the tax deductibility of interest payments from corporate earnings generates a preference for debt over equity, and corporations tend to leverage up. However, if the firm is unexpectedly required to de-leverage and it is difficult to replace debt with equity in the short term, the result could be large drops in investment and employment.

4. *Bargaining position.* A further assumption proposed in the literature is that external financing (debt/outside equity) is preferred to inside financing (entrepreneurial equity), not because of differential discounting or tax benefits, but because it affects the bargaining position of firms in the negotiation of wages and/or executive compensation. The idea is that, if the compensation of workers and managers is determined through bargaining (in the case of workers the bargaining could be with unions), high-leveraged firms would be able to bargain lower compensations simply because the bargaining surplus is reduced by the debt.

But independent of the particular modeling approach, all models with financial market frictions are characterized by the presence of at least two groups of agents—one group that would like to raise external funds and one group that provides at least some of the funds.

## 3. A SIMPLE THEORETICAL FRAMEWORK

The discussion conducted so far has provided an informal description of the basic features of models used to study the importance of financial market frictions for the business cycle. Now I provide a more analytical description using a formal model that is rich enough to capture the various ideas proposed in the literature but remains analytically tractable.

To achieve this goal, I assume that there are only two periods—period 1 and period 2—and two types of agents—a unit mass of workers and a unit mass of entrepreneurs. Variables that refer to period 2 will be indicated with a prime superscript.

The lifetime utility of workers is

$$E\left\{c - \frac{h^2}{2} + \delta c'\right\},$$

where $c$ and $h$ are consumption and labor in period 1 and $c'$ is consumption in period 2. The lifetime utility of entrepreneurs is

$$E\Big\{c + \beta c'\Big\}.$$

Thus, entrepreneurs' utility is also linear in consumption but there is no disutility from working. The assumption of risk neutrality is not essential but it simplifies the analysis. When relevant, I will comment on the importance of risk neutrality.

I now describe what happens in each of the two periods.

- *Period 1.* Entrepreneurs enter period 1 with capital $K$ and debt $B$ owed to workers. In principle, $B$ could be negative. However, as we will see, this case is not of theoretical interest.

  There are two production stages during the first period. In the first stage, intermediate goods are produced with capital and labor. In the second stage, the intermediate goods are used as inputs in the production of consumption and new capital goods.

  - *Stage 1: Production of intermediate goods.* Intermediate goods are produced by entrepreneurs with the production function

    $$y = AK^{\theta}h^{1-\theta},$$

    where $A$ is the aggregate level of productivity, $K$ is the input of capital, and $h$ is the input of labor supplied by workers.

  - *Stage 2: Production of final goods.* In this stage, intermediate goods are used as inputs in the production of consumption and new capital goods. The transformation in consumption goods is simple: One unit of intermediate goods is transformed into one

unit of consumption goods. New capital goods are produced by individual entrepreneurs using the technology

$$k^n = \omega i,$$

where $i$ is the quantity of intermediate goods used in the production of new capital goods and $\omega$ is the idiosyncratic productivity realized after the choice of $i$. The cumulative density function is denoted by $\Phi(\omega)$. Later we will consider two cases: $E\omega = 1$ and $E\omega = 0$. In the second case there is no production of investment goods and, therefore, the aggregate stock of capital in period 2 is the same as in period 1.

- *Period 2.* Second period production takes place only with the input of capital. Since this is the terminal period, only consumption goods are produced. There are two sectors of production.

  - *Sector 1: Entrepreneurial sector.* This is composed of firms owned by individual entrepreneurs with technology $y' = A'k'$, where $k'$ is the input of capital acquired by the entrepreneur in period 1.
  - *Sector 2: Residual sector.* The second sector is formed by frictionless firms directly owned by workers with technology $y' = A'G(k')$. The function $G(.)$ is strictly increasing and concave and satisfies $G'(0) = 1$.

  The key difference between the entrepreneurial sector and the residual sector is that the former is more productive than the latter, that is, $G'(k') < 1$ for $k' > 0$. As we will see, in absence of financial frictions, production will take place only in the entrepreneurial sector. With frictions, part of the production could also take place in the less productive but frictionless residual sector. For simplicity I assume that $A'$ is known in period 1 and, therefore, there is no aggregate uncertainty.

Before proceeding I impose the following conditions:

**Assumption 1** *Entrepreneurs and workers have the same discounting, $\delta = \beta$. Furthermore, $\beta A' > 1$.*

It is often assumed in the literature that $\delta > \beta$, that is, entrepreneurs (borrowers) are more impatient than workers (lenders). This is an important assumption in an infinite horizon model. With only two periods, however, the discount differential does not play an important role, which motivates the assumption $\delta = \beta$. The condition $\beta A' > 1$, instead, guarantees that postponing consumption through investment is efficient since the discounted value of the productivity of capital in period 2 is greater than 1.

### Timing Summary

The structure of the model described so far, although stylized, is fairly complex. There are important timing assumptions that are made to keep the model analytically tractable. To make sure that these assumptions are clear, it would be helpful to summarize the timing sequence.

1. Entrepreneurs start period 1 with capital $K$ and debt $B$. Workers start with wealth $B$.
2. Entrepreneurs hire workers to produce intermediate goods with the technology $y = AK^{\theta}h^{1-\theta}$. The labor market is competitive and clears at the wage rate $w$.
3. Entrepreneurs purchase intermediate goods $i$ to produce new capital goods using the technology $k^n = \omega i$. The choice of $i$ is made before observing the idiosyncratic productivity $\omega$.
4. At this point we are at the end of period 1. The idiosyncratic productivities are observed and all incomes are realized. Entrepreneurs and workers allocate their end-of-period wealth between current consumption and savings in the form of capital goods and/or financial instruments (bonds).
5. We are now in period 2. Production takes place with the capital inputs accumulated in the previous period.
6. Entrepreneurs repay the debt to workers and both agents consume their residual wealth.

### Plan for the Theoretical Analysis

I have now completed the description of preferences, technology, and timing. What is left to describe are the financial frictions that impose additional constraints on the choices of debt. These will be specified in the analysis of the various cases reviewed in this article. The presentation will be organized in four main sections:

- Section 4 characterizes the equilibrium in the frictionless model. This provides the baseline framework to which I compare the various versions of the model with financial frictions.
- Section 5 presents the costly state verification model based on information asymmetries where the financial frictions have a direct impact on investment.
- Section 6 presents the collateral/limited enforcement model. I first show the properties of this model when the frictions have a direct impact

only on investment. I then extend the analysis to the case in which the frictions also have a direct impact on the demand of labor.

- Section 7 analyzes the impact of credit shocks. I first present the model with exogenous credit shocks and then I propose one possible approach to make these shocks endogenous through a liquidity channel. In this section I also show the importance of credit shocks in an open economy framework.

## 4. BASELINE MODEL WITHOUT FINANCIAL FRICTIONS

I start with the characterization of the problem solved by workers

$$\begin{aligned} \max_{c,c',k',b'} \quad & \left\{ c - \frac{h^2}{2} + \delta c' \right\} \\ & \text{subject to:} \\ & B + wh = c + \frac{b'}{R} + qk' \\ & A'G(k') + b' = c' \\ & c \geq 0, \quad c' \geq 0, \end{aligned} \tag{1}$$

where $B$ is the initial ownership of bonds, $R$ is the gross interest rate, $w$ is the wage rate, and $q$ is the price of capital. Since $A'$ is known in period 1, workers do not face any uncertainty.

The first two constraints are the budget constraints in period 1 and 2, respectively. They equalize the available resources (left-hand side) to the expenditures (right-hand side). The problem is also subject to the non-negativity of consumption in both periods. However, thanks to Assumption 1, $c'$ will always be positive and we have to worry only about the non-negativity of consumption in period 1. Intuitively, since capital is very productive in period 2 and preferences are linear, agents may choose to maximize their savings in period 1.

The first-order conditions are

$$h = w(1+\lambda) \tag{2}$$

$$(1+\lambda)q = \delta A'G'(k') \tag{3}$$

$$1+\lambda = \delta R, \tag{4}$$

where $\lambda$ is the Lagrange multiplier associated with the non-negativity constraint on consumption in period 1.

The problem solved by entrepreneurs can be written as

$$\max_{h,i,c,k',b',c'} \quad E\left\{c + \beta c'\right\}$$

subject to:

$$AK^{\theta}h^{1-\theta} - wh + qK + (qE\omega - 1)i + \frac{b'}{R} = B + c + qk'$$

$$A'k' = b' + c'$$

$$c \geq 0, \quad c' \geq 0, \quad i \geq 0.$$

The first two constraints are the budget constraints in period 1 and 2, respectively. They equalize the available resources (left-hand side) to the expenditures (right-hand side). The terms $AK^{\theta}h^{1-\theta} - wh$ and $(qE\omega - 1)i$ are, respectively, the profit earned by the entrepreneur in the production of intermediate goods and the (expected) profit earned in the production of new capital goods.

As for workers, I do not have to worry about the non-negativity constraint on $c'$. The first-order conditions are

$$w = (1-\theta)AK^{\theta}h^{-\theta} \tag{5}$$

$$qE\omega = 1 \leq 1, \qquad (= \text{if } i > 0) \tag{6}$$

$$(1+\gamma)q = \beta A' \tag{7}$$

$$1 + \gamma = \beta R, \tag{8}$$

where $\gamma$ is the Lagrange multiplier on the non-negativity constraint on consumption in period 1. Since $\delta = \beta$ (by Assumption 1), equations (4) and (8) imply $\lambda = \gamma$. What this means is that the non-negativity of consumption in period 1 is either binding for both agents or it is not binding for both of them.

Substituting the labor supply (2) in the demand of labor (5), we get the wage equation

$$w = (1-\theta)^{\frac{1}{1+\theta}} A^{\frac{1}{1+\theta}} K^{\frac{\theta}{1+\theta}} (1+\lambda)^{\frac{-\theta}{1+\theta}}. \tag{9}$$

Substituting back in the supply of labor, working hours can be expressed as

$$h = (1-\theta)^{\frac{1}{1+\theta}} A^{\frac{1}{1+\theta}} K^{\frac{\theta}{1+\theta}} (1+\lambda)^{\frac{1}{1+\theta}}. \tag{10}$$

Entrepreneurs' income in period 1, after the production of intermediate goods is

$$Y^e = AK^{\theta}h^{1-\theta} - wh, \tag{11}$$

where $w$ and $h$ are determined in (9) and (10). Therefore, the supply of labor and entrepreneurial income depend on the multiplier $\lambda$. The value of this variable depends on the assumption about $E\omega$. When I introduce financial frictions I will consider two cases: $E\omega = 1$ and $E\omega = 0$. The first case defines an economy with capital accumulation while the second case defines an economy with fixed capital.

- *Case 1:* $E\omega = 1$. Because $\beta A' > 1$, the intermediate goods produced in period 1 are all used in the production of capital goods. Thus, current consumption is zero for both entrepreneurs and workers. This implies that the multiplier $\lambda = \gamma$ is positive. Since investment $i$ is positive, condition (6) is satisfied with equality, and therefore, $q = 1$. Then equations (7) and (8) imply that $R = A'$. Agents anticipate that the productivity of capital is high next period and it becomes convenient to save the whole income to take advantage of the higher return. The labor supply is higher than the wage since $\lambda = \gamma > 0$ (see equation [2]) and the demand of labor is determined by its marginal product (see equation [5]). The whole capital produced in period 1 is accumulated by entrepreneurs since the entrepreneurial sector is more productive than the residual sector and there are no agency problems in the repayment of the intertemporal debt $b'$.

  It is now easy to see the impact of productivity changes. An increase in current productivity $A$ generates an increase in the supply of labor and output as we can see from equations (9)–(11), after replacing $1 + \gamma = \beta A'$ from equation (7), taking into account that $\lambda = \gamma$. Since the increase in income is saved, the productivity boom also generates an investment boom. There is no impact in current consumption but this is a consequence of assuming risk neutrality. With risk aversion, consumption in period 1 is also likely to increase in response to a persistent productivity improvement.

  An increase in $A'$ also generates an increase in the current supply of labor (see equation [10] after substituting $1 + \gamma = \beta A'$), which in turn generates an increase in output and savings. Therefore, the model has the typical properties of the neoclassical business cycle model.

- *Case 2:* $E\omega = 0$. Since $E\omega = 0$, we can see from equation (6) that $i = 0$, that is, there is no capital accumulation. The whole capital $K$ is acquired by entrepreneurs because the entrepreneurial sector is more productive than the residual sector and there are no agency problems in the repayment of the intertemporal debt $b'$. Since consumption cannot be zero for both workers and entrepreneurs, $\lambda = \gamma = 0$ (in absence of investment aggregate consumption in period 1 must be equal to aggregate production in period 1). This implies that the price of capital is $q = \beta A'$ (see equations [3] and [7]). In this way both agents are indifferent between current and future consumption and the new debt $b'$ is undetermined.

  An increase in current productivity $A$ generates an increase in the supply of labor and output as we can see from (9)–(11) after substituting $\lambda = 0$. However, the productivity change in period 1 does not affect next period production since there is no capital accumulation. Similarly,

an increase in $A'$ generates an increase in next period production but it does not have any impact on production in period 1. Again, this is because of the absence of capital accumulation. As we will see, this feature of the model will change with financial frictions.

## 5. COSTLY STATE VERIFICATION MODEL

In the costly state verification model frictions derive from information asymmetry. This is the centerpiece of the financial accelerator model proposed by Bernanke and Gertler (1989). The model has been further embedded in more complex macroeconomic models with infinitely lived agents by Carlstrom and Fuerst (1997) and Bernanke, Gertler, and Gilchrist (1999).

To illustrate the key elements of the financial accelerator, I specialize the analysis to the case in which frictions are only in the production of capital goods and, in the analysis of this section, I assume that $E\omega = 1$. This guarantees that capital goods are produced and there is capital accumulation in the model.

The frictions derive from the assumption that $\omega$ is freely observable only by entrepreneurs. Other agents could observe $\omega$ but only at the cost $\mu i$. This limits the feasibility of financial contracts that are contingent on $\omega$. As it is well known from the work of Townsend (1979), the optimal contract takes the form of a standard debt contract in which the entrepreneur promises to repay an amount that is independent of the realization of $\omega$. If the entrepreneur does not repay, the lender incurs the verification cost and confiscates the residual assets.

### Optimal Contract with Costly State Verification

The central element of this model is the net worth of entrepreneurs. Before starting the production of new capital goods, entrepreneurs' net worth is $n = qK + Y^e - B$, where $Y^e$ is defined in (11). Therefore, if the entrepreneur purchases $i$ units of intermediate goods, he or she has to borrow $i - n$ units of intermediate goods on the promise to pay back $(i - n)(1 + r^k)$ units of capital goods. Notice that the interest rate $r^k$ is denominated in capital goods, which explains the different denomination of the loan (denominated in intermediate goods) and the repayment (in capital goods). The particular choice of the denomination is a simple convention that is inconsequential for the properties of the model.

After the realization of the idiosyncratic shock $\omega$, the entrepreneur defaults only if the production of new capital goods is smaller than the debt repayment, that is, $\omega i \leq (1 + r^k)(i - n)$. We can then define $\bar{\omega}$ as the shock below which

the entrepreneur defaults, which is equal to

$$\bar{\omega} = (1 + r^k)\left(\frac{i - n}{i}\right).$$

This equation makes clear that the default threshold is increasing in the leverage ratio $\frac{i-n}{i}$ and in the interest rate. Assuming competition in financial markets, the interest rate charged by the lender must satisfy the zero-profit condition

$$q\left[\int_0^{\bar{\omega}(n,i,r^k)} (\omega - \mu) i \Phi(d\omega) + \int_{\bar{\omega}(n,i,r^k)}^{\infty} (1 + r^k)(i - n)\Phi(d\omega)\right] = i - n.$$

Notice that there is no interest in the cost of funds on the right-hand side of the equation since the loan is intraperiod, that is, issued and repaid in the same period. This is different from the intertemporal debt $b'$. The equation defines implicitly the interest rate charged by the bank as a function of $n, i, q$, which I denote as $r^k(n, i, q)$. The default threshold can also be expressed as a function of the same variables, that is, $\bar{\omega}(n, i, q)$.

Since entrepreneurs are risk neutral, the production choice is independent of the consumption/saving decision. More specifically, the optimal choice of $i$ maximizes the expected entrepreneur's net worth, that is,

$$\max_i q \int_{\bar{\omega}(n,i,q)}^{\infty} \Big[\omega i - (1 + r^k(n, i, q))(i - n)\Big]\Phi(d\omega).$$

Notice that the integral starts at $\bar{\omega}$ because the entrepreneur defaults for values of $\omega < \bar{\omega}$ and the ex-post net worth is zero in the event of default.

Let $i(n, q)$ be the optimal scale chosen by the entrepreneur in the production of capital goods. We can define the net worth after production as

$$\pi(n, q, \omega) = \max\left\{0\ ,\ q\Big[\omega i(n, q) - (1 + r^k(n, i(n, q), q))(i(n, q) - n)\Big]\right\}.$$

Using this function, the consumption/saving problem solved by the entrepreneur can be written as

$$\begin{aligned} \max_{c,c',k',b'} \quad & \Big\{c + \beta c'\Big\} \\ & \text{subject to:} \\ & \pi(n, q, \omega) = c + qk' - \frac{b'}{R} \\ & c' = A'k' - b' \\ & c \geq 0, \quad c' \geq 0. \end{aligned}$$

**Equilibrium and Response to Productivity Shocks**

There are two possible equilibria depending on the net worth of entrepreneurs. In the first equilibrium, the net worth of entrepreneurs is sufficiently large that

the whole production of intermediate goods is used in the production of new capital goods. This case is similar to the baseline model without financial frictions characterized in Section 4.

The second type of equilibrium arises when the net worth of entrepreneurs is not large enough to use the whole production of intermediate goods to produce new capital goods. We have defined above $i(n, q)$ the production scale of entrepreneurs, that is, the demand of intermediate goods used in the production of new capital goods. Since there is a unit mass of entrepreneurs that are initially homogeneous, $i(n, q)$ is also aggregate investment. If $i(n, q) < AK^{\theta}h^{1-\theta}$, then only part of the production of intermediate goods is used in the production of capital goods. This implies that the consumption in period 1 of workers and/or entrepreneurs is positive. Thus, the multiplier associated with the non-negativity of consumption is $\gamma = 0$ and the equilibrium satisfies the first-order conditions

$$\begin{aligned} q &= \beta A' \\ 1 &= \beta R. \end{aligned}$$

Thus, the price of capital is equal to $\beta A'$, which is bigger than one by Assumption 1. I will focus on this particular equilibrium since this is when financial frictions matter.

I can now study the response of the economy to productivity shocks, that is, changes in $A$ and $A'$.

- *Increase in A.* The increase in $A$ raises the net worth of entrepreneurs $n = qK + Y^e - B$, where $Y^e$ is defined in (11). Since $q = \beta A'$, the price of capital $q$ does not change if $A'$ does not change. Therefore, the increase in net worth is only determined by the increase in capital income $Y^e$ earned in the first stage of production.

  The next step is to see what happens to investment in response to the higher net worth. We have already seen that investment $i$ increases with $n$. Therefore, the productivity improvement generates an investment boom and increases next period production. In this way the model generates a persistent impact of productivity shocks. This effect, however, is not necessarily bigger than the effects of a productivity shock in the baseline model without frictions characterized in Section 4. For this to be the case, the net worth $n$ has to increase proportionally more than the increase in output. This requires $qK - B < 0$, which is unlikely to be an empirically relevant condition. Therefore, the model with financial frictions could generate a lower response to nonpersistent productivity shocks.

  If the shock is persistent, that is, a higher $A$ implies a higher value of $A'$, then the model would generate an increase in net worth also through

the market value of owned capital (as we will see next). The impact on investment could then be bigger.

- *Increase in* $A'$. An anticipated increase in $A'$ generates an increase in the price of capital today since $q = \beta A'$. The price increase has two effects. First, since entrepreneurs own the capital $K$, the higher $q$ generates an increase in the entrepreneur's net worth $n = qK + Y^e - B$. Notice that the initial leverage is higher, that is, the debt $B$ relative to the owned capital $K$, and the (proportional) effect on the net worth is bigger. The increase in net worth affects investment similarly to the increase in current productivity. This first channel induces an increase in the production scale $i$ without changing the probability of default if we assume that the leverage does not change.

  The second effect derives from the impact on the intraperiod leverage. Since a higher $q$ implies higher profits from producing capital goods, entrepreneurs have an incentive to expand production proportionally more than the increase in net worth, even if this increases the cost of external financing. As a result, the probability of default, or bankruptcy rate, increases in response to an anticipated productivity shock. Thus, the model generates pro-cyclical bankruptcy rates and pro-cyclical interest rate premiums—a point emphasized, among others, by Gomes, Yaron, and Zhang (2003).

  One reason the model generates a pro-cyclical interest rate premium is because investment is very sensitive to the asset price $q$. The addition of adjustment costs as in Bernanke, Gertler, and Gilchrist (1999) could revert this property. In this case, the higher price of capital improves the net worth position of the entrepreneur, but the adjustment cost contains the expansion of the production scale. As a result, entrepreneurs could end up with a lower leverage and lower probability of default. See also Covas and Den Haan (2010).

## Quantitative Performance

In general, it is not easy for the model to generate large amplification effects in response to productivity changes. In fact, as observed above, financial frictions could dampen the impact of productivity shocks. Because of the higher profitability in the production of capital goods, entrepreneurs would like to expand the production scale. However, as they produce more, the cost of external financing increases. In a frictionless economy, instead, the cost of external finance does not increase with individual production. So the initial impact on investment is larger. In essence, financial frictions act like adjustment costs in investment, which could dampen aggregate volatility. Wang

and Wen (forthcoming) provide a formal analysis of the similarity between financial frictions and adjustment cost at the aggregate level.

Even though the model has difficulties generating large amplifications, it has the potential to generate greater persistence. In fact, higher profits earned by entrepreneurs allow them to enter the next period with higher net worth. This cannot be shown explicitly with the current model since there are only two periods. However, suppose that entrepreneurs enter period 1 with a higher $K$ made possible by the higher profits earned in the previous periods. This will reduce the external cost of financing, allowing entrepreneurs to produce more capital goods, which in turn increases production in future periods. The model could then generate a hump-shape response of output as shown in Carlstrom and Fuerst (1997).

Although quantitative applications of the financial accelerator do not find large amplification effects of productivity shocks, it could still amplify the macroeconomic response to other types of shocks. For example, Bernanke, Gertler, and Gilchrist (1999) add adjustment costs in the production of capital goods in order to generate larger fluctuations in $q$ and find that the financial accelerator could generate sizable amplifications of monetary policy shocks.

## 6. COLLATERAL CONSTRAINT MODEL

Here I illustrate the main idea of models with collateral constraints as the one studied in Kiyotaki and Moore (1997). An alternative to models with collateral constraints is the consideration of optimal contracts subject to enforcement constraints as in Kehoe and Levine (1993) and Cooley, Marimon, and Quadrini (2004). However, the business cycle implications of these two modeling approaches are similar.

To illustrate the idea of the collateral model, I assume that the frictions are not in the production of capital goods, as in the costly state verification model. Instead they derive from the ability of borrowers to repudiate their intertemporal debt. In some models, like in Kiyotaki and Moore (1997), it is even assumed that physical capital is not reproducible. Therefore, in this section I assume that $E\omega = 0$ and all intermediate goods are transformed one to one into consumption goods. I denote by $\overline{K}$ the aggregate fixed stock of capital. Since capital is not reproducible, its price fluctuates endogenously in response to changing market conditions. The price fluctuation plays a central role in the model. An alternative way to generating price fluctuations is to relax the assumption that capital is not reproducible but with the addition of adjustment costs in investment and/or risk aversion.

**Frictions on the Intertemporal Margin**

From an efficiency point of view, the stock of capital should be allocated between entrepreneurs and workers to equalize their marginal product in period 2. More specifically, given $K^{e'}$, the capital allocated to the entrepreneurial sector (that is, capital purchased by entrepreneurs), efficiency requires $A' = A'G'(\overline{K} - K^{e'})$. The first term is the expected marginal productivity in the entrepreneurial sector and the second is the marginal productivity in the residual sector. Since $G'(.)$ is strictly decreasing and $G'(0) = 1 < A'$, the equalization of marginal productivities requires $K^{e'} = \overline{K}$, that is, all the capital should be allocated to the entrepreneurial sector in period 2.

The problem is that entrepreneurs may be unable to purchase $K^{e'} = \overline{K}$ in period 1. Because of limited enforceability of debt contracts, entrepreneurs are subject to the collateral constraint

$$b' \leq \xi q' k'.$$

Here $b'$ is the new debt, $k'$ is the capital purchased by an individual entrepreneur, $q'$ is the expected price of capital in period 2, and $\xi < 1$ is a parameter that captures possible losses associated with the reallocation of capital in case of default.

The theory underlying this constraint is developed in Hart and Moore (1994) and it is based on the idea that entrepreneurs cannot be forced to produce once they renege on the debt. Thus, in case of default the lender can only recover a fraction $\xi$ of the capital that can be resold at price $q'$. Since this is the last period in the model, the price of capital would be zero in the second period. In an infinite horizon model, however, the price would not be zero because the capital can still be used in production in future periods. In our two-period model we can achieve the same outcome by assuming that a fraction $\xi$ of the liquidated capital can be reallocated to the residual sector. Therefore, the liquidation price of capital in period 2 is equal to $q' = \xi A'G'(\overline{K} - K^{e'})$. Since $G'(.) \leq 1$ and only a fraction $\xi$ can be resold, the value of capital for lenders is smaller than for entrepreneurs. This is what limits the entrepreneurs' ability to borrow.

Before continuing I should observe that, in absence of capital accumulation, period 1 consumption cannot be zero for both workers and entrepreneurs. This is because period 1 production can only be used for consumption. Thus, the first-order conditions for workers are given by (2)–(4) but with $\lambda = 0$ and the supply of labor is $h = w$.

The problem solved by entrepreneurs is

$$\max_{h,k',b'} \quad \left\{c + \beta c'\right\}$$

subject to:

$$c = q\overline{K} + A\overline{K}^{\theta}h^{1-\theta} - wh - B + \frac{b'}{R} - qk'$$

$$\xi q'k' \geq b'$$

$$c' = A'k' - b',$$

$$c \geq 0, \quad c' \geq 0, \tag{12}$$

which is deterministic since there is no capital production ($\omega = 0$) and $A'$ is perfectly anticipated.

The first-order condition for the input of labor is still given by (5), that is, the entrepreneur equalizes the marginal product of labor to the wage rate. At the center stage of the model are the choices of next period capital and debt. The first-order conditions for $k'$ and $b'$ in problem (12) are

$$(1+\gamma)q = \beta A' + \mu\xi q' \tag{13}$$

$$1+\gamma = (\beta+\mu)R, \tag{14}$$

where $\mu$ and $\gamma$ are, respectively, the Lagrange multipliers associated with the collateral constraint and the non-negativity of consumption in period 1.

I can now use equations (13)–(14) together with (3)–(4) to derive an expression for $\mu$. Using the fact that the liquidation price of capital in period 2 is $q' = A'G'(\overline{K} - K^{e'})$, we derive

$$\mu = \frac{\beta[1 - G'(\overline{K} - K^{e'})]}{(1-\xi)G'(\overline{K} - K^{e'})}. \tag{15}$$

This equation relates the multiplier $\mu$ to the capital accumulated by entrepreneurs $K^{e'}$. Since the function $G(.)$ is concave, $G'(\overline{K} - K^{e'})$ is increasing in $K^{e'}$. Therefore, if the capital accumulated by entrepreneurs is higher, $\mu$ is lower.

The equilibrium can take two configurations.

- *All the capital is accumulated by entrepreneurs.* In the first equilibrium entrepreneurs have sufficient net worth to purchase all the capital, that is, $K^{e'} = \overline{K}$. Equation (15) then implies that $\mu = 0$ since $G'(0) = 1$. In this case, entrepreneurs' consumption is positive ($\gamma = 0$) and the price of capital is $q = \beta A'$.

  This is possible only if entrepreneurs start with sufficiently high net worth, that is, small $B$. To see this, consider an entrepreneur's budget constraint when the entrepreneur borrows up to the limit and chooses zero consumption. Substituting $c = 0$ and $b' = \xi q'k'$, the budget constraint becomes $q\overline{K} + Y^e + \xi q'k'/R = B + qk'$, which can be

rearranged to

$$\left(q - \frac{\xi q'}{R}\right) k' = q\overline{K} + Y^e - B. \tag{16}$$

The term $Y^e = A\overline{K}^{\theta} h^{1-\theta} - wh$ is the entrepreneur's income earned in period 1.

Equations (3)–(4) imply $A'G'(\overline{K} - K^{e'}) = qR$. Furthermore, using $q' = A'G'(\overline{K} - K^{e'})$, the above condition can be written as

$$k'_{max} = \left(\frac{1}{1-\xi}\right)\left(\overline{K} - \frac{B - Y^e}{q}\right). \tag{17}$$

This is the maximum capital that entrepreneurs can buy given the capital price $q = \beta A'$, which I made explicit by adding the subscript. It depends negatively on $B$. Therefore, if the initial net worth is not sufficiently high, entrepreneurs will be unable to purchase $\overline{K}$ and some of the capital will be inefficiently allocated to the residual sector. In this case, $K^{e'} = k'_{max} < \overline{K}$. We are then in the second type of equilibrium configuration.

- *Only part of the capital is accumulated by entrepreneurs.* In the second equilibrium, entrepreneurs choose zero consumption and the collateral constraint is binding. Since entrepreneurs cannot purchase enough capital, $G'(\overline{K} - K^{e'}) < 1$. Then equation (15) tells us that $\mu > 0$ and equation (14) implies that $\gamma > 0$ since $\beta R = 1$ (from [4] if workers' consumption is positive, implying $\lambda = 0$). Therefore, the entrepreneur borrows up to the limit and the non-negativity constraint on consumption is binding.

  Using the binding collateral constraint and zero consumption, the budget constraint can be rewritten again as in (16). This expression provides a simple intuition for the key mechanism of the model. The cost of one unit of capital, $q$, can be financed with $\frac{\xi q'}{R}$ units of debt and the rest must be financed with owned wealth. Therefore, $q - \frac{\xi q'}{R}$ is the minimum down payment required on each unit of capital. Multiplied by $k'$ we get the total down payment necessary to purchase $k'$ units of capital. In order to make the down payment, the entrepreneur needs to have enough net worth, which is the term on the right-hand side of (16). Therefore, the lower is the entrepreneurs' net worth, the lower is the amount of capital allocated to entrepreneurs. Since entrepreneurs are more productive than producers in the residual sector of the economy, lower net worth in period 1 implies lower production in period 2.

  As equation (16) makes clear, the capital allocated to the entrepreneurial sector depends crucially on the equilibrium prices $R$, $q$, and $q'$. Although all three prices contribute to the equilibrium outcome, it will

be helpful to focus on $q$ and $q'$ to see the importance of asset prices. There are several effects induced by changes in these prices.

- *Current price*: An increase in the current price, $q$, has two effects. On the one hand, it increases the entrepreneur's net worth $q\overline{K} + Y^e - B$. On the other hand, it increases the cost of purchasing new capital. The first effect has a positive impact on $k'$, while the impact of the second effect is negative.
- *Next period price*: An increase in the (expected) next period price, $q'$, allows entrepreneurs to issue more debt. Therefore, for a given net worth, more capital can be purchased.

Following Kiyotaki and Moore (1997), suppose that $q$ and $q'$ both increase by the same proportion. For example they both increase by 1 percent.[2] Provided that $B > Y^e$, this generates an increase in the capital purchased by entrepreneurs, which, in the next period, increases output. The condition $B > Y^e$ is a leverage condition. Therefore, if entrepreneurs enter the period with a high leverage, a persistent increase in prices generates an output boom.

How would the response change if contracts were enforceable? This is equivalent to the equilibrium in which the collateral constraint is not binding. In particular, all the capital is purchased by entrepreneurs since they can borrow without limit. Then a change in price would not affect the allocation of $\overline{K}$ and would not have any additional impact on aggregate production beyond the direct impact of the factors that cause the price change.

***Response to Productivity Shocks***

I will now focus on the equilibrium in which the collateral constraint is binding, that is, the equilibrium that prevails if entrepreneurs are highly leveraged. In a general model with infinitely lived agents this would arise in the long run if entrepreneurs have some incentives to take on more debt. As discussed in Section 2, there are different assumptions made in the literature to have this property. For example, a common assumption is that entrepreneurs (borrowers) are more impatient than workers (lenders). In the simple two-period model considered here, however, we can simply take the initial leverage to be sufficiently high.

[2] To facilitate the intuition, I take a partial equilibrium approach here and assume that the prices change exogenously.

If the collateral constraint is binding, the capital acquired by entrepreneurs is given by equation (17), which for convenience I rewrite here:

$$K^{e'} = \left(\frac{1}{1-\xi}\right)\left(\overline{K} - \frac{b - Y^e}{q}\right). \tag{18}$$

We now consider the impact of an increase in current and (anticipated) future productivity.

- *Increase in A.* The higher value of $A$ increases entrepreneurs' income $Y^e$ in period 1 (see equation [11]). We see from equation (18) that this induces an increase in $K^{e'}$. Essentially, entrepreneurs earn higher capital income in period 1 and this allows them to purchase more capital for period 2.

  In addition to this direct effect, there is an indirect effect induced by the price of capital. Since $K^{e'}$ increases, equation (3) implies that the current price of capital $q$ also increases. As long as $B > Y^e$, that is, entrepreneurs are sufficiently leveraged, the increase in $q$ induces a further increase in $K^{e'}$. Since entrepreneurs are more productive, that is, $G'(.) < 1$ for $K^{e'} < \overline{K}$, the reallocation of productive capital to the entrepreneurial sector generates an output boom in period 2. This second effect comes from the endogeneity of the collateral constraint, which depends on the market price $q$. Since the value of capital depends on $q$ while the value of debt is fixed, the change in price has a large impact on the net worth if entrepreneurs are highly leveraged. This is the celebrated "amplification" effect of productivity shocks induced by endogenous asset prices.

- *Increase in $A'$.* Suppose that $A'$ increases, that is, in period 1 we expect a higher productivity in period 2. We can think of this as a "news" shock. In this way it relates to the recent literature that investigates the impact of anticipated future productivity changes on the macroeconomy. See, for example, Beaudry and Portier (2006) and Jiamovich and Rebelo (2009). Here I show that financial markets could be an important transmission of these news shocks.

  From equation (2) we see that an increase in $A'$ generates an increase in the price of capital $q$. Then, equation (18) shows that the increase in $q$ induces a reallocation of capital to the entrepreneurial sector, further increasing $q$. This implies that production in period 2 increases more than the increase in productivity. We thus have an "amplification" effect. As far as current production is concerned, however, output does not change. We will see in the next section that, with the addition of working capital, the anticipated news can also affect employment in the current period. Therefore, in addition to generating an immediate

asset price boom, the news shock also generates an immediate macroeconomic boom. This mechanism has been explored in Jermann and Quadrini (2007) and Chen and Song (2009).

Although we have considered only the case of nonreproducible capital, similar results apply when there is capital accumulation together with adjustment costs on investment. With investment adjustment costs, the price of capital is not always one. An increase in future productivity raises the demand of capital, inducing an asset price boom, which in turn amplifies the impact of the initial productivity improvement. Sometimes the adjustment costs can be in the form of capital irreversibility as in Caggese (2007).

***Quantitative Performance***

There are many quantitative applications of the collateral model. Sometimes the borrowers are households engaged in real estate investments as in Iacoviello (2005). Other studies consider firms to be in need of funds for productive investments. However, the quantitative amplification induced by collateral constraints is often weak. This point has been emphasized in Cordoba and Ripoll (2004).

There are two reasons for the weak amplification. Similar to the simple model described above, for a group of models proposed in the literature, the "direct" effect of the frictions is on investment, not on the input of labor. Although this has the potential to generate large fluctuations in investments, the production inputs—capital and labor—are only marginally affected by this mechanism. As a result, output fluctuations are not affected in important ways by the financial frictions. I would also like to point out that the consideration of risk-averse agents will further reduce the amplification effects since savings, and therefore investments, will become more stable (see Kocherlakota [2000] and Cordoba and Ripoll [2004]). For the financial frictions to generate large output fluctuations that are in line with the data, they need to have a direct impact on labor. This point will be further developed in the next section.

The second reason for the weak amplification is that typical macromodels do not generate large asset price fluctuations even with the addition of binding marginal requirements (see Coen-Pirani [2005]). The centerpiece of the amplification mechanism induced by the collateral constraint model is the fact that the availability of credit, and therefore investment, depends on the price of assets, that is,

$$b' \leq \xi q' k'.$$

In economic expansions $q'$ increases and this allows for more capital investment thanks to the relaxation of the borrowing constraint. However, for this mechanism to be quantitatively important, the model should generate sizable fluctuations in $q'$, which is typically not the case in standard macromodels. In

this regard, the inability of the model to generate large amplification effects is more a consequence of the poor asset price performance of macromodels (which generate much lower asset price fluctuations than in the data) than the weakness of the collateral or financial accelerator mechanisms.

This suggests that an improvement in the asset price performance of macromodels could also enhance the amplification effect induced by financial frictions. In making this conjecture, however, we should use some caution. If the model generates large asset price fluctuations, borrowing up to the limit becomes riskier. Thus, agents may choose to stay away from the limit, that is, they will act in a precautionary manner. As a result, it is not obvious whether large asset price fluctuations will generate large macroeconomic fluctuations since, as shown in the simple model studied above, this requires the collateral constraint to be binding. But with precautionary behavior, the borrowing limit is only occasionally binding.

Unfortunately, exploring the quantitative importance of occasionally binding constraints cannot be done with local approximation techniques, which is the dominant approach used to study quantitative general equilibrium models. It is only recently that the importance of occasionally binding constraints for business cycle fluctuations has been fully recognized. Mendoza (2010) is one of the first articles that explores this issue quantitatively. I will return to the issue of occasionally binding collateral constraints later.

### Working Capital Model

The financial mechanisms presented so far affect the transmission of productivity shocks through the investment channel. For example, in the costly state verification model, the entrepreneur's net worth affects the production of new capital goods, which in turn affects next period production. In the model with collateral constraints, the net worth of entrepreneurs also plays a central role. Higher net worth allows entrepreneurs to purchase more capital. As a result, a larger fraction of productive assets are used in the more productive entrepreneurial sector enhancing aggregate output. In both models the price of capital $q$ plays a central role. However, this mechanism has a limited impact on labor.

The intuition for the weak impact on labor is simple. If we use a Cobb-Douglas production function $y = AK^{\theta}h^{1-\theta}$, an increase in the input of capital increases the demand of labor because $h$ is complementary to $K$. However, even though investment is highly volatile, the volatility of capital is small. Thus, changes in investment that are quantitatively plausible are unlikely to generate large fluctuations in labor. Empirically, however, labor input fluctuations are an important driver of output volatility. So in general, having financial frictions that primarily affect investment may not be enough for the frictions to play a central role in labor and output fluctuations. A more direct impact can be obtained if financial frictions directly affect the demand of labor.

One way to achieve this is by assuming that employers need working capital, which is complementary to labor.

The idea of working capital is not new in macroeconomics. For example, the limited participation models of monetary policy are based on the idea that producers need to finance working capital. See, for example, Christiano and Eichenbaum (1992); Fuerst (1992); Christiano, Eichenbaum, and Evans (1997); and Cooley and Quadrini (1999, 2004). See also Neumeyer and Perri (2005) for the modeling of working capital in a nonmonetary model. On one hand, besides the need of working capital, there are not other financial frictions in these models. On the other hand, business cycle models with financial frictions have mostly focused on investment, posing little importance on working capital. Jermann and Quadrini (2006), Mendoza (2010), and Jermann and Quadrini (forthcoming) are attempts at merging the two ideas: working capital needs with financially constrained borrowers.

To show how working capital interacts with financial constraints, I consider again the collateral model studied in the previous section. The only additional assumption is that entrepreneurs also need working capital in the first period of production. Specifically, they need to pay wages before the realization of revenues. To make these payments, entrepreneurs must borrow $wh$. This is an intraperiod loan, and therefore, there are no interest payments. The collateral constraint becomes

$$b' + wh \leq \xi q' k'. \tag{19}$$

The left-hand side is the total debt: intertemporal debt that will be paid back next period and the intraperiod debt that needs to be repaid at the end of period 1. The right-hand side is the collateral value of assets.

The problem solved by the entrepreneur is similar to (12) but with the new collateral constraint, that is,

$$\begin{aligned} \max_{h,k',b'} \quad & \left\{ c + \beta c' \right\} \\ & \text{subject to:} \\ & c = q\overline{K} + A\overline{K}^{\theta} h^{1-\theta} - wh - B + \frac{b'}{R} - qk' \\ & \xi q' k' \geq b' + wh \\ & c' = A' k' - b' \\ & c \geq 0, \quad c' \geq 0. \end{aligned} \tag{20}$$

The first-order conditions are also similar with the exception of the optimality condition for the input of labor, which becomes

$$(1-\theta) A \overline{K}^{\theta} h^{-\theta} = w(1+\mu). \tag{21}$$

The variable $\mu$ is the Lagrange multiplier associated with the collateral constraint as in the model without working capital. The multiplier creates a

wedge in the demand for labor.[3] When the collateral constraint is tighter, $\mu$ increases and the demand for labor declines.

Using the supply of labor, $h = w$, the wage rate is

$$w(\mu) = \left(\frac{1-\theta}{1+\mu}\right)^{\frac{1}{1+\theta}} A^{\frac{1}{1+\theta}} \overline{K}^{\frac{\theta}{1+\theta}}.$$

We can see that the wage depends negatively on the multiplier $\mu$, which I made explicit in the notation. This also implies that the entrepreneur's income, $Y^e(\mu) = A\overline{K}^{\theta} h^{1-\theta} - wh$, depends on $\mu$.

The budget constraint for the entrepreneur under a binding collateral constraint (and zero consumption) is

$$\left(q - \frac{\xi q'}{R}\right) k' = q\overline{K} + Y^e(\mu) - B. \tag{22}$$

From this equation I can derive the maximum capital that entrepreneurs can acquire as

$$k'_{max} = \left(\frac{1}{1-\xi}\right)\left(k - \frac{B - Y^e(\mu)}{q}\right). \tag{23}$$

The actual capital acquired in equilibrium by entrepreneurs is $K^{e'} = \min\left\{k'_{max}, \overline{K}\right\}$.

***Response to Productivity Shocks***

I now consider the impact of changes in current and future productivity.

- *Increase in A.* Keeping constant $\mu$, the higher productivity induces an increase in entrepreneurial income $Y^e(\mu)$. This implies that the net worth of entrepreneurs increases and, as we can see in (23), more capital will be allocated to the entrepreneurial sector.

  The next step is to see what happens to the price of capital, $q$, and to the multiplier $\mu$. From equation (3) we see that the higher $K^{e'}$ (smaller capital $k'$ accumulated by workers) must be associated with an increase in the price of capital $q$. As long as $B > Y^e(\mu)$, that is, entrepreneurs are sufficiently leveraged, the increase in $q$ further increases $K^{e'}$.

  We can now see what happens to the Lagrange multiplier $\mu$. According to equation (15), an increase in $K^{e'}$ must be associated with a decline in $\mu$. Going back to the first-order condition for labor—equation (21)—we observe that this reduces the labor wedge and generates an increase in the demand for labor, busting current production.

[3] It is common in the literature to use the phrase "labor wedge" to refer to terms that modify the optimality condition for the input of labor that we would have without frictions. Later I will discuss in more detail the issue of the labor wedge and provide a more precise definition.

To summarize, the model with working capital can generate an amplification of productivity shocks also in the current period, in addition to next period output. A key element of the amplification mechanism is the endogeneity of the asset price $q$. Because of the asset price boom, the borrowing constraint is relaxed and firms can borrow more. They will use the higher borrowing to increase both current employment and next period capital.

- *Increase in $A'$.* Let's consider now the impact of an anticipated productivity improvement (news shock). From equation (3) we see that an increase in $A'$ generates an increase in the price of capital $q$. Then, from equation (23) we observe that the increase in $q$ must be associated with a reallocation of capital to the entrepreneurial sector, further increasing $q$ (again from equation [23]). This implies that the increase in next period production is bigger than the increase in next period productivity (amplification).

  As we have seen earlier, the amplification result for period 2 is also obtained in the model without working capital. With working capital, however, the news shock also generates an output boom in the current period. Therefore, news shocks affect current employment and production even if there is no productivity change in the current period. This mechanism has been studied in Jermann and Quadrini (2007) and Chen and Song (2009) and it is consistent with the findings of Beaudry and Portier (2006) based on the estimation of structural vector autoregressions.

***Labor Wedge***

Financial frictions have the ability to generate a labor wedge if wages or other costs that are complementary to labor require advance financing (working capital). Since there is an extensive literature studying the importance of the labor wedge for business cycle fluctuations, it will be helpful to relate the properties of the wedge generated by financial frictions with the labor wedge discussed in the literature.

The labor wedge is defined in the literature as a deviation from the optimality condition for the supply of labor we would have in an economy without frictions. Without frictions the optimality condition equalizes two terms: (i) the marginal rate of substitution between consumption and leisure; and (ii) the marginal product of labor. Thus, the labor wedge is defined as the difference between these two terms. If the difference is zero, we have the same optimality condition as in the frictionless model and, therefore, there is no wedge. If the difference is not zero, we have a labor wedge since we are deviating from the optimality condition without frictions.

Using a constant elasticity of substitution utility and a Cobb-Douglas production function, the wedge can be written as

$$Wedge \equiv mrs - mpl = \frac{\phi C}{1-H} - (1-\theta)\frac{Y}{H}, \tag{24}$$

where $C$ is consumption, $H$ is hours worked, $Y$ is output, and $\phi$ and $\theta$ are, respectively, preferences and technology parameters. With the special utility function for workers used here, the wedge is

$$Wedge \equiv mrs - mpl = H - (1-\theta)\frac{Y}{H}.$$

Besides the fact that consumption does not enter the equation, the wedge generated by the model is very similar to the wedge derived from a more standard model. Since the labor supply is $H = w$ and the demand of labor satisfies $(1-\theta)\frac{Y}{H} = w(1+\mu)$, the wedge is equal to $-w\mu$.

Gali, Gertler, and López-Salido (2007) conduct a decomposition of the labor wedge in two components. The first component is the wedge between the marginal rate of substitution ($mrs$) and the wage rate ($w$). The second component is the wedge between the wage rate ($w$) and the marginal product of labor ($mpl$). More specifically,

$$Wedge \equiv mrs - w + w - mpl \equiv Wedge_1 + Wedge_2.$$

Using postwar data for the United States (although excluding the period of the recent crisis), Gali, Gertler, and López-Salido (2007) show that the first component of the wedge ($Wedge_1$) has played a predominant role in the dynamics of the whole wedge. In the version of the model studied here, however, the opposite is true since financial frictions generate only a wedge between the wage rate and the marginal product of labor ($Wedge_2$). In the model presented here, wages are fully flexible and the $mrs$ is always equal to the wage rate. Therefore, $Wedge_1 = 0$.

At first, this finding may seem to cast doubts on the empirical relevance of financial frictions for the dynamics of labor. However, it is important to recognize that the problem arises because wages are assumed to be fully flexible. To make this point, suppose that there is some wage rigidity. For example, we could assume that workers update their wages only with some probability (like in Calvo pricing). Then a change in the labor demand would lead to a change in the labor supply but with a small change in the wage. As a result, $Wedge_1$ is no longer zero.

To show this point more clearly, suppose that the wage is fixed at $\bar{w}$. The first component of the wedge is equal to $Wedge_1 = H - \bar{w}$. Eliminating $H$ using the first-order condition of firms $(1-\theta)A\overline{K}^{\theta}H^{-\theta} = (1+\mu)\bar{w}$, we get

$$Wedge_1 = \frac{(1-\theta)A\overline{K}^{\theta}}{(1+\mu)\bar{w}} - \bar{w}.$$

Therefore, the first component of the wedge is now dependent on $\mu$, which in turn depends on the shock. So, in principle, by adding wage rigidities the model could capture some of the movements in the two components of the wedge.

***Quantitative Performance***

As discussed above, the addition of working capital gives an extra kick to the amplification potential of the model. As far as productivity shocks are concerned, the amplification effect remains weak. As for the collateral model without working capital, large amplification effects require sizable fluctuations in asset prices $q'$. However, I have already observed that standard macroeconomic models, even with the addition of financial frictions, find it difficult to generate large fluctuations in asset prices. As a result, the amplification effect remains weak.

The analysis of the amplification of other shocks, besides productivity, has not received much attention in the literature. An exception is Bernanke, Gertler, and Gilchrist (1999). They embed the financial accelerator in a New Keynesian monetary model and find that the amplification effects on monetary policy shocks could be sizable: Based on their calibration, the impulse response of output to a monetary policy shock is about 50 percent larger with financial frictions.

## 7. MODEL WITH CREDIT SHOCKS

In the analysis conducted in the previous sections, I have focused on the propagation of productivity shocks, that is, shocks that arise in the real sector of the economy. Although the analysis of real shocks is clearly important for business cycle fluctuations, less attention has been devoted to studying the macroeconomic impact of shocks that arise in the financial sector of the economy—in particular, shocks that directly impact the ability of entrepreneurs or other borrowers to raise debt. Of course, we would like to have a theory of why the ability to borrow could change independently of changes that arise in the real sector of the economy. I will describe one possible theory later. For the moment, however, I start with a reduced form approach where the shocks are exogenous.

### Model with Exogenous Credit Shocks

Consider the model with working capital analyzed in the previous section where entrepreneurs face the collateral constraint (19). Now, however, I assume that the constraint factor $\xi$ is stochastic. I will call the stochastic changes in $\xi$ "credit" shocks since they affect the borrowing capability of entrepreneurs.

Given the analysis conducted in the previous section, it is easy to see how the economy responds to these shocks. The impact is similar to the response to an asset price boom induced by a productivity improvement: By changing the tightness of the collateral constraint, the shock has an immediate impact on the multiplier $\mu$, and, therefore, on the labor wedge. The change in $\xi$ also affects the price of capital and this interacts with the exogenous change in the borrowing limit. Thus, the price mechanism described in the previous section also acts as an amplification mechanism for the credit shock.

To show this in more detail, consider again equation (23) derived from the budget constraint of entrepreneurs when the collateral constraint is binding (and consumption is zero). For simplicity I rewrite the equation here,

$$K^{e'} = \left(\frac{1}{1-\xi}\right)\left(k - \frac{b - Y^e(\mu)}{q}\right). \tag{25}$$

This equation makes clear that, keeping constant the multiplier $\mu$, an increase in $\xi$ (positive credit shock) increases the capital allocated to the entrepreneurial sector. As a result of this reallocation, we can see from equation (3) that the price of capital $q$ increases. As long as $B > Y^e(\mu)$, that is, entrepreneurs are sufficiently leveraged, the increase in $q$ further increases $K^{e'}$. Thus, the positive credit shock generates a reallocation of capital to the entrepreneurial sector, which in turn increases next period output.

The reallocation of capital also affects the multiplier $\mu$ and, therefore, the labor wedge. From equation (15) we can see that an increase in $K^{e'}$ generates a decline in $\mu$. The multiplier also depends positively on $\xi$. However, if the negative effect from the increase in $K^{e'}$ dominates the positive effect from $\xi$, the multiplier $\mu$ and the labor wedge both decline in response to the positive credit shock. Therefore, credit shocks also have a positive impact on current employment and production. This is the channel explored in Jermann and Quadrini (forthcoming).

***More on Credit Shocks***

There are several articles that consider shocks to collateral or enforcement constraints. Some examples are Kiyotaki and Moore (2008), Del Negro et al. (2010), and Gertler and Karadi (2011). In the latter article, the shock arises in the financial intermediation sector. Mendoza and Quadrini (2010) also consider a financial shock to the intermediation sector but in the form of losses on outstanding loans. The impact of these shocks is very similar to a change in $\xi$.

Christiano, Motto, and Rostano (2008) propose a different way of modeling a credit shock. They use a version of the costly state verification model described in Section 5 and assume that the "volatility" of the idiosyncratic shock $\omega$ is time-variant. Thus, the financial shock is associated with greater investment risks. Since the risk is idiosyncratic and entrepreneurs are risk

neutral, the transmission mechanism is similar to shocks that affect the verification cost. Furthermore, once we recognize that a higher verification cost reduces the liquidation value of assets, this is not that different from the collateral model in which $\xi$ falls because a lower portion of the capital can be recovered. The importance of time-varying risk when there are financial frictions is also studied in Arellano, Bai, and Kehoe (2010) and Gilchrist, Sim, and Zakrajsek (2010).

***Alterative Specification of the Collateral Constraint***

From a quantitative point of view and abstracting from changes in $\xi$, the collateral constraint (19) may have an undesirable quantitative property. In particular, if this constraint is binding, the model generates a volatility of debt that is similar or higher than the volatility of the market price of capital $q$. The reason is because $k'$ co-moves positively with $q'$ in the model. Then, the linear relation between $q'k'$ and the debt $b'$ implies that the volatility of $b'$ is bigger than the volatility of $q'$. In the data, however, asset prices are much more volatile than debt. Thus, if the model can generate plausible fluctuations in asset prices, it also generates excessive fluctuations in the stock of debt.

This problem does not arise if we use an enforcement constraint in which the liquidation value of capital is related to the book value, that is,

$$b' + wh \leq \xi k'. \tag{26}$$

Conceptually, this could derive from the fact that, once the firm goes in the liquidation stage, the capital ends up being reallocated to alternative uses and the price is different from $q'$.

With this specification the model could generate plausible fluctuations in both asset prices and debt. Recognizing this, Perri and Quadrini (2011) and Jermann and Quadrini (forthcoming) use a specification of the collateral constraint where the liquidation value of capital does not depend on $q'$. Of course, by eliminating $q'$ in the collateral constraint we no longer have the amplification mechanism generated by the price of capital. However, once we focus on credit shocks, the amplification mechanism becomes secondary since these shocks can already generate significant macroeconomic volatility.

***Asset Price Bubbles and Financial Frictions***

In various versions of the model presented so far, we have seen that the price of assets plays an important role when there are financial market frictions. Whatever makes the price of assets move, it can affect the real sector of the economy by changing the tightness of the borrowing constraint. One factor that could generate movement in asset prices is bubbles. Traditionally we think of bubbles as situations in which the price of assets keeps growing over time even if nothing "fundamental" changes in the economy. Independent of

what can generate and sustain a bubble, it is easy to see the macroeconomic implications in the context of the simple model studied here.

Consider the version of the collateral model with working capital studied in Section 6. In this model, the fundamental price of capital in period 2 is $A'G'(\overline{K} - K^{e'})$. In the presence of a bubble, the price of capital would be higher. Without going into the details of whether the bubble is rational or not, the asset price with a bubble will be $A'G'(\overline{K} - K^{e'}) + \mathcal{B}'$, where $\mathcal{B}'$ is the bubble component. The macroeconomic effects are similar to the ones we have already examined when the change in asset prices was driven by productivity.

The modeling of rational bubbles is often challenging, especially in models with infinitely lived agents. To avoid this problem, Jermann and Quadrini (2007) design a mechanism that looks like a bubble, that is, it generates asset price movements, but it is based on fundamentals. The idea is that the economy can experience different rates of growth and switches from one growth regime to the other with some probability. When the "believed" probability of switching to a higher growth regime increases, current asset prices increase and the model generates a macroeconomic expansion. Even if the mechanism is not technically a bubble, it generates similar macroeconomic effects.

An alternative approach is to work with models where agents have limited life spans. These models allow for rational bubbles if certain conditions about discounting and population growth are met. Examples of these studies are Farhi and Tirole (2011) and Martin and Ventura (2011).

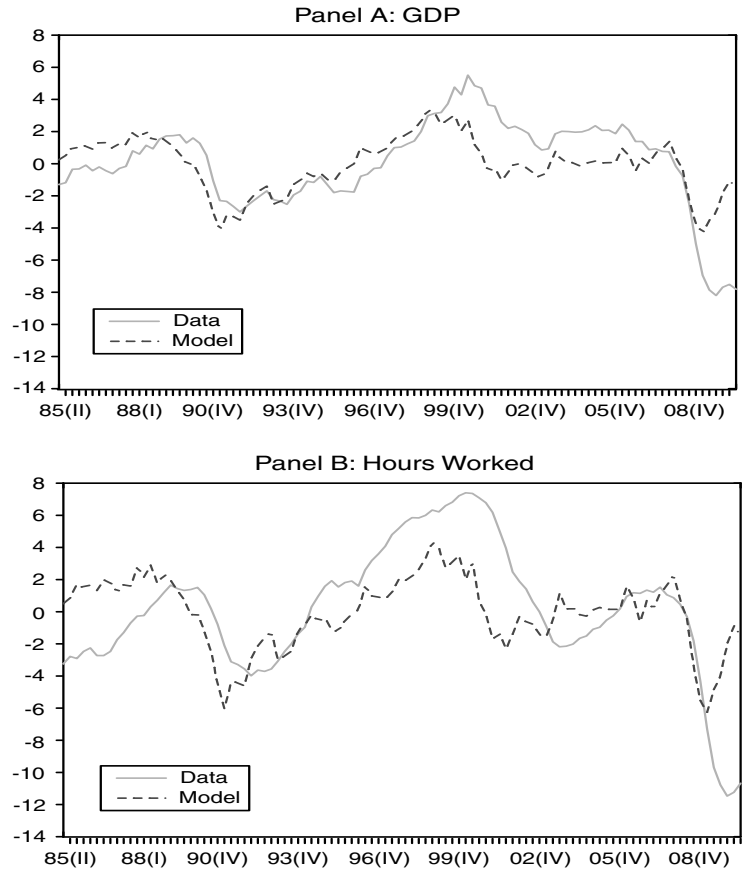
A third approach is based on the idea of multiple equilibria as in Kocherlakota (2009). This study is inspired by the study of Kiyotaki and Moore (2008), who develop a model with two monetary equilibria. In the first equilibrium, money is valued because there is the expectation that agents are willing to accept money, while in the second equilibrium money has no value because agents are not willing to accept it. Kocherlakota (2009) reinterprets money more generally as a nonproductive asset that could be used as a collateral. For example, housing. He then considers sunspot equilibria in which the economy switches stochastically from one equilibrium to the other. The switch is associated with asset price fluctuations, which have an impact on the real sector of the economy.

***Quantitative Performance***

The study of the quantitative implications of credit shocks is relatively recent but the findings suggest that these shocks play an important role for the business cycle. This is especially true if they directly affect the demand of labor.

An important issue in conducting a quantitative exploration of these shocks is their identification. Jermann and Quadrini (forthcoming) propose two approaches. The first approach uses a strategy that is reminiscent of the Solow

**Figure 2 Responses to Financial Shocks from Jermann and Quadrini (Forthcoming)**




residual procedure to construct productivity shocks. Consider the enforcement constraint specified in equation (26). If this constraint is always binding, we can use empirical time series for debt, $b'$, wages, $wh$, and capital, $k'$, to construct time series for the credit variable $\xi$ as residuals from this equation. Once we have the time series for $\xi$ we can feed the constructed series into the (calibrated) model and study the response of the variables of interest.

Figure 2 shows the empirical and simulated series of output and labor generated by the model studied in Jermann and Quadrini (forthcoming). According to the simulation, credit shocks have played an important role in capturing the dynamics of labor and output in the U.S. economy during the last two and a half decades.

Another approach used to evaluate the importance of the credit shocks is to conduct a structural estimation of the model. This, however, requires the consideration of many more shocks because, effectively, a structural estimation has the flavor of a horse race among the shocks included in the model. For that reason Jermann and Quadrini (forthcoming) extend the basic model by adding more frictions and shocks. The estimated model is similar to Smets and Wouters (2007) but with financial frictions and financial shocks. Through the structural estimation they find that credit shocks contributed at least one-third to the variance of U.S. output and labor. Christiano, Motto, and Rostagno (2008) and Liu, Wang, and Zha (2011) also conduct a structural estimation of a model with financial frictions and financial shocks and they find that these shocks contributed significantly to the volatility of aggregate output.

## Model with Endogenous Liquidity and Multiple Equilibria

So far the analysis has focused on equilibria in which entrepreneurs face binding collateral constraints. This is typically the case when there is no uncertainty. However, in the presence of uncertainty and especially with credit shocks, the enforcement constraint may not be binding in some contingencies. The possibility of "occasionally" binding constraints allows us to think about the issue of liquidity and the emergence of multiple equilibria.[4]

I continue to use the collateral constraint specified in (19) but with further assumptions about the liquidation value of capital.

Following Perri and Quadrini (2011), I assume that in the event of debt repudiation, the liquidated capital can be sold not only to the residual sector (as in the previous model) but also to other nondefaulting entrepreneurs. However, if the capital is sold to the residual sector, only a fraction $\xi$ is usable. Instead, if the capital is sold to other entrepreneurs, there is no loss of capital. Since the marginal productivity of capital for entrepreneurs in the next period is $A'$, this is also the price that nondefaulting entrepreneurs would be willing to pay for the liquidated capital. The price obtained by selling capital to the residual sector, instead, is $A'G'(\overline{K} - K^{e'})$. Because $\xi < 1$ and $G'(\overline{K} - K^{e'}) \leq 1$, the resale to the entrepreneurial sector is the preferred option.

Notice that the default decision is made after all entrepreneurs have decided their borrowing $b'$. If there were no limits to the ability of nondefaulting entrepreneurs to purchase liquidated capital, then the residual sector would

[4] Occasionally binding constraints is a feature of models studied in Brunnermeier and Sannikov (2010) and Mendoza (2010), although they abstract from credit shocks and there are not multiple equilibria. Boz and Mendoza (2010) also consider occasionally binding constraints with credit shocks but not multiple equilibria. See also Guerrieri and Lorenzoni (2009).

be irrelevant. Thus, I now introduce an additional assumption that in some contingencies limits the ability of the entrepreneurial sector to purchase the liquidated capital.

The assumption is that entrepreneurs can purchase the capital of liquidated firms only if they have the liquidity to do so. In this context, the liquidity is determined by the credit ability of entrepreneurs, which in turn depends on their borrowing decision. More specifically, if the collateral constraint binds, entrepreneurs will not be able to purchase the capital of liquidated firms since they no longer have access to additional credit. In this case, the only available option for the lender is to sell the liquidated capital to the residual sector at a lower price. However, if entrepreneurs do not borrow up to the limit, they still have access to credit that can be used in the event of an investment opportunity. In this case the capital of a liquidated firm can be sold to entrepreneurs at a higher price.

We now have all the elements to show that the model has the potential to generate multiple equilibria. Suppose that the initial state $B$ is such that the enforcement constraint (19) is binding if the residual sector is the only option for the liquidated capital but it is not binding if the liquidated capital can be sold to entrepreneurs. In the first case the collateral value is $\xi q'k' = \xi A'G'(\overline{K} - K^{e'})$, while in the second it is $\xi q'k' = \xi A'$. Since the second is bigger than the first, it is possible that the collateral constraint is binding in the first case but not in the second. Under these conditions the model admits multiple self-fulfilling equilibria.

- *Bad equilibrium.* Suppose that agents expect that the unit value of the liquidated capital is $\xi q' = \xi A'G'(\overline{K} - K^{e'})$. This imposes a tight constraint on entrepreneurs and, as a result, they borrow up to the limit. But then, if an entrepreneur defaults, the lender is unable to sell the liquidated capital to other entrepreneurs since there are no entrepreneurs capable of purchasing the capital. The recovery value is $\xi A'G'(\overline{K} - K^{e'})$ per each unit of capital. Therefore, the expectation of a lower liquidation price is ex-post validated by the lack of "liquidity" available to entrepreneurs.

- *Good equilibrium.* Suppose that agents expect that the unit value of the liquidated capital is $q' = A'$. This relaxes the borrowing constraint on entrepreneurs and allows them to borrow more than required to purchase $k' = \overline{K}$. Thus, the collateral constraint is not binding. But then, if an entrepreneur defaults, the lender is able to sell the liquidated capital to other entrepreneurs and the recovery value is $A'$. Therefore, the expectation of high liquidation prices is ex-post validated by the "liquidity" available to entrepreneurs.

The possibility of multiple equilibria introduces an endogenous mechanism for fluctuations in $\xi$. More specifically, the value of $\xi$ is low if the

enforcement constraint is binding, which in turn generates a low value of $\xi$. Instead, if the value of $\xi$ is high, the enforcement constraint is not binding, which in turn generates a high value of $\xi$. In this way the credit shock $\xi$ becomes endogenous and could fluctuate in response to the states of the economy. This provides a concept of liquidity-driven crisis: Expectations of high prices increase liquidity, which in turn sustains high prices. Instead, expectations of low prices generate a contraction in liquidity, which in turn induces a downfall in the liquidation price. The transmission of "endogenous" credit shocks to the real sector of the economy in a closed economy is similar to the model with "exogenous" credit shocks already described in the previous section.

## The International Transmission of Credit Shocks

The 2007–2009 crisis has been characterized by a high degree of international synchronization in which most of the industrialized countries experienced large macroeconomic contractions. There are two main explanations for the synchronization. The first explanation is that country-specific shocks are internationally correlated. The second explanation is that shocks that arise in one or few countries are transmitted to other countries because of economic and financial integration.

The first hypothesis is not truly an explanation: If shocks are correlated across countries, we would like to understand why they are correlated. Although this is obvious for certain shocks, think for example to oil shocks, it is less intuitive for others. For instance, if we think that shocks to the labor wedge are important drivers of the business cycle, it is not obvious why they should be correlated across countries. The second hypothesis—international transmission of country-specific shocks—is a more interesting line of research.

In this section I show that "credit" shocks that hit one or few countries could generate large macroeconomic spillovers to other countries if they are financially integrated. Therefore, these shocks are possible candidates to account for the international co-movement in macroeconomic aggregates.

To show this, I will consider a two-country version of the collateral model described earlier. The only additional feature I need to specify is the meaning of financial integration. One obvious implication of financial integration is that borrowing and lending can be done internationally. This also implies that the interest rate is equalized across countries (law of one price). In the simple model studied here, however, this is inconsequential because agents are risk neutral and the interest rates are constant and equal across countries even if they operate in a regime of financial autarky. Therefore, this is not the important dimension of international integration that matters here.

Another possible implication of financial integration is that investors (in our case entrepreneurs) hold domestic and foreign firms. Effectively, it is as

if each firm has two units: one operating in country 1 and the other operating in country 2. The problem solved by the entrepreneur can be written as

$$\max_{\{h_j,k'_j,b'_j\}_{j=1}^2} \quad \left\{c + \beta c'\right\}$$

subject to:

$$c = \sum_{j=1}^{2}\left[q_j\overline{K} + A_j\overline{K}^{\theta}h_j^{1-\theta} - w_j h_j - B + \frac{b'_j}{R} - q_j k'_j\right]$$

$$\xi_j q'_j k'_j \geq b'_j + w_j h_j, \qquad j = 1, 2$$

$$c' = \sum_{j=1}^{2}(A'_j k'_j - b'_j)$$

$$c \geq 0, \quad c' \geq 0, \tag{27}$$

where the index $j = 1, 2$ identifies the country. Since entrepreneurs have operations at home and abroad, they make production and investment decisions in both countries. Notice, however, that they face a consolidated budget constraint. Also notice that the variable $\xi$ is indexed by $j$ since credit shocks could be country-specific. This would be the case, for example, if there are financial problems in the banking system of one country but not in the other.

I now show how a credit shock in country 1 (changes in $\xi_1$) affects the economies of both countries. This can easily be seen from the first-order conditions with respect to $k'_1$, $b'_1$, $k'_2$, and $b'_2$,

$$(1+\gamma)q_1 = \beta A'_1 + \mu_1\xi_1 Eq'_1 \tag{28}$$

$$1+\gamma = (\beta+\mu_1)R \tag{29}$$

$$(1+\gamma)q_2 = \beta A'_2 + \mu_2\xi_2 Eq'_2 \tag{30}$$

$$1+\gamma = (\beta+\mu_2)R. \tag{31}$$

Equations (29) and (31) imply that the Lagrange multipliers are equalized across countries, that is, $\mu_1 = \mu_2$.

Now consider the first-order conditions with respect to labor,

$$(1-\theta)A_1\overline{K}^{\theta}h_1^{-\theta} = w_1(1+\mu_1)$$

$$(1-\theta)A_2\overline{K}^{\theta}h_2^{-\theta} = w_2(1+\mu_2).$$

Since $\mu_1 = \mu_2$, a credit shock in country 1 (change in $\xi_1$) has the same impact on the demand of labor of both countries. Therefore, a country-specific credit shock gets propagated to other countries through the labor wedge. This mechanism is emphasized in Perri and Quadrini (2008, 2011).

The impact on the accumulation of capital is not perfectly symmetric in the two countries, as we can see from equations (28) and (30). However, $k'_1$ and $k'_2$ move in the same direction.

***Endogenous Credit Shocks***

Perri and Quadrini (2011) go beyond "exogenous" credit shocks and, adopting a framework with occasionally binding constraints similar to the model described in the previous subsection, they study the implications of endogenous $\xi_j$ in an international environment.

The emergence of multiple equilibria characterized by different degrees of liquidity also arises in the two-country model. What is interesting is that, if countries are financially integrated, then *bad* and *good* equilibria outcomes become perfectly correlated across countries. Thus, the model provides not only a mechanism for the international transmission of country-specific credit shocks, but also a mechanism in which "endogenous" credit shocks are internationally correlated. It is important to emphasize that the international correlation of $\xi_j$ is not an assumption but an equilibrium property.

To see this, consider again the two-country model studied in the previous section. The first-order conditions with respect to $k_1'$, $b_1'$, $k_2'$, and $b_2'$ are still given by (28)–(30). Therefore, $\mu_1 = \mu_2$. This means that, if the collateral constraint is binding in one country, it must also be binding in the other country. But then we cannot have that in one country the liquidation price of capital is determined by the marginal product in the entrepreneurial sector while in the other country the price is determined by the marginal product in the residual sector. If the collateral constraints are binding in both countries, entrepreneurs lack the liquidity to purchase the capital of liquidated firms and the collateral value will be low in both countries. This makes the collateral constraints tighter and entrepreneurs borrow up to the limit (bad equilibrium outcome). However, if the collateral constraints are not binding in both countries, then entrepreneurs have the liquidity to purchase the liquidated capital in both countries. The collateral value is high in both countries and firms do not borrow up to the limit.

To summarize, either both countries end up in a bad equilibrium or both countries end up in a good equilibrium. In this way self-fulfilling equilibria (endogenous shocks) become perfectly correlated across countries.

***Quantitative Performance***

To the best of my knowledge, the quantitative properties of the international model with endogenous credit shocks have been explored only in Perri and Quadrini (2011). This article emphasizes four properties. First, the response to credit shocks is highly asymmetric. Negative credit shocks generate large- and short-lived macroeconomic contractions while credit expansions generate gradual and long-lasting macroeconomic booms.

The second finding is that credit contractions (negative credit shocks) have larger macroeconomic effects if they arise after long periods of credit expansions. Therefore, long credit expansions create the conditions for highly

disrupting financial crises. A similar prediction is obtained in Gorton and Ordoñez (2011) but with a mechanism that is based on the information quality of collateral assets.

The third finding relates to the difference between exogenous versus endogenous credit shocks. While exogenous credit shocks can generate macroeconomic co-movement, they do not generate cross-country co-movement in financial flows or leverages, which is a strong empirical regularity. The model with endogenous credit shocks, however, is also capable of generating co-movement in financial flows since $\xi_j$ are endogenously correlated across countries.

The last quantitative feature of the model I would like to emphasize is that it can generate sizable fluctuations in asset prices. For this feature, however, the risk aversion of entrepreneurs becomes important, which we have abstracted from in the simple version of the model presented here. Assuming that there is market segmentation and firms cannot be purchased by workers, a negative credit shock induces firms to pay lower dividends, which in turn reduces the consumption of entrepreneurs. This implies that the discount rate of entrepreneurs, $\beta U'(c_{t+1})/U'(c_t)$, falls. As a result, their valuation of future dividends falls, leading to an immediate drop in the market value of firms. Since the impact of the credit shocks on entrepreneurs' consumption is large, the model generates sizable drops in asset prices.

## 8. CONCLUSION

The key principles for adding financial market frictions in general equilibrium models are not new in the macroliterature. However, it is only with the recent crisis that the profession has fully recognized the importance of financial markets for business cycle fluctuations. Thus, more effort has been devoted to the construction of models that can capture the role of financial markets for macroeconomic dynamics.

This article has reviewed the most common and popular ideas proposed in the literature. Using a stylized model with only two periods and two types of agents, I have shown that the modeling of financial market frictions is useful for understanding several dynamic features of the macroeconomy in general and of the business cycle in particular.

The ideas reviewed in this article are all based on the transmission of shocks through the "credit channel," that is, conditions that limit the availability of funds or increase the cost of funds needed to make investment and hiring decisions. Some authors have also proposed models in which the credit channel and adverse selection in credit markets could generate economic fluctuations even in absence of exogenous shocks—an example is Suarez and Sussman (1997). Less attention has been devoted in the literature to studying alternative mechanisms through which financial frictions have an impact on

the macroeconomic dynamics. One of these mechanisms is studied in Monacelli, Quadrini, and Trigari (2010), who embed financial market frictions in a matching model of the labor market with wage bargaining. In this article, collateral constraints affect employment not because they limit the amount of funds available to firms for hiring workers, but because they affect the bargaining of wages. One interesting feature of this mechanism is that the impact of credit shocks on employment is much more *persistent* than the impact generated by the typical credit channel reviewed in this article.[5]

## REFERENCES

Arellano, Cristina, Yan Bai, and Patrick Kehoe. 2010. "Financial Markets and Fluctuations in Uncertainty." Research Department Staff Report, Federal Reserve Bank of Minneapolis (April).

Beaudry, Paul, and Franck Portier. 2006. "Stock Prices, News, and Economic Fluctuations." *American Economic Review* 96 (September): 1,293–307.

Bernanke, Ben, and Mark Gertler. 1989. "Agency Costs, Net Worth, and Business Fluctuations." *American Economic Review* 79 (March): 14–31.

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist. 1999. "The Financial Accelerator in a Quantitative Business Cycle Framework." In *Handbook of Macroeconomics*, Vol. 1C, edited by J. B. Taylor and M. Woodford. Amsterdam: Elsevier Science; 1,341–93.

Bewley, Truman F. 1986. "Stationary Monetary Equilibrium with a Continuum of Independent Fluctuating Consumers." In *Contributions to Mathematical Economics in Honor of Gérard Debreu*, edited by Werner Hildenbrand and Andreu Mas-Colell. Amsterdam: North-Holland.

Boz, Emine, and Enrique G. Mendoza. 2010. "Financial Innovation, the Discovery of Risk, and the U.S. Credit Crisis." Working Paper 16020. Cambridge, Mass.: National Bureau of Economic Research (May).

Brunnermeier, Markus K., and Yuliy Sannikov. 2010. "A Macroeconomic Model with a Financial Sector." Manuscript, Princeton University.

[5] Other contributions that embed financial market frictions in models with searching and matching frictions are Weil and Wasmer (2004), Chugh (2009), Petrosky-Nadeau (2009), and Petrosky-Nadeau and Wasmer (2010). In these articles, however, the main transmission mechanism is still based on the "credit channel."

Caggese, Andrea. 2007. "Financing Constraints, Irreversibility and Investment Dynamics." *Journal of Monetary Economics* 54 (October): 2,102–30.

Carlstrom, Charles T., and Timothy S. Fuerst. 1997. "Agency Costs, Net Worth, and Business Fluctuations: A Computable General Equilibrium Analysis."*American Economic Review* 87 (December): 893–910.

Chen, Kaiji, and Zheng Song. 2009. "Financial Frictions on Capital Allocation: A Transmission Mechanism of TFP Fluctuations." Manuscript, Emory University.

Christiano, Lawrence J., and Martin Eichenbaum. 1992. "Liquidity Effects and the Monetary Transmission Mechanism." *American Economic Review* 82 (May): 346–53.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 1997. "Sticky Price and Limited Participation Models of Money: A Comparison." *European Economic Review* 41 (June): 1,201–49.

Christiano, Lawrence J., Roberto Motto, and Massimo Rostagno. 2008. "Financial Factors in Economic Fluctuations." Manuscript, Northwestern University and European Central Bank.

Chugh, Sanjay K. 2009. "Costly External Finance and Labor Market Dynamics." Manuscript, University of Maryland.

Coen-Pirani, Daniele. 2005. "Margin Requirements and Equilibrium Asset Prices." *Journal of Monetary Economics* 52 (March): 449–75.

Cooley, Thomas F., Ramon Marimon, and Vincenzo Quadrini. 2004. "Aggregate Consequences of Limited Contract Enforceability." *Journal of Political Economy* 112 (August): 817–47.

Cooley, Thomas F., and Vincenzo Quadrini. 1999. "A Neoclassical Model of the Phillips Curve Relation." *Journal of Monetary Economics* 44 (October): 165–93.

Cooley, Thomas F., and Vincenzo Quadrini. 2004. "Optimal Monetary Policy in a Phillips-Curve World." *Journal of Economic Theory* 118 (October): 174–208.

Cordoba, Juan Carlos, and Marla Ripoll. 2004. "Credit Cycles Redux." *International Economic Review* 45 (November): 1,011–46.

Covas, Francisco, and Wouter J. Den Haan. 2010. "The Role of Debt and Equity Finance over the Business Cycle." Manuscript, University of Amsterdam.

Covas, Francisco, and Wouter J. Den Haan. 2011. "The Cyclical Behavior of Debt and Equity Finance." *American Economic Review* 101 (April): 877–99.

Del Negro, Marco, Gauti Eggertsson, Andrea Ferrero, and Nobuhiro Kiyotaki. 2010. "The Great Escape? A Quantitative Evaluation of the Fed's Non-Standard Policies." Manuscript, Federal Reserve Bank of New York.

Farhi, Emmanuel, and Jean Tirole. 2011. "Bubbly Liquidity." Working Paper 16750. Cambridge, Mass.: National Bureau of Economic Research (January).

Fuerst, Timothy S. 1992. "Liquidity, Loanable Funds, and Real Activity." *Journal of Monetary Economics* 29 (February): 3–24.

Gali, Jordi, Mark Gertler, and J. David López-Salido. 2007. "Markups, Gaps, and the Welfare Costs of Business Fluctuations." *The Review of Economics and Statistics* 89 (November): 44–59.

Gertler, Mark, and Peter Karadi. 2011. "A Model of Unconventional Monetary Policy." *Journal of Monetary Economics* 58 (January): 17–34.

Gilchrist, Simon, Jae W. Sim, and Egon Zakrajsek. 2010. "Uncertainty, Financial Frictions, and Investment Dynamics." Manuscript, Boston University.

Gilchrist, Simon, Vladimir Yankov, and Egon Zakrajsek. 2009. "Credit Market Shocks and Economic Fluctuations: Evidence from Corporate Bond and Stock Markets." *Journal of Monetary Economics* 56 (May): 471–93.

Gomes, Joao F., Amir Yaron, and Lu Zhang. 2003. "Asset Prices and Business Cycles with Costly External Finance." *Review of Economic Dynamics* 6 (October): 767–88.

Gorton, Gary, and Guillermo Ordoñez. 2011. "Collateral Crises." Manuscript, Yale University.

Guerrieri, Veronica, and Guido Lorenzoni. 2009. "Liquidity and Trading Dynamics." *Econometrica* 77 (November): 1,751–90.

Guerrieri, Veronica, and Guido Lorenzoni. 2010. "Credit Crises, Precautionary Savings and the Liquidity Trap." Manuscript, University of Chicago Booth and Massachusetts Institute of Technology.

Hart, Oliver, and John Moore. 1994. "A Theory of Debt Based on the Inalienability of Human Capital." *Quarterly Journal of Economics* 109 (November): 841–79.

Iacoviello, Matteo. 2005. "House Prices, Borrowing Constraints, and Monetary Policy in the Business Cycle." *American Economic Review* 95 (June): 739–64.

Jaimovich, Nir, and Sergio Rebelo. 2009. "Can News about the Future Drive the Business Cycle?" *American Economic Review* 99 (September): 1,097–118.

Jermann, Urban, and Vincenzo Quadrini. 2006. "Financial Innovations and Macroeconomic Volatility." Working Paper 12308. Cambridge, Mass.: National Bureau of Economic Research (June).

Jermann, Urban, and Vincenzo Quadrini. 2007. "Stock Market Boom and the Productivity Gains of the 1990s." *Journal of Monetary Economics* 54 (March): 413–32.

Jermann, Urban, and Vincenzo Quadrini. Forthcoming. "Macroeconomic Effects of Financial Shocks." *American Economic Review*.

Kehoe, Timothy J., and David K. Levine. 1993. "Debt Constrained Asset Markets." *Review of Economic Studies* 60 (October): 865–88.

Khan, Aubhik, and Julia K. Thomas. 2011. "Credit Shocks and Aggregate Fluctuations in an Economy with Production Heterogeneity." Manuscript, Ohio State University.

Kiyotaki, Nobuhiro, and John H. Moore. 1997. "Credit Cycles." *Journal of Political Economy* 105 (April): 211–48.

Kiyotaki, Nobuhiro, and John H. Moore. 2008. "Liquidity, Business Cycles, and Monetary Policy." Manuscript, Princeton University and Edinburgh University.

Kocherlakota, Narayana R. 2000. "Creating Business Cycles Through Credit Constraints." Federal Reserve Bank of Minneapolis *Quarterly Review* 24 (Summer): 2–10.

Kocherlakota, Narayana R. 2009. "Bursting Bubbles: Consequences and Cures." Manuscript, Federal Reserve Bank of Minneapolis.

Krusell, Per, Anthony A. Smith, Jr. 1998. "Income and Wealth Heterogeneity in the Macroeconomy." *Journal of Political Economy* 106 (October): 867–96.

Liu, Zheng, Pengfei Wang, and Tao Zha. 2011. "Land-Price Dynamics and Macroeconomic Fluctuations." Working Paper 17045. Cambridge, Mass.: National Bureau of Economic Research (May).

Martin, Alberto, and Jaume Ventura. 2011. "Economic Growth with Bubbles." Manuscript, Pompeu Fabra University.

Mendoza, Enrique G. 2010. "Sudden Stops, Financial Crises, and Leverage." *American Economic Review* 100 (December): 1,941–66.

Mendoza, Enrique G., and Vincenzo Quadrini. 2010. "Financial Globalization, Financial Crises and Contagion." *Journal of Monetary Economics* 57 (January): 24–39.

Miao, Jianjun, and Pengfei Wang. 2010. "Credit Risk and Business Cycles." Manuscript, Boston University and Hong Kong University of Science and Technology.

Modigliani, Franco, and Merton H. Miller. 1958. "The Cost of Capital, Corporate Finance and the Theory of Investment." *American Economic Review* 48 (June): 261–97.

Monacelli, Tommaso, Vincenzo Quadrini, and Antonella Trigari. 2010. "Financial Markets and Unemployment." Manuscript, Bocconi University and University of Southern California.

Neumeyer, Pablo A., and Fabrizio Perri. 2005. "Business Cycles in Emerging Economies: The Role of Interest Rates." *Journal of Monetary Economics* 52 (March): 345–80.

Perri, Fabrizio, and Vincenzo Quadrini. 2008. "Understanding the International Great Moderation." Manuscript, University of Minnesota and University of Southern California.

Perri, Fabrizio, and Vincenzo Quadrini. 2011. "International Recessions." Working Paper 17201. Cambridge, Mass.: National Bureau of Economic Research (July).

Petrosky-Nadeau, Nicolas. 2009. "Credit, Vacancies and Unemployment Fluctuations." Manuscript, Carnegie-Mellon University.

Petrosky-Nadeau, Nicolas, and Etienne Wasmer. 2010. "The Cyclical Volatility of Labor Markets under Frictional Financial Markets." Manuscript, Carnegie-Mellon University and Science-Po Paris.

Smets, Frank, and Rafael Wouters. 2007. "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach." *American Economic Review* 97 (June): 586–606.

Suarez, Javier, and Oren Sussman. 1997. "Endogenous Cycles in a Stiglitz-Weiss Economy." *Journal of Economic Theory* 76 (September): 47–71.

Townsend, Robert M. 1979. "Optimal Contracts and Competitive Markets with Costly State Verification." *Journal of Economic Theory* 21 (October): 265–93.

Wang, Pengfei, and Yi Wen. Forthcoming. "Hayashi Meets Kiyotaki and Moore: A Theory of Capital Adjustment Costs." *Review of Economic Dynamics*.

Weil, Philippe, and Etienne Wasmer. 2004. "The Macroeconomics of Labor and Credit Market Imperfections." *American Economic Review* 94 (September): 944–63.

*Economic Quarterly—Volume 97, Number 3—Third Quarter 2011—Pages 255–326*

# Macroeconomics with Heterogeneity: A Practical Guide

Fatih Guvenen

> What is the origin of inequality among men and is it authorized by natural law?
>
> —Academy of Dijon, 1754 (Theme for essay competition)

The quest for the origins of inequality has kept philosophers and scientists occupied for centuries. A central question of interest—also highlighted in Academy of Dijon's solicitation for its essay competition[1]—is whether inequality is determined solely through a natural process or through the interaction of innate differences with man-made institutions and policies. And, if it is the latter, what is the precise relationship between these origins and socioeconomic policies?

While many interesting ideas and hypotheses have been put forward over time, the main impediment to progress came from the difficulty of scientifically testing these hypotheses, which would allow researchers to refine ideas that were deemed promising and discard those that were not. Economists, who grapple with the same questions today, have three important advantages that can allow us to make progress. First, modern quantitative economics provides a wide set of powerful tools, which allow researchers to build "laboratories" in which various hypotheses regarding the origins and consequences of

For helpful discussions, the author thanks Dean Corbae, Cristina De Nardi, Per Krusell, Serdar Ozkan, and Tony Smith. Special thanks to Andreas Hornstein and Kartik Athreya for detailed comments on the draft. David Wiczer and Cloe Ortiz de Mendivil provided excellent research assistance. The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Chicago, the Federal Reserve Bank of Richmond, or the Federal Reserve System. Guvenen is affiliated with the University of Minnesota, the Federal Reserve Bank of Chicago, and NBER. E-mail: guvenen@umn.edu.

[1] The competition generated broad interest from scholars of the time, including Jean-Jacques Rousseau, who wrote his famous *Discourse on the Origins of Inequality* in response, but failed to win the top prize.

inequality can be studied. Second, the widespread availability of rich micro-data sources—from cross-sectional surveys to panel data sets from administrative records that contain millions of observations—provides fresh input into these laboratories. Third, thanks to Moore's law, the cost of computation has fallen radically in the past decades, making it feasible to numerically solve, simulate, and estimate complex models with rich heterogeneity on a typical desktop workstation available to most economists.

There are two broad sets of economic questions for which economists might want to model heterogeneity. First, and most obviously, these models allow us to study cross-sectional, or distributional, phenomena. The U.S. economy today provides ample motivation for studying distributional issues, with the top 1 percent of households owning almost half of all stocks and one-third of all net worth in the United States, and wage inequality having risen virtually without interruption for the last 40 years. Not surprisingly, many questions of current policy debate are inherently about their distributional consequences. For example, heated disagreements about major budget issues—such as reforming Medicare, Medicaid, and the Social Security system—often revolve around the redistributional effects of such changes. Similarly, a crucial aspect of the current debate on taxation is about "who should pay what?" Answering these questions would begin with a sound understanding of the fundamental determinants of different types of inequality.

A second set of questions for which heterogeneity could matter involves aggregate phenomena. This second use of heterogeneous-agent models is less obvious than the first, because various aggregation theorems as well as numerical results (e.g., Ríos-Rull [1996] and Krusell and Smith [1998]) have established that certain types of heterogeneity do not change (many) implications relative to a representative-agent model.[2]

To understand this result and its ramifications, in Section 1, I start by reviewing some key theoretical results on aggregation (Rubinstein 1974; Constantinides 1982). Our interest in these theorems comes from a practical concern: Basically, a subset of the conditions required by these theorems are often satisfied in heterogeneous-agent models, making the aggregate implications of such models closely mimic those from a representative-agent economy. For example, an important theorem proved by Constantinides (1982) establishes the existence of a representative agent if markets are complete.[3] This central role of complete markets turned the spotlight since the late 1980s onto its testable implications for perfect risk sharing (or "full insurance"). As

[2] These aggregation results do *not* imply that all aspects of a representative-agent model will be the same as those of the underlying individual problem. I discuss important examples to the contrary in Section 6.

[3] (Financial) markets are "complete" when agents have access to a sufficiently rich set of assets that allows them to transfer their wealth/resources across any two dates and/or states of the world.

I review in Section 2, these implications have been tested by an extensive literature using data sets from all around the world—from developed countries such as the United States to village economies in India, Thailand, Uganda, and so on. While this literature delivered a clear *statistical* rejection, it also revealed a surprising amount of "partial" insurance, in the sense that individual consumption growth (or, more generally, marginal utility growth) does not seem to respond to many seemingly large shocks, such as long spells of unemployment, strikes, and involuntary moves (Cochrane [1991] and Townsend [1994], among others).

This raises the more practical question of "how *far* are we from the complete markets benchmark?" To answer this question, researchers have recently turned to directly measuring the degree of partial insurance, defined for our purposes as the degree of consumption smoothing over and above what an individual can achieve on her own via "self-insurance" in a permanent income model (i.e., using a single risk-free asset for borrowing and saving). Although this literature is quite new—and so a definitive answer is still not on hand—it is likely to remain an active area of research in the coming years.

The empirical rejection of the complete markets hypothesis launched an enormous literature on incomplete markets models starting in the early 1990s, which I discuss in Section 3. Starting with Imrohoroglu (1989), Huggett (1993), and Aiyagari (1994), this literature has been addressing issues from a very broad spectrum, covering diverse topics such as the equity premium and other puzzles in finance; important life-cycle choices, such as education, marriage/divorce, housing purchases, fertility choice, etc.; aggregate and distributional effects of a variety of policies ranging from capital and labor income taxation to the overhaul of Social Security, reforming the health care system, among many others. An especially important set of applications concerns trends in wealth, consumption, and earnings inequality. These are discussed in Section 4.

A critical prerequisite for these analyses is the disentangling of "ex ante heterogeneity" from "risk/uncertainty" (also called ex post heterogeneity)—two sides of the same coin, with potentially very different implications for policy and welfare. But this is a challenging task, because inequality often arises from a mixture of heterogeneity and idiosyncratic risk, making the two difficult to disentangle. It requires researchers to carefully combine cross-sectional information with sufficiently long time-series data for analysis. The state-of-the-art methods used in this field increasingly blend the set of tools developed and used by quantitative macroeconomists with those used by structural econometricians. Despite the application of these sophisticated tools, there remains significant uncertainty in the profession regarding the magnitudes of idiosyncratic risks as well as whether or not these risks have increased since the 1970s.

The Imrohoroglu-Huggett-Aiyagari framework sidestepped a difficult issue raised by the lack of aggregation—that aggregates, including prices, depend on the entire wealth distribution. This was accomplished by abstracting from aggregate shocks, which allowed them to focus on stationary equilibria in which prices (the interest rate and the average wage) were simply some constants to be solved for in equilibrium. A far more challenging problem with incomplete markets arises in the presence of aggregate shocks, in which case equilibrium prices become *functions* of the entire wealth distribution, which varies with the aggregate state. Individuals need to know these equilibrium functions so that they can forecast how prices will evolve in the future as the aggregate state evolves in a stochastic manner. Because the wealth distribution is an infinite-dimensional object, an exact solution is typically not feasible. Krusell and Smith (1998) proposed a solution whereby one approximates the wealth distribution with a finite number of its moments (inspired by the idea that a given probability distribution can be represented by its moment-generating function). In a remarkable finding, they showed that the first moment (the mean) of the wealth distribution was all individuals needed to track in this economy for predicting all future prices. This result—generally known as "approximate aggregation"—is a double-edged sword. On the one hand, it makes feasible the solution of a wide range of interesting models with incomplete markets and aggregate shocks. On the other hand, it suggests that ex post heterogeneity does not often generate aggregate implications much different from a representative-agent model. So, the hope that some aggregate phenomena that were puzzling in representative-agent models could be explained in an incomplete markets framework is weakened with this result. While this is an important finding, there are many examples where heterogeneity *does* affect aggregates in a significant way. I discuss a variety of such examples in Section 6.

Finally, I turn to computation and calibration. First, in Section 5, I discuss some details of the Krusell-Smith method. A number of potential pitfalls are discussed and alternative checks of accuracy are studied. Second, an important practical issue that arises with calibrating/estimating large and complex quantitative models is the following. The objective function that we minimize often has lots of jaggedness, small jumps, and/or deep ridges because of a variety of reasons that have to do with approximations, interpolations, binding constraints, etc. Thus, local optimization methods are typically of little help on their own, because they very often get stuck in some local minima. In Section 7, I describe a global optimization algorithm that is simple yet powerful and is fully parallelizable without requiring any knowledge of MPI, OpenMP, and so on. It works on any number of computers that are connected to the Internet and have access to a synchronization service like DropBox. I provide a discussion of ways to customize this algorithm with different options to experiment.

## 1. AGGREGATION

Even in a simple static model with no uncertainty we need a way to deal with consumer heterogeneity. Adding dynamics and risk into this environment makes things more complex and requires a different set of conditions to be imposed. In this section, I will review some key theoretical results on various forms of aggregation. I begin with a very simple framework and build up to a fully dynamic model with idiosyncratic (i.e., individual-specific) risk and discuss what types of aggregation results one can hope to get and under what conditions.

Our interest in aggregation is not mainly for theoretical reasons. As we shall see, some of the conditions required for aggregation are satisfied (sometimes inadvertently!) by commonly used heterogeneous-agent frameworks, making them behave very much like a representative-agent model. Although this often makes the model easier to solve numerically, at the same time it can make its implications "boring"—i.e., too similar to a representative-agent model. Thus, learning about the assumptions underlying the aggregation theorems can allow model builders to choose the features of their models carefully so as to avoid such outcomes.

### A Static Economy

Consider a finite set $\mathcal{I}$ (with cardinality $I$) of consumers who differ in their preferences (over $l$ types of goods) and wealth in a static environment. Consider a particular good and let $x_i(p, w_i)$ denote the demand function of consumer $i$ for this good, given prices $p \in R^l$ and wealth $w_i$. Let $(w_1, w_2, ..., w_I)$ be the vector of wealth levels for all $I$ consumers. "Aggregate demand" in this economy can be written as

$$x\,(p, w_1, w_2, ..., w_I) = \sum_{i=1}^{I} x_i(p, w_i).$$

As seen here, the aggregate demand function $x$ depends on the entire wealth distribution, which is a formidable object to deal with. The key question then is, when can we write $x(p, w_1, w_2, ..., w_n) \equiv x(p, \sum w_i)$? For the wealth distribution to not matter, we need aggregate demand to not change for any redistribution of wealth that keeps aggregate wealth constant ($\sum dw_i = 0$). Taking the total derivative of $x$, and setting it to zero yields

$$\frac{\partial x\left(p, \sum w_i\right)}{\partial w_i} = 0 \Rightarrow \sum_{i=1}^{n} \frac{\partial x_i\,(p, w_i)}{\partial w_i} dw_i = 0$$

for all possible redistributions. This will only be true if

$$\frac{\partial x_i\,(p, w_i)}{\partial w_i} = \frac{\partial x_j\,(p, w_j)}{\partial w_j} \quad \forall i, j \in \mathcal{I}.$$

Thus, the key condition for aggregation is that individuals have the same marginal propensity to consume (MPC) out of wealth (or linear Engel curves). In one of the earliest works on aggregation, Gorman (1961) formalized this idea via restrictions on consumers' indirect utility function, which delivers the required linearity in Engel curves.

**Theorem 1 (Gorman 1961)** *Consider an economy with $N < \infty$ commodities and a set $\mathcal{I}$ of consumers. Suppose that the preferences of each consumer $i \in I$ can be represented by an indirect utility function*[4] *of the form*

$$v_i\,(p, w_i) = a_i\,(p) + b\,(p)\,w_i,$$

*and that each household $i \in \mathcal{I}$ has a positive demand for each commodity, then these preferences can be aggregated and represented by those of a representative household, with indirect utility*

$$v\,(p, w) = a\,(p) + b\,(p)\,w,$$

*where $a(p) = \sum_i a_i(p)$ and $w = \sum_i w_i$ is aggregate income.*

As we shall see later, the importance of linear Engel curves (or constant MPCs) for aggregation is a key insight that carries over to much more general models, all the way up to the infinite-horizon incomplete markets model with aggregate shocks studied in Krusell and Smith (1998).

## A Dynamic Economy (*No* Idiosyncratic Risk)

Rubinstein (1974) extends Gorman's result to a dynamic economy where individuals consume out of wealth (no income stream). Linear Engel curves are again central in this context.

Consider a frictionless economy in which each individual solves an intertemporal consumption-savings/portfolio allocation problem. That is, every period current wealth $w_t$ is apportioned between current consumption $c_t$ and a portfolio of a risk-free and a risky security with respective (gross) returns $R_t^f$ and $R_t^s$.[5] Let $\alpha_t$ denote the portfolio share of the risk-free asset at time $t$, and $\delta$ denote the subjective time discount factor. Individuals solve

[4] Denoting the consumer's utility function over goods with $U$, the indirect utility function is simply $v_i(p, w_i) \equiv U(x_i(p, w_i))$—that is, the maximum utility of a consumer who has wealth $w_i$ and faces price vector $p$.

[5] We can easily allow for multiple risky securities at the expense of complicating the notation.

$$\max_{\{c_t,\alpha_t\}} E\left(\sum_{t=1}^{T}\delta^t U\left(c_t\right)\right)$$
$$\text{s.t. } w_{t+1} = \left(w_t - c_t\right)\left(\alpha_t R_t^f + \left(1-\alpha_t\right) R_t^s\right).$$

Furthermore, assume that the period utility function, $U$, belongs to the hyperbolic absolute risk aversion (HARA) class, which is defined as utility functions that have linear risk tolerance: $T(c) \equiv -U(c)'/U(c)'' = \rho + \gamma c$ and $\gamma < 1$.[6] This class encompasses three utility functions that are well-known in economics: $U(c) = (\gamma-1)^{-1}(\rho+\gamma c)^{1-\gamma^{-1}}$ (generalized power utility; standard constant relative risk aversion [CRRA] form when $\rho \equiv 0$); $U(c) = -\rho \times \exp(-c/\rho)$ if $\gamma \equiv 0$ (exponential utility); and $U(c) = 0.5(\rho - c)^2$ defined for values $c < \rho$ (quadratic utility).

The following theorem gives six sets of conditions under which aggregation obtains.[7]

**Theorem 2 (Rubenstein 1974)** *Consider the following homogeneity conditions:*

*1. All individuals have the same resources $w_0$, and tastes $\delta$ and $U$.*

*2. All individuals have the same $\delta$ and taste parameters $\gamma \neq 0$.*

*3. All individuals have the same taste parameters $\gamma = 0$.*

*4. All individuals have the same resources $w_0$ and taste parameters $\rho = 0$ and $\gamma = 1$.*

*5. A complete market exists and all individuals have the same taste parameter $\gamma = 0$.*

*6. A complete market exists and all individuals have the same resources $w_0$ and taste $\delta$, $\rho = 0$, and $\gamma = 1$.*

*Then, all equilibrium rates of return are determined in case (1) as if there exist only composite individuals each with resources $w_0$ and tastes $\delta$ and $U$; and equilibrium rates of return are determined in cases (2)–(6) as if there exist only composite individuals each with the following economic characteristics: (i) resources: $w_0 = \sum w_0^i/I$; (ii) tastes: $\sigma = \Pi(\sigma^i)^{(\rho_i/\sum\rho_i)}$ (where $\sigma \equiv 1/\delta - 1$) or $\delta = \sum \delta^i/I$; and (iv) preference parameters: $\rho = \sum \rho_i/I$, and $\gamma$.*

Several remarks are in order.

---

[6] "Risk tolerance" is the reciprocal of the Arrow-Pratt measure of "absolute risk aversion," which measures consumers' willingness to bear a fixed amount of consumption risk. See, e.g., Pratt (1964).

[7] The language of Theorem 2 differs from Rubinstein's original statement by assuming rational expectations and combines results with the extension to a multiperiod setting in his footnote 5.

***Demand Aggregation***

An important corollary to this theorem is that whenever a composite consumer can be constructed, in equilibrium, rates of return are insensitive to the distribution of resources among individuals. This is because the aggregate demand functions (for both consumption and assets) depend only on total wealth and not on its distribution. Thus, we have "demand aggregation."

***Aggregation and Heterogeneity in Relative Risk Aversion***

Notice that all six cases that give rise to demand aggregation in the theorem require individuals to have the same curvature parameter, $\gamma$. To see why this is important, note that (with HARA preferences) the optimal holdings of the risky asset are a linear function of the consumer's wealth: $\kappa_1 + \kappa_2 w_t/\gamma$, where $\kappa_1$ and $\kappa_2$ are some constants that depend on the properties of returns. It is easy to see that with identical slopes, $\frac{\kappa_2}{\gamma}$, it does not matter who holds the wealth. In other words, redistributing wealth between any two agents would cause changes in total demand for assets that will cancel out each other, because of linearity and same slopes. Notice also that while identical curvature is a necessary condition, it is not sufficient for demand aggregation: Each of the six cases adds more conditions on top of this identical curvature requirement.[8]

## A Dynamic Economy (*With* Idiosyncratic Risk)

While Rubinstein's (1974) theorem delivers a strong aggregation result, it achieves this by abstracting from a key aspect of dynamic economies: uncertainty that evolves over time. Almost every interesting economy that we discuss in the coming sections will feature some kind of idiosyncratic risk that individuals face (coming from labor income fluctuations, shocks to health, shocks to housing prices and asset returns, among others). Rubinstein's (1974) theorem is silent about how the aggregate economy behaves under these scenarios.

This is where Constantinides (1982) comes into play: He shows that if markets are complete, under much weaker conditions (on preferences, beliefs, discount rates, etc.) one can replace heterogeneous consumers with a planner who maximizes a weighted sum of consumers' utilities. In turn, the central planner can be replaced by a composite consumer who maximizes a utility function of aggregate consumption.

To show this, consider a private ownership economy with production as in Debreu (1959), with $m$ consumers, $n$ firms, and $l$ commodities. As in Debreu

[8] Notice also that, because in some cases (such as [2]) heterogeneity in $\rho$ is allowed, individuals will exhibit different relative risk aversions (if they have different $w_t$), for example in the generalized CRRA case, and still allow aggregation.

(1959), these commodities can be thought of as date-event labelled goods (and concave utility functions, $U_i$, as being defined over these goods), allowing us to map these results into an intertemporal economy with uncertainty. Consumer $i$ is endowed with wealth $(w_{i1}, w_{i2}, ..., w_{il})$ and shares of firms $(\theta_{i1}, \theta_{i2}, ..., \theta_{in})$ with $\theta_{ij} \geq 0$ and $\sum_m \theta_{ij} = 1$. Let the vectors $C_i$ and $Y_j$ denote, respectively, individual $i$'s consumption set and firm $j$'s production set.

An equilibrium is an $(m+n+1)$-tuple $((\mathbf{c}_i^*)_{i=1}^m, (\mathbf{y}_j^*)_{j=1}^n, p^*)$ such that, as usual, consumers maximize utility, firms maximize their profits, and markets clear. Under standard assumptions, an equilibrium exists and is Pareto optimal.

Optimality implies that there exist positive numbers $\lambda_i$, $i = 1, ..., m$, such that the solution to the following problem (P1),

$$\begin{aligned} & \max_{\mathbf{c},\mathbf{y}} \sum_{i=1}^{m} \lambda_i U_i(\mathbf{c}_i) \qquad \text{(P1)} \\ \text{s.t. } \mathbf{y}_j &\in Y_j, \quad j = 1, 2, ...n; \\ \mathbf{c}_i &\in C_i, \quad i = 1, 2, ..., m; \\ \sum_{i=1}^{m} c_{ih} &= \sum_{j=1}^{n} y_{jh} + \sum_{i=1}^{m} w_{ih,} \quad h = 1, 2, ...l, \end{aligned}$$

(where $h$ indexes commodities) is given by $(\mathbf{c}_i) = (\mathbf{c}_i^*)$ and $(\mathbf{y}_j) = (\mathbf{y}_j^*)$. Let aggregate consumption be $\mathbf{z} \equiv (z_1, \cdots, z_l)$, $z_h \equiv \sum_{i=1}^m c_{ih}$. Now, for a given $\mathbf{z}$, consider the problem (P2) of efficiently allocating it across consumers:

$$\begin{aligned} U(\mathbf{z}) &\equiv \max_{\mathbf{c}} \sum_{i=1}^{m} \lambda_i U_i(\mathbf{c}_i) \qquad \text{(P2)} \\ \text{s.t. } \mathbf{c}_i &\in C_i, \quad i = 1, 2, ..., m, \\ \sum_{i=1}^{m} c_{ih} &= z_h, \quad h = 1, 2, ..., l. \end{aligned}$$

Now, given the production sets of each firm and the aggregate endowments of each commodity, consider the optimal production decision (P3):

$$\begin{aligned} & \max_{\mathbf{y},\mathbf{z}} U(\mathbf{z}) \qquad \text{(P3)} \\ \text{s.t. } \mathbf{y}_j &\in Y_j, \forall j; \quad z_h = \sum_j y_{jh} + w_h, \forall_h. \end{aligned}$$

**Theorem 3 (Constantinides [1982, Lemma 1])** *(a) The solution to (P3) is* $(\mathbf{y}_j) = (\mathbf{y}_j^*)$ *and* $z_h = \sum_{j=1}^n y_{jh}^* + w_h, \forall_h$.

*(b)* $U(\mathbf{z})$ *is increasing and concave in* $\mathbf{z}$.

*(c) If* $z_h = \sum y_{jh}^* + w_h, \forall_h$, *then the solution to (P2) is* $(x_i) = (x_i^*)$.

*(d) Given* $\lambda_i, i = 1, 2, \cdots, m$*, then if the consumers are replaced by one composite consumer with utility* $U(\mathbf{z})$*, with endowment equal to the sum of* $m$ *consumers' endowments and shares the sum of their shares, then the* $(1 + n + 1)$*-tuple* $(\sum_{i=1}^{m} \mathbf{c}_i^*, (\mathbf{y}_j^*)_{j=1}^{n}, p^*)$ *is an equilibrium.*

***Constantinides versus Rubinstein***

Constantinides allows for much more generality than Rubinstein by relaxing two important restrictions. First, no conditions are imposed on the homogeneity of preferences, which was a crucial element in every version of Rubinstein's theorem. Second, Constantinides allows for both exogenous endowment as well as production at every date and state. In contrast, recall that, in Rubinstein's environment, individuals start life with a wealth stock and receive no further income or endowment during life. In exchange, Constantinides requires complete markets and does not get demand aggregation. Notice that the existence of a composite consumer does not imply demand aggregation, for at least two reasons. First, composite demand depends on the weights in the planner's problem and, thus, depends on the distribution of endowments. Second, the composite consumer is defined at equilibrium prices and there is no presumption that its demand curve is identical to the aggregate demand function.

Thus, the usefulness of Constantinides's result hinges on (i) the degree to which markets are complete, (ii) whether we want to allow for idiosyncratic risk and heterogeneity in preferences (which are both restricted in Rubinstein's theorem), and (iii) whether or not we need demand aggregation. Below I will address these issues in more detail. We will see that, interestingly, even when markets are not complete, in certain cases, we will not only get close to a composite consumer representation, but we can also get quite close to the much stronger result of demand aggregation! An important reason for this outcome is that many heterogeneous-agent models assume identical preferences, which eliminates an important source of heterogeneity, satisfying Rubinstein's conditions for preferences. While these models do feature idiosyncratic risk, as we shall see, when the planning horizon is long such shocks can often be smoothed effectively using even a simple risk-free asset. More on this in the coming sections.

***Completing Markets by Adding Financial Assets***

It is useful to distinguish between "physical" assets—those in positive net supply (e.g., equity shares, capital, housing, etc.)—and "financial" assets—those in zero net supply (bonds, insurance contracts, etc.). The latter are simply some contracts written on a piece of paper that specify the conditions under which one agent transfers resources to another. In principle, it can be created with little cost. Now suppose that we live in a world with $J$ physical assets and

that there are $S(> J)$ states of the world. In this general setting, markets are incomplete. However, if consumers have homogenous tastes, endowments, and beliefs, then markets are (effectively) complete by simply adding enough *financial* assets (in zero net supply). There is no loss of optimality and nothing will change by this action, because in equilibrium identical agents will not trade with each other. The bottom line is that the more "homogeneity" we are willing to assume among consumers, the less demanding the complete markets assumption becomes. This point should be kept in mind as we will return to it later.

## 2. EMPIRICAL EVIDENCE ON INSURANCE

Dynamic economic models with heterogeneity typically feature individual-specific uncertainty that evolves over time—coming from fluctuations in labor earnings, health status, portfolio returns, among others. Although this structure does not fit into Rubinstein's environment, it is covered by Constantinides's theorem, which requires complete markets. Thus, a key empirical question is *the extent to which complete markets can serve as a useful benchmark* and a good approximation to the world we live in. As we shall see in this section, the answer turns out to be more nuanced than a simple yes or no.

To explain the broad variety of evidence that has been brought to bear on this question, this section is structured in the following way. First, I begin by discussing a large empirical literature that has tested a key prediction of complete markets—that marginal utility growth is equated across individuals. This is often called "perfect" or "full" insurance, and it is soundly rejected in the data. Next, I discuss an alternative benchmark, inspired by this rejection. This is the permanent income model, in which individuals have access to only borrowing and saving—or "self-insurance." In a way, this is the other extreme end of the insurance spectrum. Finally, I discuss studies that take an intermediate view—"partial insurance"—and provide some evidence to support it. We now begin with the tests of full insurance.

### Benchmark 1: Full Insurance

To develop the theoretical framework underlying the empirical analyses, start with an economy populated by agents who derive utility from consumption $c_t$ as well as some other good(s) $d_t$ : $U^i\left(c^i_{t+1}, d^i_{t+1}\right)$, where $i$ indexes individuals. These other goods can include leisure time (of husband and wife if the unit of analysis is a household), children, lagged consumption (as in habit formation models), and so on.

The key implication of perfect insurance can be derived by following two distinct approaches. The first environment assumes a social planner who pools

all individuals' resources and maximizes a social welfare function that assigns a positive weight to every individual. In the second environment, allocations are determined in a competitive equilibrium of a frictionless economy where individuals are able to trade in a complete set of financial securities. Both of these frameworks make the following strong prediction for the growth rate of individuals' marginal utilities:

$$\delta^i \frac{U_c^i\left(c_{t+1}^i, d_{t+1}^i\right)}{U_c^i\left(c_t^i, d_t^i\right)} = \frac{\Lambda_{t+1}}{\Lambda_t}, \tag{1}$$

where $U_c$ denotes the marginal utility of consumption and $\Lambda_t$ is the aggregate shock.[9] Thus, this condition says that every individual's marginal utility must grow in locksteps with the aggregate and, hence, with each other. No individual-specific term appears on the right-hand side, such as idiosyncratic income shocks, unemployment, sickness, and so on. All these idiosyncratic events are perfectly insured in this world. From here one can introduce a number of additional assumptions for empirical tractability.

### *Complete Markets and Cross-Sectional Heterogeneity: A Digression*

So far we have focused on what market completeness implies for the study of aggregate phenomena in light of Constantinides's theorem. However, complete markets also imposes restrictions on the evolution of the *cross-sectional distribution*, which can be seen in (1). For a given specification of $U$, (1) translates into restrictions on the evolutions of $c_t$ and $d_t$ (possibly a vector). Although it is possible to choose $U$ to be sufficiently general and flexible (e.g., include preference shifters, assume non-separability) to generate rich dynamics in cross-sectional distributions, this strategy would attribute all the action to preferences, which are essentially unobservable. Even in that case, models that are not bound by (1)—and therefore have idiosyncratic shocks affect individual allocations—can generate a much richer set of cross-sectional distributions. Whether that extra richness is necessary for explaining salient features of the data is another matter and is not always obvious (see, e.g., Caselli and Ventura [2000], Badel and Huggett [2007], and Guvenen and Kuruscu [2010]).[10]

---

[9] Alternatively stated, $\Lambda_t$ is the Lagrange multiplier on the aggregate resource constraint at time $t$ in the planner's problem or the state price density in the competitive equilibrium interpretation.

[10] Caselli and Ventura (2000) show that a wide range of distributional dynamics and income mobility patterns can arise in the Cass-Koopmans optimal savings model and in the Arrow-Romer model of productivity spillovers. Badel and Huggett (2007) show that life-cycle inequality patterns (discussed later) that have been viewed as evidence of incomplete markets can in fact be generated using a complete markets model. Guvenen and Kuruscu (2010) show that a human capital model with heterogeneity in learning ability and skill-biased technical change generates rich nonmonotonic

Now I return back to the empirical tests of (1).

In a pioneering article, Altug and Miller (1990) were the first to formally test the implications of (1). They considered households as their unit of analysis and specified a rich Beckerian utility function that included husbands' and wives' leisure times as well as consumption (food expenditures), and adjusted for demographics (children, age, etc.). Using data from the Panel Study of Income Dynamics (PSID), they could not reject full insurance. Hayashi, Altonji, and Kotlikoff (1996) revisited this topic a few years later and, using the same data set, they rejected perfect risk sharing.[11] Given this rejection in the whole population, they investigated if there might be better insurance *within* families, who presumably have closer ties with each other than the population at large and could therefore provide insurance to the members in need. They found that this hypothesis too was statistically rejected.[12]

In a similar vein, Guvenen (2007a) investigates how the extent of risk sharing varies across different wealth groups, such as stockholders and nonstockholders. This question is motivated by the observation that stockholders (who made up less than 20 percent of the population for much of the 20th century) own about 80 percent of net worth and 90 percent of financial wealth in the U.S. economy, and therefore play a disproportionately large role in the determination of macroeconomic aggregates. On the one hand, these wealthy individuals have access to a wide range of financial securities that can presumably allow better risk insurance; on the other hand, they are exposed to different risks not faced by the less-wealthy nonstockholders. Using data from the PSID, he strongly rejects perfect risk sharing among stockholders, but, perhaps surprisingly, does not find evidence against it among nonstockholders. This finding suggests further focus on risk factors that primarily affect the wealthy, such as entrepreneurial income risk that is concentrated at the top of the wealth distribution.

A number of other articles impose further assumptions before testing for risk sharing. A very common assumption is the separability between $c_t$ and $d_t$ (for example, leisure), which leads to an equation that only involves consumption (Cochrane 1991, Nelson 1994, Attanasio and Davis 1996).[13] Assuming power utility in addition to separability, we can take the logs of both sides of

---

dynamics consistent with the U.S. data since the 1970s, despite featuring no idiosyncratic shocks (and thus has complete markets).

[11] Data sets such as the PSID are known to go through regular revisions, which might be able to account for the discrepancy between the two articles' results.

[12] This finding has implications for the modeling of the household decision-making process as a unitary model as opposed to one in which there is bargaining between spouses.

[13] Non-separability, for example between consumption and leisure, can be allowed for *if* the planner is assumed to be able to transfer leisure freely across individuals. While transfers of consumption are easier to implement (through taxes and transfers), the transfer of leisure is harder to defend on empirical grounds.

equation (1) and then time-difference to obtain

$$\Delta C_{i,t} = \Delta \Lambda_t, \tag{2}$$

where $C_t \equiv \log(c_t)$ and $\Delta C_t \equiv C_t - C_{t-1}$. Several articles have tested this prediction by running a regression of the form

$$\Delta C_{i,t} = \Delta \Lambda_t + \Psi' \mathbf{Z}_t^i + \epsilon_{i,t}, \tag{3}$$

where the vector $\mathbf{Z}_t^i$ contains factors that are idiosyncratic to individual/household/group $i$. Perfect insurance implies that all the elements of the vector $\Psi$ are equal to zero.

Cochrane (1991), Mace (1991), and Nelson (1994) are the early studies that exploit this simple regression structure. Mace (1991) focuses on whether or not consumption responds to idiosyncratic wage shocks, i.e., $\mathbf{Z}_t^i = \Delta W_t^i$.[14] While Mace fails to reject full insurance, Nelson (1994) later points out several issues with the treatment of data (and measurement error in particular) that affect Mace's results. Nelson shows that a more careful treatment of these issues results in strong rejection.

Cochrane (1991) raises a different point. He argues that studies such as Mace's, that test risk sharing by examining the response of consumption growth to income, may have low power if income changes are (at least partly) anticipated by individuals. He instead proposes to use idiosyncratic events that are arguably harder to predict, such as plant closures, long strikes, long illnesses, and so on. Cochrane rejects full insurance for illness or involuntary job loss but not for long spells of unemployment, strikes, or involuntary moves. Notice that a crucial assumption in all of the work of this kind is that none of these shocks can be correlated with unmeasured factors that determine marginal utility growth.

Townsend (1994) tests for risk sharing in village economies of India and concludes that, although the model is statistically rejected, full insurance provides a surprisingly good benchmark. Specifically, he finds that individual consumption co-moves with village-level consumption and is not influenced much by own income, sickness, and unemployment.

Attanasio and Davis (1996) observe that equation (2) must also hold for multiyear changes in consumption and when aggregated across groups of individuals.[15] This implies, for example, that even if one group of individuals experiences faster income growth relative to another group during a 10-year period, their consumption growth must be the same. The substantial rise in the education premium in the United States (i.e., the wages of college graduates

[14] Because individual wages are measured with (often substantial) error in microsurvey data sets, an ordinary least squares estimation of this regression would suffer from attenuation bias, which may lead to a failure to reject full insurance even when it is false. The articles discussed here employ different approaches to deal with this issue (such as using an instrumental variables regression or averaging across groups to average out measurement error).

[15] Hayashi, Altonji, and Kotlikoff (1996) also use multiyear changes to test for risk sharing.

relative to high school graduates) throughout the 1980s provided a key test of perfect risk sharing. Contrary to this hypothesis, Attanasio and Davis (1996) find that the consumption of college graduates grows much faster than that of high school graduates during the same period, violating the premise of perfect risk sharing.

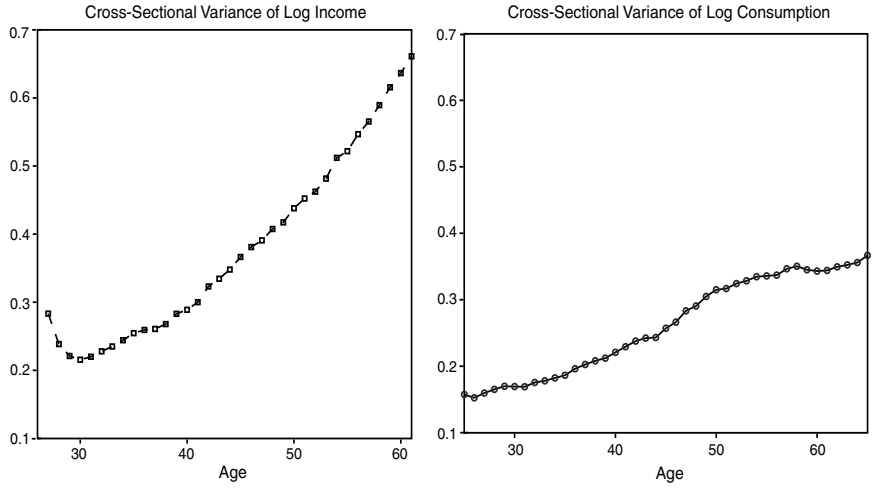
Finally, Schulhofer-Wohl (2011) sheds new light on this question. He argues that if more risk-tolerant individuals self-select into occupations with more (aggregate) income risk, then the regressions in (3) used by Cochrane (1991), Nelson (1994), and others (which incorrectly assume away such correlation) will be biased toward rejecting perfect risk sharing. By using self-reported measures of risk attitudes from the Health and Retirement Survey, Schulhofer-Wohl establishes such a correlation. Then he develops a method to deal with this bias and, applying the corrected regression, he finds that consumption growth responds very weakly to idiosyncratic shocks, implying much larger risk sharing than can be found in these previous articles. He also shows that the coefficients estimated from this regression can be mapped into a measure of "partial insurance."

### *Taking Stock*

As the preceding discussion makes clear, with few exceptions, all empirical studies agree that perfect insurance in the whole population is strongly rejected in a statistical sense. However, this statistical rejection per se is not sufficient to conclude that complete markets is a *poor benchmark* for economic analysis for two reasons. First, there seems to be a fair deal of insurance against certain types of shocks, as documented by Cochrane (1991) and Townsend (1994), and among certain groups of households, such as in some villages in less developed countries (Townsend 1994), or among nonstockholders in the United States (Guvenen 2007a). Second, the reviewed empirical evidence arguably documents statistical tests of an extreme benchmark (equation [1]) that we should not expect to hold precisely—for every household, against every shock. Thus, with a large enough sample, statistical rejection should not be surprising.[16] What these tests do not do is tell us how "far" the economy is from the perfect insurance benchmark. In this sense, analyses such as in Townsend (1994)—that identify the types of shocks that are and are not insured—are somewhat more informative than those in Altug and Miller (1990), Hayashi, Altonji, and Kotlikoff (1996), and Guvenen (2007a), which rely on model misspecification-type tests of risk sharing.

[16] One view is that hypothesis tests without an explicit alternative (such as the ones discussed here) often "degenerate into elaborate rituals designed to measure the sample size (Leamer 1983, 39)."

**Figure 1 Within-Cohort Inequality over the Life Cycle**




### Benchmark 2: Self-Insurance

The rejection of full consumption insurance led economists to search for other benchmark frameworks for studying individual choices under uncertainty. One of the most influential studies of this kind has been Deaton and Paxson (1994), who bring a different kind of evidence to bear. They begin by documenting two empirical facts. Using microdata from the United States, United Kingdom, and Taiwan, they first document that within-cohort inequality of labor income (as measured by the variance of log income) increases substantially and almost linearly over the life cycle. Second, they document that within-cohort consumption inequality shows a very similar pattern and also rises substantially as individuals age. The two empirical facts are replicated in Figure 1 from data in Guvenen (2007b, 2009a).

To understand what these patterns imply for the market structure, first consider a complete markets economy. As we saw in the previous section, if consumption is separable from leisure and other potential determinants of marginal utility, consumption growth will be equalized across individuals, independent of any idiosyncratic shock (equation [2]). Therefore, while consumption level may differ across individuals because of differences in permanent lifetime resources, this dispersion should *not* change as the cohort ages.[17] Therefore, Deaton and Paxson's (1994) evidence has typically been

[17] There are two obvious modifications that preserve complete markets and would be consistent with rising consumption inequality. The first one is to introduce heterogeneity in time

interpreted as contradicting the complete markets framework. I now turn to the details.

***The Permanent Income Model***

The canonical framework for self-insurance is provided by the permanent income life-cycle model, in which individuals only have access to a risk-free asset for borrowing and saving. Therefore, as opposed to full insurance, there is only "self-insurance" in this framework. Whereas the complete markets framework represents the maximum amount of insurance, the permanent income model arguably provides the lower bound on insurance (to the extent that we believe individuals have access to a savings technology, and borrowing is possible subject to some constraints).

It is instructive to develop this framework in some detail as the resulting equations will come in handy in the subsequent exposition. The framework here closely follows Hall and Mishkin (1982) and Deaton and Paxson (1994). Start with an income process with permanent and transitory shocks:

$$\begin{aligned} y_t &= y_t^P + \varepsilon_t, \\ y_t^P &= y_{t-1}^P + \eta_t. \end{aligned} \tag{4}$$

Suppose that individuals discount the future at the rate of interest and define: $\delta = 1/(1+r)$. Preferences are of quadratic utility form:

$$\begin{aligned} &\max E_0 \left[ -\frac{1}{2} \sum_{t=1}^{T} \delta_t \left( c^* - c_t \right)^2 \right] \\ &\text{s.t.} \sum_{t=1}^{T} \delta_t \left( y_t - c_t \right) + A_0 = 0, \end{aligned} \tag{5}$$

where $c^*$ is the bliss level and $A_0$ is the initial wealth level (which may be zero). This problem can be solved in closed form to obtain a consumption function. First-differencing this consumption rule yields

$$\Delta c_t = \eta_t + \gamma_t \varepsilon_t, \tag{6}$$

where $\gamma_t \equiv 1/\left( \sum_{\tau=0}^{T-t} \delta^\tau \right)$ is the annuitization factor.[18] This term is close to zero when the horizon is long and the interest rate is not too high, the

discounting. This is not very appealing because it "explains" by entirely relying on unobservable preference heterogeneity. Second, one could question the assumption of separability: If leisure is non-separable and wage inequality is rising over the life cycle—which it does—then consumption inequality would also rise to keep marginal utility growth constant (even under complete markets). But this explanation also predicts that hours inequality should also rise over the life cycle, a prediction that does not seem to be borne out in the data—although see Badel and Huggett (2007) for an interesting dissenting take on this point.

[18] Notice that the derivation of (6) requires two more pieces in addition to the Euler equation: It requires us to explicitly specify the budget constraint (5) as well as the stochastic process for income (4).

well-understood implication being that the response of consumption to transitory shocks is very weak given their low annuitized value. More importantly: Consumption responds to permanent shocks one-for-one. Thus, consumption changes reflect permanent income changes.

For the sake of this discussion, assume that the horizon is long enough so that $\gamma_t \approx 0$ and thus $\Delta c_t \cong \eta_t$. If we further assume that $\text{cov}_i\left(c^i_{t-1}, \eta^i_t\right) = 0$ (where $i$ indexes individuals and the covariance is taken cross-sectionally), we get

$$\text{var}_i\left(c^i_t\right) \cong \text{var}_i\left(c^i_{t-1}\right) + \text{var}\left(\eta_t\right).$$

So the rise in consumption inequality from age $t-1$ to $t$ is a measure of the variance of the permanent shock between those two ages. Since, as seen in Figure 1, consumption inequality rises significantly and almost linearly, this figure is consistent with permanent shocks to income that are fully accommodated as predicted by the permanent income model.

***Deaton and Paxson's Striking Conclusion***

Based on this evidence, Deaton and Paxson (1994) argue that the permanent income model is a better benchmark for studying individual allocations than is complete markets. Storesletten, Telmer, and Yaron (2004a) go one step further and show that a calibrated life-cycle model with incomplete markets can be quantitatively consistent with the rise in consumption inequality as long as income shocks are sufficiently persistent ($\rho \gtrsim 0.90$). In his presidential address to the American Economic Association, Robert Lucas (2003, 10) succinctly summarized this view: "The fanning out over time of the earnings and consumption distributions within a cohort that Deaton and Paxson [1994] document is striking evidence of a sizable, uninsurable random walk component in earnings." This conclusion was shared by the bulk of the profession in the 1990s and 2000s, giving a strong impetus to the development of incomplete markets models featuring large and persistent shocks that are uninsurable. I review many of these models in Sections 3 and 4. However, a number of recent articles have revisited the original Deaton-Paxson finding and have reached a different conclusion.

***Reassessing the Facts: An Opposite Conclusion***

Four of these articles, by and large, follow the same methodology as described and implemented by Deaton and Paxson (1994), but each uses a data set that extends the original Consumer Expenditure Survey (CE) sample used by these authors (that covered 1980–1990) and differ somewhat in their sample selection strategy. Specifically, Primiceri and van Rens (2009, Figure 2) use data from 1980–2000; Heathcote, Perri, and Violante (2010, Figure 14) use the 1980–1998 sample; Guvenen and Smith (2009, Figure 11) use the 1980–1992

sample and augment it with the 1972–73 sample; and Kaplan (2010, Figure 2) uses data from 1980–2003. Whereas Deaton and Paxson (1994, Figures 4 and 8) and Storesletten, Telmer, and Yaron (2004a, Figure 1) document a rise in consumption inequality of about 30 log points (between ages 25 and 65), these four articles find a much smaller rise of about 5–7 log points.

***Taking Stock***

Taken together, these re-analyses of CE data reveal that Deaton and Paxson's (1994) earlier conclusion is not robust to small changes in the sample period studied. Although more work on this topic certainly seems warranted,[19] these recent studies raise substantial concerns on one of the key pieces of empirical evidence on the extent of market incompleteness. A small rise in consumption inequality is hard to reconcile with the combination of large permanent shocks *and* self-insurance. Hence, if this latter view is correct, either income shocks are not as permanent as we thought or there is insurance above and beyond self-insurance. Both of these possibilities are discussed next.

## An Intermediate Case: Partial Insurance

A natural intermediate case to consider is an environment between the two extremes of full insurance and self-insurance. That is, perhaps individuals have access to various sources of insurance (e.g., through charities, help from family and relatives, etc.) *in addition* to borrowing and saving, but these forms of insurance still fall short of full insurance. If this is the case, is there a way to properly measure the degree of this "partial insurance?"

To address this question, Blundell, Pistaferri, and Preston (2008) examine the response of consumption to innovations in income. They start with equation (6) derived by Hall and Mishkin (1982) that links consumption change to income innovations, and modify it by introducing two parameters—$\theta$ and $\phi$—to encompass a variety of different scenarios:

$$\Delta c_t = \theta \eta_t + \phi \gamma_t \varepsilon_t. \tag{7}$$

Now, at one extreme is the self-insurance model (i.e., the permanent income model): $\theta = \phi = 1$; at the other extreme is a model with full insurance: $\theta = \phi = 0$. Values of $\theta$ and $\phi$ between zero and one can be interpreted as the degree of partial insurance—the lower the value, the more insurance there

[19] For example, as Attanasio, Battistin, and Ichimura (2007) show, the facts regarding the rise in consumption inequality over time are sensitive to whether one uses the "recall survey" or the "diary survey" in the CE data set. All the articles discussed in this section (on consumption inequality over the life cycle, including Deaton and Paxson [1994]) use the recall survey data. It would be interesting to see if the diary survey alters the conclusions regarding consumption inequality over the life cycle.

is. In their baseline analysis, Blundell, Pistaferri, and Preston (2008) estimate $\theta \approx \frac{2}{3}$ and find that it does not vary significantly over the sample period.[20] They interpret the estimate of $\theta$ to imply that about $\frac{1}{3}$ of permanent shocks are insured above and beyond what can be achieved through self-insurance.[21]

A couple of remarks are in order. First, the derivation of equation (6) that forms the basis of the empirical analysis here requires quadratic preferences. Indeed, this was the maintained assumption in Hall and Mishkin (1982) and Deaton and Paxson (1994). Blundell, Pistaferri, and Preston (2008) show that one can derive, as an approximation, an analogous equation (7) with CRRA utility and self-insurance, but now $\theta = \phi \approx \pi_{i,t}$, where $\pi_{i,t}$ is the ratio of human wealth to total wealth. In other words, the coefficients $\theta$ and $\phi$ are both equal to one under self-insurance only if preferences are of quadratic form; generalizing to CRRA predicts that even with self-insurance the response to permanent shocks, given by $\pi_{i,t}$, will be less than one-for-one if non-human wealth is positive. Thus, accumulation of wealth because of precautionary savings or retirement can dampen the response of consumption to permanent shocks and give the appearance of partial insurance. Blundell, Pistaferri, and Preston (2008) examine if younger individuals (who have less non-human wealth and thus have a higher $\pi_{i,t}$ than older individuals) have a higher response coefficient to permanent shocks. They do find this to be the case.

### *Insurance or Advance Information?*

Primiceri and van Rens (2009) conduct an analysis similar to Blundell, Pistaferri, and Preston (2008) and also find a small response of consumption to permanent income movements. However, they adopt a different interpretation for this finding—that income movements are largely "anticipated" by the individuals as opposed to being genuine permanent "shocks." As has been observed as far back as Hall and Mishkin (1982), this alternative interpretation illustrates a fundamental challenge with this kind of analysis: Advance information and partial insurance are difficult to disentangle by simply examining the response of consumption to income.

### *Insurance or Less Persistent Shocks?*

Kaplan and Violante (2010) raise two more issues regarding the interpretation of $\theta$. First, they ask, what if income shocks are persistent but not permanent?

---

[20] They also find $\phi\gamma_t = 0.0533$ (0.0435), indicating very small transmission of transitory shocks to consumption. This is less surprising since it would also be implied by the permanent income model.

[21] The parameter $\phi$ is of lesser interest given that transitory shocks are known to be smoothed quite well even in the permanent income model and the value of $\phi$ one estimates depends on what one assumes about $\gamma_t$—hence, the interest rates.

This is a relevant question because, as I discuss in the next section, nearly all empirical studies that estimate the persistence coefficient (of an AR(1) or ARMA(1,1)) find it to be 0.95 or lower—sometimes as low as 0.7. To explore this issue, they simulate data from a life-cycle model with self-insurance only, in which income shocks follow an AR(1) process with a first-order autocorrelation of 0.95. They show that when they estimate $\theta$ as in Blundell, Pistaferri, and Preston (2008), they find it to be close to the $\frac{2}{3}$ figure reported by these authors.[22] Second, they add a retirement period to the life-cycle model, which has the effect that now even a unit root shock is not permanent, given that its effect does not translate one-for-one into the retirement period. Thus, individuals have even more reason not to respond to permanent shocks, especially when they are closer to retirement. Overall, their findings suggest that the response coefficient of consumption to income can be generated in a model of pure self-insurance to the extent that income shocks are allowed to be slightly less than permanent.[23] One feature this model misses, however, is the age profile of response coefficients, which shows no clear trend in the data according to Blundell, Pistaferri, and Preston (2008), but is upward sloping in Kaplan and Violante's (2010) model.

### *Taking Stock*

Before the early 1990s, economists typically appealed to aggregation theorems to justify the use of representative-agent models. Starting in the 1990s, the widespread rejections of the full insurance hypothesis (necessary for Constantinides's [1982] theorem), combined with the findings of Deaton and Paxson (1994), led economists to adopt versions of the permanent income model as a benchmark to study individual's choices under uncertainty (Hubbard, Skinner, and Zeldes [1995], Carroll [1997], Carroll and Samwick [1997], Blundell and Preston [1998], Attanasio et al. [1999], and Gourinchas and Parker [2002], among many others). The permanent income model has two key assumptions: a single risk-free asset for self-insurance *and* permanent—or very persistent—shocks, typically implying substantial idiosyncratic risk. The more recent evidence, discussed in this subsection, however, suggests that a more appropriate benchmark needs to incorporate either more opportunities for partial insurance or idiosyncratic risk that is smaller than once assumed.

---

[22] The reason is simple. Because the AR(1) shock decays exponentially, this shock loses 5 percent of its value in one year, but $1 - 0.95^{10} \approx 40$ percent in 10 years and 65 percent in 20 years. Thus, the discounted lifetime value of such a shock is significantly lower than a permanent shock, which retains 100 percent of its value at all horizons.

[23] Another situation in which $\theta < 1$ with self-insurance alone is if permanent and transitory shocks are not separately observable and there is estimation risk.

## 3. INCOMPLETE MARKETS IN GENERAL EQUILIBRIUM

This section and the next discuss incomplete markets models in general equilibrium without aggregate shocks. Bringing in a general equilibrium structure allows researchers to jointly analyze aggregate and distributional issues. As we shall see, the two are often intertwined, making such models very useful. The present section discusses the key ingredients that go into building a general equilibrium incomplete markets model (e.g., types of risks to consider, borrowing limits, modeling individuals versus households, among others). The next section presents three broad questions that these models have been used to address: the cross-sectional distributions of consumption, earnings, and wealth. These are substantively important questions and constitute an entry point into broader literatures. I now begin with a description of the basic framework.

### The Aiyagari (1994) Model

In one of the first quantitative models with heterogeneity, Imrohoroglu (1989) constructed a model with liquidity constraints and unemployment risk that varied over the business cycle. She assumed that interest rates were constant to avoid the difficulties with aggregate shocks, which were subsequently solved by Krusell and Smith (1998). She used this framework to re-assess Lucas's (1987) earlier calculation of the welfare cost of business cycles. She found only a slightly higher figure than Lucas, mainly because of her focus on unemployment risk, which typically has a short duration in the United States.[24] Regardless of its empirical conclusions, this article represents an important early effort in this literature.

In what has become an important benchmark model, Aiyagari (1994) studies a version of the deterministic growth framework, with a Neoclassical production function and a large number of infinitely lived consumers (dynasties). Consumers are ex ante identical, but there is ex post heterogeneity because of idiosyncratic shocks to labor productivity, which are not directly insurable (via insurance contracts). However, consumers can accumulate a (conditionally) risk-free asset for self-insurance. They can also borrow in this asset, subject to a limit determined in various ways. At each point in time, consumers may differ in the history of productivities experienced, and hence in accumulated wealth.

[24] There is a large literature on the costs of business cycles following Lucas's original calculation. I do not discuss these articles here for brevity. Lucas's (2003) presidential address to the American Economic Association is an extensive survey of this literature that also discusses how Lucas's views on this issue evolved since the original 1987 article.

More concretely, an individual solves the following problem:

$$\max_{\{c_t\}} E_0 \left[ \sum_{t=0}^{\infty} \delta^t U(c_t) \right]$$
$$\text{s.t. } c_t + a_{t+1} = w l_t + (1 + r) a_t,$$
$$a_t \geq -B_{\min}, \tag{8}$$

and $l_t$ follows a finite-state first-order Markov process.[25]

There are (at least) two ways to embed this problem in general equilibrium. Aiyagari (1994) considers a production economy and views the single asset as the capital in the firm, which obviously has a positive net supply. In this case, aggregate production is determined by the savings of individuals, and both $r$ and the wage rate $w$, must be determined in general equilibrium. Huggett (1993) instead assumes that the single asset is a household bond in zero net supply. In this case, the aggregate amount of goods in the economy is exogenous (exchange economy), and the only aggregate variable to be determined is $r$.

The borrowing limit $B_{\min}$ can be set to the "natural" limit, which is defined as the loosest possible constraint consistent with certain repayment of debt: $B_{\min} = w l_{\min}/r$. Note that if $l_{\min}$ is zero, this natural limit will be zero. Some authors have used this feature to rule out borrowing (e.g., Carroll [1997] and Gourinchas and Parker [2002]). Alternatively, it can be set to some ad hoc limit stricter than the natural one. More on this later.

The main substantive finding in Aiyagari (1994) is that with incomplete markets, the aggregate capital stock is higher than it is with complete markets, although the difference is not quantitatively very large. Consequently, the interest rate is lower (than the time preference rate), which is also true in Huggett's (1993) exchange economy version. This latter finding initially led economists to conjecture that these models could help explain the equity premium puzzle,[26] which is also generated by a low interest rate. It turns out that while this environment helps, it is neither necessary nor sufficient to generate a low interest rate. I return to this issue later. Aiyagari (1994) also shows that the model generates the right ranking between different types of inequality: Wealth is more dispersed than income, which is more dispersed than consumption.

---

[25] Prior to Aiyagari, the decision problem described here was studied in various forms by, among others, Bewley (undated), Schechtman and Escudero (1977), Flavin (1981), Hall and Mishkin (1982), Clarida (1987, 1990), Carroll (1991), and Deaton (1991). With the exceptions of Bewley (undated) and Clarida (1987, 1990), however, most of these earlier articles did not consider general equilibrium, which is the main focus here.

[26] The equity premium puzzle of Mehra and Prescott (1985) is the observation that, in the historical data, stocks yield a much higher return than bonds over long horizons, which has turned out to be very difficult to explain by a wide range of economic models.

The frameworks analyzed by Huggett (1993) and Aiyagari (1994) contain the bare bones of a canonical general equilibrium incomplete markets model. As such, they abstract from many ingredients that would be essential today for conducting serious empirical/quantitative work, especially given that almost two decades have passed since their publication. In the next three subsections, I review three main directions the framework can be extended. First, the nature of idiosyncratic risk is often crucial for the implications generated by the model. There is a fair bit of controversy about the precise nature and magnitude of such risks, which I discuss in some detail. Second, and as I alluded to above, the treatment of borrowing constraints is very reduced form here. The recent literature has made significant progress in providing useful microfoundations for a richer specification of borrowing limits. Third, the Huggett-Aiyagari model considers an economy populated by bachelor(ette)s as opposed to families—this distinction clearly can have a big impact on economic decisions, which is also discussed.

### Nature of Idiosyncratic Income Risk[27]

The rejection of perfect insurance brought to the fore idiosyncratic shocks as important determinants of economic choices. However, after three decades of empirical research (since Lillard and Willis [1978]), a consensus among researchers on the nature of labor income risk still remains elusive. In particular, the literature in the 1980s and 1990s produced two—quite opposite—views on the subject. To provide context, consider this general specification for the wage process:

$$y_t^i = \underbrace{g\,(t, \text{observables}, ...)}_{\text{common systematic component}} + \underbrace{\left[\alpha^i + \beta^i t\right]}_{\text{profile heterogeneity}} + \underbrace{\left[z_t^i + \varepsilon_t^i\right]}_{\text{stochastic component}} \tag{9}$$

$$z_t^i = \rho z_{t-1}^i + \eta_t^i, \tag{10}$$

where $\eta_t^i$ and $\varepsilon_t^i$ are zero mean innovations that are i.i.d. over time and across individuals.

The early articles on income dynamics estimate versions of the process given in (9) from labor income data and find: $0.5 < \rho < 0.7$, and $\sigma_\beta^2 \gg 0$ (Lillard and Weiss 1979; Hause 1980). Thus, according to this first view, which I shall call the "heterogeneous income profiles" (HIP) model, individuals are subject to shocks with modest persistence, while facing life-cycle profiles that

[27] The exposition here draws heavily on Guvenen (2009a).

are individual-specific (and hence vary significantly across the population). As we will see in the next section, one theoretical motivation for this specification is the human capital model, which implies differences in income profiles if, for example, individuals differ in their ability level.

In an important article, MaCurdy (1982) casts doubt on these findings. He tests the null hypothesis of $\sigma_\beta^2 = 0$ and fails to reject it. He then proceeds by imposing $\sigma_\beta^2 \equiv 0$ before estimating the process in (9), and finds $\rho \approx 1$ (see, also, Abowd and Card [1989], Topel [1990], Hubbard, Skinner, and Zeldes [1995], and Storesletten, Telmer, and Yaron [2004b]). Therefore, according to this alternative view, which I shall call the "restricted income profiles" (RIP) model, individuals are subject to extremely persistent—nearly random walk—shocks, while facing similar life-cycle income profiles.

***MaCurdy's (1982) Test***

More recently, two articles have revived this debate. Baker (1997) and Guvenen (2009a) have shown that MaCurdy's test has low power and therefore the lack of rejection does not contain much information about whether or not there is growth rate heterogeneity. MaCurdy's test was generally regarded as the strongest evidence against the HIP specification, and it was repeated in different forms by several subsequent articles (Abowd and Card 1989; Topel 1990; and Topel and Ward 1992), so it is useful to discuss in some detail.

To understand its logic, notice that, using the specification in (9) and (10), the $n$th autocovariance of income growth can be shown to be

$$\operatorname{cov}\left(\Delta y_t^i, \Delta y_{t+n}^i\right) = \sigma_\beta^2 - \rho^{n-1}\left(\frac{1-\rho}{1+\rho}\sigma_\eta^2\right), \tag{11}$$

for $n \geq 2$. The idea of the test is that for *sufficiently large* $n$, the second term will vanish (because of exponential decay in $\rho^{n-1}$), leaving behind a positive autocovariance equal to $\sigma_\beta^2$. Thus, *if* HIP is indeed important—$\sigma_\beta^2$ is positive—then higher order autocovariances must be positive.

Guvenen (2009a) raises two points. First, he asks how large $n$ must be for the second term to be negligible. He shows that for the value of persistence he estimates with the HIP process ($\rho \cong 0.82$), the autocovariances in (11) do not even turn positive before the 13th lag (because the second term dominates), whereas MaCurdy only studies the first 5 lags. Second, he conducts a Monte Carlo analysis in which he simulates data using equation (9) with substantial heterogeneity in growth rates.[28] The results of this analysis are reproduced here in Table 1. MaCurdy's test does not reject the false null hypothesis of $\sigma_\beta^2 = 0$ for any sample size smaller than 500,000 observations (column 3)!

[28] More concretely, the estimated value of $\sigma_\beta^2$ used in his Monte Carlo analysis implies that at age 55 more than 70 percent of wage inequality is because of profile heterogeneity.

**Table 1 How Informative is MaCurdy's (1982) Test?**

| | | Autocovariances | | | Autocorrelations | |
|---|---|---|---|---|---|---|
| | | Data | HIP Process | | Data | HIP Process |
| Lag ↓ | $N \longrightarrow$ | 27,681 | 27,681 | 500,00 | 27,681 | 27,681 |
| 0 | | .1215 | .1136 | .1153 | 1.00 | 1.00 |
| | | (.0023) | (.00088) | (.00016) | (.000) | (.000) |
| 1 | | −.0385 | −.04459 | −.04826 | −.3174 | −.3914 |
| | | (.0011) | (.00077) | (.00017) | (.010) | (.0082) |
| 2 | | −.0031 | −.00179 | −.00195 | −.0261 | −.0151 |
| | | (.0010) | (.00075) | (.00018) | (.008) | (.0084) |
| 3 | | −.0023 | −.00146 | −.00154) | −.0192 | −.0128 |
| | | (.0008) | (.00079) | (.00020) | (.009) | (.0087) |
| 4 | | −.0025 | −.00093 | −.00120 | −.0213 | −.0080 |
| | | (.0007) | (.00074) | (.00019) | (.010) | (.0083) |
| 5 | | −.0001 | −.00080 | −.00093 | −.0012 | −.0071 |
| | | (.0008) | (.00081) | (.00020) | (.007) | (.0090) |
| 10 | | −.0017 | −.00003 | −.00010 | −.0143 | −.0003 |
| | | (.0006) | (.00072) | (.00019) | (.009) | (.0081) |
| 15 | | .0053 | .00017 | .00021 | .0438 | .0015 |
| | | (.0007) | (.00076) | (.00020) | (.008) | (.0086) |
| 18 | | .0012 | .00036 | .00030 | .0094 | .0032 |
| | | (.0009) | (.00076) | (.00018) | (.011) | (.0087) |

Notes: The table is reproduced from Guvenen (2009a, Table 3). $N$ denotes the sample size (number of individual-years) used to compute the statistics. Standard errors are in parentheses. The statistics in the "data" columns are calculated from a sample of 27,681 males from the PSID as described in that article. The counterparts from simulated data are calculated using the same number of individuals and a HIP process fitted to the covariance matrix of income residuals.

Even in that case, only the 18th autocovariance is barely significant (with a $t$-statistic of 1.67). For comparison, MaCurdy's (1982) data set included around 5,000 observations. Even the more recent PSID data sets typically contain fewer than 40,000 observations.

In light of these results, imposing the *a priori* restriction of $\sigma_\beta^2 = 0$ on the estimation exercise seems a risky route to follow. Baker (1997), Haider (2001), Haider and Solon (2006), and Guvenen (2009a) estimate the process in (9) without imposing this restriction and find substantial heterogeneity in $\beta^i$ and a low persistence, confirming the earlier results of Lillard and Weiss (1979) and Hause (1980). Baker and Solon (2003) use a large panel data set drawn from Canadian tax records and allow for both permanent shocks and profile heterogeneity. They find statistically significant evidence of both components.

In an interesting recent article, Browning, Ejrnaes, and Alvarez (2010) estimate an income process that allows for "lots of" heterogeneity. The authors use a simulated method of moments estimator and match a number of

moments whose economic significance is more immediate than the covariance matrix of earnings residuals, which has typically been used as the basis of a generalized method of moments estimation in the bulk of the extant literature. They uncover a lot of interesting heterogeneity, for example, in the innovation variance as well as in the persistence of AR(1) shocks. Moreover, they "find strong evidence against the hypothesis that any worker has a unit root." Gustavsson and Österholm (2010) use a long panel data set (1968–2005) from administrative wage records on Swedish individuals. They employ local-to-unity techniques on individual-specific time series and reject the unit root assumption.

***Inferring Risk versus Heterogeneity from Economic Choices***

Finally, a number of recent articles examine the response of consumption to income shocks to infer the nature of income risk. In an important article, Cunha, Heckman, and Navarro (2005) measure the fraction of individual-specific returns to education that are predictable by individuals by the time they make their college decision versus the part that represents uncertainty. Assuming a complete markets structure, they find that slightly more than half of the returns to education represent known heterogeneity from the perspective of individuals.

Guvenen and Smith (2009) study the joint dynamics of consumption and labor income (using PSID data) in order to disentangle "known heterogeneity" from income risk (coming from shocks as well as from uncertainty regarding one's own income growth rate). They conclude that a moderately persistent income process ($\rho \approx 0.7$–$0.8$) is consistent with the joint dynamics of income and consumption. Furthermore, they find that individuals have significant information about their own $\beta^i$ at the time they enter the labor market and hence face little uncertainty coming from this component. Overall, they conclude that with income shocks of modest persistence and largely predictable income growth rates, the income risk perceived by individuals is substantially smaller than what is typically assumed in calibrating incomplete markets models (many of which borrow their parameter values from MaCurdy [1982], Abowd and Card [1989], and Meghir and Pistaferri [2004], among others). Along the same lines, Krueger and Perri (2009) use rich panel data on Italian households and conclude that the response of consumption to income suggests low persistence for income shocks (or a high degree of partial insurance).[29]

Studying economic choices to disentangle risk from heterogeneity has many advantages. Perhaps most importantly, it allows researchers to bring a

[29] A number of important articles have also studied the response of consumption to income, such as Blundell and Preston (1998) and Blundell, Pistaferri, and Preston (2008). These studies, however, assume the persistence of income shocks to be constant and instead focus on what can be learned about the sizes of income shocks over time.

much broader set of data to bear on the question. For example, many dynamic choices require individuals to carefully weigh the different future risks they perceive against predictable changes before making a commitment. Decisions on home purchases, fertility, college attendance, retirement savings, and so on are all of this sort. At the same time, this line of research also faces important challenges: These analyses need to rely on a fully specified economic model, so the results can be sensitive to assumptions regarding the market structure, specification of preferences, and so on. Therefore, experimenting with different assumptions is essential before a definitive conclusion can be reached with this approach. Overall, this represents a difficult but potentially very fruitful area for future research.

## Wealth, Health, and Other Shocks

One source of idiosyncratic risk that has received relatively little attention until recently comes from shocks to wealth holdings, resulting for example from fluctuations in housing prices and stock returns, among others. A large fraction of the fluctuations in housing prices are because of local or regional factors and are substantial (as the latest housing market crash showed once again). So these fluctuations can have profound effects on individuals' economic choices. In one recent example, Krueger and Perri (2009) use panel data on Italian households' income, consumption, and wealth. They study the response of consumption to income and wealth shocks and find the latter to be very important. Similarly, Mian and Sufi (2011) use individual-level data from 1997–2008 and show that housing price boom leads to significant equity extraction—about 25 cents for every dollar increase in prices—which in turn leads to higher leverage and personal default during this time. Their "conservative" estimate is that home equity-based borrowing added $1.25 trillion in household debt and accounted for about 40 percent of new defaults from 2006–2008.

Another source of idiosyncratic shocks is out-of-pocket medical expenditures (hospital bills, nursing home expenses, medications, etc.), which can potentially have significant effects on household decisions. French and Jones (2004) estimate a stochastic process for health expenditures, modeled as a normal distribution adjusted to capture the risk of catastrophic health care costs. Simulating this process, they show that 0.1 percent of households every year receive a health cost shock with a present value exceeding $125,000. Hubbard, Skinner, and Zeldes (1994, 1995) represent the earliest efforts to introduce such shocks into quantitative incomplete markets models. The 1995 article shows that the interaction of such shocks with means-tested social insurance programs is especially important to account for in order to understand the very low savings rate of low-income individuals.

De Nardi, French, and Jones (2010) ask if the risk of large out-of-pocket medical expenditures late in life can explain the savings behavior of the elderly. They examine a new and rich data set called AHEAD, which is part of the Health and Retirement Survey conducted by the University of Michigan, which allows them to characterize medical expenditure risk for the elderly (even for those in their 90s) more precisely than previous studies, such as Hubbard, Skinner, and Zeldes (1995) and Palumbo (1999).[30] De Nardi, French, and Jones (2010) find out-of-pocket expenditures to rise dramatically at very old ages, which (in their estimated model) provides an explanation for the lack of significant dissaving by the elderly. Ozkan (2010) shows that the life-cycle profile of medical costs (inclusive of the costs paid by private and public insurers to providers) differs significantly between rich and poor households. In particular, on average, the medical expenses of the rich are higher than those of the poor until mid-life, after which the expenses of the poor exceed those of the rich—by 25 percent in absolute terms. Further, the expenses of the poor have thick tails—lots of individuals with zero expenses and many with catastrophically high costs. He builds a model in which individuals can invest in their health (i.e., preventive care), which affects the future distribution of health shocks and, consequently, the expected lifetime. High-income individuals do precisely this, which explains their higher spending early on. Low-income individuals do the opposite, which ends up costing more later in life. He concludes that a reform of the health care system that encourages use of health care for low-income individuals has positive welfare gains, even when fully accounting for the increase in taxes required to pay for them.

## Endogenizing Credit Constraints

The basic Aiyagari model features a reduced-form specification for borrowing constraints (8), and does not model the lenders' problem that gives rise to such constraints. As such, it is silent about potentially interesting variations in borrowing limits across individuals and states of the economy. A number of recent articles attempt to close this gap.

In one of the earliest studies of this kind, Athreya (2002) constructs a general equilibrium model of unsecured household borrowing to quantify the welfare effects of the Bankruptcy Reform Act of 1999 in the United States. In

[30] Palumbo's estimates of medical expenditures are quite a bit smaller than those in De Nardi, French, and Jones (2010), which are largely responsible for the smaller effects he quantifies. De Nardi, French, and Jones (2010) argue that one reason for the discrepancy could be the fact that Palumbo used data from the National Medical Care Expenditure Survey, which, unlike the AHEAD data set, does not contain direct measures of nursing home expenses. He did imputations from a variety of sources, which may be missing the large actual magnitude of such expenses found in the AHEAD data set.

the *pooling* equilibrium of this model (which is what Athreya focuses on), the competitive lending sector charges a higher borrowing rate than the market lending rate to break even (i.e., zero-profit condition), accounting for the fraction of households that will default. This framework allows him to study different policies, such as changing the stringency of means testing as well as eliminating bankruptcy altogether.

In an important article, Chatterjee et al. (2007) build a model of personal default behavior and endogenous borrowing limits. The model features (i) several types of shocks—to earnings, preferences, and liabilities (e.g., hospital and lawsuit bills, which precede a large fraction of defaults in the United States), (ii) a competitive banking sector, and (iii) post-bankruptcy legal treatment of defaulters that mimics the U.S. Chapter 7 bankruptcy code. The main contribution of Chatterjee et al. (2007) is to show that a *separating* equilibrium exists in which banks offer a *menu* of debt contracts to households whose interest rates vary optimally with the level of borrowing to account for the changing default probability. Using a calibrated version of the model, they quantify the separate contributions of earnings, preferences, and liability shocks to debt and default. Chatterjee and Eyigungor (2011) introduce collateralized debt (i.e., mortgage debt) into this framework to examine the causes of the run-up in foreclosures and crash in housing prices after 2007.

Livshits, MacGee, and Tertilt (2007) study a model similar to Chatterjee et al. (2007) in order to quantify the advantages to a "fresh start" bankruptcy system (e.g., U.S. Chapter 7) against a European style system in which debtors cannot fully discharge their debt in bankruptcy. The key tradeoff is that dischargeable debts add insurance against bad shocks, helping to smooth across states, but the inability to commit to future repayment increases interest rates and limits the ability to smooth across time. Their model is quite similar to Chatterjee et al. (2007), except that they model an explicit overlapping generations structure. They calibrate the model to the age-specific bankruptcy rate and debt-to-earnings ratio. For their baseline parameterization, they find that fresh-start bankruptcy is welfare improving, but that result is sensitive to the process for expenditure and income shocks, the shape of the earnings profile, and household size. Livshits, MacGee, and Tertilt (2010) build on this framework to evaluate several theories for the rise in personal bankruptcies since the 1970s. Finally, Glover and Short (2010) use the model of personal bankruptcy to understand the incorporation of entrepreneurs. Incorporation protects the owners' personal assets and their access to credit markets in case of default, but by increasing their likelihood of default, incorporation also implies a risk premium is built into their borrowing rate.

### From Bachelor(ette)s to Families

While the framework described above can shed light on some interesting distributional issues (e.g., inequality in consumption, earnings, and wealth), it is completely silent on a crucial source of heterogeneity—the household structure. In reality, individuals marry, divorce, have kids, and make their decisions regarding consumption, savings, labor supply, and so on jointly with these other life choices. For many economic and policy questions, the interaction between these domestic decisions and economic choices in an incomplete markets world can have a first-order effect on the answers we get. Just to give a few examples, consider these facts: Men and women are well-known to exhibit different labor supply elasticities; the tax treatment of income varies depending on whether an individual is single, married, and whether he/she has kids, etc.; the trends in the labor market participation rate in the United States since the 1960s have been markedly different for single and married women; the fractions of individuals who are married and divorced have changed significantly, again since the 1960s; and so on.

A burgeoning literature works to bring a richer household structure into macroeconomics. For example, in an influential article, Greenwood, Seshadri, and Yorukoglu (2005) study the role of household technologies (the widespread availability of washing machines, vacuum cleaners, refrigerators, etc.) in leading women into the labor market. Greenwood and Guner (2009) extend the analysis to study the marriage and divorce patterns since World War II. Jones, Manuelli, and McGrattan (2003) explore the role of the closing gender wage gap for married women's rising labor supply. Knowles (2007) argues that the working hours of men are too long when viewed through the lens of a unitary model of the household in which the average wage of females rises as in the data. He shows that introducing bargaining between spouses into the model reconciles it with the data. Guner, Kaygusuz, and Ventura (2010) study the effects of potential reforms in the U.S. tax system in a model of families with children and an extensive margin for female labor supply. Guvenen and Rendall (2011) study the insurance role of education for women against divorce risk and the joint evolution of education trends with those in marriage and divorce.

## 4. INEQUALITY IN CONSUMPTION, WEALTH, AND EARNINGS

A major use of heterogeneous-agent models is to study inequality or dispersion in key economic outcomes, most notably in consumption, earnings, and wealth. The Aiyagari model—as well as its aggregate-shock augmented version, the Krusell-Smith model presented in the next section—takes earnings dispersion to be exogenous and makes predictions about inequality in consumption and wealth. The bulk of the incomplete markets literature follows

this lead in their analysis. Some studies introduce an endogenous labor supply choice and instead specify the wage process to be exogenous, delivering earnings dispersion as an endogenous outcome (Pijoan-Mas [2006], Domeij and Floden [2006], Heathcote, Storesletten, and Violante [2008], among others). While this is a useful step forward, a lot of the dispersion in earnings before age 55 is because of wages and not hours, so the assumption of an exogenous wage process still leaves quite a bit to be understood. Other strands of the literature attempt to close this gap by writing models that also generate wage dispersion as an endogenous outcome in the model—for example, because of human capital accumulation (e.g., Guvenen and Kuruscu [2010, forthcoming], and Huggett, Ventura, and Yaron [2011]) or because of search frictions.[31]

## Consumption Inequality

Two different dimensions of consumption inequality have received attention in the literature. The first one concerns how much within-cohort consumption inequality increases over the life cycle. The different views on this question have been summarized in Section 2.[32] The second one concerns whether, and by how much, (overall) consumption inequality has risen in the United States since the 1970s, a question whose urgency was raised by the substantial rise in wage inequality during the same time. In one of the earliest articles on this topic, Cutler and Katz (1992) use data from the 1980s on U.S. households from the CE and find that the evolution of consumption inequality closely tracks the rise in wage inequality during the same time. This finding serves as a rejection of earlier claims in the literature (e.g., Jencks 1984) that the rise of means-tested in-kind transfers starting in the 1970s had improved the material well-being of low-income households relative to what would be judged by their income statistics.

Interest in this question was reignited more recently by a thought-provoking article by Krueger and Perri (2006), who conclude from an analysis of CE data that, from 1980–2003, *within-group* income inequality increased substantially more than within-group consumption inequality (in contrast, they find

[31] The search literature is very large with many interesting models to cover. I do not discuss these models here because I cannot do justice to this extensive body of work in this limited space. For an excellent survey, see Rogerson, Shimer, and Wright (2005). Note, however, that as Hornstein, Krusell, and Violante (2011) show, search models have trouble generating the magnitudes of wage dispersion we observe in the data.

[32] Another recent article of interest is Aguiar and Hurst (2008), who examine the life-cycle mean and variance profiles of the subcomponents of consumption—housing, utility bills, clothing, food at home, food away from home, etc. They show rich patterns that vary across categories, whereby the variance rises monotonically for some categories, while being hump-shaped for others, and yet declining monotonically for some others. The same patterns are observed for the mean profile. These disaggregated facts provide more food for thought to researchers.

that *between-group* income and consumption inequality tracked each other).[33] They then propose an explanation based on the premise that the development of financial services in the U.S. economy has helped households smooth consumption fluctuations relative to income variation.

To investigate this story, they apply a model of endogenous debt constraints as in Kehoe and Levine (1993). In this class of models, what is central is not the ability of households to pay back their debt, but rather it is their incentive or willingness to pay back. To give the right incentives, lenders can punish a borrower that defaults, for example, by banning her from financial markets forever (autarky). However, if the individual borrows too much or if autarky is not sufficiently costly, it may still make sense to default in certain states of the world. Thus, given the parameters of the economic environment, lenders will compute the optimal state-contingent debt limit, which will ensure that the borrower never defaults in equilibrium. Krueger and Perri (2006) notice that if income shocks are really volatile, then autarky is a really bad outcome, giving borrowers less incentive to default. Lenders who know this, in turn, are more willing to lend, which endogenously loosens the borrowing constraints. This view of the last 30 years therefore holds that the rise in the volatility of income shocks gave rise to the development of financial markets (more generous lending), which in turn led to a smaller rise in consumption inequality.[34]

Heathcote, Storesletten, and Violante (2007) argue that the small rise in consumption inequality can be explained simply if the rise in income shocks has been of a more transitory nature, since such shocks are easier to smooth through self-insurance. Indeed, Blundell and Preston (1998) earlier made the same observation and concluded that in the 1980s the rise in income shock variance was mostly permanent in nature (as evidenced by the observation that income and consumption inequality grew together), whereas in the 1990s it was mostly transitory given that the opposite was true. Heathcote, Storesletten, and Violante (2007) calibrate a fully specified model and show that it can go a long way toward explaining the observed trends in consumption inequality. One point to keep in mind is that these articles take as given that the volatility of income shocks rose during this period, a conclusion that is subject to uncertainty in light of the new evidence discussed above.

---

[33] Attanasio, Battistin, and Ichimura (2007) question the use of the CE interview survey and argue that some expenditure items are poorly measured in the survey relative to another component of CE, called the diary survey. They propose an optimal way of combining the two survey data and find that consumption inequality, especially in the 1990s has increased more than what is revealed by the interview survey alone.

[34] Aguiar and Bils (2011) take a different approach and construct a measure of CE consumption by using data on income and (self-reported) savings rate by households. They argue that consumption inequality tracked income inequality closely in the past 30 years. Although this is still preliminary work, the article raises some interesting challenges.

Before concluding, a word of caution about measurement. The appropriate price deflator for consumption may have trended differently for households in different parts of the income distribution (i.e., the "Walmart effect" at the lower end). To the extent that this effect is real, the measured trend in consumption inequality could be overstating the actual rise in the dispersion of material well-being. This issue still deserves a fuller exploration.

## Wealth Inequality

The main question about wealth inequality is a cross-sectional one: Why do we observe such enormous disparities in wealth, with a Gini coefficient of about 0.80 for net worth and a Gini exceeding 0.90 for financial wealth?

Economists have developed several models that can generate highly skewed wealth distributions (see, for example, Huggett [1996], Krusell and Smith [1998], Quadrini [2000], Castañeda, Díaz-Giménez, and Ríos-Rull [2003], Guvenen [2006], and Cagetti and De Nardi [2006]). These models typically use one (or more) of three mechanisms to produce this inequality: (1) dispersion in luck in the form of large and persistent shocks to labor productivity: the rich are luckier than the poor; (2) dispersion in patience or thriftiness: the rich save more than the poor; and (3) dispersion in rates of return: the rich face higher asset returns than the poor. This subsection describes a baseline model and variations of it that incorporate various combinations of the three main mechanisms that economists have used to generate substantial inequality in general equilibrium models.[35]

### *Dispersion in Luck*

Huggett (1996) asks how much progress can be made toward understanding wealth inequality using (i) a standard life-cycle model with (ii) Markovian idiosyncratic shocks, (iii) uncertain lifetimes, and (iv) a Social Security system. He finds that although the model can match the Gini coefficient for wealth in the United States, this comes from low-income households holding too little wealth, rather than the extreme concentration of wealth at the top in the U.S. economy. Moreover, whereas in the U.S. data the dispersion of wealth within each cohort is nearly as large as the dispersion across cohorts, the model understates the former significantly.

Castañeda, Díaz-Giménez, and Ríos-Rull (2003) study an enriched model that combines elements of Aiyagari (1994) and Huggett (1996). Specifically, the model (i) has a Social Security system, (ii) has perfectly altruistic bequests,

[35] Some of the models discussed in this section contain aggregate shocks in addition to idiosyncratic ones. While aggregate shocks raise some technical issues that will be addressed in the next section, they pose no problems for the exposition in this section.

(iii) allows for intergenerational correlation of earnings ability, (iv) has a progressive labor and estate tax system as in the United States, and (v) allows a labor supply decision. As for the stochastic process for earnings, they do not calibrate its properties based on microeconometric evidence on income dynamics as is commonly done, but rather they choose its features (the $4 \times 4$ transition matrix and four states of a Markov process) so that the model matches the cross-sectional distribution of earnings and wealth. To match the extreme concentration of wealth at the upper tail, this calibration procedure implies that individuals must receive a large positive shock (about 1,060 times the median income level) with a small probability. This high income level is also very fleeting—it lasts for about five years—which leads these high income individuals to save substantially (for consumption smoothing) and results in high wealth inequality.

***Dispersion in Patience***

Laitner's (1992, 2002) original insight was that wealth inequality could result from a combination of: (1) random heterogeneity in lifetime incomes across generations, and (2) altruistic bequests, which are constrained to be non-negative. Each newly born consumer in Laitner's model receives a permanent shock to his lifetime income and, unlike in the Aiyagari model, faces no further shocks to income during his lifetime. In essence, in Laitner's model only households that earn higher than average lifetime income want to transfer some amount to their offspring, who are not likely to be as fortunate. This altruistic motive makes these households effectively more thrifty (compared to those that earn below average income) since they also care about future utility. Thus, even small differences in lifetime income can result in large differences in savings rates—a fact empirically documented by Carroll (2000)—and hence in wealth accumulation.

The stochastic-beta model of Krusell and Smith (1998) is a variation on this idea in a dynastic framework, where heterogeneity in thrift (i.e., in the time-discount rate) is imposed exogenously.[36] Being more parsimonious, the stochastic-beta model also allows for the introduction of aggregate shocks. Krusell and Smith show that even small differences in the time discount factor that are sufficiently persistent are sufficient to generate the extreme skewness of the U.S. wealth distribution. The intuition for this result will be discussed in a moment.

[36] Notice that in this article I use $\delta$ to denote the time discount factor and $\beta$ was used to denote the income growth rate. I will continue with this convention, except when I specifically refer to the Krusell-Smith model, which has come to be known as a stochastic-beta model.

***Dispersion in Rates of Return***

Guvenen (2006) introduces return differentials into a standard stochastic growth model (i.e., in which consumers have identical, time-invariant discount factors and idiosyncratic shocks do not exist). He allows all households to trade in a risk-free bond, but restricts one group of agents from accumulating capital. Quadrini (2000) and Cagetti and De Nardi (2006) study models of inequality with entrepreneurs and workers, which can also generate skewed wealth distributions. The mechanisms have similar flavors: Agents who face higher returns end up accumulating a substantial amount of wealth.

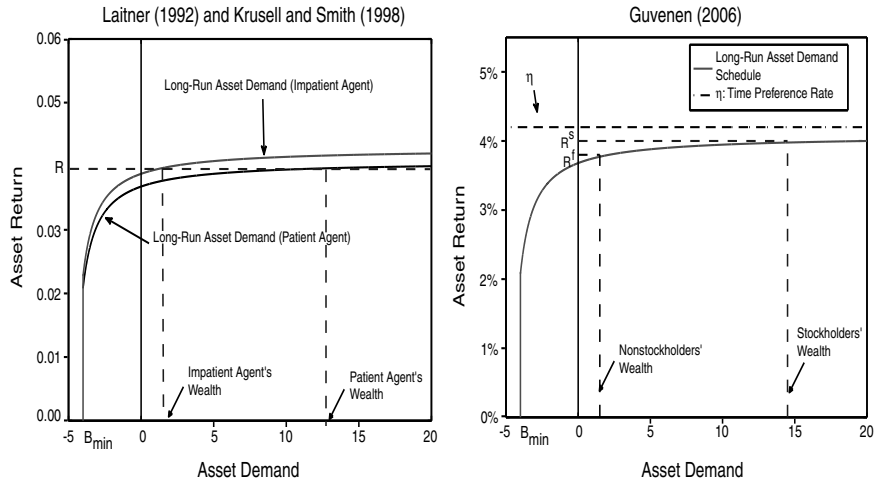
The basic mechanism in Guvenen (2006) can be described as follows. Nonstockholders have a precautionary demand for wealth (bonds), but the only way they can save is if stockholders are willing to borrow. In contrast, stockholders have access to capital accumulation, so they could smooth consumption even if the bond market was completely shut down. Furthermore, nonstockholders' asset demand is even more inelastic because they are assumed to have a lower elasticity of intertemporal substitution (consistent with empirical evidence) and therefore have a strong desire for consumption smoothing. Therefore, trading bonds for consumption smoothing is more important for nonstockholders than it is for stockholders. As a result, stockholders will only trade in the bond market if they can borrow at a low interest rate. This low interest rate in turn dampens nonstockholders' demand for savings further, and they end up with little wealth in equilibrium (and stockholders end up borrowing very little). Guvenen (2009b) shows that a calibrated version of this model easily generates the extremely skewed distribution of the relative wealth of stockholders to nonstockholders in the U.S. data.

***Can We Tell Them Apart?***

The determination of wealth inequality in the three models discussed so far can be explained using variations of a diagram used by Aiyagari (1994). The left panel of Figure 2 shows how wealth inequality is determined in Laitner's model and, given their close relationship, in the Krusell-Smith model. The top solid curve originating from "$-B_{\min}$" plots the long-run asset demand schedule for the impatient agent; the bottom curve is for the patient agent. A well-known feature of incomplete markets models is that the asset demand schedule is very flat for values of returns that are close to the time preference rate, $\eta$ (so $\delta \equiv 1/(1+\eta)$). Thus, both types of individuals' demand schedules asymptote to their respective time preference rates (with $\eta_{\text{patient}} < \eta_{\text{impatient}}$).[37] If the equilibrium return (which must be lower than $\eta_{\text{patient}}$ for an equilibrium to exist) is sufficiently close to $\eta_{\text{patient}}$, the high sensitivity of asset demands

[37] See, for example, Aiyagari (1994) and references therein. This result also holds when asset returns are stochastic (Chamberlain and Wilson 2000).

**Figure 2 Determination of Wealth Inequality in Various Models**




to interest rates will generate substantial wealth inequality between the two types of agents.

Similarly, the right panel shows the mechanism in the limited participation model, which has a similar flavor. For simplicity, let us focus on the case where stockholders and nonstockholders have the same preferences and face the same portfolio constraints. We have $\eta > R^S > R^f$. Again, given the sensitivity of asset demand to returns near $\eta$, even a small equity premium generates substantial wealth inequality. It should be stressed, however, that a large wealth inequality is not a foregone conclusion in any of these models. If returns were too low relative to $\eta$, individuals would be on the steeper part of their demand curves, which could result in smaller differences in wealth holdings.

While the mechanics described here may appear quite similar for the three models, their substantive implications differ in crucial ways. For example, consider the effect of eliminating aggregate shocks from all three models. In Guvenen (2006), there will be no equity premium without aggregate shocks and, consequently, no wealth inequality. In Krusell and Smith (1998), wealth inequality will increase as the patient agent holds more of the aggregate wealth (and would own all the wealth if there were no idiosyncratic shocks). In Laitner (1992), wealth inequality will remain unchanged, since it is created by idiosyncratic lifetime income risk. These dramatically different implications suggest that one can devise methods to bring empirical evidence to bear on the relevance of these different mechanisms.

***Cagetti and De Nardi (2006)***

Cagetti and De Nardi (2006) introduce heterogeneity across individuals in both work and entrepreneurial ability. Entrepreneurial firms operate decreasing returns to scale production functions, and higher entrepreneurial ability implies a higher optimal scale. Because debt contracts are not perfectly enforceable due to limited commitment, business owners need to put up some of their assets as collateral, a portion of which would be confiscated in case of default. Thus, entrepreneurs with very promising projects have more to lose from default, which induces them to save more for collateral, borrow more against it, and reach their larger optimal scale. The model is able to generate the extreme concentration of wealth at the top of the distribution (among households, many of whom are entrepreneurs).

Although this model differs from the limited participation framework in many important ways, the differential returns to saving is a critical element for generating wealth inequality in both models. This link could be important because many individuals in the top 1 percent and 5 percent of the U.S. wealth distribution hold significant amounts of stocks but are not entrepreneurs (hold no managerial roles), which the Cagetti/De Nardi model misses. The opposite is also true: Many very rich entrepreneurs are not stockholders (outside of their own company), which does not fit well with Guvenen's model (see Heaton and Lucas [2000] on the empirical facts about wealthy entrepreneurs and stockholders). The view that perhaps the very high wealth holdings of these individuals is driven by the higher returns that they enjoy—either as a stockholder or as an entrepreneur—can offer a unified theory of savings rate differences.

## Wage and Earnings Inequality

Because the consumption-savings decision is the cornerstone of the incomplete markets literature, virtually every model has implications for consumption and wealth inequality. The same is not true for earnings inequality. Many models assume that labor supply is inelastic and the stochastic process for wages is exogenous, making the implications for wage and earnings inequality to be mechanical reflections of the assumptions of the model. Even if labor supply is assumed to be endogenous, many properties of the earnings distribution (exceptions noted below) mimic those of the wage distribution. For things to get more interesting, it is the latter that needs to be endogenized.

In this subsection, I first review the empirical facts regarding wage inequality—both over the life cycle and over time. These facts are useful for practitioners since they are commonly used as exogenous inputs into incomplete markets models. Unless specifically mentioned, all the facts discussed here pertain to male workers, because the bulk of the existing work is

available consistently for this group.[38] Second, I discuss models that attempt to endogenize wages and explain the reported facts and trends.

***Inequality Over the Life Cycle***

The main facts about the evolution of (within-cohort) earnings inequality over the life cycle were first documented by Deaton and Paxson (1994) and shown in the left panel of Figure 1. The same exercise has been repeated by numerous authors using different data sets or time periods (among others, Storesletten, Telmer, and Yaron [2004a], Guvenen [2009a], Heathcote, Perri, and Violante [2010], and Kaplan [2010]). While the magnitudes differ somewhat, the basic fact that wage and earnings inequality rise substantially over the life cycle is well-established.

One view is that this fact does not require an elaborate explanation, because wages follow a very persistent, perhaps permanent, stochastic process as implied by the RIP model. Thus, the rising life-cycle inequality is simply a reflection of the accumulation of such shocks, which drive up the variance of log wages in a linear fashion (in the case of permanent shocks). I will continue to refer to this view as the RIP model because of its emphasis on persistent "shocks."[39]

An alternative perspective, which has received attention more recently, emphasizes systematic factors—heterogeneity as opposed to random shocks. This view is essentially in the same spirit as the HIP model of the previous section. But it goes one step further by *endogenizing* the wage distribution based on the human capital framework of Becker (1964) and Ben-Porath (1967), among others. In an influential article, Huggett, Ventura, and Yaron (2006) study the distributional implications of the standard Ben-Porath (1967) model by asking about the types of heterogeneity that one needs to introduce to generate patterns consistent with the U.S. data. They find that too much heterogeneity in initial human capital levels results in the counterfactual implication that wage inequality *should fall* over the life cycle. In contrast, heterogeneity in learning ability generates a rise in wage inequality consistent with the data. A key implication of this finding is that the rise in wage inequality can be generated without appealing to idiosyncratic shocks of any kind. Instead, it is the systematic fanning out of wage profiles, resulting from different investment rates, that generates rising inequality over the life cycle. Guvenen, Kuruscu, and Ozkan (2009) and Huggett, Ventura, and Yaron (2011) introduce

[38] Some of the empirical trends discussed also apply to women, while others do not. See Autor, Katz, and Kearney (2008) for a comparative look at wage trends for males and females during the period.

[39] Of course, one can write deeper economic models that generate the observation that wages follow a random walk process, such as the learning model of Jovanovic (1979) in a search and matching environment, or the optimal contracts in the limited commitment model of Harris and Holmstrom (1982).

idiosyncratic shocks into the Ben-Porath framework. Both articles find that heterogeneous growth rates continue to play the dominant role for the rise in wage inequality. The Ben-Porath formulation is also central for wage determination in Heckman, Lochner, and Taber (1998) and Kitao, Ljungqvist, and Sargent (2008).

***Inequality Trends Over Time***

A well-documented empirical trend since the 1970s is the rise in wage inequality among male workers in the United States. This trend has been especially prominent above the median of the wage distribution: For example, the log wage differential between the 90th and the 50th percentiles has been expanding in a secular fashion for the past four decades. The changes at the bottom have been more episodic, with the log 50-10 wage differential strongly expanding until the late 1980s and then closing subsequently (see Autor, Katz, and Kearney [2008] for a detailed review of the evidence). Acemoglu (2002) contains an extensive summary of several related wage trends, as well as a review of proposed explanations. Here I only discuss the subset of articles that are more closely relevant for the incomplete markets macroliterature.

***Larger Shocks or Increasing Heterogeneity?***

Economists' interpretations of the rise in wage inequality over the life cycle and over time are intimately related. The RIP view that was motivated by analyses of life-cycle wages was dominant in the 1980s and 1990s, so it was natural for economists to interpret the rise in wage inequality *over time*, through the same lens. Starting with Gottschalk and Moffitt (1994) and Moffitt and Gottschalk (1995), this trend has been broadly interpreted as reflecting a rise in the variances of idiosyncratic shocks, either permanent or transitory (Meghir and Pistaferri 2004; Heathcote, Storesletten, and Violante 2008; etc.). This approach remains the dominant way to calibrate economic models that investigate changes in economic outcomes from the 1970s to date.

However, some recent articles have documented new evidence that seems hard to reconcile with the RIP view. The first group of articles revisits the econometric analyses of wage and earnings data. Among these, Sabelhaus and Song (2009, 2010) use panel data from Social Security records covering millions of American workers, in contrast to the long list of previous studies that use survey data (e.g., the PSID).[40] While this data set has the potential drawback of under-reporting (because it is based on income reported to the Internal Revenue Service), it has three important advantages: (i) a much larger sample size (on the order of 50+ million observations, compared to at

[40] These include, among others, Meghir and Pistaferri (2004), Dynan, Elmendorf, and Sichel (2007), Heathcote, Storesletten, and Violante (2008), and Shin and Solon (2011).

most 50,000 in the PSID), (ii) no survey response error, and (iii) no attrition. Sabelhaus and Song find that the volatility of annual earnings growth increased during the 1970s, but that it *declined* monotonically during the 1980s and 1990s. Furthermore, applying the standard permanent-transitory decomposition as in Moffitt and Gottschalk (1995) and Meghir and Pistaferri (2004) reveals that permanent shock variances were stable and transitory shocks became *smaller* from 1980 into the 2000s. A separate study conducted by the Congressional Budget Office (2008), also using wage earnings from Social Security records from 1980–2003, reached the same conclusion.[41] Finally, Kopczuk, Saez, and Song (2010) document (also using Social Security data) that both long-run and short-run mobility have stayed remarkably stable from the 1960s into the 2000s. But this finding seems difficult to reconcile with Moffitt and Gottschalk (1995) and the subsequent literature that found permanent and transitory shock variances to have risen in different subperiods from 1970 to the 2000s. If true, the latter would result in fluctuations in mobility patterns over these subperiods, which is not borne out in Kopczuk, Saez, and Song's (2010) analysis.

Another piece of evidence from income data is offered by Haider (2001), who estimates a stochastic process for wages similar to the one in Moffitt and Gottschalk (1995) and others, but with one key difference. He allows for individual-specific wage growth rates (HIP) and he also allows for the dispersion of growth rates to vary over time. The stochastic component is specified as an ARMA(1,1). With this more flexible specification, he finds no evidence of a rise in the variance of income shocks after the 1970s, but instead finds a large increase in the dispersion of systematic wage growth rates.

A second strand of the literature studies the trends in labor market flows in the United States. These articles do not find any evidence of rising job instability or churning, which one might expect to see in conjunction with larger idiosyncratic shocks. In contrast, these studies document an across-the-board moderation in labor market flows. For example, Gottschalk and Moffitt (1999) focus on male workers between the ages of 20 and 62 and conclude their analysis as follows:

> [W]e believe that a consistent picture is emerging on changes in job stability and job security in the 1980s and 1990s. Job instability does

[41] Sabelhaus-Song attribute the reason why some earlier studies found rising variances of wage shocks (e.g., Moffitt and Gottschalk 2008) to the inclusion of individuals with self-employment income and those who earn less than the Social Security minimum. Even though there are few of these households, Sabelhaus and Song show that they make a big difference in the computed statistics. Similarly, Shin and Solon (2011, 978–80) use PSID data and also do not find a trend in the volatility of wage earnings changes during the 1980s and 1990s. They argue that the increasing volatility found in earlier studies, such as Dynan, Elmendorf, and Sichel (2007), seems to be coming from the inclusion of some auxiliary types of income (from business, farming, etc.) whose treatment has been inconsistent in the PSID over the years.

> not seem to have increased, and the consequences of separating from an employer do not seem to have worsened.[42]

Shimer (2005, 2007) and Davis et al. (2010) extend this analysis to cover the 2000s and use a variety of data sets to reach the same conclusion. Further, both articles show that expanding the sample of individuals to include women and younger workers shows a *declining* trend in labor market flows and an *increase* in job security.

***Taking Stock***

To summarize, the seeming consensus of the 1990s—that rising wage inequality was driven by an increase in idiosyncratic shock variances—is being challenged by a variety of new evidence, some of which comes from data sets many orders of magnitude larger than the surveys used in previous analyses. In addition, the evidence from labor market flows described above—while perhaps more indirect—raises questions about the sources of larger idiosyncratic shocks in a period where labor market transitions seem to have moderated. Although, it would be premature to conclude that the alternative view is the correct one—more evidence is needed to reach a definitive conclusion. Having said that, if this alternative view *is* true, and income shock variances have not increased, this new "fact" would require economists to rethink a variety of explanations put forward for various trends, which assumed a rise in shock variances during this period.

## 5. HETEROGENEITY WITH AGGREGATE SHOCKS

Krusell and Smith (1998) add two elements to the basic Aiyagari framework. First, they introduce aggregate technology shocks. Second, they assume that the cumulative discount factor at time $t$ (which was assumed to be $\delta^t$ before), now follows the stochastic process $\delta_t = \tilde{\delta}\delta_{t-1}$, where $\tilde{\delta}$ is a finite-state Markov chain. The stochastic evolution of the discount factors within a dynasty captures some elements of an explicit overlapping-generations structure with altruism and less-than-perfect correlation in genes between parents and children, as in Laitner (1992). With this interpretation in mind, the evolution of $\tilde{\delta}$ is calibrated so that the average duration of any particular value of the discount factor is equal to the lifetime of a generation. (Krusell and Smith [1997] study a version of this model where consumers are allowed to hold a risk-free bond in addition to capital.)

The specifics of the model are as follows. There are two types of shocks: (i) idiosyncratic employment status: $(\epsilon_e, \epsilon_u) \equiv$ (employed, unemployed);

[42] They also say, "Almost all studies based on the various Current Population Surveys (CPS) supplements...show little change in the overall separation rates through the early 1990s."

and (ii) aggregate productivity: $(z_g, z_b) \equiv$ (expansion, recession). Employment status and aggregate productivity jointly evolve as a first-order Markov process. Assume that $\epsilon$ is i.i.d. conditional on $z$, so the fraction of employed workers (and hence $l$) only depends on $z$. Competitive markets imply

$$w(K, L, z) = (1-\alpha) z (K/L)^{-\alpha}, \text{ and } r(K, L, z) = \alpha z (K/L)^{\alpha-1}. \quad (12)$$

Finally, the entire wealth distribution, which I denote with $\Gamma$ is a state variable for this model, and let $\Gamma' = H(\Gamma, z; z')$ denote its endogenous transition function (or law of motion).

**Krusell-Smith Algorithm**

A key equilibrium object in this class of models is the law of motion, $H$. In principle, computing this object is a formidable task since the distribution of wealth is infinite-dimensional. Krusell and Smith (1997, 1998) show, however, that this class of models, when reasonably parameterized, exhibits "approximate aggregation": Loosely speaking, to predict future prices consumers need to forecast only a small set of statistics of the wealth distribution rather than the entire distribution itself. This result makes it possible to use numerical methods to analyze this class of models. Another key feature of the Krusell-Smith algorithm is that it solves the model by simulating it. Specifically, the basic version of the algorithm works as follows:

1. Approximate $\Gamma$ with a finite number ($I$) of moments. ($H$ reduces to a function mapping the $I$ moments today into the $I$ moments tomorrow depending on $z$ today.)

   (a) We will start by selecting one moment—the mean—so $I = 1$.[43]

2. Select a family of functions for $H$. I will choose a log-linear function following Krusell and Smith.

$$\begin{aligned} V(k, \epsilon; \Gamma, z) &= \max_{c, k'} \left[ U(c) + \delta E \left[ V\left(k', \epsilon'; \Gamma', z'\right) \mid z, \epsilon \right] \right] \\ c + k' &= w(K, L, z) \times l \times \epsilon + r(K, L, z) \times k, \quad k' \geq 0 \\ \log K' &= a_0 + a_1 \log K \quad \text{for } z = z_b \\ \log K' &= b_0 + b_1 \log K \quad \text{for } z = z_g \end{aligned}$$

3. Make an (educated) initial guess about $(a_0, a_1, b_0, b_1) \Longrightarrow$ yields initial guess for $H_0$. Make also an initial guess for $\Gamma_0$.

[43] When we add more moments, we do not have to proceed as mean, variance, skewness, and so on. We can include, say, the wealth holdings of the top 10 percent of population, mean-to-median wealth ratio, etc.

4. Solve the consumer's dynamic program. Using only the resulting decision rules, simulate $\{k_{n,t}\}_{n=1,t=1}^{N,T}$ for $(N, T)$ large.

5. Update $H$ by estimating (where $\tilde{K} = \frac{1}{N}\sum_{n=1}^{N} k_n$):

$$\begin{aligned} \log \tilde{K}' &= a_0 + a_1 \log \tilde{K} \quad \text{for } z = z_b \\ \log \tilde{K}' &= b_0 + b_1 \log \tilde{K} \quad \text{for } z = z_g \end{aligned}$$

6. Iterate on 4–5, until the $R^2$ of this regression is "sufficiently high" and the forecast variance is "small."

   (a) If accuracy remains insufficient, go back to step 1 and increase $I$.

## Details

### *Educated Initial Guess*

As with many numerical methods, a good initial guess is critical. More often than not, the success or failure of a given algorithm will depend on the initial guess. One idea (used by Krusell and Smith) is to first solve a standard representative-agent real business cycle (RBC) model with the same parameterization. Then estimate the coefficients $(a_0, a_1, b_0, b_1)$ using capital series simulated from this model to obtain an initial guess for step 1 above.[44] More generally, a good initial guess can often be obtained by solving a simplified version of the full model. Sometimes this simplification involves ignoring certain constraints, sometimes by shutting down certain shocks, and so on.

### *Discretizing an AR(1) Process*

Oftentimes, the exogenous driving force in incomplete markets models is assumed to be generated by an AR(1) process, which is discretized and converted into a Markov chain. One popular method for discretization is described in Aiyagari (1993) and has been used extensively in the literature. However, an alternative method by Rouwenhorst (1995) (and which received far less attention until recently) is far superior in the quality of the approximation that it provides, especially when the process is very persistent, which is often the case. Moreover, it is very easy to implement. Kopecky and Suen (2010) and Galindev and Lkhagvasuren (2010) provide comprehensive comparisons of different discretization methods, which reveal the general superiority of

[44] Can't we update $H$ without simulating? Yes, we can. Den Haan and Rendahl (2009) propose a method where they use the policy functions for capital holdings and integrate them over distribution $\Lambda(k, \epsilon)$ of households across capital and employment status: $K' = H_j(K, z) = \int k'_j(k, \epsilon; \Gamma, z) d\Lambda(k, \epsilon)$. This works well when the decision rules are parameterized in a particular way. See den Haan and Rendahl (2009).

Rouwenhorst's (1995) method. They also show how this method can be extended to discretize more general processes.

### *Non-Trivial Equilibrium Pricing Function*

One simplifying feature of Krusell and Smith (1998) is that equilibrium prices (wages and interest rates) are determined trivially by the marginal product conditions (12). Thus, they depend only on the aggregate capital stock and not on its distribution. Some models do not have this structure—instead pricing functions must be determined by equilibrium conditions—such as market-clearing or zero-profit conditions—that explicitly depend on the wealth distribution. This would be the case, for example, if a household bond is traded in the economy. Its price must be solved for using a market-clearing condition, which is a challenging task. Moreover, if there is an additional asset, such as a stock, two prices must be determined simultaneously, and this must be done in such a way that avoids providing arbitrage opportunities—along the iterations of the solution algorithm. Otherwise, individuals' portfolio choices will go haywire (in an attempt to take advantage of perceived arbitrage), wreaking havoc with the solution algorithm. Krusell and Smith (1997) solve such a model and propose an algorithm to tackle these issues. I refer the interested reader to that article for details.

### *Checking for Accuracy of Solution*

Many studies with aggregate fluctuations and heterogeneity use two simple criteria to asses the accuracy of the law of motion in their limited information approximation. If agents are using the law of motion

$$\log K' = \alpha_1 + \alpha_2 z + \alpha_3 \log K + u, \tag{13}$$

a perfectly solved model should find $u = 0$. Thus, practitioners will continue to solve the model until either the $R^2$ of this regression is larger than some minimum or $\sigma_u$ falls below some threshold (step 6 in the algorithm above).

However, one should *not* rely solely on $R^2$ and $\sigma_u$. There are at least three reasons for this (den Haan 2010). First, both measures average over all periods of the simulation. Thus, infrequent but large deviations from the forecast rule can be hidden in the average. These errors may be very important to agents and their decision rule. For example, the threat of a very large recession may increase buffer stock saving, but the approximation may understate the movement of capital in such a case. Second, and more importantly, these statistics only measure one-step-ahead forecasts. The regression only considers the dynamics from one period to the next, so errors are only the deviations between the actual next-period capital and the expected amount in the next period. This misses potentially large deviations between the *long-term* forecast for capital and its actual level. Aware of this possibility, Krusell and Smith (1998) also

check the $R^2$ for forecasting prices 25 years ahead (100 model periods) and find it to be extremely high as well! (They also check the maximum error in long-term forecasts, which is very small.) Third, $R^2$ is scaled by the left-hand side of the regression. An alternative is to check the $R^2$ of

$$\log K' - \log K = \alpha_1 + \alpha_2 z + (\alpha_3 - 1) \log K.$$

As a particularly dramatic demonstration, den Haan (2010) uses a savings decision rule that solves the Krusell and Smith (1998) model, simulates it for $T$ periods, and estimates a law of motion in the form of equation (13). He then manipulates $\alpha_1, \alpha_3$ such that $T^{-1} \sum u_t = 0$ but the $R^2$ falls from 0.9999 to 0.999 and then 0.99. This has economically meaningful consequences: The time series standard deviation of the capital stock simulated from the perturbed versions of equation (13) falls to 70 percent and then 46 percent of the true figure.

Finally, den Haan (2010) proposes a useful test that begins with the approximated law of motion,

$$\log K' = \hat{\alpha}_1 + \hat{\alpha}_2 z + \hat{\alpha}_3 \log K + u, \tag{14}$$

to generate a sequence of realizations of $\left\{\tilde{K}_{t+1}\right\}_{t=0}^{T}$ and then compares these to the sequence generated by aggregating from decision rules, the true law of motion. Because $\left\{\tilde{K}_{t+1}\right\}_{t=0}^{T}$ is obtained by repeatedly applying equation (14) starting from $\tilde{K}_0$, errors can accumulate. This is important because, in the true model, today's choices depend on expectations about the future state, which in turn depends on the future's future expectations and so errors cascade. To systematically compare $\tilde{K}_t$ to $K_t$, den Haan proposes an "essential accuracy plot." For a sequence of shocks (*not* those originally used when estimating $\alpha$ to solve the model), generate a sequence of $\tilde{K}_t$ and $K_t$. One can then compare moments of the two simulated sequences. The "main focus" of the accuracy test is the errors calculated by $\tilde{u}_t = \left|\log \tilde{K}_t - \log K_t\right|$, whose maximum should be made close to zero.

***Prices are More Sensitive than Quantities***

The accuracy of the numerical solution becomes an even more critical issue if the main focus of analysis is (asset) prices rather than quantities. This is because prices are much more sensitive to approximation errors (see, e.g., Christiano and Fisher [2000] and Judd and Guu [2001]). The results in Christiano and Fisher (2000) are especially striking. These authors compare a variety of different implementations of the "parameterized expectations" method and report the approximation errors resulting from each. For the standard deviation of output, consumption, and investment (i.e., "quantities"), the approximation errors range from less than 0.1 percent of the true value to 1

percent to 2 percent in some cases. For the stock and bond return and the equity premium, the errors regularly exceed 50 percent and are greater than 100 percent in several cases. The bottom line is that the computation of asset prices requires *extra* care.

### *Pros and Cons*

An important feature of the Krusell-Smith method is that it is a "local" solution around the stationary recursive equilibrium. In other words, this method relies on simulating a very long time series of data (e.g., capital series) and making sure that after this path has converged to the ergodic set, the predictions of agents are accurate for behavior inside that set. This has some advantages and some disadvantages. One advantage is the efficiency gain compared to solving a full recursive equilibrium, which enforces the equilibrium conditions at every point of a somewhat arbitrary grid, regardless of whether or not a given state is ever visited in the stationary equilibrium.

One disadvantage is that if you take a larger deviation—say by setting $K_0$ to a value well below the steady-state value—your "equilibrium functions" are likely to be inaccurate, and the behavior of the solution may differ significantly from the true solution. Why should we care about this? Suppose you solve your model and then want to study a policy experiment where you eliminate taxes on savings. You would need to write a separate program from the "transition" between the two stationary equilibria. Instead, if you solve for the full recursive equilibrium over a grid that contains both the initial and final steady states, you would not need to do this. However, solving for the full equilibrium is often much harder and, therefore, is often overkill.

### *An Alternative to Krusell-Smith: Tracking History of Shocks*

Some models have few exogenous state variables, but a large number of endogenous states. In such cases, using a formulation that keeps track of all these state variables can make the numerical solution extremely costly or even infeasible. An alternative method begins with the straightforward observation that all current endogenous state variables are nothing more than functions of the infinite history of exogenous shocks. So, one could replace these endogenous states with the infinite history of exogenous states. Moreover, many models turn out to have "limited memory" in the sense that only the recent history of shocks matters in a quantitatively significant way, allowing us to only track a truncated history of exogenous states. The first implementation of this idea I have been able to find is in Veracierto (1997), who studied a model with plant-level investment irreversibilities, which give rise to S-s type policies. He showed that it is more practical to track a short history of the S-s thresholds instead of the current-period endogenous state variables.

As another example, consider an equilibrium model of the housing market where the only exogenous state is the interest rate, which evolves as a Markov process. Depending on the precise model structure, the individual endogenous state variables can include the mortgage debt outstanding, the time-to-maturity of the mortgage contract, etc., and the aggregate endogenous state can include the entire distribution of agents over the individual states. This is potentially an enormously large state space! Arslan (2011) successfully solves a model of this sort with realistic fixed-rate mortgage contracts, a life-cycle structure, and stochastic interest rates, using four lags of interest rates. Other recent examples that employ this basic approach include Chien and Lustig (2010), who solve an asset pricing model with aggregate and idiosyncratic risk in which agents are subject to collateral constraints arising from limited commitment, and Lorenzoni (2009), who solves a business cycle model with shocks to individuals' expectations. In all these cases, tracking a truncated history turns out to provide computational advantages over choosing endogenous state variables that evolve in a Markovian fashion. One advantage of this approach is that it is often less costly to add an extra endogenous state variable relative to the standard approach, because the same number of lags may still be sufficient. One drawback is that if the Markov process has many states or the model has long memory, the method may not work as well.

## 6. WHEN DOES HETEROGENEITY MATTER FOR AGGREGATES?

As noted earlier, Krusell and Smith (1998) report that the time series behavior of the aggregated incomplete markets model, by and large, looks very similar to the corresponding representative-agent model. A similar result was reported in Ríos-Rull (1996). However, it is important to interpret these findings correctly and to not overgeneralize them. For example, even if a model aggregates *exactly*, modeling heterogeneity can be very important for aggregate problems. This is because the problem solved by the representative agent can look dramatically different from the problem solved by individuals (for example, have very different preferences). Here, I discuss some important problems in macroeconomics where introducing heterogeneity yields conclusions quite different from a representative-agent model.

### The Curse of Long Horizon

It is useful to start by discussing why incomplete markets do not matter in many models. Loosely speaking, this outcome follows from the fact that a long horizon makes individuals' savings function approximately linear in wealth (i.e., constant MPC out of wealth). As we saw in Section 1, the exact linearity of savings rule delivers exact demand aggregation in both Gorman (1961)

and Rubinstein's (1974) theorems. As it turns out, even with idiosyncratic shocks, this near-linearity holds for wealth levels that are not immediately near borrowing constraints. Thus, even though markets are incomplete, redistributing wealth would matter little, and we have something that looks like demand aggregation!

Is there a way to get around this result? It is instructive to look at a concrete example. Mankiw (1986) was one of the first articles in the literature on the equity premium puzzle and one that gave prominence to the role of heterogeneity. Mankiw (1986) shows that, in a two-period model with incomplete markets and idiosyncratic risk of the right form, one can generate an equity premium as large as desired. However, researchers who followed up on this promising lead (Telmer 1993, Heaton and Lucas 1996, and Krusell and Smith 1997) quickly came to a disappointing conclusion: Once agents in these models are allowed to live for multiple periods, trading a single risk-free asset yields sufficient consumption insurance, which in turn results in a tiny equity premium.

This result—that a sufficiently long horizon can dramatically weaken the effects of incomplete markets—is quite general. In fact, Levine and Zame (2002) prove that if, in a single good economy with no aggregate shocks, (i) idiosyncratic income shocks follow a Markov process, (ii) marginal utility is convex, and (iii) all agents have access to a single risk-free asset, then, as individuals' subjective time discount factor ($\delta$) approaches unity, incomplete markets allocations (and utilities) converge to those from a complete markets economy with the same aggregate resources. Although Levine and Zame's result is theoretical for the limit of such economies (as $\delta \longrightarrow 1$), it still sounds a cautionary note to researchers building incomplete markets models: Unless shocks are extremely persistent and/or individuals are very impatient, these models are unlikely to generate results much different from a representative-agent model.

Constantinides and Duffie (1996) show one way to get around the problem of a long horizon, which is also consistent with the message of Levine-Zame's theorem. Essentially, they assume that individuals face permanent shocks, which eliminate the incentives to smooth such shocks. Therefore, they behave as if they live in a static world and choose not to trade. Constantinides and Duffie also revive another feature of Mankiw's (1986) model: Idiosyncratic shocks must have larger variance in recessions (i.e., countercyclical variances) to generate a large equity premium. With these two features, they show that Mankiw's original insight can be made to work once again in an infinite horizon model. Storesletten, Telmer, and Yaron (2007) find that a calibrated model along the lines suggested by Constantinides and Duffie can generate about $\frac{1}{4}$ of the equity premium observed in the U.S. data.

The bottom line is that, loosely speaking, if incomplete markets matter in a model mainly through its effect on the consumption-saving decision, a

long horizon can significantly weaken the bite of incomplete markets. With a long enough horizon, agents accumulate sufficient wealth and end up on the nearly linear portion of their savings function, delivering results not far from a complete markets model. This is also the upshot of Krusell and Smith's (1998) analysis.

## Examples Where Heterogeneity Does Matter

There are many examples in which heterogeneity does matter for aggregate phenomena. Here, I review some examples.

First, aggregating heterogeneous-agent models can give rise to preferences for the representative agent that may have nothing to do with the preferences in the underlying model. A well-known example of such a transformation is present in the early works of Hansen (1985) and Rogerson (1988), who show that in a model in which individuals have no intensive margin of labor supply (i.e., zero Frisch labor supply elasticity), one can aggregate the model to find that the representative agent has linear preferences in leisure (i.e., infinite Frisch elasticity!). This conclusion challenges one of the early justifications for building models with microfoundations, which was to bring evidence from microdata to bear on the calibration of macroeconomic models. In an excellent survey article, Browning, Hansen, and Heckman (1999) issue an early warning, giving several examples where this approach is fraught with danger.

Building on the earlier insights of Hansen and Rogerson, Chang and Kim (2006) construct a model in which aggregate labor-supply elasticity depends on the reservation-wage distribution in the population. The economy is populated by households (husband and wife) that each supply labor only along the extensive margin: they either work full time or stay home. Workers are hit by idiosyncratic productivity shocks, causing them to move in and out of the labor market. The aggregate labor-supply elasticity of such an economy is around one, greater than a typical microestimate and much greater than the Frisch elasticity one would measure at the intensive margin (which is zero) in this model. The model thus provides a reconciliation between the micro- and macro-labor-supply elasticities. In a similar vein, Chang, Kim, and Schorfheide (2010) show that preference shifters that play an important role in discussions of aggregate policy are not invariant to policies if they are generated from the aggregation of a heterogeneous-agent model. Such a model also generates "wedges" at the aggregate level that do not translate into any well-defined notion of preference shifters at the microlevel.[45] Finally, Erosa, Fuster, and Kambourov (2009) build a life-cycle model of labor supply, by combining and extending the ideas introduced in earlier articles, such

[45] See Chari, Kehoe, and McGrattan (2007) for the definition of business-cycle wedges.

as Chang and Kim (2006) and Rogerson and Wallenius (2009). Their goal is to build a model with empirically plausible patterns of hours over the life cycle and examine the response elasticities of labor supply to various policy experiments.

In another context, Guvenen (2006) asks why macroeconomic models in the RBC tradition typically need to use a high elasticity of intertemporal substitution (EIS) to explain output and investment fluctuations, whereas Euler equation regressions (such as in Hall [1988] and Campbell and Mankiw [1990]) that use aggregate consumption data estimate a much smaller EIS (close to zero) to fit the data. He builds a model with two types of agents who differ in their EIS. The model generates substantial wealth inequality and much smaller consumption inequality, both in line with the U.S. data. Consequently, capital and investment fluctuations are mainly driven by the rich (who hold almost all the wealth in the economy) and thus reflect the high EIS of this group. Consumption fluctuations, on the other hand, reflect an average that puts much more weight on the EIS of the poor, who contribute significantly to aggregate consumption. Thus, a heterogeneous-agent model is able to explain aggregate evidence that a single representative-agent model has trouble fitting.

In an asset pricing context, Constantinides and Duffie (1996), Chan and Kogan (2002), and Guvenen (2011) show a similar result for risk aversion. Constantinides and Duffie (1996) show theoretically how the cross-sectional distribution of consumption in a heterogeneous-agent model gets translated into a higher risk aversion for the representative agent. Guvenen (2011) shows that, in a calibrated model with limited stock market participation that matches several asset pricing facts, the aggregate risk aversion is measured to be as high as 80, when the individuals' risk aversion is only two. These results, as well as the articles discussed above, confirm and amplify the concerns originally highlighted by Browning, Hansen, and Heckman (1999). The conclusion is that researchers must be very careful when using microeconomic evidence to calibrate representative-agent models.

## 7. COMPUTATION AND CALIBRATION

Because of their complexity, the overwhelming majority of models in this literature are solved on a computer using numerical methods.[46] Thus, I now turn to a discussion of computational issues that researchers often have to confront when solving models with heterogeneity.

[46] There are a few examples of analytical solutions and theoretical results established with heterogeneous-agent models. See, e.g., Constantinides and Duffie (1996), Heathcote, Storesletten, and Violante (2007), Rossi-Hansberg and Wright (2007), Wang (2009), and Guvenen and Kuruscu (forthcoming).

**Calibration and Estimation: Avoid Local Search**

Economists often need to minimize an objective function of multiple variables that has lots of kinks, jaggedness, and deep ridges. Consequently, the global minimum is often surrounded by a large number of local minima. A typical example of such a problem arises when a researcher tries to calibrate several structural parameters of an economic model by matching some data moments. Algorithms based on local optimization methods (e.g., Newton-Raphson style derivative-based methods or Nelder-Mead simplex style routines) very often get stuck in local minima because the objective surface is typically very rough (non-smooth).

It is useful to understand some of the sources of this roughness. For example, linear interpolation that is often used in approximating value functions or decision rules generates an interpolated function that is non-differentiable (i.e., has kinks) at every knot point. Similarly, problems with (borrowing, portfolio, etc.) constraints can create significant kinks. Because researchers use a finite number of individuals to simulate data from the model (to compute moments), a small change in the parameter value (during the minimization of the objective) can move some individuals across the threshold—from being constrained to unconstrained or vice versa—which can cause small jumps in the objective value. And sometimes, the moments that the researcher decides to match would be inherently discontinuous in the underlying parameters (with a finite number of individuals), such as the median of a distribution (e.g., wealth holdings). Further compounding the problems, if the moments are not jointly sufficiently informative about the parameters to be calibrated, the objective function would be flat in certain directions. As can be expected, trying to minimize a relatively flat function with lots of kinks, jaggedness, and even small jumps can be a very difficult task indeed.[47]

While the algorithm described here can be applied to the calibration of any model, it is especially useful in models with heterogeneous agents—since such models are time consuming to solve even once, an exhaustive search of the parameter space becomes prohibitively costly (which could be feasible in simpler models).

[47] One simple, but sometimes overlooked, point is that when minimizing an objective function of moments to calibrate a model, one should use the same "seeds" for the random elements of the model that are used to simulate the model in successive evaluations of the objective function. Otherwise, some of the change in objective value will be because of the inherent randomness in different draws of random variables. This can create significant problems with the minimization procedure.

## A Simple Fully Parallelizable Global Optimization Algorithm

Here, I describe a global optimization algorithm that I regularly use for calibrating models and I have found it to be very practical and powerful. It is relatively straightforward to implement, yet allows full parallelization across any number of central processing unit (CPU) cores as well as across any number of computers that are connected to the Internet. It requires no knowledge of MPI, OpenMP, or related tools, and no knowledge of computer networking other than using some commercially available synchronization tools (such as DropBox, SugarSync, etc.).

A broad outline of the algorithm is as follows. As with many global algorithms, this procedure combines a global stage with a local search stage that is restarted at various locations in the parameter space. First, we would like to search the parameter space as thoroughly as possible, but do so in as efficient a way as possible. Thoroughness is essential because we want to be sure that we found the true global minimum, so we are willing to sacrifice some speed to ensure this. The algorithm proceeds by taking an initial starting point (chosen in a manner described momentarily) and conducting a local search from that point on until the minimization routine converges as specified by some tolerance. For local search, I typically rely on the Nelder-Mead's downhill simplex algorithm because it does not require derivative information (that may be inaccurate given the approximation errors in the model's solution algorithm).[48] The minimum function value as well as the parameter combination that attained that minimum are recorded in a file saved on the computer's hard disk. The algorithm then picks the next "random" starting point and repeats the previous step of local minimization. The results are then added to the previous file, which records all the local minima found up to that point.

Of course, the most obvious algorithm would be to keep doing a very large number of restarts of this sort and take the minimum of all the minima found in the process. But this would be very time consuming and would not be particularly efficient. Moreover, in many cases, the neighborhood of the global minimum can feature many deep ridges and kinks nearby, which requires more extensive searching near the global minimum, whereas the proposed approach would devote more time to points far away from the true global minimum and to the points near it. Further, if the starting points are chosen literally randomly, this would also create potentially large efficiency losses, because these points have a non-negligible chance of falling near points previously tried. Because those areas have been previously searched, devoting more time is not optimal.

[48] An alternative that can be much faster but requires a bit more tweaking for best performance is the trust region method of Zhang, Conn, and Scheinberg (2010) that builds on Powell's (2009) BOBYQA algorithm.

A better approach is to use "quasi-random" numbers to generate the starting points. Quasi-random numbers (also called low-discrepancy sequences) are sequences of deterministic numbers that spread to any space in the maximally separated way. They avoid the pitfall of random draws that may end up being too close to each other. Each draw in the sequence "knows" the location of previous points drawn and attempts to fill the gaps as evenly as possible.[49] Among a variety of sequences proposed in the literature, the Sobol' sequence is generally viewed to be superior in most practical applications, having a very uniform filling of the space (i.e., maximally separated) even when a small number of points is drawn, as well as a very fast algorithm that generates the sequence.[50]

Next, how do we use the accumulated information from previous restarts? As suggested by genetic algorithm heuristics, I combine information from previous best runs to adaptively direct the new restarts to areas that appear more promising. This is explained further below. Now for the specifics of the algorithm.

**Algorithm 1** *Let* $\boldsymbol{p}$ *be a* $J$*-dimensional parameter vector with generic element* $p^j$, $j = 1, ..., J$.

- *Step 0.* ***Initialization****:*
  - *Determine bounds for each parameter, outside of which the objective function should be set to a high value.*
  - *Generate a sequence of Sobol' numbers with a sequence length of* $I_{max}$ *(the maximum anticipated number of restarts in the global stage). Set the global iteration number* $i = 1$.
- *Step 1.* ***Global Stage****:*
  - *Draw the* $i^{th}$ *(vector) value in the Sobol' sequence:* $\boldsymbol{s}_i$.
  - *If* $i > 1$*, open and read from the text file* "saved_parameters.dat" *the function values (and corresponding parameter vectors) of previously found local minima. Denote the lowest function value found as of iteration* $i-1$ *as* $f_{i-1}^{low}$ *and the corresponding parameter vector as* $\boldsymbol{p}_{i-1}^{low}$.
  - *Generate a starting point for the local stage as follows:*

[49] Another common application of low-discrepancy sequences is in quasi-Monte Carlo integration, where they have been found to improve time-to-accuracy by several orders of magnitude.

[50] In a wide range of optimization problems, Kucherenko and Sytsko (2005) and Liberti and Kucherenko (2005) find that Sobol' sequences outperform Holton sequences, both in terms of computation time and probability of finding the global optimum. The Holton sequence is particularly weak in high dimensional applications.

  - *If $i < I_{\min}(< I_{\max})$, then use $\mathbf{s}_i$ as the initial guess: $\mathbf{S}_i = \mathbf{s}_i$. Here, $I_{\min}$ is the threshold below which we use fully quasi-random starting points in the global stage.*
  - *If $i \geq I_{\min}$, take the initial guess to be a convex combination of $\mathbf{s}_i$ and the parameter value that generated the best local minima so far: $\mathbf{S}_i = (1-\theta_i)\mathbf{s}_i + \theta \mathbf{p}_{i-1}^{low}$. The parameter $\theta_i \in [0, \bar{\theta}]$ with $\bar{\theta} < 1$, and increases with $i$. For example, I found that a convex increasing function, such as $\theta_i = \min[\bar{\theta}, (i/I_{\max})^2]$, works well in some applications. An alternative heuristic is given later.*
  - *As $\theta_i$ is increased, local searches are restarted from a narrower part of the parameter space that yielded the lowest local minima before.*

- *Step 2:* ***Local Stage****:*
  - *Using $\mathbf{S}_i$ as a starting point, use the downhill simplex algorithm to search for a local minimum. (For the other vertices of the simplex, randomly draw starting points within the bounds of the parameter space.)*
  - *Stop when either (i) a certain tolerance has been achieved, (ii) function values do not improve more than a certain amount, or (iii) the maximum iteration number is reached.*
  - *Open* saved_parameters.dat *and record the local minimum found (function value and parameters).*

- *Step 3.* ***Stopping Rule****:*
  - *Stop if the termination criterion described below is satisfied. If not go to Step 1.*

***Termination Criterion***

One useful heuristic criterion relies on a Bayesian procedure that estimates the probability that the next local search will find a new local minimum based on the rate at which new local minima have been located in past searches. More concretely, if $W$ different local minima have been found after $K$ local searches started from a set of uniformly distributed points, then the expectation of the number of local minima is

$$W_{\exp} = W\,(K-1)\,/\,(K-W-2)\,,$$

provided that $K > W+2$. The searching procedure is terminated if $W_{\exp} < W+0.5$. The idea is that, after a while of searching, if subsequent restarts keep

finding one of the same local minima found before, the chances of improvement in subsequent searches is not worth the additional time cost. Although this is generally viewed as one of the most reliable heuristics, care must be applied as with any heuristic.

Notice also that $W_{\text{exp}}$ can be used to adaptively increase the value of $\theta_i$ in the global stage (Step 1 [3] above). The idea is that, as subsequent global restarts do not yield a new local minimum with a high enough probability, it is time to narrow the search and further explore areas of promising local minima. Because jaggedness and deep ridges cause local search methods to often get stuck, we want to explore promising areas more thoroughly.

One can improve on this basic algorithm in various ways. I am going to mention a few that seem worth exploring.

***Refinements: Clustering and Pre-Testing***

First, suppose that in iteration $k$, the proposed starting point $\mathbf{S}_k$ ends up being "close" to one of the previous minima, say $\mathbf{p}_n^{low}$, for $n < k$. Then it is likely that the search starting from $\mathbf{S}_k$ will end up converging to $\mathbf{p}_n^{low}$. But then we have wasted an entire cycle of local search without gaining anything. To prevent this, one heuristic (called "clustering methods") proceeds by defining a "region of attraction" (which is essentially a $J$-dimensional ball centered) around each one of the local minima found so far.[51] Then the algorithm would discard a proposed restarting point if it falls into the region attraction of any previous local minima. Because the local minimization stage is the most computationally intensive step, this refinement of restarting the local search only once in a given region of attraction can result in significant computational gains. Extensive surveys of clustering methods can be found in Rinnooy Kan and Timmer (1987a, 1987b).

Second, one can add a "pre-test" stage where $N$ points from the Sobol' sequence are evaluated before any local search (i.e., in Step 0 above), and only a subset of $N^* < N$ points with lowest objective values are used as seeds in the local search. The remaining points, as well as regions of attraction around them are ignored as not promising. Notice that while this stage can improve speed, it trades off reliability in the process.

***Narrowing Down the Search Area***

The file `saved_parameters.dat` contains a lot of useful information gathered in each iteration to the global stage, which can be used more efficiently as follows. As noted, the Nelder-Mead algorithm requires $J + 1$ candidate

[51] While different formulas have been proposed for determining the optimal radius, these formulas contain some undetermined coefficients, making the formulas less than useful in real life applications.

points as inputs (the vertices of the $J$-dimensional simplex). One of these points is given by $\mathbf{S}_i$, chosen as described above; the other vertices were drawn randomly. But as we accumulate more information with every iteration on the global stage, if we keep finding local minima that seem to concentrate in certain regions, it makes sense to narrow the range of values from which we pick the vertices. One way to do this is as follows: After a sufficiently large number of restarts have been completed, rank all the function values and take the lowest $x$ percent of values (e.g., 10 percent or 20 percent). Then for each dimension, pick the minimum ($p^j_{min}$) and maximum parameter value ($p^j_{max}$) within this set of minima. Then, to generate vertices, take randomly sampled points between $p^j_{min}$ and $p^j_{max}$ in each dimension $j$. This allows the simplex algorithm to search more intensively in a narrower area, which can improve results very quickly when there are ridges or jaggedness in the objective function that make the algorithm get stuck.

**Parallelizing the Algorithm**

The algorithm can be parallelized in a relatively straightforward manner.[52] The basic idea is to let each CPU core perform a separate local search in a different part of the parameter space, which is a time-consuming process. If we can do many such searches simultaneously, we can speed up the solution dramatically. One factor that makes parallelization simple is the fact that the CPU cores do not need to communicate with each other during the local search stage. In between the local stages, each CPU core will contribute its findings (the last local minimum it found along with the corresponding parameter vector) to the collective wisdom recorded in `saved_parameters.dat` and also get the latest updated information about the best local minimum found so far from the same file. Thus, as long as all CPU cores have access to the same copy of the file `saved_parameters.dat`, parallelization requires no more than a few lines for housekeeping across CPUs. Here are some more specifics.

Suppose that we have a workstation with $N$ CPU cores (for example, $N = 4$, 6, or 12). The first modification we need to make is to change the program to distinguish between the different "copies" of the code, running on different CPU cores. This can be done by simply having the program ask the user (only once, upon starting the code) to input an integer value, $n$, between 1 and $N$, which uniquely identifies the "sequence number" of the particular instance of the program running. Then open $N$ terminal windows and launch a copy of the program in each window. Then for each one, enter a unique sequence number $n = 1, 2, ..., N$.

[52] I am assuming here that a compiled language, such as Fortran or C, is used to write the program. So multiple parallel copies of the same code can be run in different terminal windows.

Upon starting, each program will first simulate the same quasi-random sequence regardless of $n$, but each run will pick a different element of this sequence as its own seed. For simplicity, suppose run $n$ chooses the $n$th element of the sequence as its seed and launches a local search from that point. After completion, each run will open the same file `saved_parameters.dat` and record the local minimum and parameter value it finds.[53]

Now suppose that all copies of the program complete their respective first local searches, so there are $N$ lines, each written by a different CPU core, in the file `saved_parameters.dat`. Then each run will start its second iteration and pick as its next seed the $(N+n)$th element of the quasi-random sequence. When the total number of iterations across all CPUs exceed some threshold $I_{\min}$, then we would like to combine the quasi-random draw with the previous best local minima as described in Step 1 (3) above. This is simple since all runs have access to the same copy of `saved_parameters.dat`.[54]

Notice that this parallelization method is completely agnostic about whether the CPU cores are on the same personal computer (PC) or distributed across many PCs *as long as* all computers keep synchronized copies of `saved_parameters.dat`. This can be achieved by using a synchronization service like DropBox. This feature easily allows one to harness the computational power of many idle PCs distributed geographically with varying speeds and CPU cores.

## 8. FUTURE DIRECTIONS AND FURTHER READING

This article surveys the current state of the heterogeneous-agent models literature and draws several conclusions. First, two key ingredients in such models are (i) the magnitudes and types of risk that the model builder feeds into the model and (ii) the insurance opportunities allowed in the economy. In many cases, it is difficult, if not impossible, to measure each component separately. In other words, the assumptions a researcher makes regarding insurance opportunities will typically affect the inference drawn about the magnitudes of risks and vice versa. Further complicating the problem is the measurement of risk: Individuals often have more information than the econometrician about

[53] Because this opening and writing stage takes a fraction of a second, the likelihood that two or more programs access the file simultaneously and create a run-time error is negligible.

[54] It is often useful for each run to keep track of the *total* number of local searches completed by all CPUs—call this $N_{Last}$. For example, sometimes the increase in $\theta_i$ can be linked to $N_{Last}$. This number can be read as the total number of lines recorded up to that point in `saved_parameters.dat`. Another use of this index is for determining which point in the sequence to select as the next seed point. So as opposed to running $n$ by selecting the $(kN+n)$th point in the sequence where $k$ is the number of local searches completed by run $n$, it could just pick the $(N_{Last}+1)$th number in the sequence. This avoids leaving gaps in the sequence for seeds, in case some CPUs are much faster than others and hence finish many more local searches than others.

future changes in their lives. So, for example, a rise or fall in income that the econometrician may view as a "shock" may in fact be partially or completely anticipated by the individual. This suggests that equating income movements observed in the data with risk (as is often done in the literature) is likely to overstate the true magnitude. This entanglement of "risk," "anticipated changes," and "insurance" presents a difficult challenge to researchers in this area. Although some recent progress has been made, more work remains.

A number of surveys contain very valuable material that are complementary to this article. First, Heathcote, Storesletten, and Violante (2009) is a recent survey of quantitative macroeconomics with heterogeneous households that is complementary to this article. Second, Browning, Hansen, and Heckman (1999) contains an extensive review of microeconomic models that are often used as the foundations of heterogeneous-agent models. It highlights several pitfalls in trying to calibrate macroeconomic models using microevidence. Third, Meghir and Pistaferri (2011) provides a comprehensive treatment of how earnings dynamics affect life-cycle consumption choice, which is closely related to the issues discussed in Section 3 of this survey. Finally, because heterogeneous-agent models use microeconomic survey data in increasingly sophisticated ways, a solid understanding of issues related to measurement error (which is pervasive in microdata) is essential. Failure to understand such problems can wreak havoc with the empirical analysis. Bound, Brown, and Mathiowetz (2001) is an extensive and authoritative survey of the subject.

The introduction of families into incomplete markets models represents an exciting area of current research. For many questions of empirical relevance, the interactions taking place within a household (implicit insurance, bargaining, etc.) can have first-order effects on how individuals respond to idiosyncratic changes. To give a few examples, Gallipoli and Turner (2011) document that the labor supply responses to disability shocks of single workers are larger and more persistent than those of married workers. They argue that an important part of this difference has to do with the fact that couples are able to optimally change their time (and task) allocation within households in response to disability, an option not available to singles. This finding suggests that modeling households would be important for understanding the design of disability insurance policies. Similarly, Guner, Kaygusuz, and Ventura (2010) show that to quantify the effects of alternative tax reforms, it is important to take into account the joint nature of household labor supply. In fact, it is hard to imagine any distributional issue for which the household structure does not figure in an important way.

Another promising area is the richer modeling of household finances in an era of ever-increasing sophistication in financial services. The Great Recession, which was accompanied by a housing market crash and soaring personal bankruptcies, home foreclosures, and so on, has created a renewed sense of

urgency for understanding household balance sheets. Developments on two fronts—advances in theoretical modeling as discussed in Section 3, combined with richer data sources on credit histories and mortgages that are increasingly becoming available to researchers—will make faster progress feasible in this area.

## REFERENCES

Abowd, John M., and David E. Card. 1989. "On the Covariance Structure of Earnings and Hours Changes." *Econometrica* 57 (March): 411–45.

Acemoglu, Daron. 2002. "Technical Change, Inequality, and the Labor Market." *Journal of Economic Literature* 40 (March): 7–72.

Aguiar, Mark, and Erik Hurst. 2008. "Deconstructing Lifecycle Expenditures." Working Paper, University of Rochester.

Aguiar, Mark, and Mark Bils. 2011. "Has Consumption Inequality Mirrored Income Inequality?" Working Paper, University of Rochester.

Aiyagari, S. Rao. 1993. "Uninsured Idiosyncratic Risk and Aggregate Saving." Federal Reserve Bank of Minneapolis Working Paper 502.

Aiyagari, S. Rao. 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving." *The Quarterly Journal of Economics* 109 (August): 659–84.

Altug, Sumru, and Robert A. Miller. 1990. "Household Choices in Equilibrium." *Econometrica* 58 (May): 543–70.

Arslan, Yavuz. 2011. "Interest Rate Fluctuations and Equilibrium in the Housing Market." Working Paper, Central Bank of the Republic of Turkey.

Athreya, Kartik B. 2002. "Welfare Implications of the Bankruptcy Reform Act of 1999." *Journal of Monetary Economics* 49 (November): 1,567–95.

Attanasio, Orazio, and Steven J. Davis. 1996. "Relative Wage Movements and the Distribution of Consumption." *Journal of Political Economy* 104 (December): 1,227–62.

Attanasio, Orazio, Erich Battistin, and Hidehiko Ichimura. 2007. "What Really Happened to Consumption Inequality in the United States?" In *Hard-to-Measure Goods and Services: Essays in Honour of Zvi Griliches*, edited by E. Berndt and C. Hulten. Chicago: University of Chicago Press.

Attanasio, Orazio P., James Banks, Costas Meghir, and Guglielmo Weber. 1999. "Humps and Bumps in Lifetime Consumption." *Journal of Business & Economic Statistics* 17 (January): 22–35.

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. 2008. "Trends in U.S. Wage Inequality: Revising the Revisionists." *The Review of Economics and Statistics* 90 (2): 300–23.

Badel, Alejandro, and Mark Huggett. 2007. "Interpreting Life-Cycle Inequality Patterns as an Efficient Allocation: Mission Impossible?" Working Paper, Georgetown University.

Baker, Michael. 1997. "Growth-Rate Heterogeneity and the Covariance Structure of Life-Cycle Earnings." *Journal of Labor Economics* 15 (April): 338–75.

Baker, Michael, and Gary Solon. 2003. "Earnings Dynamics and Inequality among Canadian Men, 1976–1992: Evidence from Longitudinal Income Tax Records." *Journal of Labor Economics* 21 (April): 267–88.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago: University of Chicago Press.

Ben-Porath, Yoram. 1967. "The Production of Human Capital and the Life Cycle of Earnings." *Journal of Political Economy* 75 (4): 352–65.

Bewley, Truman F. Undated. "Interest Bearing Money and the Equilibrium Stock of Capital." Working Paper.

Blundell, Richard, and Ian Preston. 1998. "Consumption Inequality And Income Uncertainty." *The Quarterly Journal of Economics* 113 (May): 603–40.

Blundell, Richard, Luigi Pistaferri, and Ian Preston. 2008. "Consumption Inequality and Partial Insurance." *American Economic Review* 98 (December): 1,887–921.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, edited by J. J. Heckman and E. E. Leamer. Amsterdam: Elsevier; 3,705–843.

Browning, Martin, Lars Peter Hansen, and James J. Heckman. 1999. "Micro Data and General Equilibrium Models." In *Handbook of Macroeconomics*, edited by J. B. Taylor and M. Woodford. Amsterdam: Elsevier; 543–633.

Browning, Martin, Mette Ejrnaes, and Javier Alvarez. 2010. "Modelling Income Processes with Lots of Heterogeneity." *Review of Economic Studies* 77 (October): 1,353–81.

Cagetti, Marco, and Mariacristina De Nardi. 2006. "Entrepreneurship, Frictions, and Wealth." *Journal of Political Economy* 114 (October): 835–70.

Campbell, John Y., and N. Gregory Mankiw. 1990. "Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence." Working Paper 2924. Cambridge, Mass.: National Bureau of Economic Research (May).

Carroll, Christopher. 2000. "Why Do the Rich Save So Much?" In *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, edited by Joel Slemrod. Boston: Harvard University Press, 466–84.

Carroll, Christopher D. 1991. "Buffer Stock Saving and the Permanent Income Hypothesis." Board of Governors of the Federal Reserve System Working Paper Series/Economic Activity Section 114.

Carroll, Christopher D. 1997. "Buffer-Stock Saving and the Life Cycle/Permanent Income Hypothesis." *The Quarterly Journal of Economics* 112 (February): 1–55.

Carroll, Christopher D., and Andrew A. Samwick. 1997. "The Nature of Precautionary Wealth." *Journal of Monetary Economics* 40 (September): 41–71.

Caselli, Francesco, and Jaume Ventura. 2000. "A Representative Consumer Theory of Distribution." *American Economic Review* 90 (September): 909–26.

Castañeda, Ana, Javier Díaz-Giménez, and José-Víctor Ríos-Rull. 2003. "Accounting for the U.S. Earnings and Wealth Inequality." *The Journal of Political Economy* 111 (August): 818–57.

Chamberlain, Gary, and Charles A. Wilson. 2000. "Optimal Intertemporal Consumption Under Uncertainty." *Review of Economic Dynamics* 3 (July): 365–95.

Chan, Yeung Lewis, and Leonid Kogan. 2002. "Catching up with the Joneses: Heterogeneous Preferences and the Dynamics of Asset Prices." *Journal of Political Economy* 110 (December): 1,255–85.

Chang, Yongsung, and Sun-Bin Kim. 2006. "From Individual to Aggregate Labor Supply: A Quantitative Analysis based on a Heterogeneous-Agent Macroeconomy." *International Economic Review* 47 (1): 1–27.

Chang, Yongsung, Sun-Bin Kim, and Frank Schorfheide. 2010. "Labor Market Heterogeneity, Aggregation, and the Lucas Critique." Working Paper 16401. Cambridge, Mass.: National Bureau of Economic Research.

Chari, V. V., Patrick J. Kehoe, and Ellen R. McGrattan. 2007. "Business Cycle Accounting." *Econometrica* 75 (3): 781–836.

Chatterjee, Satyajit, and Burcu Eyigungor. 2011. "A Quantitative Analysis of the U.S. Housing and Mortgage Markets and the Foreclosure Crisis." Federal Reserve Bank of Philadelphia Working Paper 11-26 (July).

Chatterjee, Satyajit, Dean Corbae, Makoto Nakajima, and José-Víctor Ríos-Rull. 2007. "A Quantitative Theory of Unsecured Consumer Credit with Risk of Default." *Econometrica* 75 (November): 1,525–89.

Chien, YiLi, and Hanno Lustig. 2010. "The Market Price of Aggregate Risk and the Wealth Distribution." *Review of Financial Studies* 23 (April): 1,596–650.

Christiano, Lawrence J., and Jonas D. M. Fisher. 2000. "Algorithms for Solving Dynamic Models with Occasionally Binding Constraints." *Journal of Economic Dynamics and Control* 24 (July): 1,179–232.

Clarida, Richard H. 1987. "Consumption, Liquidity Constraints, and Asset Accumulation in the Presence of Random Income Fluctuations." *International Economic Review* 28 (June): 339–51.

Clarida, Richard H. 1990. "International Lending and Borrowing in a Stochastic, Stationary Equilibrium." *International Economic Review* 31 (August): 543–58.

Cochrane, John H. 1991. "A Simple Test of Consumption Insurance." *Journal of Political Economy* 99 (October): 957–76.

Congressional Budget Office. 2008. "Recent Trends in the Variability of Individual Earnings and Family Income." Washington, D.C.: CBO (June).

Constantinides, George M. 1982. "Intertemporal Asset Pricing with Heterogeneous Consumers and Without Demand Aggregation." *Journal of Business* 55 (April): 253–67.

Constantinides, George M., and Darrell Duffie. 1996. "Asset Pricing with Heterogeneous Consumers." *Journal of Political Economy* 104 (April): 219–40.

Cunha, Flavio, James Heckman, and Salvador Navarro. 2005. "Separating Uncertainty from Heterogeneity in Life Cycle Earnings." *Oxford Economic Papers* 57 (2): 191–261.

Cutler, David M., and Lawrence F. Katz. 1992. "Rising Inequality? Changes in the Distribution of Income and Consumption in the 1980's." *American Economic Review* 82 (May): 546–51.

Davis, Steven J., R. Jason Faberman, John Haltiwanger, Ron Jarmin, and Javier Miranda. 2010. "Business Volatility, Job Destruction, and Unemployment." Working Paper 14300. Cambridge, Mass.: National Bureau of Economic Research (September).

De Nardi, Mariacristina, Eric French, and John B. Jones. 2010. "Why Do the Elderly Save? The Role of Medical Expenses." *Journal of Political Economy* 118 (1): 39–75.

Deaton, Angus. 1991. "Saving and Liquidity Constraints." *Econometrica* 59 (September): 1,221–48.

Deaton, Angus, and Christina Paxson. 1994. "Intertemporal Choice and Inequality." *Journal of Political Economy* 102 (June): 437–67.

Debreu, Gerard. 1959. *Theory of Value*. New York: John Wiley and Sons.

den Haan, Wouter J. 2010. "Assessing the Accuracy of the Aggregate Law of Motion in Models with Heterogeneous Agents." *Journal of Economic Dynamics and Control* 34 (January): 79–99.

den Haan, Wouter J., and Pontus Rendahl. 2009. "Solving the Incomplete Markets Model with Aggregate Uncertainty Using Explicit Aggregation." Working Paper, University of Amsterdam.

Domeij, David, and Martin Floden. 2006. "The Labor-Supply Elasticity and Borrowing Constraints: Why Estimates are Biased." *Review of Economic Dynamics* 9 (April): 242–62.

Dynan, Karen E., Douglas W. Elmendorf, and Daniel E. Sichel. 2007. "The Evolution of Household Income Volatility." Federal Reserve Board Working Paper 2007-61.

Erosa, Andrés, Luisa Fuster, and Gueorgui Kambourov. 2009. "The Heterogeneity and Dynamics of Individual Labor Supply over the Life Cycle: Facts and Theory." Working Paper, University of Toronto.

Flavin, Marjorie A. 1981. "The Adjustment of Consumption to Changing Expectations About Future Income." *Journal of Political Economy* 89 (October): 974–1,009.

French, Eric, and John Bailey Jones. 2004. "On the Distribution and Dynamics of Health Care Costs." *Journal of Applied Econometrics* 19 (6): 705–21.

Galindev, Ragchaasuren, and Damba Lkhagvasuren. 2010. "Discretization of Highly Persistent Correlated AR(1) Shocks." *Journal of Economic Dynamics and Control* 34 (July): 1,260–76.

Gallipoli, Giovanni, and Laura Turner. 2011. "Household Responses to Individual Shocks: Disability and Labour Supply." Working Paper, University of British Columbia.

Glover, Andrew, and Jacob Short. 2010. "Bankruptcy, Incorporation, and the Nature of Entrepreneurial Risk." Working Paper, University of Western Ontario.

Gorman, William M. 1961. "On a Class of Preference Fields." *Metroeconomica* 13 (June): 53–6.

Gottschalk, Peter, and Robert Moffitt. 1994. "The Growth of Earnings Instability in the U.S. Labor Market." *Brookings Papers on Economic Activity* 25 (2): 217–72.

Gottschalk, Peter, and Robert Moffitt. 1999. "Changes in Job Instability and Insecurity Using Monthly Survey Data." *Journal of Labor Economics* 17 (October): S91–126.

Gourinchas, Pierre-Olivier, and Jonathan A. Parker. 2002. "Consumption over the Life Cycle." *Econometrica* 70 (January): 47–89.

Greenwood, Jeremy, Ananth Seshadri, and Mehmet Yorukoglu. 2005. "Engines of Liberation." *Review of Economic Studies* 72 (1): 109–33.

Greenwood, Jeremy, and Nezih Guner. 2009. "Marriage and Divorce since World War II: Analyzing the Role of Technological Progress on the Formation of Households." In *NBER Macroeconomics Annual*, Vol. 23. Cambridge, Mass.: National Bureau of Economic Research, 231–76.

Guner, Nezih, Remzi Kaygusuz, and Gustavo Ventura. 2010. "Taxation and Household Labor Supply." Working Paper, Arizona State University.

Gustavsson, Magnus, and Pär Österholm. 2010. "Does the Labor-Income Process Contain a Unit Root? Evidence from Individual-Specific Time Series." Working Paper, Uppsala University.

Guvenen, Fatih. 2006. "Reconciling Conflicting Evidence on the Elasticity of Intertemporal Substitution: A Macroeconomic Perspective." *Journal of Monetary Economics* 53 (October): 1,451–72.

Guvenen, Fatih. 2007a. "Do Stockholders Share Risk More Effectively than Nonstockholders?" *The Review of Economics and Statistics* 89 (2): 275–88.

Guvenen, Fatih. 2007b. "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?" *American Economic Review* 97 (June): 687–712.

Guvenen, Fatih. 2009a. "An Empirical Investigation of Labor Income Processes." *Review of Economic Dynamics* 12 (January): 58–79.

Guvenen, Fatih. 2009b. "A Parsimonious Macroeconomic Model for Asset Pricing." *Econometrica* 77 (November): 1,711–50.

Guvenen, Fatih. 2011. "Limited Stock Market Participation Versus External Habit: An Intimate Link." University of Minnesota Working Paper 450.

Guvenen, Fatih, and Anthony A. Smith. 2009. "Inferring Labor Income Risk from Economic Choices: An Indirect Inference Approach." Working Paper, University of Minnesota.

Guvenen, Fatih, and Burhanettin Kuruscu. 2010. "A Quantitative Analysis of the Evolution of the U.S. Wage Distribution, 1970–2000." In *NBER Macroeconomics Annual 2009* 24 (1): 227–76.

Guvenen, Fatih, and Burhanettin Kuruscu. Forthcoming. "Understanding the Evolution of the U.S. Wage Distribution: A Theoretical Analysis." *Journal of the European Economic Association*.

Guvenen, Fatih, and Michelle Rendall. 2011. "Emancipation Through Education." Working Paper, University of Minnesota.

Guvenen, Fatih, Burhanettin Kuruscu, and Serdar Ozkan. 2009. "Taxation of Human Capital and Wage Inequality: A Cross-Country Analysis." Working Paper 15526. Cambridge, Mass.: National Bureau of Economic Research (November).

Haider, Steven J. 2001. "Earnings Instability and Earnings Inequality of Males in the United States: 1967–1991." *Journal of Labor Economics* 19 (October): 799–836.

Haider, Steven, and Gary Solon. 2006. "Life-Cycle Variation in the Association between Current and Lifetime Earnings." *American Economic Review* 96 (September): 1,308–20.

Hall, Robert E. 1988. "Intertemporal Substitution in Consumption." *Journal of Political Economy* 96 (April): 339–57.

Hall, Robert E., and Frederic S. Mishkin. 1982. "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households." *Econometrica* 50 (March): 461–81.

Hansen, Gary D. 1985. "Indivisible Labor and the Business Cycle." *Journal of Monetary Economics* 16 (November): 309–27.

Harris, Milton, and Bengt Holmstrom. 1982. "A Theory of Wage Dynamics." *Review of Economic Studies* 49 (July): 315–33.

Hause, John C. 1980. "The Fine Structure of Earnings and the On-the-Job Training Hypothesis." *Econometrica* 48 (May): 1,013–29.

Hayashi, Fumio, Joseph Altonji, and Laurence Kotlikoff. 1996. "Risk-Sharing between and within Families." *Econometrica* 64 (March): 261–94.

Heathcote, Jonathan, Fabrizio Perri, and Giovanni L. Violante. 2010. "Unequal We Stand: An Empirical Analysis of Economic Inequality in the United States, 1967–2006." *Review of Economic Dynamics* 13 (January): 15–51.

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. 2007. "Consumption and Labour Supply with Partial Insurance: An Analytical Framework." CEPR Discussion Papers 6280 (May).

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. 2008. "The Macroeconomic Implications of Rising Wage Inequality in the United States." Working Paper 14052. Cambridge, Mass.: National Bureau of Economic Research (June).

Heathcote, Jonathan, Kjetil Storesletten, and Giovanni L. Violante. 2009. "Quantitative Macroeconomics with Heterogeneous Households." *Annual Review of Economics* 1 (1): 319–54.

Heaton, John, and Deborah J. Lucas. 1996. "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing." *Journal of Political Economy* 104 (June): 443–87.

Heaton, John, and Deborah Lucas. 2000. "Portfolio Choice and Asset Prices: The Importance of Entrepreneurial Risk." *Journal of Finance* 55 (June): 1,163–98.

Heckman, James, Lance Lochner, and Christopher Taber. 1998. "Explaining Rising Wage Inequality: Explanations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents." *Review of Economic Dynamics* 1 (January): 1–58.

Hornstein, Andreas, Per Krusell, and Giovanni L. Violante. 2011. "Frictional Wage Dispersion in Search Models: A Quantitative Assessment." *American Economic Review* 101 (December): 2,873–98.

Hubbard, R. Glenn, Jonathan Skinner, and Stephen P. Zeldes. 1994. "The Importance of Precautionary Motives in Explaining Individual and Aggregate Saving." *Carnegie-Rochester Conference Series on Public Policy* 40 (June): 59–125.

Hubbard, R. Glenn, Jonathan Skinner, and Stephen P. Zeldes. 1995. "Precautionary Saving and Social Insurance." *Journal of Political Economy* 103 (April): 360–99.

Huggett, Mark. 1993. "The Risk-Free Rate in Heterogeneous-Agent Incomplete-Insurance Economies." *Journal of Economic Dynamics and Control* 17: 953–69.

Huggett, Mark. 1996. "Wealth Distribution in Life-cycle Economies." *Journal of Monetary Economics* 38 (December): 469–94.

Huggett, Mark, Gustavo Ventura, and Amir Yaron. 2006. "Human Capital and Earnings Distribution Dynamics." *Journal of Monetary Economics* 53 (March): 265–90.

Huggett, Mark, Gustavo Ventura, and Amir Yaron. 2011. "Sources of Lifetime Inequality." *American Economic Review* 101 (December): 2,923–54.

Imrohoroglu, Ayse. 1989. "Cost of Business Cycles with Indivisibilities and Liquidity Constraints." *Journal of Political Economy* 97 (December): 1,364–83.

Jencks, Christopher. 1984. "The Hidden Prosperity of the 1970s." *Public Interest* 77: 37–61.

Jones, Larry E., Rodolfo E. Manuelli, and Ellen R. McGrattan. 2003. "Why are Married Women Working So Much?" Federal Reserve Bank of Minneapolis Staff Report 317 (June).

Jovanovic, Boyan. 1979. "Job Matching and the Theory of Turnover." *Journal of Political Economy* 87 (October): 972–90.

Judd, Kenneth L., and Sy-Ming Guu. 2001. "Asymptotic Methods for Asset Market Equilibrium Analysis." *Economic Theory* 18 (1): 127–57.

Kaplan, Greg. 2010. "Inequality and the Life Cycle." Working Paper, University of Pennsylvania.

Kaplan, Greg, and Giovanni L. Violante. 2010. "How Much Consumption Insurance Beyond Self-Insurance?" *American Economic Journal: Macroeconomics* 2 (October): 53–87.

Kehoe, Timothy J., and David K. Levine. 1993. "Debt-Constrained Asset Markets." *Review of Economic Studies* 60 (October): 865–88.

Kitao, Sagiri, Lars Ljungqvist, and Thomas J. Sargent. 2008. "A Life Cycle Model of Trans-Atlantic Employment Experiences." Working Paper, University of Southern California and New York University.

Knowles, John. 2007. "Why Are Married Men Working So Much? The Macroeconomics of Bargaining Between Spouses." Working Paper, University of Pennsylvania.

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song. 2010. "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data Since 1937." *Quarterly Journal of Economics* 125 (February): 91–128.

Kopecky, Karen A., and Richard M. H. Suen. 2010. "Finite State Markov-chain Approximations to Highly Persistent Processes." *Review of Economic Dynamics* 13 (July): 701–14.

Krueger, Dirk, and Fabrizio Perri. 2006. "Does Income Inequality Lead to Consumption Inequality? Evidence and Theory." *Review of Economic Studies* 73 (1): 163–93.

Krueger, Dirk, and Fabrizio Perri. 2009. "How Do Households Respond to Income Shocks?" Working Paper, University of Minnesota.

Krusell, Per, and Anthony A. Smith. 1997. "Income and Wealth Heterogeneity, Portfolio Choice, and Equilibrium Asset Returns." *Macroeconomic Dynamics* 1 (June): 387–422.

Krusell, Per, and Anthony A. Smith, Jr.. 1998. "Income and Wealth Heterogeneity in the Macroeconomy." *Journal of Political Economy* 106 (October): 867–96.

Kucherenko, Sergei, and Yury Sytsko. 2005. "Application of Deterministic Low-Discrepancy Sequences in Global Optimization." *Computational Optimization and Applications* 30: 297–318.

Laitner, John. 1992. "Random Earnings Differences, Lifetime Liquidity Constraints, and Altruistic Intergenerational Transfers." *Journal of Economic Theory* 58 (December): 135–70.

Laitner, John. 2002. "Wealth Inequality and Altruistic Bequests." *American Economic Review* 92 (May): 270–3.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (March): 31–43.

Levine, David K., and William R. Zame. 2002. "Does Market Incompleteness Matter?" *Econometrica* 70 (September): 1,805–39.

Liberti, Leo, and Sergei Kucherenko. 2005. "Comparison of Deterministic and Stochastic Approaches to Global Optimization." *International Transactions in Operations Research* 12: 263–85.

Lillard, Lee A., and Robert J. Willis. 1978. "Dynamic Aspects of Earnings Mobility." *Econometrica* 46 (September): 985–1,012.

Lillard, Lee A., and Yoram Weiss. 1979. "Components of Variation in Panel Earnings Data: American Scientists 1960–70." *Econometrica* 47 (March): 437–54.

Livshits, Igor, James MacGee, and Michèle Tertilt. 2007. "Consumer Bankruptcy—A Fresh Start." *American Economic Review* 97 (March): 402–18.

Livshits, Igor, James MacGee, and Michèle Tertilt. 2010. "Accounting for the Rise in Consumer Bankruptcies." *American Economic Journal: Macroeconomics* 2 (April): 165–93.

Lorenzoni, Guido. 2009. "A Theory of Demand Shocks." *American Economic Review* 99 (December): 2,050–84.

Lucas, Jr., Robert E. 1987. *Models of Business Cycles*. New York: Basil Blackwell.

Lucas, Jr., Robert E. 2003. "Macroeconomic Priorities." *American Economic Review* 93 (March): 1–14.

Mace, Barbara J. 1991. "Full Insurance in the Presence of Aggregate Uncertainty." *Journal of Political Economy* 99 (October): 928–56.

MaCurdy, Thomas E. 1982. "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis." *Journal of Econometrics* 18 (January) 83–114.

Mankiw, N. Gregory. 1986. "The Equity Premium and the Concentration of Aggregate Shocks." *Journal of Financial Economics* 17 (September): 211–9.

Meghir, Costas, and Luigi Pistaferri. 2004. "Income Variance Dynamics and Heterogeneity." *Econometrica* 72 (1): 1–32.

Meghir, Costas, and Luigi Pistaferri. 2011. "Earnings, Consumption, and Life Cycle Choices." In *Handbook of Labor Economics*, Vol 4B, edited by Orley Ashenfelter and David Card. Amsterdam: Elsevier, 773–854.

Mehra, Rajnish, and Edward C. Prescott. 1985. "The Equity Premium: A Puzzle." *Journal of Monetary Economics* 15 (March): 145–61.

Mian, Atif, and Amir Sufi. 2011. "House Prices, Home Equity-Based Borrowing, and the U.S. Household Leverage Crisis." *American Economic Review* 101 (August): 2,132–56.

Moffitt, Robert, and Peter Gottschalk. 1995. "Trends in the Covariance Structure of Earnings in the United States: 1969–1987." Institute for Research on Poverty Discussion Papers 1001-93, University of Wisconsin Institute for Research and Poverty.

Moffitt, Robert, and Peter Gottschalk. 2008. "Trends in the Transitory Variance of Male Earnings in the U.S., 1970–2004." Working Paper, Johns Hopkins University.

Nelson, Julie A. 1994. "On Testing for Full Insurance Using Consumer Expenditure Survey Data: Comment." *Journal of Political Economy* 102 (April): 384–94.

Ozkan, Serdar. 2010. "Income Differences and Health Care Expenditures over the Life Cycle." Working Paper, University of Pennsylvania.

Palumbo, Michael G. 1999. "Uncertain Medical Expenses and Precautionary Saving Near the End of the Life Cycle." *Review of Economic Studies* 66 (April): 395–421.

Pijoan-Mas, Josep. 2006. "Precautionary Savings or Working Longer Hours?" *Review of Economic Dynamics* 9 (April): 326–52.

Powell, Michael J. D. 2009. "The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives." Numerical Analysis Papers NA06, Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, Cambridge (August).

Pratt, John W. 1964. "Risk Aversion in the Small and in the Large." *Econometrica* 32 (1/2): 122–36.

Primiceri, Giorgio E., and Thijs van Rens. 2009. "Heterogeneous Life-Cycle Profiles, Income Risk and Consumption Inequality." *Journal of Monetary Economics* 56 (January): 20–39.

Quadrini, Vincenzo. 2000. "Entrepreneurship, Saving, and Social Mobility." *Review of Economic Dynamics* 3 (January): 1–40.

Rinnooy Kan, Alexander, and G. T. Timmer. 1987a. "Stochastic Global Optimization Methods Part I: Clustering Methods." *Mathematical Programming* 39: 27–56.

Rinnooy Kan, Alexander, and G. T. Timmer. 1987b. "Stochastic Global Optimization Methods Part II: Multilevel Methods." *Mathematical Programming* 39: 57–78.

Ríos-Rull, José Victor. 1996. "Life-Cycle Economies and Aggregate Fluctuations." *Review of Economic Studies* 63 (July): 465–89.

Rogerson, Richard. 1988. "Indivisible Labor, Lotteries and Equilibrium." *Journal of Monetary Economics* 21 (January): 3–16.

Rogerson, Richard, and Johanna Wallenius. 2009. "Micro and Macro Elasticities in a Life Cycle Model With Taxes." *Journal of Economic Theory* 144 (November): 2,277–92.

Rogerson, Richard, Robert Shimer, and Randall Wright. 2005. "Search-Theoretic Models of the Labor Market: A Survey." *Journal of Economic Literature* 43 (December): 959–88.

Rossi-Hansberg, Esteban, and Mark L. J. Wright. 2007. "Establishment Size Dynamics in the Aggregate Economy." *American Economic Review* 97 (December): 1,639–66.

Rouwenhorst, K. Geert. 1995. "Asset Pricing Implications of Equilibrium Business Cycle Models." In *Frontiers of Business Cycle Research*. Princeton, N.J.: Princeton University Press, 294–330.

Rubinstein, Mark. 1974. "An Aggregation Theorem for Securities Markets." *Journal of Financial Economics* 1 (September): 225–44.

Sabelhaus, John, and Jae Song. 2009. "Earnings Volatility Across Groups and Time." *National Tax Journal* 62 (June): 347–64.

Sabelhaus, John, and Jae Song. 2010. "The Great Moderation in Micro Labor Earnings." *Journal of Monetary Economics* 57 (May): 391–403.

Schechtman, Jack, and Vera L. S. Escudero. 1977. "Some Results on an 'Income Fluctuation Problem.'" *Journal of Economic Theory* 16 (December): 151–66.

Schulhofer-Wohl, Sam. 2011. "Heterogeneity and Tests of Risk Sharing." Federal Reserve Bank of Minneapolis Staff Report 462 (September).

Shimer, Robert. 2005. "The Cyclicality of Hires, Separations, and Job-to-Job Transitions." Federal Reserve Bank of St. Louis *Review* 87 (4): 493–507.

Shimer, Robert. 2007. "Reassessing the Ins and Outs of Unemployment." Working Paper 13421. Cambridge, Mass.: National Bureau of Economic Research (September).

Shin, Donggyun, and Gary Solon. 2011. "Trends in Men's Earnings Volatility: What Does the Panel Study of Income Dynamics Show?" *Journal of Public Economics* 95 (August): 973–82.

Storesletten, Kjetil, Christopher I. Telmer, and Amir Yaron. 2004a. "Consumption and Risk Sharing Over the Life Cycle." *Journal of Monetary Economics* 51 (April): 609–33.

Storesletten, Kjetil, Christopher I. Telmer, and Amir Yaron. 2004b. "Cyclical Dynamics in Idiosyncratic Labor Market Risk." *Journal of Political Economy* 112 (June): 695–717.

Storesletten, Kjetil, Christopher I. Telmer, and Amir Yaron. 2007. "Asset Pricing with Idiosyncratic Risk and Overlapping Generations." *Review of Economic Dynamics* 10 (October): 519–48.

Telmer, Christopher I. 1993. "Asset Pricing Puzzles and Incomplete Markets." *Journal of Finance* 48 (December): 1,803–32.

Topel, Robert H. 1990. "Specific Capital, Mobility, and Wages: Wages Rise with Job Seniority." Working Paper 3294. Cambridge, Mass.: National Bureau of Economic Research (March).

Topel, Robert H., and Michael P. Ward. 1992. "Job Mobility and the Careers of Young Men." *Quarterly Journal of Economics* 107 (May): 439–79.

Townsend, Robert M. 1994. "Risk and Insurance in Village India." *Econometrica* 62 (May): 539–91.

Veracierto, Marcelo. 1997. "Plant-Level Irreversible Investment and Equilibrium Business Cycles." Federal Reserve Bank of Minneapolis Discussion Paper 115 (March).

Wang, Neng. 2009. "Optimal Consumption and Asset Allocation with Unknown Income Growth." *Journal of Monetary Economics* 56 (May): 524–34.

Zhang, Hongchao, Andrew R. Conn, and Katya Scheinberg. 2010. "A Derivative-Free Algorithm for Least-Squares Minimization." *SIAM Journal on Optimization* 20 (6): 3,555–76.

*Economic Quarterly—Volume 97, Number 3—Third Quarter 2011—Pages 329–357*

# Recent Developments in Economic Growth

Diego Restuccia

A fundamental question in the field of economic growth and development is why some countries are rich and others poor. Both the longer term historical experience of individual countries and the more recent data for a large number of countries show periods of marked increases in income inequality across countries, as well as episodes where individual countries catch up with the leading country. What determines when countries start the process of modern economic growth? Why do some countries sustain positive economic growth for long periods of time while others countries seem to fail to catch up with the leading country and even fall behind other countries that are able to catch up? Understanding the factors driving income inequality has potentially enormous welfare consequences and the design of effective economic policy hinges on answers to these and related questions.

I start this survey article by first describing a broad set of facts from international data on gross domestic product (GDP) per capita as a measure of welfare across countries. These facts motivate most of the inquiry in the field of growth economics. The main facts can be summarized as follows. First, not only are there remarkable differences in per capita income across countries, but also inequality has increased over the last 30 years. To be more concrete, while average GDP per capita of the richest countries was about 25 times that of the poorest countries in 1960, it was about 65 times that of the poorest countries in 2005. Second, the international evidence presents numerous episodes of countries catching up, stagnating, or falling behind in relative income over time.

The author would like to thank Tasso Adamopoulos, Margarida Duarte, Andreas Hornstein, and Pierre Sarte for very useful and detailed comments. The author would also like to thank Baran Doda for excellent research assistance. All remaining errors and misinterpretations are the author's. The opinions expressed in this article do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. Restuccia is affiliated with the University of Toronto. E-mail: diego.restuccia@utoronto.ca.

Next, I review the recent literature in growth economics. I take a narrow view of the field with a focus on quantitative explorations.[1] I discuss the literature that directly or indirectly addresses the facts on income differences across countries and over time. Essentially, this literature emphasizes that cross-country differences for aggregate outcomes arise from cross-country differences in the allocation of factors of production and productivity across heterogeneous production units where those units can generically refer to sectors/industries or establishments within sectors. I begin my survey with the literature that focuses on the structural transformation of the economy—broadly described as systematic changes in the allocation of factors of production across sectors in the economy. I emphasize the role of agriculture for the early stages of development and for the current income differences between rich and poor countries. I also emphasize the reallocation of factors to the service sector in determining recent patterns of aggregate productivity growth across countries. I then discuss models that focus on understanding differences in measured aggregate total factor productivity (TFP) arising from the allocation of factors of production across establishments with heterogenous productivity levels. Substantial work remains to be done on identifying the fundamental determinants of productivity and resource allocation across productive units.

The article is organized as follows. In the next section, I lay out the main facts in economic growth and development that organize the ultimate objectives of the recent quantitative literature in growth economics. Section 2 surveys models emphasizing the role of the structural transformation in the economy—changes in the allocation of factors of production across sectors. In Section 3, I discuss the literature that relates measured TFP differences across countries to distortions that misallocate factors of production across heterogeneous establishments. I conclude in Section 4.

## 1. FACTS

In this article, I focus on documenting a narrow set of facts using the recent data on GDP per capita from Heston, Summers, and Aten (2009). The data is often referred to as the Penn World Table (PWT). To provide a broader perspective, I complement the description of the facts from this data with references to the literature where refinements of the basic facts have been made. Let me first describe the data. I use GDP per capita as a measure of welfare in each country.[2] A critical element of the data is that the measure of GDP

[1] Even with a narrow focus, the survey is bound to leave out the discussion of many important contributions for which I preemptively apologize.

[2] Clearly, GDP per capita is a limited measure of welfare in an economy as cross-country differences in life expectancy, education, work hours, and inequality, among others, are also relevant

reported in the PWT is adjusted for price differences across countries (purchasing power parity adjusted) and, hence, represents a measure of income in units that are comparable across countries.[3] The data spans from 1950–2007 for 189 countries in the world. Since I am interested in assessing the evolution of cross-country incomes over time, I restrict attention to a sample of 101 countries that have data for each year from 1960–2007 and that have a population of more than 1 million people in 2007. I emphasize two sets of facts from this data. First, income differences across rich and poor countries are not only large at any point in time between 1960–2007, but also have increased quite substantially in the last two decades. Second, while the dispersion in income per capita has either stayed constant or increased in the last two decades, the data reveal remarkable episodes of individual countries catching up, stagnating, and declining in per capita income relative to that of the United States. I now elaborate on the description of these basic facts.

## Income Differences

To start, for each year between 1960–2007, I rank countries by their GDP per capita relative to that of the United States. I use the United States as a benchmark country for comparison since it is a large, stable, and diverse country that has been at the frontier of the world's production technology during the sample period. As a result, changes in income in the United States roughly approximate changes in the world state of knowledge that, in principle, should be available for adoption elsewhere. I then calculate the average GDP per capita for the richest 5 percent of the countries and the poorest 5 percent of the countries (i.e., I calculate the average of the richest and poorest 5 countries in the sample). The ratio of the average GDP per capita of the richest and poorest 5 percent of countries is reported in Figure 1.[4] Income per capita differences across countries are large. GDP per capita in the richest countries is, on average, 40 times that of the poorest countries. Moreover, income differences, while relatively stable between 1960 to about 1985, have been increasing since then such that in 2007 GDP per capita in the richest countries was, on average, 66 times that of the poorest countries.[5] The increase in income

---

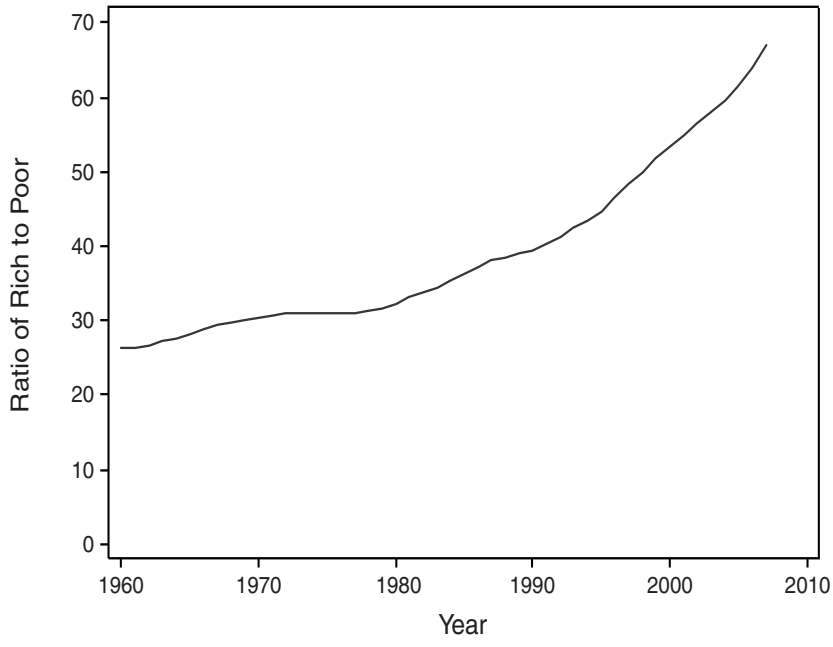
measures in a country's welfare. I follow the standard practice in the literature of focusing on GDP per capita as the main determinant of welfare in a country. See Jones and Klenow (2011) for an analysis of welfare across countries and time that includes measures of consumption, leisure, inequality, and mortality.

[3] In the version of the PWT I use, international prices refer to world prices of 2005.

[4] Parente and Prescott (1993) emphasize the ratio of the richest and poorest 5 percent of countries in GDP per capita as a measure of dispersion in income across countries at a point in time and across time. Duarte and Restuccia (2006) emphasize similar statistics but for measures of labor productivity such as GDP per worker.

[5] Note that while there is substantial persistence in cross-country income differences over time, the set of poor and rich countries may be changing over time.

**Figure 1 GDP per Capita Ratio of Rich to Poor**




Notes: GDP per capita from Heston, Summers, and Aten (2009). The ratio refers to the average of the richest 5 percent of countries to the poorest 5 percent of countries in each year. Since the sample contains 101 countries, these are averages of 5 countries.

inequality between the rich and poor countries is mainly driven by a fall in relative income in the poorest countries, which is not necessarily a decline in absolute incomes of poor countries, but a failure of poor countries to grow as fast as the United States. This fact is not a curiosity of the poorest countries alone in this sample, which happen to be mostly in Africa, but continues to hold even when focusing on larger groups of poor countries or on different subgroups of the poorest countries. To illustrate this fact, Table 1 summarizes the evolution of GDP per capita across countries relative to that of the United States for deciles of the income distribution in selected years. The richest 10 percent of countries (Decile 10) gained on average, increasing relative GDP per capita from 0.87 in 1960 to 0.91 in 2007. The poorest 10 percent of countries (Decile 1) failed to keep up with the United States, losing half of the relative income position, from a relative income of 0.04 in 1960 to less than 0.02 in 2007. But relative income also declined for most of the other groups

**Table 1 GDP per Capita Relative to the United States (Percent)**

| | Year | | | | | |
|---|---|---|---|---|---|---|
| **Decile** | **1960** | **1970** | **1980** | **1990** | **2000** | **2007** |
| 1 | 4.3 | 3.9 | 3.5 | 2.8 | 2.2 | 1.9 |
| 2 | 6.3 | 6.1 | 5.2 | 3.9 | 3.3 | 3.6 |
| 3 | 8.7 | 7.5 | 7.0 | 5.9 | 5.1 | 5.0 |
| 4 | 11.4 | 9.9 | 10.3 | 9.1 | 8.3 | 7.8 |
| 5 | 15.0 | 15.0 | 15.4 | 14.2 | 12.4 | 12.7 |
| 6 | 20.4 | 18.9 | 21.3 | 17.9 | 17.1 | 17.9 |
| 7 | 27.3 | 28.9 | 28.4 | 26.8 | 25.0 | 24.9 |
| 8 | 39.3 | 43.3 | 45.3 | 42.4 | 47.3 | 52.7 |
| 9 | 57.6 | 64.5 | 67.8 | 68.1 | 70.4 | 72.0 |
| 10 | 86.8 | 86.2 | 87.7 | 87.0 | 87.0 | 91.4 |

Notes: GDP per capita from Heston, Summers, and Aten (2009). Countries are ranked according to GDP per capita in each year and divided into groups, with Decile 1 being the poorest countries and Decile 10 being the richest countries. As a result, countries in each decile may vary from year to year.

of poor countries, such as Deciles 2–7, even though their relative decline is not as dramatic as in the poorest countries.

One explanation for the large differences in income per capita observed across countries today attributes them to the countries' timing of the start of industrialization: Poor countries are slowly catching up to rich countries that started the process of modern growth much earlier. In particular, Lucas (2000, 2002) describes the cross-country differences in the timing of takeoff in growth in income per capita by looking at the historical time series of GDP per capita from 1500 to today.[6] Lucas shows that prior to 1800, differences in income per capita were moderate (about a factor of 2 between rich and poor countries), but that the differences quickly expanded when, starting with the process of industrialization, GDP per capita no longer remained stagnant for a group of initially western countries and started to increase at positive rates. Lucas conjectures that if today's income differences across countries result from differences in the timing when modern growth takes off in a country, then the distribution of per capita income may shrink again to pre-industrial levels once all countries have made the transition. This interpretation of the historical relevance of today's income differences seems difficult to reconcile with the expanding income differences observed in most deciles of the income distribution in the cross-country data reported in Table 1. I will return to this issue in Section 2, where I review the related literature.

[6] In related work, Buera, Monge-Naranjo, and Primiceri (2011) study the evolution of state-intervention and market-oriented policies across countries and time in the context of a learning model where past experiences (including those of countries' neighbors) determine policy choices.

In addition to documenting the large income differences across countries, the development accounting literature has established that differences in income per capita are mostly driven by differences in labor productivity (often measured as either GDP per worker or GDP per labor hour) since differences in labor supply (measured as either employment to population ratio or total hours of work per capita) are not large enough to explain a substantial portion of the differences in per capita income across countries. In turn, differences in labor productivity are mostly accounted for by differences in TFP. That is, differences in income per capita are not explained by measurable factors such as employment, physical capital, or human capital.[7]

**Country Experiences over Time**

The reported evolution of the income distribution across countries hides tremendous variation in country experiences over time. In the data, there are numerous episodes of catch up, catch up followed by a slowdown, stagnation, and even decline. While reporting time series for 101 countries is impractical, Table 2 attempts to summarize country experiences by reporting the evolution of average GDP per capita relative to that of the United States for 20 groups, each comprising 5 percent of countries in the sample. Unlike in Table 1 and Figure 1, the countries in each group in Table 2 remain the same over time and represent the ranking of countries according to relative GDP per capita in 1960.

Focusing on the richest and poorest 5 percent of countries in 1960, I find that inequality in GDP per capita actually declined from a factor of 26 in 1960 to 16 in 2007, as a result of the richest countries in 1960 declining relative to the leading country (from .95 in 1960 to .81 in 2007) and the poorest countries in 1960 catching up relative to the leading country (from .037 in 1960 to .049 in 2007). Table 2 also shows that episodes of catch up and decline occur throughout the income distribution in 1960, with countries in Group 7 almost tripling their relative income (from .11 in 1960 to .30 in 2007). Table 2 does not identify individual countries featuring catch up or decline in relative income. To complement the summary in Table 2, Figure 2 documents the time series of GDP per capita for selected countries with remarkable growth experiences in the sample period. I emphasize the episodes of remarkable catch up in per capita income by highlighting Singapore, Botswana, and more recently China and India. I also note the growing gap in per capita income between the United

[7] See, for instance, Klenow and Rodríguez-Clare (1997), Hall and Jones (1999), Caselli (2005), and Hsieh and Klenow (2010). A critical element in establishing the relative importance of TFP and factors of production in explaining income differences across countries is the treatment of human capital. There is a recent literature addressing the importance of human capital in amplifying differences in TFP across countries; for instance, Manuelli and Seshadri (2006) and Erosa, Koreshkova, and Restuccia (2010).

**Table 2 GDP per Capita Relative to the United States (Percent)**

| | Year | | | | | |
|---|---|---|---|---|---|---|
| **GR5pc** | **1960** | **1970** | **1980** | **1990** | **2000** | **2007** |
| 1 | 3.7 | 3.7 | 3.6 | 3.5 | 3.9 | 4.9 |
| 2 | 5.1 | 5.0 | 4.7 | 3.9 | 3.6 | 3.8 |
| 3 | 5.8 | 6.2 | 6.8 | 7.3 | 7.1 | 7.4 |
| 4 | 6.8 | 6.0 | 6.3 | 6.4 | 6.3 | 6.3 |
| 5 | 7.9 | 7.7 | 8.5 | 8.8 | 8.5 | 8.7 |
| 6 | 9.4 | 7.4 | 6.8 | 6.6 | 6.4 | 6.7 |
| 7 | 10.7 | 12.5 | 18.0 | 21.9 | 26.0 | 29.5 |
| 8 | 12.1 | 10.7 | 9.2 | 7.6 | 5.9 | 5.7 |
| 9 | 14.1 | 12.2 | 12.7 | 11.5 | 11.6 | 12.6 |
| 10 | 15.9 | 14.6 | 15.3 | 14.0 | 13.6 | 14.5 |
| 11 | 18.8 | 17.5 | 17.8 | 13.2 | 10.8 | 10.3 |
| 12 | 22.1 | 24.9 | 25.9 | 21.2 | 17.5 | 14.7 |
| 13 | 25.9 | 26.8 | 33.7 | 39.1 | 41.9 | 45.7 |
| 14 | 28.7 | 32.3 | 31.2 | 28.8 | 30.2 | 31.3 |
| 15 | 36.6 | 37.9 | 37.9 | 35.6 | 34.3 | 34.7 |
| 16 | 41.9 | 47.7 | 47.3 | 45.0 | 46.6 | 49.1 |
| 17 | 51.8 | 55.1 | 56.2 | 50.0 | 54.9 | 62.3 |
| 18 | 63.4 | 68.6 | 72.2 | 67.6 | 64.3 | 65.5 |
| 19 | 78.7 | 79.9 | 81.3 | 80.8 | 83.3 | 85.2 |
| 20 | 94.9 | 91.8 | 88.4 | 83.2 | 79.0 | 80.6 |

Notes: GDP per capita from Heston, Summers, and Aten (2009). Countries are ranked according to GDP per capita in 1960 and divided into groups. The country groups remain constant across years. For instance, Group 1 refers to the poorest countries in 1960 whose GDP per capita relative to the United States was 3.7 percent in 1960 and 4.9 percent in 2007.

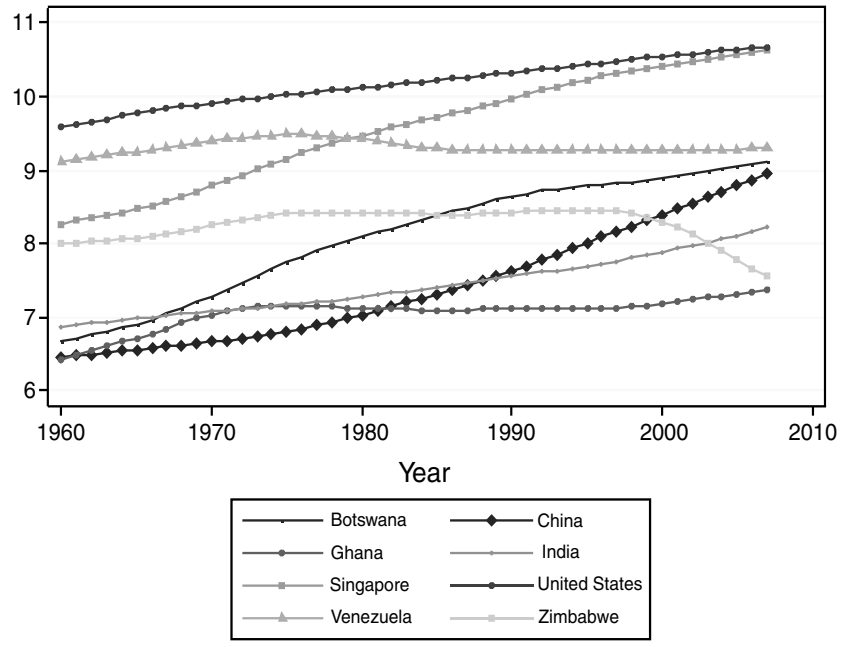
States and Venezuela, Ghana, and Zimbabwe. Explaining these remarkable growth and collapse episodes is a challenging and exciting task for the field of quantitative growth economics.

## 2. STRUCTURAL TRANSFORMATION

In this section I discuss the recent quantitative literature that emphasizes the role of factor reallocation across sectors in explaining income and growth differences across countries.[8] The process of economic development is associated with a systematic reallocation of factors of production across sectors—

[8] The literature on structural transformation is too large to be fairly recognized in this article; please see the recent survey in Herrendorf, Rogerson, and Valentinyi (2011) for references. I note, however, that the literature considers several approaches in driving reallocation across sectors. For example, some models emphasize non-homotetic preferences, such as Echevarria (1997) and Kongsamut, Rebelo, and Xie (2001), while other models emphasize non-unitary elasticity of substitution across goods and differential productivity growth across sectors such as Baumol (1967) and Ngai and Pissarides (2007).

**Figure 2 GDP per Capita, Selected Countries (in logs)**




Notes: GDP per capita data is from Heston, Summers, and Aten (2009).

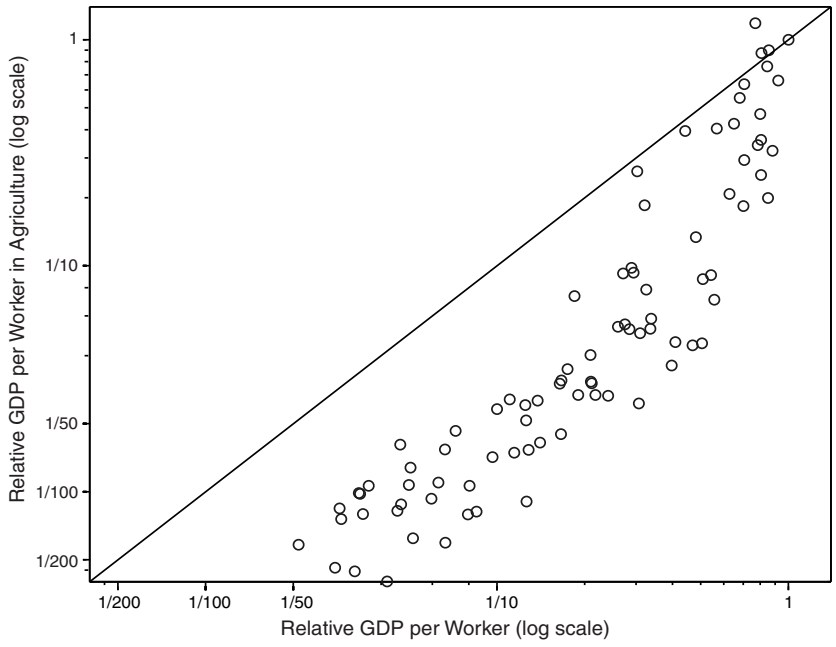
the structural transformation—whereby factors are reallocated initially from agriculture to industry and services and later from agriculture and industry to services. There is a growing literature, following Kuznets (1966), emphasizing the importance of sectoral reallocation for aggregate outcomes.

## The Role of Agriculture

An important development in the understanding of income differences across countries has been the recognition that agriculture plays a crucial role. Progress in this area has been enhanced by the availability of comparable data on agricultural output across countries, allowing a quantitative characterization of the magnitude of agricultural productivity differences, and by quantitative assessments of plausible hypotheses using sectoral models.[9] To start, let me motivate why agriculture is important. From a historical perspective, the reallocation

[9] See, for instance, Rao (1993) and Restuccia, Yang, and Zhu (2008).

**Figure 3 Labor Productivity in Agriculture across Countries**

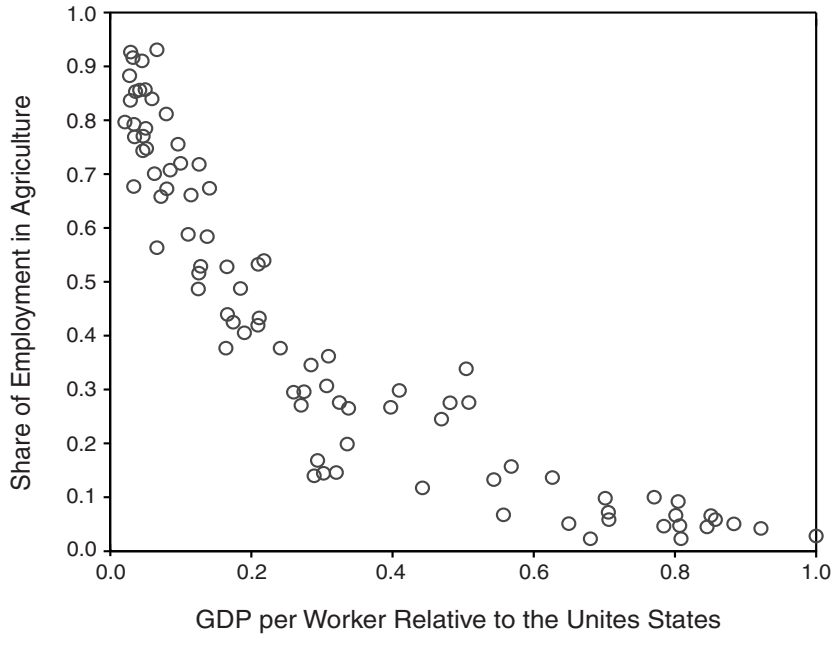



Notes: Data from Restuccia, Yang, and Zhu (2008). Data for 1985.

process away from agriculture—hence, the process of industrialization—has been associated with improvements in agricultural productivity (see, for instance, Kuznets [1966]). In addition, in the more recent cross-country data, we observe that agriculture plays a critical role since, relative to rich countries, labor productivity in agriculture in poor countries is much lower than in the rest of the economy (see Figure 3) and most of their labor is allocated to agriculture. Whereas poor countries allocate more than 85 percent of the labor force to agriculture, rich countries only allocate 4 percent (see Figure 4). Noting that aggregate labor productivity is the sum of labor productivity across sectors weighted by the share of employment in each sector, and using labor productivity and employment data for rich and poor countries, I find that agriculture accounts for 85 percent of the difference in aggregate labor productivity across rich and poor countries.[10] Recalling that the bulk of the

[10] The data reported in Restuccia, Yang, and Zhu (2008) suggests that if poor countries were to have the same share of employment and labor productivity in agriculture as the rich countries, then the aggregate labor productivity factor between rich and poor countries would be

**Figure 4 Share of Employment in Agriculture**




Notes: This is Figure 1 in Restuccia, Yang, and Zhu (2008). Data for 1985.

differences in income per capita across countries are explained by differences in labor productivity, the literature concludes that understanding productivity and labor allocation in agriculture may be at the core of income differences among rich and poor countries. The recognition that agriculture is central in understanding low productivity in poor countries is important in seeking the factors that account for this outcome, whether these factors are policy driven or institutional.

There are two broad branches of this literature. The first branch can be roughly summarized as emphasizing the timing of industrialization in explaining current differences in income. The focus is on the delay in the process of structural transformation—broadly described as the process of resource reallocation from agriculture to other sectors in the economy. The second branch focuses on explaining the factors behind the low productivity in

approximately 5-fold instead of the actual 34-fold difference. Hence, agriculture accounts for 85 percent $(100 - 5/34 \times 100)$ of the difference in aggregate labor productivity between rich and poor countries in the data.

agriculture in poor countries observed in the cross-country data at a point in time. The two branches are closely connected as they seek to assess the relevance of the sectoral structure (agriculture versus non-agriculture in particular) in cross-country income differences. The two branches differ in terms of the relevance of the information that can be extracted from time-series variations in the sectoral structure across countries. I expand on this issue below.

While there is an old and extensive literature in development on the role of agriculture and structural transformation, only recently has the literature provided a quantitative assessment. Gollin, Parente, and Rogerson (2002) provide a model that rationalizes delays in the process of structural transformation and rising income inequality over long periods of time.[11] The model formalizes many ideas in the traditional development literature and provides a quantitative assessment of the importance of the timing of the adoption of modern agricultural technology in explaining current international income differences. The model in Gollin, Parente, and Rogerson (2002) is quite simple. The economy is populated by homogeneous individuals that derive utility from consuming agricultural and non-agricultural goods and there is a subsistence need for agricultural goods. Thus, at low levels of income, individuals spend a bigger fraction of their income on agricultural goods than at high levels of income. There is strong empirical evidence in support of these types of preferences. Agricultural goods can be produced with two alternative technologies: a traditional production function that is linear in labor and features no labor productivity growth, and a modern technology, also linear in labor, that features positive labor productivity growth. The technology for producing non-agricultural goods is standard, featuring capital and labor inputs and positive labor productivity growth. The economy is characterized as follows. When the productivity of the modern agricultural technology is low—below that of the traditional technology—all labor is allocated to agriculture and income per capita is low and stagnant—essentially people are consuming close to their subsistence needs. This characterization resembles economies prior to 1800, where income per capita was roughly constant over time. Because of positive productivity growth in the modern agricultural technology, at some point in time the modern technology becomes more productive than the traditional technology and the adoption of the modern technology in agriculture starts the process of industrialization and modern growth. With productivity growth in modern agriculture, labor is systematically reallocated from agriculture to non-agriculture over time. In the long run, the economy features properties that are consistent with the characterization of modern growth—a positive and stable per capita income growth.

[11] Closely related is the work of Lucas (2000) and Hansen and Prescott (2002), although these articles do not explicitly consider the agricultural sector.

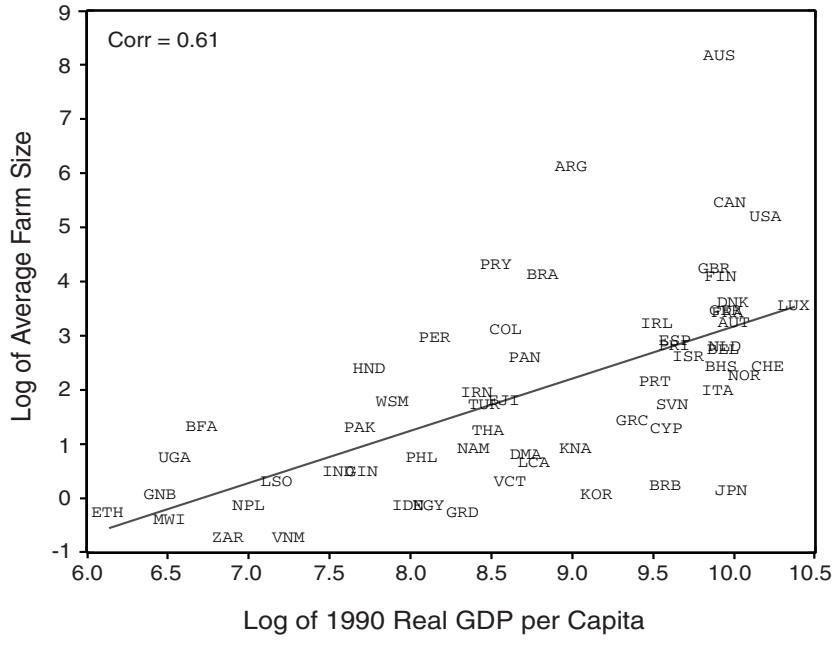
Gollin, Parente, and Rogerson (2002) calibrate a benchmark economy to U.K. data for about 250 years and show that the model reproduces very well the reallocation of labor out of agriculture over time, as well as the growth in output per capita. Then, the authors use the model to conduct experiments where the productivity of modern agriculture is lowered relative to the level in the United Kingdom. Different productivity levels imply different dates at which the modern technology in agriculture becomes more productive than the traditional technology and, hence, the date at which industrialization and modern growth starts. Interestingly, reasonable differences in the timing of adoption of the modern agricultural technology imply large current differences in output per capita across economies. Moreover, the differences in income per capita persist for long periods of time. One conclusion from this study is that, as argued by Lucas (2002), a large portion of today's income differences across countries result from differences in the timing of the adoption of modern technologies.[12] There are two issues with this interpretation of the results. First, the persistence of income gaps over time in the model is related to the assumption that the process of reallocation of employment out of agriculture is common across counties. Cross-country data indicate, however, that countries that started the process of industrialization later than the United States or United Kingdom have accomplished a comparable transformation in a much shorter time (see Duarte and Restuccia [2007] for the case of Portugal). Second, the model implies that income gaps should diminish over time, which is not observed in the recent cross-country data in Section 1. I conclude that this branch of the literature is useful in understanding cross-country differences in the timing of industrialization and the related transition, but it is unlikely to explain the current differences in agricultural productivity observed between rich and poor countries.

The second branch of the literature focuses on the factors behind low productivity in agriculture in poor countries. The focus is on understanding cross-country differences in the agricultural sector at a point in time as opposed to cross-country differences over time. Restuccia, Yang, and Zhu (2008) develop a two-sector model of agriculture and non-agriculture emphasizing economy-wide differences in productivity and barriers to intermediate input use and labor mobility in agriculture. Empirical evidence suggests there is a strong systematic relationship between the level of development of a country and two forms of barriers in agriculture: a wedge between wages in agriculture and non-agriculture (barriers due to limited labor mobility), and a high relative price of non-agricultural intermediate inputs such as fertilizers and pesticides (interpreted broadly as a barrier to intermediate input use). These empirical regularities suggest that inefficiencies in agriculture may contribute to low

[12] See Ngai (2004) for a related study of the importance of barriers to investment in physical capital in the delay of the adoption of modern technologies.

agricultural productivity in poor countries and, as a consequence, a large share of employment in agriculture. Restuccia, Yang, and Zhu (2008) embed these features in a model where preferences for consumption goods feature a subsistence level requirement for food. Furthermore, producing non-agricultural goods requires only labor while producing agricultural goods requires land, labor, and non-agricultural intermediate inputs. The spirit of the exercise conducted in Restuccia, Yang, and Zhu (2008) is as follows. Since the technology for producing non-agricultural goods is linear in the labor input, data on labor productivity in non-agriculture pins down the level of economy-wide productivity in each country. This level of productivity is assumed to be exogenous in the analysis but standard explanations of technology adoption and capital accumulation can be applied for this factor. Importantly, these explanations are not specific to the agricultural sector. Restuccia, Yang, and Zhu (2008) also take as given the differences across countries in the land-to-population ratio, the barriers to intermediate input use in agriculture, and the barriers to labor mobility. These objects are directly pinned down by country observations. Then the question becomes: How important are all these factors (and each in isolation) in explaining low productivity in agriculture and high agricultural employment in poor countries? There are several results worth emphasizing. First, if the model could reproduce the low productivity in agriculture observed in poor countries (by, for example, lowering an agriculture-specific productivity parameter in poor countries), then the model can rationalize the observed large share of employment in agriculture in these countries. Hence, understanding low productivity in agriculture in poor countries is key, with the ensuing reallocation of labor acting as a transmission mechanism to aggregate productivity differences. Second, exogenous differences in economy-wide productivity (measured as differences in non-agricultural productivity) and barriers are important in explaining low productivity in agriculture in poor countries, whereas differences in land endowments are of second-order importance. In particular, the model with exogenous differences in economy-wide productivity, barriers, and land endowments, can explain two-thirds of the differences in labor productivity in agriculture between rich and poor countries, still leaving an important factor unexplained (about one-third). Third, inefficiencies in agriculture are not the only determinant of low productivity in agriculture in poor countries. If productivity in non-agriculture in poor countries were to be equalized to that of rich countries—even keeping productivity and barriers in agriculture the same—the model would imply levels of productivity and employment in poor countries much closer to that of rich countries compared to the baseline model, for instance, a share of employment in agriculture of 30 percent versus 68 percent in the baseline model, a factor difference in labor productivity in agriculture of 10-fold versus 23-fold in the baseline model, and an aggregate productivity difference of 1.4-fold versus 10.8-fold in the baseline model. This result suggests that not all problems

**Figure 5 Average Farm Size across Countries**




Source: Adamopoulos and Restuccia (2011).

lie in agriculture; instead, solving the problems that prevent non-agricultural productivity in poor countries to rise to the level of developed countries can help in eliminating a substantial portion of the large differences in income among rich and poor countries.

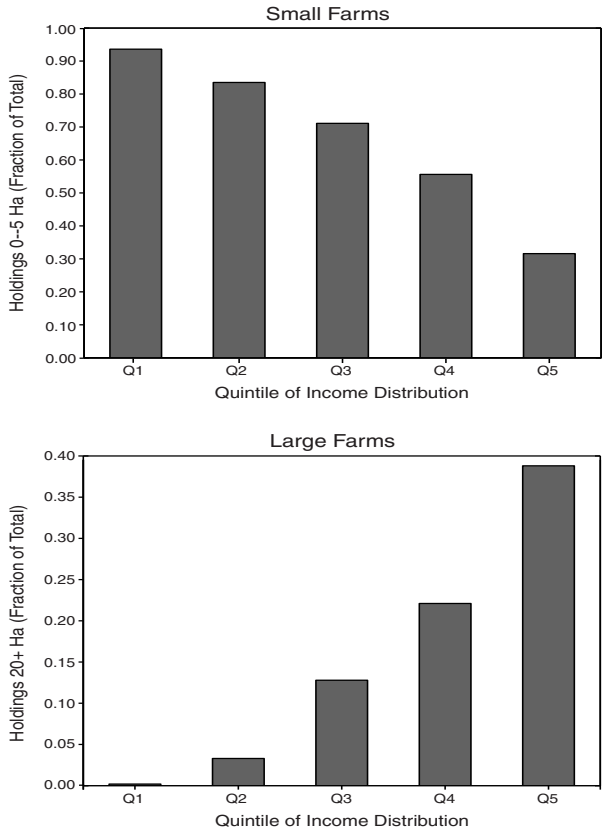
Since there is still a large unexplained gap in labor productivity in agriculture, understanding low productivity in agriculture in poor countries has remained an active area of research. Four recent contributions have emphasized the role of transportation infrastructure (Adamopoulos 2011), the role of ability selection into agriculture (Lagakos and Waugh 2011), the role of farm size (Adamopoulos and Restuccia 2011), and the role of trade restrictions for importing food (Tombe 2011).[13] In this article, I only summarize the findings on the importance of farm-size differences across countries. Adamopoulos and

[13] See also the recent accounting exercises of the productivity gap between agriculture and non-agriculture in Herrendorf and Schoellman (2011), who emphasize the differences across U.S. states, and in Gollin, Lagakos, and Waugh (2011), who emphasize the differences across developing countries.

Restuccia (2011) develop a model of farm size to investigate its importance in understanding the low productivity problem in agriculture. The motivation for why farm size may matter is twofold. First, there are striking differences in average farm sizes and farm-size distributions across countries. Whereas average farm size is 54 Hectares (Ha) in the richest 20 percent of countries, average farm size is only 1.6 Ha in the poorest 20 percent of countries, a 34-fold difference. Figure 5 documents the positive relationship between the level of development and average farm size across countries. Cross-country differences in farm-size distributions are systematic. Whereas in poor countries, more than 90 percent of the farms are small (less than 5 Ha), only around 30 percent of the farms in rich countries are small. In poor countries, none of the farms are large (more than 20 Ha), while almost 40 percent of the farms in rich countries are large. (See Figure 6 for a documentation of the share of small and large farms across quintiles of the income distribution.) Second, labor productivity is much higher in large than in small farms. For instance, in the data from the U.S. Census of Agriculture, average labor productivity in farms greater than 800 Ha relative to farms less than 4 Ha is a factor between 14- and 34-fold depending on how operators and hired labor are treated in the measure of labor in farms. The question addressed by Adamopoulos and Restuccia (2011) is what explains farm-size differences across countries and whether or not these differences help explain the productivity problem in agriculture in poor countries.

Adamopoulos and Restuccia (2011) consider a model of farm size that is based on the span-of-control model of Lucas (1978) embedded into a standard sectoral model of agriculture and non-agriculture. The production unit in agriculture is a farm that requires the input of a farmer (labor), capital, and land. Farmers differ in their productivity of managing a farm and the farming technology is such that for each type of farmer there is an optimal farm size where more productive farmers demand more capital and land and, hence, manage larger farms. While reallocation between agriculture and non-agriculture in the model depends on the same fundamental channels described in the previous literature (e.g., Gollin, Parente, and Rogerson [2002] and Restuccia, Yang, and Zhu [2008]), productivity in agriculture is also determined by the allocation of factors (capital and land) across farmers. There are three main findings. First, farm-size distortions, such as land reforms that cap the size of farms and progressive land taxes, are the most likely explanation for differences in farm-size distributions. There is overwhelming evidence for these distortions in cross-country data and measured distortions can account quantitatively for most of the differences in farm-size distributions across countries. Other potential explanations such as cross-country differences in aggregate factor endowments (land, capital, and economy-wide productivity) can account for, at most, one-fourth of the cross-country farm-size differences. Second, calibrating farm-size distortions to account for the observed farm-size differences helps explain

**Figure 6 Farm Size Distribution across Countries**




Source: Adamopoulos and Restuccia (2011).

three-fourths of the differences in agricultural and aggregate labor productivity across countries, with the remaining one-fourth being explained by differences in

aggregate factors. Third, specific distortionary policies in individual countries such as a land reform in the Philippines and progressive land taxation reform in Pakistan are found to generate substantial drops in size and productivity in these countries. Moreover, other factors occurring at the same time or over time in these countries are found to potentially mask the negative effects of distortionary policies on size and productivity in the agricultural sector, making empirical characterizations of these distortionary policies difficult.

**Reallocation to Services**

As emphasized earlier, models of structural transformation, that is the reallocation of labor across sectors in an economy over time, have featured prominently in historical perspectives of growth and the timing of industrialization such as in Lucas (2000, 2002) and Gollin, Parente, and Rogerson (2002).[14] Duarte and Restuccia (2010) argue that structural transformation is also closely connected with the set of facts emphasized in Section 1 about the diversity of growth patterns in the time series for individual countries, the patterns of catch up, slowdown, stagnation, and decline in labor productivity that are observed even for more developed countries. For these countries, agriculture is less important in the economy and the more relevant transformation involves a substantial shift to services rather than a shift out of agriculture.[15]

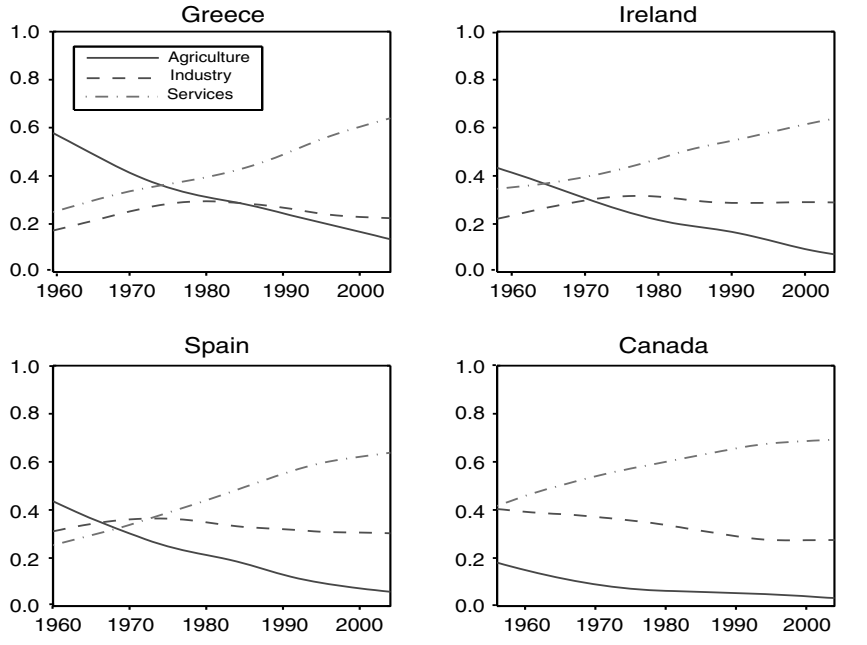
Duarte and Restuccia (2010) develop a tractable model of the structural transformation to quantitatively assess the contribution of sectoral labor productivity growth in understanding the evolution of aggregate productivity across countries. The model consists of three sectors: agriculture, industry, and services, with linear technologies in labor in each sector. Structural transformation is driven in the model by two factors: non-homothetic preferences for agriculture and services goods (with income elasticity less than one for agriculture and more than one for services) and an elasticity of substitution less than one for industry and services so that differential productivity growth in industry and services also generates reallocation across these sectors. Hence, a poor country in the model featuring low productivity in all sectors allocates a large share of labor to agriculture, a low share of labor to services, and the remaining labor to industry. With positive productivity growth in all sectors, labor is reallocated away from agriculture toward industry and services. With faster productivity growth in manufacturing than in services—as documented in the cross-country data by Duarte and Restuccia (2010)—there is further reallocation of labor from industry to services. Further, faster productivity growth in agriculture produces a speedier transformation out of agriculture. The framework is used with two purposes. The first purpose is to infer from the model comparable measures of labor productivity across sectors and countries. These sectoral measures of labor productivity are not generally available for a large cross-section of countries. The second purpose is to assess quantitatively the relevance of sectoral labor productivity growth in driving labor reallocation across sectors and aggregate productivity over time across countries.

---

[14] See also the recent survey article by Herrendorf, Rogerson, and Valentinyi (2011) on models of structural transformation.

[15] For example, notice in Figure 7 how, in the earlier stages of structural transformation in Greece, Ireland, and Spain, labor reallocated from agriculture to both industry and services, but in a later stage (and throughout Canada) reallocation also occurs from industry to services, with the agricultural sector representing in a small fraction of total hours.

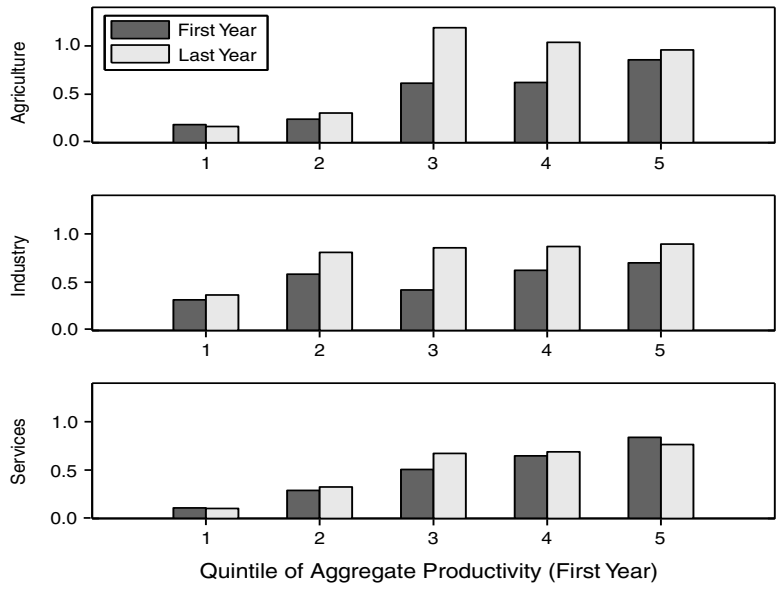
**Figure 7 Share of Hours across Sectors, Selected Countries**




Notes: This is Figure 2 in Duarte and Restuccia (2010).

Two key findings emerge from this framework. The first finding is that labor productivity differences across countries at a point in time are largest in agriculture and services and smaller in industry. These findings have the following mechanical and intuitive implication. Suppose for the moment that labor productivity differences across sectors and countries remain constant over time, that is, assume that growth in labor productivity in each sector is equal across countries. Then, with positive productivity growth in all sectors, the process of structural transformation implies that countries are reallocating labor from agriculture to manufacturing and to services. Since labor productivity is lower in agriculture relative to industry in poor countries compared to rich countries, the reallocation of labor from agriculture to manufacturing can explain an increase (catch up) in relative productivity for the poor countries. As the process of structural transformation continues with reallocation from manufacturing (and to a lesser extent agriculture) to services, a lower ratio of labor productivity in services relative to industry in poor countries compared to rich ones may imply episodes of slowdown, stagnation, and decline in relative aggregate productivity. The cross-country growth pattern across sectors gets a

**Figure 8 Relative Labor Productivity across Sectors and Countries**




Notes: This is Figure 6 in Duarte and Restuccia (2010).

bit more complicated when, in addition, labor productivity gaps are changing over time. In fact, the evidence suggests that there has been substantial cross-country catch up in labor productivity in agriculture and manufacturing over time but not in services, and that this process is important in understanding the evolution of aggregate productivity across countries. Figure 8 shows the implications of the model in Duarte and Restuccia (2010) for the first year in the sample (1956 for most countries) and the last year in the sample (2005 for most countries). Countries in the second, third, and fourth quintiles of the income distribution managed to achieve substantial catch up in relative sectoral productivity for agriculture and industry, but in general there is a lack of catch up in productivity in services.

The second finding is that the patterns of sectoral productivity across sectors and countries just emphasized account for most of the labor reallocation observed across countries.[16] Moreover, the catch up in manufacturing

[16] Duarte and Restuccia (2010) emphasize that, for some countries, sectoral productivity growth generates labor reallocation that is different from the data, suggesting that distortions/frictions may be important for some individual-country experiences.

productivity accounts for 50 percent of the catch up in aggregate productivity across countries and the lack of catch up in services explains all the experiences of slowdown, stagnation, and decline in aggregate productivity across countries. These findings point to the importance of the service sector in current growth experiences and present a challenge for economic policy in disentangling the relevant policies/regulations that affect the evolution of service-sector and aggregate productivity across countries.

## 3. REALLOCATION ACROSS ESTABLISHMENTS

A recurrent finding of the development accounting literature such as in Klenow and Rodríguez-Clare (1997) and Prescott (1998) is that TFP is the most important factor in explaining income differences across countries. Most of the analysis in explaining productivity differences across countries was done in the context of frameworks with a stand-in or representative firm featuring constant returns to scale of production. The result was then an emphasis on aggregate factors that explain the lack of technology adoption in poor countries. For instance, Parente and Prescott (1994, 2000) develop a framework emphasizing barriers to technology adoption in poor countries.

Complementing this work, the evidence from microeconomic studies, such as Baily, Hulten, and Campbell (1992) and Foster, Haltiwanger, and Syverson (2008), suggests that the reallocation of factors of production—from failing to entering firms, and especially from less to more productive firms—accounts for a substantial portion of aggregate productivity growth in the data. For this reason, Restuccia and Rogerson (2008) consider a model of heterogeneous production units where reallocation across these units is at the core of measured productivity in the economy.[17]

### Misallocation and Productivity

The model in Restuccia and Rogerson (2008) embeds an industry equilibrium model of Hopenhayn (1992) into a standard one-sector growth model.[18] Production takes place in establishments. The technology at the establishment level differs in TFP and features decreasing returns to scale in capital and labor inputs. The implication of these two features is that there is an optimal size of establishments, i.e., an optimal amount of capital, labor, and output for each productivity type and the size of an establishment is positively related to productivity. In other words, the efficient allocation of factors given

[17] See also Banerjee and Duflo (2005) for a survey of closely related literature in microeconomic development.

[18] An early analysis of the importance of reallocation is in Hopenhayn and Rogerson (1993), who focus on the effect of firing taxes on employment differences across countries.

these assumptions is such that capital and labor are allocated according to productivity, and the amount of aggregate resources determines the number of establishments. The aggregate production function then features constant returns to scale in the sense that if capital and labor were to double in the economy, then the number of establishments and output would double too. A critical feature of the model is that policies or institutions that affect the prices paid or received by establishments (what Restuccia and Rogerson [2008] call idiosyncratic distortions) generate a reallocation across establishments that lowers productivity. The list of institutions and policies that create such reallocation is large and is a prevalent feature of poor countries. For example, non-competitive banking systems offering below-market interest rate loans to selected producers based on non-economic factors, governments exempting certain producers of regulations or taxes, public enterprises often associated with low productivity receiving large subsidies from the government for their operation (financed through taxes on other producers), are the type of distortions that affect the size of certain establishments inducing a misallocation of factors of production. Labor market regulation and trade restrictions may also lead to idiosyncratic distortions. The approach in Restuccia and Rogerson (2008) is to represent all these potential sources of distortions through a generic form of tax/subsidy schemes and to assess their potential impact on aggregate productivity.

Restuccia and Rogerson (2008) study policy configurations whereby a fraction of establishments is taxed at a specified rate and the remaining fraction of establishments is subsidized. The subsidy rate is such that the aggregate capital stock remains the same. The reason for this approach is that the elements that affect capital accumulation are well understood and research has shown that capital accumulation is not a crucial factor in accounting for income differences (see, for instance, Klenow and Rodríguez-Clare [1997]).[19] To make a quantitative assessment, Restuccia and Rogerson (2008) calibrate a benchmark economy with no distortions to data for the United States. The key components in calibrating the model are the elements that allow the model to reproduce the distribution of establishments and their size in the data. Experiments are conducted assuming that all countries are identical to the benchmark economy except on a configuration of idiosyncratic distortions. Even though the experiments are such that aggregate resources and the distribution of production efficiencies are the same as in the benchmark economy, idiosyncratic distortions are shown to have substantial negative effects on measured TFP and output. In particular, a policy configuration where 50 percent of the

[19] More generally though, idiosyncratic distortions to establishments can also lead to substantial effects on aggregate capital accumulation, which may be of importance for individual-country experiences. See, for instance, Bello, Blyde, and Restuccia (2011) for an assessment of idiosyncratic distortions on capital accumulation in Venezuela.

most productive establishments are taxed at 40 percent implies a drop in TFP and output of 30 percent. Drops in TFP and output can be larger if more establishments are taxed, for instance if 90 percent of establishments were taxed and only 10 percent subsidized, measured TFP and output would drop by 50 percent.[20]

While the policy experiments that Restuccia and Rogerson (2008) implement are hypothetical, there is substantial evidence on the types of policies that create idiosyncratic distortions. In related work, Hsieh and Klenow (2009) use microeconomic data of plants in the manufacturing sector for China, India, and the United States to measure the size of policy distortions and evaluate their aggregate impact. They find that eliminating misallocation in China and India (relative to that of the United States) can increase measured TFP between 30 percent and 60 percent. Roughly speaking, the intuition for how the microeconomic data can uncover the size of policy distortions is that in an economy without distortions, establishments with access to the same technology (except for TFP) and facing the same prices for output and factor inputs would equalize the marginal product of factors to the aggregate prices. With underlying differences in productivity across establishments, the more productive establishments are larger than less productive establishments. Idiosyncratic policy distortions affect the prices faced by individual establishments and, hence, prevent establishments from equalizing their marginal products. Data on establishment-level output, factor inputs, and input payments permit an evaluation of the price distortions that must be in place for the data to be an equilibrium of the distorted economy. Therefore, given the distortions, an evaluation can be made of the productivity gains from eliminating them.[21]

Interestingly, Hsieh and Klenow's (2009) empirical work also uncovers important differences between China, India, and the United States in the distribution of establishment-level productivity. The distribution of productivity across establishments is assumed to be the same across countries in Restuccia and Rogerson's (2008) experiments as the focus is on reallocation across these units. Differences in the distribution of productivity are also abstracted from in the gains from reallocation in Hsieh and Klenow's (2009) calculations.[22] The differences in the distribution of productivity across establishments can

[20] I note that Restuccia and Rogerson (2008) also look at other potential policy configurations whereby distortionary policies are either random (some establishments are subsidized and others taxed but which establishment is taxed/subsidized is not related to productivity) or the more productive establishments are subsidized. While less damaging, these alternative policy configurations also have a negative impact on aggregate productivity as the size of establishments is distorted.

[21] Much work has followed Hsieh and Klenow's (2009) approach using microeconomic data on firms to uncover distortions and productivity gains from reallocation in many countries. See, for instance, Pagés (2010) for applications in Latin American countries.

[22] Hsieh and Klenow (2009) calculate the gains from reallocation as the ratio of efficient output to actual output for each country, where efficient output is produced by assuming factors of production are assigned efficiently to the establishments in the country.

potentially be the result of distortionary policies and can be studied jointly, for example, by allowing the policy distortions to have an impact on the selection of establishments through entry/exit and on productivity investment by establishments. Recent work has started to allow for an interaction between policy distortions and the distribution of establishments. In these frameworks, the shift in the distribution of establishment-level productivity is a consequence of changes in the amount of investment by establishments on their level of productivity in the face of idiosyncratic distortions that may discourage higher efficiency and barriers to entry and doing business, which are quite prevalent in poor countries.[23] In this regard, Restuccia (2011) and Bello, Blyde, and Restuccia (2011) study variants of the Restuccia and Rogerson (2008) model, where policy distortions shift the distribution of productivity across establishments in the economy toward the lower productivity units.[24]

## Specific Policies and Institutions

A limitation of the empirical measures of idiosyncratic distortions in Hsieh and Klenow (2009) is that they don't directly connect with specific policies and institutions. Such connection is critical in the determination of policy prescriptions for poor countries. Recent studies have tried to provide a quantitative assessment of specific policies or institutions in accounting for misallocation and low productivity in poor countries. This literature cannot be described in much detail in this article.[25] Broadly speaking, the applications span issues that include: the importance of financial development such as in Buera, Kaboski, and Shin (2011), Greenwood, Sanchez, and Wang (2010, 2011), and Midrigan and Xu (2010); the relevance of size-dependent policies that discourage large-scale operation through heavier regulation and taxes such as in Guner, Ventura, and Xu (2008); the importance of restrictions to foreign direct investment such as in Burstein and Monge-Naranjo (2009); the relevance of specific policies such as land reforms and progressive land taxes that discourage large-scale operation in farming in Adamopoulos and Restuccia (2011), among many others.[26] Focusing on the role of specific factors reduces the

[23] See, for instance, the empirical measures of cost of entry in *Doing Business* 2011 from the World Bank (2011).

[24] See also the interesting work in Ranasinghe (2011a, 2011b) and a related literature in trade that emphasize a shift in the distribution of productivities, e.g., Atkeson and Burstein (2010), and Rubini (2010).

[25] The growing literature on misallocation and productivity will be the subject of a special issue of the *Review of Economic Dynamics* to be published in January 2013.

[26] There is also a growing empirical literature assessing the importance of policies on specific experiences, but often the empirical studies are limited by the availability of good-quality microeconomic data and by the difficulty of accessing the data. Two interesting examples of cases where good microeconomic data is available are the study of trade reform in Colombia in Eslava et al. (2011), where the data includes quantity and price information for each producer, allowing for a real measure of productivity at the establishment level as opposed to the typical revenue

scope of potential impact on aggregate productivity and often still involves difficult issues of measurement. As a result, much work remains to be done in identifying and measuring specific policies and institutions and assessing their quantitative significance on the allocation of resources across productive units and, hence, on understanding aggregate productivity differences across countries.

## 4. CONCLUSIONS

Differences in income across nations are large. Moreover, the data shows remarkable episodes of growth catch up and collapse. In this article, I reviewed the recent literature in quantitative growth economics, broadly addressing these facts. In a nutshell, substantial progress has been made by studying the determinants of resource allocation across heterogeneous productive units, whether across sectors or across establishments within sectors. Much more work remains to be done in determining the fundamental factors in resource allocation across productive units.

To be more concrete, while agriculture has been shown to be important in explaining the income differences between rich and poor countries, further advances are needed in accounting for the low productivity problem in agriculture in poor countries. For instance, what specific policies and institutions explain the small-scale operations in agriculture in poor countries? Is the lack of well-defined property rights important? Are price distortions or other specific policies that discriminate against large operational scales important? What sort of barriers prevent trade in agricultural goods in low productivity countries? Similarly, while differences in labor productivity across sectors and countries are found to be important in accounting for the patterns of aggregate labor productivity growth across countries, it remains to be analyzed in detail what factors/policies/institutions explain the observed differences in labor productivity levels and growth rates across sectors and countries. For example, what determines the large gap in labor productivity in the service sector even among relatively developed countries? How do regulations and market structure affect productivity in services across countries? Closely related, misallocation of resources across heterogenous production units are also found to generate substantial negative effects on measured aggregate TFP. But empirical measures of misallocation have so far been addressed in a relatively small number of countries, and these measures need to be linked with specific policies and institutions. Better measurement of individual policies and institutions affecting productivity at the establishment level, as well as better measurement of productivity at the microeconomic level, are likely to yield

---

measure of productivity, and the study of the increase in dispersion in tariffs associated with the Smoot-Hawley Tariff in the United States during the Great Depression in Bond et al. (2011).

important returns in terms of our understanding of productivity differences across countries. These advances are likely to allow for the design of effective policies addressing frictions and market imperfections that prevent an optimal allocation of resources, as well as the removal of barriers that prevent poor countries from operating closer to the technological frontier.

## REFERENCES

Adamopoulos, Tasso. 2011. "Transportation Costs, Agricultural Productivity, and Cross-Country Income Differences." *International Economic Review* 52 (2): 489–521.

Adamopoulos, Tasso, and Diego Restuccia. 2011. "The Size Distribution of Farms and International Productivity Differences." Manuscript, University of Toronto.

Atkeson, Andrew, and Ariel Tomás Burstein. 2010. "Innovation, Firm Dynamics, and International Trade." *Journal of Political Economy* 118 (3): 433–84.

Baily, Martin Neal, Charles Hulten, and David Campbell. 1992. "Productivity Dynamics in Manufacturing Plants." *Brooking Papers on Economic Activity: Microeconomics*, 187–267.

Banerjee, Abhijit V., and Esther Duflo. 2005. "Growth Theory through the Lens of Development Economics." In *Handbook of Economic Growth*, Vol. 1A, edited by Philippe Aghion and Steven Durlauf. New York: North Holland, 473–552.

Baumol, William J. 1967. "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis." *American Economic Review* 57 (June): 415–26.

Bello, Omar D., Juan S. Blyde, and Diego Restuccia. 2011. "Venezuela's Growth Experience." *Latin American Journal of Economics* 48 (2): 199–226.

Bond, Rick, Mario Crucini, Tristan Potter, and Joel Rodrigue. 2011. "Misallocation and Productivity Effects of the Hawley-Smoot Tariff of 1930." Manuscript, Vanderbilt University.

Buera, Francisco J., Alexander Monge-Naranjo, and Giorgio E. Primiceri. 2011. "Learning the Wealth of Nations." *Econometrica* 79 (1): 1–45.

Buera, Francisco J., Joseph Kaboski, and Yongseok Shin. 2011. "Finance and Development: A Tale of Two Sectors." *American Economic Review* 101 (August): 1,964–2,002.

Burstein, Ariel T., and Alexander Monge-Naranjo. 2009. "Foreign Know-How, Firm Control, and the Income of Developing Countries." *Quarterly Journal of Economics* 124 (1): 149–95.

Caselli, Francesco. 2005. "Accounting for Cross-Country Income Differences." In *Handbook of Economic Growth*, Vol. 1A, edited by Philippe Aghion and Steven Durlauf. New York: North Holland, 679–741.

Duarte, Margarida, and Diego Restuccia. 2006. "The Productivity of Nations." Federal Reserve Bank of Richmond *Economic Quarterly* 92 (Summer): 195–223.

Duarte, Margarida, and Diego Restuccia. 2007. "The Structural Transformation and Aggregate Productivity in Portugal." *Portuguese Economic Journal* 6 (April): 26-46.

Duarte, Margarida, and Diego Restuccia. 2010. "The Role of the Structural Transformation in Aggregate Productivity." *Quarterly Journal of Economics* 125 (February): 129–73.

Echevarria, Cristina. 1997. "Changes in Sectoral Composition Associated with Economic Growth." *International Economic Review* 38 (May): 431–52.

Erosa, Andres, Tatyana Koreshkova, and Diego Restuccia. 2010. "How Important is Human Capital: A Quantitative Theory Assessment of World Income Inequality." *Review of Economic Studies* 77 (October): 1,421–49.

Eslava, Marcela, John Haltiwanger, Adriana Kugler, and Maurice Kugler. 2011. "Trade, Technical Change and Market Selection: Evidence from Manufacturing Plants in Colombia." Manuscript, University of Maryland.

Foster, Lucia, John Haltiwanger, and Chad Syverson. 2008. "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review* 98 (1): 394–425.

Gollin, Doug, David Lagakos, and Mike Waugh. 2011. "The Agricultural Productivity Gap in Developing Countries." Manuscript, Arizona State University.

Gollin, Doug, Stephen L. Parente, and Richard Rogerson. 2002. "The Role of Agriculture in Development." *American Economic Review* 92 (May): 160–4.

Greenwood, Jeremy, Juan M. Sanchez, and Cheng Wang. 2010. "Financing Development: The Role of Information Costs." *American Economic Review* 100 (September): 1,875–91.

Greenwood, Jeremy, Juan M. Sanchez, and Cheng Wang. 2011. "Quantifying the Impact of Financial Development on Economic Development." Manuscript, University of Pennsylvania.

Guner, Nezih, Gustavo Ventura, and Yi Xu. 2008. "Macroeconomic Implications of Size-Dependent Policies." *Review of Economic Dynamics* 11 (October): 721–44.

Hall, Robert E., and Charles I. Jones. 1999. "Why Do Some Countries Produce so Much More Output per Worker than Others." *The Quarterly Journal of Economics* 114 (February): 83–116.

Hansen, Gary D., and Edward C. Prescott. 2002. "Malthus to Solow." *American Economic Review* 92 (September): 1,205–17.

Herrendorf, Berthold, and Todd Schoellman. 2011. "Why is Labor Productivity so Low in Agriculture." Manuscript, Arizona State University.

Herrendorf, Berthold, Richard Rogerson, and Akos Valentinyi. 2011. "Growth and Structural Transformation." Manuscript, Princeton University. Forthcoming in the *Handbook of Economic Growth*.

Heston, Alan, Robert Summers, and Bettina Aten. 2009. "Penn World Table Version 6.3." Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania (August).

Hopenhayn, Hugo A. 1992. "Entry, Exit, and Firm Dynamics in Long Run Equilibrium." *Econometrica* 60 (September): 1,127–50.

Hopenhayn, Hugo A., and Richard Rogerson. 1993. "Job Turnover and Policy Evaluation: A General Equilibrium Analysis." *Journal of Political Economy* 101 (October): 915–38.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *The Quarterly Journal of Economics* 124 (November): 1,403–48.

Hsieh, Chang-Tai, and Peter J. Klenow. 2010. "Development Accounting." *American Economic Journal: Macroeconomics* 2 (January): 207–23.

Jones, Charles I., and Peter J. Klenow. 2011. "Beyond GDP? Welfare across Countries and Time." Manuscript, Stanford University.

Klenow, Peter J., and Andrés Rodríguez-Clare. 1997. "The Neoclassical Revival in Growth Economics: Has It Gone Too Far?" In *NBER Macroeconomics Annual 1997*, Vol. 12, edited by Ben S. Bernanke and Julio Rotemberg. Cambridge, Mass.: MIT Press, 73–114.

Kongsamut, Piyabha, Sergio Rebelo, and Danyang Xie. 2001. "Beyond Balanced Growth." *Review of Economic Studies* 68 (October): 869–82.

Kuznets, S. 1966. *Modern Economic Growth*. New Haven, Conn.: Yale University Press.

Lagakos, David, and Michael Waugh. 2011. "Specialization, Economic Development, and Aggregate Productivity Differences." Mimeo, New York University.

Lucas, Jr., Robert E. 1978. "On the Size Distribution of Business Firms." *Bell Journal of Economics* 9 (Autumn): 508–23.

Lucas, Jr., Robert E. 2000. "Some Macroeconomics for the 21st Century." *Journal of Economic Perspectives* 14 (Winter): 159–68.

Lucas, Jr., Robert E. 2002. "The Industrial Revolution: Past and Future." In *Lectures on Economic Growth*. Cambridge, Mass.: Harvard University Press, 109–90.

Manuelli, Rodolfo, and Ananth Seshadri. 2006. "Human Capital and the Wealth of Nations." Manuscript, University of Wisconsin.

Midrigan, Virgiliu, and Daniel Yi Xu. 2010. "Finance and Misallocation: Evidence from Plant-level Data." Manuscript, New York University.

Ngai, L. Rachel. 2004. "Barriers and the Transition to Modern Growth." *Journal of Monetary Economics* 51 (October): 1,353–83.

Ngai, L. Rachel, and Christopher A. Pissarides. 2007. "Structural Change in a Multi-Sector Model of Growth." *American Economic Review* 97 (March): 429–43.

Pagés, Carmen. 2010. *The Age of Productivity: Transforming Economies from the Bottom Up*. New York: Palgrave MacMillan.

Parente, Stephen L., and Edward C. Prescott. 1993. "Changes in the Wealth of Nations." Federal Reserve Bank of Minneapolis *Quarterly Review* 17 (Spring): 3–16.

Parente, Stephen L., and Edward C. Prescott. 1994. "Barriers to Technology Adoption and Development." *Journal of Political Economy* 102 (April): 298–321.

Parente, Stephen L., and Edward C. Prescott. 2000. *Barriers to Riches*. Cambridge, Mass.: MIT Press.

Prescott, Edward C. 1998. "Needed: A Theory of Total Factor Productivity." *International Economic Review* 39 (August): 525–51.

Ranasinghe, Ashantha. 2011a. "Property Rights, Extortion and the Misallocation of Talent." Manuscript, University of Toronto.

Ranasinghe, Ashantha. 2011b. "Impact of Policy Distortions on Plant-level Innovation, Productivity Dynamics and TFP." Manuscript, University of Toronto.

Rao, D. S. Prasada. 1993. *Intercountry Comparisons of Agricultural Output and Productivity*. Rome: Food and Agriculture Organization of the United Nations.

Restuccia, Diego, and Richard Rogerson. 2008. "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments." *Review of Economic Dynamics* 11 (October): 707–20.

Restuccia, Diego, Dennis Tao Yang, and Xiaodong Zhu. 2008. "Agriculture and Aggregate Productivity: A Quantitative Cross-Country Analysis." *Journal of Monetary Economics* 55 (March): 234–50.

Restuccia, Diego. 2011. "The Latin American Development Problem." Manuscript, University of Toronto.

Rubini, Loris. 2010. "Innovation and the Elasticity of Trade Volumes to Tariff Reductions." Manuscript, Universidad Carlos III de Madrid.

Tombe, Trevor. 2011. "The Missing Food Problem: How Low Agricultural Imports Contribute to International Income Differences." Manuscript, University of Toronto.

World Bank. 2011. *Doing Business 2011*. Prepared by the Doing Business Unit. Washington, D.C.: World Bank (November).