Journal of Economic Perspectives—Volume 8, Number 1—Winter 1994–Pages 3–22

The Origins of Endogenous Growth

Paul M. Romer

The phrase "endogenous growth" embraces a diverse body of theoretical and empirical work that emerged in the 1980s. This work distinguishes itself from neoclassical growth by emphasizing that economic growth is an endogenous outcome of an economic system, not the result of forces that impinge from outside. For this reason, the theoretical work does not invoke exogenous technological change to explain why income per capita has increased by an order of magnitude since the industrial revolution. The empirical work does not settle for measuring a growth accounting residual that grows at different rates in different countries. It tries instead to uncover the private and public sector choices that cause the rate of growth of the residual to vary across countries. As in neoclassical growth theory, the focus in endogenous growth is on the behavior of the economy as a whole. As a result, this work is complementary to, but different from, the study of research and development or productivity at the level of the industry or firm.

This paper recounts two versions that are told of the origins of work on endogenous growth. The first concerns what has been called the convergence controversy. The second concerns the struggle to construct a viable alternative to perfect competition in aggregate-level theory. These accounts are not surveys. They are descriptions of the scholarly equivalent to creation myths, simple stories that economists tell themselves and each other to give meaning and structure to their current research efforts. Understanding the differences between these two stories matters because they teach different lessons about the relative importance of theoretical work and empirical work in economic analysis and they suggest different directions for future work on growth.

■ *Paul M. Romer is Professor of Economics, University of California, Berkeley, California.*

Version #1: The Convergence Controversy

The question that has attracted the most attention in recent work on growth is whether per capita income in different countries is converging. A crucial stimulus to work on this question was the creation of new data sets with information on income per capita for many countries and long periods of time (Maddison, 1982; Heston and Summers, 1991).

In his analysis of the Maddison data, William Baumol (1986) found that poorer countries like Japan and Italy substantially closed the per capita income gap with richer countries like the United States and Canada in the years from 1870 to 1979. Two objections to his analysis soon became apparent. First, in the Maddison data set, convergence takes place only in the years since World War II. Between 1870 and 1950, income per capita tended to diverge (Abramovitz, 1986). Second, the Maddison data set included only those economies that had successfully industrialized by the end of the sample period. This induces a sample selection bias that apparently accounts for most of the evidence in favor of convergence (De Long, 1988).

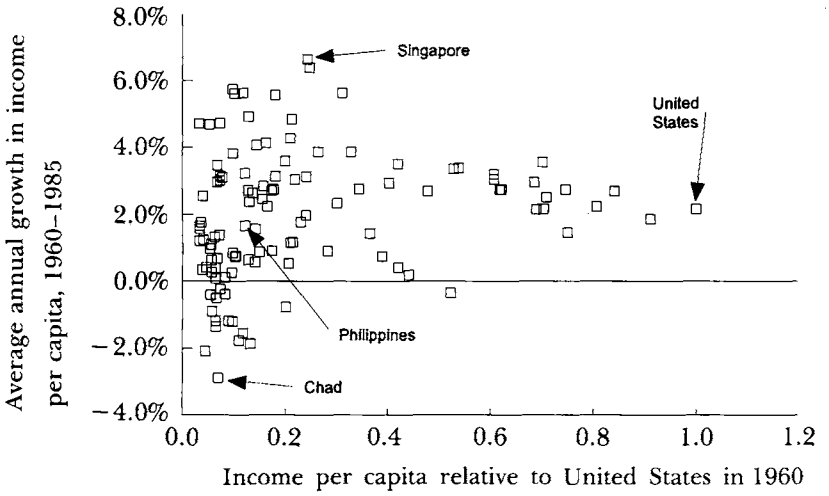
As a result, attention then shifted to the broad sample of countries in the Heston-Summers data set. As Figure 1 shows, convergence clearly fails in this broad sample of countries. Income per capita in 1960 is plotted on the horizontal axis. The average annual rate of growth of income per capita from 1960 to 1985 is plotted on the vertical axis.[1] On average, poor countries in this sample grow no faster than the rich countries.

Figure 1 poses one of the central questions in development. Why is it that the poor countries as a group are not catching up with the rich countries in the same way that, for example, the low income states in the United States have been catching up with the high income states? Both Robert Lucas (1988) and I (Romer, 1986) cited the failure of cross-country convergence to motivate models of growth that drop the two central assumptions of the neoclassical model: that technological change is exogenous and that the same technological opportunities are available in all countries of the world.

To see why Figure 1 poses a problem for the conventional analysis, consider a very simple version of the neoclassical model. Let output take the simple Cobb-Douglas form $Y = A(t)K^{1-\beta}L^{\beta}$. In this expression, Y denotes net national product, K denotes the stock of capital, L denotes the stock of labor, and A denotes the level of technology. The notation indicating that A is a function of time signals the standard assumption in neoclassical or exogenous growth models: the technology improves for reasons that are outside the model. Assume that a constant fraction of net output, s, is saved by consumers each year. Because the model assumes a closed economy, s is also the ratio of net investment to net national product. Because we are working with net

[1]The data here are taken from version IV of the Penn World Table. The income measure is RGDP2. See Summers and Heston (1988) for details.

Figure 1
Testing for Convergence



(rather than gross) national product and investment, sY is the rate of growth of the capital stock. Let $y = Y/L$ denote output per worker and let $k = K/L$ denote capital per worker. Let n denote the rate of growth of the labor force. Finally, let a "^" over a variable denote its exponential rate of growth. Then the behavior of the economy can be summarized by the following equation:

$$\hat{y} = (1-\beta)\hat{k} + \hat{A}$$

$$= (1-\beta)\left[sA(t)^{1/(1-\beta)}y^{(-\beta)/(1-\beta)} - n\right] + \hat{A}$$

The first line in this equation follows by dividing total output by the stock of labor and then calculating rates of growth. This expression specifies the procedure from growth accounting for calculating the technology residual. Calculate the growth in output per worker, then subtract the rate of growth of the capital-labor ratio times the share of capital income in total income from the rate of growth of output per worker. The second line follows by substituting in an expression for the rate of growth of the stock of capital per worker, as a function of the savings rate s, the growth rate of the labor force n, the level of the technology $A(t)$, and the level of output per worker, y.

Outside of the steady state, the second line of the equation shows how variation in the investment rate and in the level of output per worker should translate into variation in the rate of growth. The key parameter is the exponent β on labor in the Cobb-Douglas expression for output. Under the neoclassical assumption that the economy is characterized by perfect competition, β is equal to the share of total income that is paid as compensation to

labor, a number that can be calculated directly from the national income accounts. In the sample as a whole, a reasonable benchmark for β is 0.6. (In industrialized economies, it tends to be somewhat larger.) This means that in the second line of the equation, the exponent $(-\beta)/(1-\beta)$ on the level of output per worker y should be on the order of about -1.5.

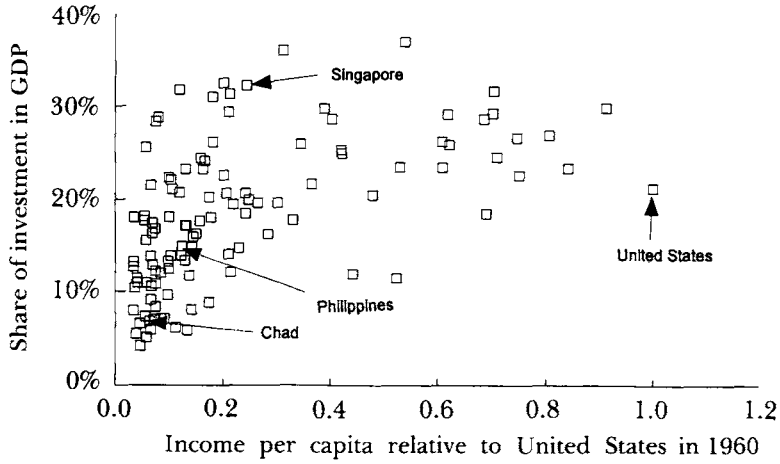
We can now perform the following calculation. Pick a country like the Philippines that had output per worker in 1960 that was equal to about 10 percent of output per worker in the United States. Because $0.1^{-1.5}$ is equal to about 30, the equation suggests that the United States would have required a savings rate that is about 30 times larger than the savings rate in the Philippines for these two countries to have grown at the same rate. If we use 2/3 instead of .6 as the estimate of β, the required savings rate in the United States would be 100 times larger than the savings rate in the Philippines. The evidence shows that these predicted saving rates for the United States are orders of magnitude too large.

A key assumption in this calculation is that the level of the technology $A(t)$ is the same in the Philippines and the United States. (The possibility that $A(t)$ might differ is considered below.) If they have the same technology, the only way to explain why workers in the Philippines were only 10 percent as productive as workers in the United States is to assume that they work with about $0.1^{1/(1-\beta)}$ or between 0.3 percent and 0.1 percent as much capital per worker. Because the marginal product of capital depends on the capital stock raised to the power $-\beta$, the marginal product of an additional unit of capital is $0.1^{-\beta/(1-\beta)}$ times larger in the Philippines than it is in the United States, so a correspondingly higher rate of investment is needed in the United States to get the same effect on output.

Figure 2 plots the level of per capita income against the ratio of gross investment to gross domestic product for the Heston-Summers sample of countries. The correlation in this figure at least has the correct sign to explain why poor countries on average are not growing faster than the rich countries—that is, a higher level of income is associated with a higher investment rate. But if β is between 0.6 and 0.7, the variation in investment between rich and poor countries is at least an order of magnitude too small to explain why the rich and poor countries seem to grow at about the same rate. In concrete terms, the share of investment in the United States is not 30 or 100 times the share in the Philippines. At most, it is twice as large.

Of course, the data in Figures 1 and 2 are not exactly what the theory calls for, but the differences are not likely to help resolve the problem here. For example, the display equation depends on the net investment rate instead of the gross investment rate. Because we do not have reliable data on depreciation for this sample of countries, it is not possible to construct a net investment ratio. A reasonable conjecture, however, is that depreciation accounts for a larger share of GDP in rich countries than it does in poor countries, so the difference between the net investment rate in rich and poor countries will be even smaller

Figure 2
Per Capita Income and Investment



than the difference between the gross investment rates illustrated in the figure. The display equation also calls for output per worker rather than output per capita, but for a back-of-the-envelope calculation, variation in income per capita should be close enough to variation in output per worker to show that a simple version of the neoclassical model will have trouble fitting the facts.

The way to reconcile the data with the theory is to reduce β so that labor is relatively less important in production and diminishing returns to capital accumulation set in more slowly. The theoretical challenge in constructing a formal model with a smaller value for β lies in justifying why labor is paid more than its marginal product and capital is paid less. To explain these divergences between private and social returns, I proposed a model in which A was determined locally by knowledge spillovers (Romer, 1987a). I followed Arrow's (1962) treatment of knowledge spillovers from capital investment and assumed that each unit of capital investment not only increases the stock of physical capital but also increases the level of the technology for all firms in the economy through knowledge spillovers. I also assumed that an increase in the total supply of labor causes negative spillover effects because it reduces the incentives for firms to discover and implement labor-saving innovations that also have positive spillover effects on production throughout the economy.

This leads to a functional relationship between the technology in a country and the other variables that can be written as $A(K, L)$. Then output for firm j can be written as $Y_j = A(K, L)K_j^{1-\alpha}L_j^{\alpha}$, where variables with subscripts are ones that firm j can control, and variables without subscripts represent economy-wide totals. Because the effect that a change in a firm's choice of K or L has on A is an external effect that any individual firm can ignore, the exponent α measures the private effect of an increase in employment on output. A

1 percent increase in the labor used by a firm leads to an α percent increase in its output. As a result, α will be equal to the fraction of output that is paid as compensation to labor. Suppose, purely for simplicity, that the expression linking the stock of A to K and L takes the form $A(K, L) = K^{\gamma}L^{-\gamma}$ for some γ greater than zero. Then the reduced form expression for aggregate output as a function of K and L would be $Y = K^{1-\beta}L^{\beta}$ where β is equal to $\alpha - \gamma$. This exponent β represents the aggregate effect of an increase in employment. It captures both the private effect α and the external effect $-\gamma$. In the calculation leading up to the equation displayed above, it is this aggregate or social effect that matters. According to this model, β can now be smaller than labor's share in national income.

Using a simple cross-country regression based on an equation like the display equation, I found that the effect of the investment rate on growth was positive and the effect of initial income on growth was negative. Many other investigators have found this kind of negative coefficient on initial income in a growth regression. This result has received special attention, particularly in light of the failure of overall convergence exhibited in Figure 1. It suggests that convergence or regression to the mean would have taken place if all other variables had been held constant.

After imposing the constraint implied by the equation, I estimated the value of β to be in the vicinity of 0.25 (Romer, 1987a, Table 4). With this value, it would only take a doubling of the investment rate—rather than a 30- or 100-fold increase—to offset the negative effect that a ten-fold increase in the level of output per worker would have on the rate of growth. These figures are roughly consistent with the numbers for the United States and the Philippines. For the sample as a whole, the small negative effect on growth implied by higher levels of output per worker are offset by higher investment rates in richer countries.

Robert Barro and Xavier Sala i Martin (1992) subsequently showed that the conclusions about the size of what I am calling β (they use different notation) were the same whether one looked across countries or between states in the United States. They find that a value for β on the order of 0.2 is required to reconcile the convergence dynamics of the states with the equation presented above. Convergence takes place, but at a very slow rate. They also observe that this slow rate of convergence would be even harder to explain if one introduced capital mobility into the model.

As a possible explanation of the slow rate of convergence, Barro and Sala i Martin (1992) propose an alternative to the neoclassical model that is somewhat less radical than the spillover model that I proposed. As in the endogenous growth models, they suggest that the level of the technology $A(t)$ can be different in different states or countries and try to model its dynamics. They take the initial distribution of differences in $A(t)$ as given by history and suggest that knowledge about A diffuses slowly from high A to low A regions. This would mean that across the states, there is underlying variation in $A(t)$ that

causes variation in both k and y. As a result, differences in output per worker do not necessarily signal large differences in the marginal product of capital. In fact, free mobility of capital can be allowed in this model and the rate of return on capital can be equalized between the different regions. Because the flow of knowledge from the technology leader makes the technology grow faster in the follower country, income per capita will grow faster in the follower as diffusion closes what has been called a technology gap.[2] The speed of convergence will be determined primarily by the rate of diffusion of knowledge, so the convergence dynamics tell us nothing about the exponents on capital and labor.

The assumption that the level of technology can be different in different regions is particularly attractive in the context of an analysis of the state data, because it removes the prediction of the closed-economy, identical-technology neoclassical model that the marginal productivity of capital can be many times larger in poorer regions than in rich regions.[3] According to the data reported by Barro and Sala i Martin (1992), in 1880, income per capita in states such as North Carolina, South Carolina, Virginia, and Georgia was about one-third of income per capita in industrial states such as New York, Massachusetts, and Rhode Island. If β is equal to 0.6, $-\beta/(1-\beta)$ is equal to -1.5 and $(1/3)^{-1.5}$ is equal to about 5. This means that the marginal product of capital should have been about five times higher in the South than it was in New England. It is difficult to imagine barriers to flows of capital between the states that could have kept these differences from rapidly being arbitraged away. In particular, it would be difficult to understand why any capital investment at all took place in New England after 1880. But if there were important differences in the technology in use in the two regions, the South may not have offered higher returns to capital investment.

In a third approach to the analysis of cross country data, Greg Mankiw, David Romer, and David Weil (1992) took the most conservative path, showing that it is possible to justify a low value for β even in a pure version of the closed economy, neoclassical model which assumes that the level of technology is the same in each country in the world. The only change they make is to extend the usual two-factor neoclassical model by allowing for human capital H as well as physical capital K. They use the fraction of the working age population that attends secondary school as a measure of the rate of investment in human capital that is analogous to the share of physical capital investment in total GDP.

They conclude from their cross-country growth regressions that $Y = A(t)K^{1/3}H^{1/3}L^{1/3}$ is a reasonable specification for aggregate output. In this

[2]Nelson and Phelps (1966) give a theoretical model that allows for diffusion of the technology between countries. Fagerberg (1987) interprets cross-country growth regressions in the context of a technology gap model instead of a neoclassical model or a spillover model. For further discussion of diffusion, see also Barro and Sala i Martin (forthcoming 1994) and Jovanovic and Lach (1993).

[3]See King and Rebelo (1993) for a fuller discussion of both the price and quantity implications of the neoclassical model.

model, the exponent β on the fixed factor of production L has been reduced from 0.6 to 0.33. This lower value of β is consistent with the data on income shares because total wage payments consist of payments to both human capital and unskilled labor. If K and H vary together across countries, this specification implies that it takes about a three-fold increase in investment (an increase by the factor $0.1^{-0.5}$ to be precise) to offset a 10-fold increase in output per worker in a comparison across nations. Once one takes account of variation in investment in schooling as well as in investment in physical capital, a factor of three is roughly consistent with the variation in total investment rates observed in the Summer-Heston sample of countries.

Although Mankiw, Romer and Weil do not examine the state data, it is clear what their style of explanation would suggest. They would assume that the same technology was available in the North and the South. Suppose that Northern states had levels of both human capital and physical capital that were higher than those in the Southern states in the same ratio. A value of β equal to 1/3, together with the fact that output per capita was about one-third as large in the South in 1880, would imply that rate of return on physical capital and the wage for human capital were both about $(1/3)^{-0.5}$ (or about 1.7) times higher in the Southern states than they were in the New England states. Compared to the factor of 5 implied by the model without human capital, these parameters would imply much smaller incentives to shift all capital investment to the South. (They would imply, however, that human capital would tend to migrate from the North to the South.)

The implication from this work is that if you are committed to the neoclassical mode, the kind of data exhibited in Figures 1 and 2 cannot be used to make you recant. They do not compel you to give up the convenience of a model in which markets are perfect. They cannot force you to address the complicated issues that arise in the economic analysis of the production and diffusion of technology, knowledge, and information.

An Evaluation of the Convergence Controversy

The version of the development of endogenous growth theory outlined above skips lots of detail and smooths over many complications that made this seem like a real controversy at the time. In retrospect, what is striking is how little disagreement there is about the basic facts. Everyone agrees that a conventional neoclassical model with an exponent of about one-third on capital and about two-thirds on labor cannot fit the cross-country or cross-state data. Everyone agrees that the marginal product of investment cannot be orders of magnitudes smaller in rich countries than in poor countries. The differences between the different researchers concern the inferences about models that we should draw from these facts. As is usually the case in macroeconomics, many different inferences are consistent with the same regression statistics.

This history has many elements in common with other stories about the development of economics. The story starts with the emergence of new data. These present anomalies that lead to new theoretical models, some of which differ markedly from previous, well-accepted models. Then a more conservative interpretation emerges that accommodates the new evidence and preserves much of the structure of the old body of theory. In the end, we have refined the set of alternatives somewhat, but seem to be left in about the same position where we started, with too many theories that are consistent with the same small number of facts.

But economists who accept this interpretation come to the conclusion that we do not have enough data only because they restrict attention to the kind of statistical evidence illustrated in Figures 1 and 2. They fail to take account of all the other kinds of evidence that are available. My original work on growth (Romer, 1983; 1986) was motivated primarily by the observation that in the broad sweep of history, classical economists like Malthus and Ricardo came to conclusions that were completely wrong about prospects for growth. Over time, growth rates have been increasing, not decreasing.[4] Lucas (1988) emphasized the fact that international patterns of migration and wage differentials are very difficult to reconcile with a neoclassical model. If the same technology were available in all countries, human capital would not move from places where it is scarce to places where it is abundant and the same worker would not earn a higher wage after moving from the Philippines to the United States.

The main message of this paper is that the convergence controversy captures only part of what endogenous growth has been all about. It may encompass a large fraction of the recently published papers, but it nevertheless represents a digression from the main story behind endogenous growth theory. The story told about the convergence controversy also tends to reinforce a message that I think is seriously misleading—that data are the only scarce resource in economic analysis.

Version #2: The Passing of Perfect Competition

The second version of the origins of endogenous growth starts from the observation that we had enough evidence to reject all the available growth models throughout the 1950s, 1960s, and 1970s. What we lacked were good aggregate-level models. This version of the origins of endogenous growth is therefore concerned with the painfully slow progress we have made in constructing formal economic models at the aggregate level. It suggests that progress in economics does not come merely from the mechanical application of hypothesis tests to data sets. There is a creative act associated with the construction of new models that is also crucial to the process.

[4]See Kremer (1993) for a stimulating look at this question from a very long-run point of view.

The evidence about growth that economists have long taken for granted and that poses a challenge for growth theorists can be distilled to five basic facts.

Fact #1: There are many firms in a market economy. The fact is so obvious that we often do not bother to state it, but it clearly will not do to have a model in which there are overwhelming forces that tend to concentrate all output in the hands of a single, economy-wide monopolist.

Fact #2: Discoveries differ from other inputs in the sense that many people can use them at the same time. The idea behind the transistor, the principles behind internal combustion, the organizational structure of a modern corporation, the concepts of double entry bookkeeping—all these pieces of information and many more like them have the property that it is technologically possible for everybody and every firm to make use of them at the same time. In the language of public finance, ordinary goods are rival goods, but information is nonrival.

Fact #3: It is possible to replicate physical activities. Replication implies that the aggregate production function representing a competitive market should be characterized by homogeneity of degree one in all of its conventional (that is, rival) inputs. If we represent output in the form $Y = AF(K, H, L)$, then doubling all three of K, H, and L should allow a doubling of output. There is no need to double the nonrival inputs represented by A because the existing pieces of information can be used in both instances of the productive activity at the same time. (The assumption that the market is competitive means that the existing activity already operates at the minimum efficient scale, so there are no economies of scale from building a single plant that is twice as large as the existing one.)

If farming were the relevant activity instead of manufacturing, we would clearly need to include land as an input in production, and in the economy as a whole, it is not possible to double the stock of land. This does not change the fundamental implication of the replication argument. If aggregate output is homogeneous of degree 1 in the rival inputs and firms are price takers, Euler's theorem implies that the compensation paid to the rival inputs must exactly equal the value of output produced. This fact is part of what makes the neoclassical model so simple and makes growth accounting work. The only problem is that this leaves nothing to compensate any inputs that were used to produce the discoveries that lead to increases in A.

Fact #4: Technological advance comes from things that people do. No economist, so far as I know, has ever been willing to make a serious defense of the proposition that technological change is literally a function of elapsed calendar time. Being explicit about the issues here is important nevertheless, because it can help untangle a link that is sometimes made between exogeneity and randomness. If I am prospecting for gold or looking for a change in the DNA of a bacterium that will let it eat the oil from an oil spill, success for me will be dominated by chance. Discovery will seem to be an exogenous event in the

sense that forces outside of my control seem to determine whether I succeed. But the aggregate rate of discovery is endogenous. When more people start prospecting for gold or experimenting with bacteria, more valuable discoveries will be found. This will be true even if discoveries are accidental side effects of some other activity (finding gold as a side effect of ditch-digging) or if market incentives play no role in encouraging the activity (as when discoveries about basic molecular biology were induced by government research grants). The aggregate rate of discovery is still determined by things that people do.

Fact #5: Many individuals and firms have market power and earn monopoly rents on discoveries. Even though the information from discoveries is nonrival (as noted in fact 2), economically important discoveries usually do not meet the other criterion for a public good; they typically are partially excludable, or excludable for at least some period of time. Because people and firms have some control over the information produced by most discoveries, it cannot be treated as a pure public good. This information is not like a short-wave radio broadcast that everyone can access without the permission of the sender. But if a firm can control access to a discovery, it can charge a price that is higher than zero. It therefore earns monopoly profits because information has no opportunity cost.

The neoclassical model that was developed and applied by Robert Solow (1956, 1967) and others constituted a giant first step forward in the process of constructing a formal model of growth. The discussion of the convergence controversy, framed as it was almost entirely in terms of the neoclassical model, illustrates the model's power and durability. Like any model, the neoclassical model is a compromise between what we would like from a model and what is feasible given the state of our modeling skills. The neoclassical model captured facts 1, 2, and 3, but postponed consideration of facts 4 and 5. From a theoretical point of view, a key advantage of the model is its treatment of technology as a pure public good. This makes it possible to accommodate fact 2 —that knowledge is a nonrival good—in a model that retains the simplicity of perfect competition. The public good assumption also implies that knowledge is nonexcludable, and this is clearly inconsistent with the evidence summarized in fact 5—that individuals and firms earn profits from their discoveries. This assumption was useful, nevertheless, as part of an interim modeling strategy that was adopted until models with nonrivalry and excludability could be formulated.

Endogenous growth models try to take the next step and accommodate fact 4. Work in this direction started in the 1960s. For example, Karl Shell (1966) made the point about replication noted above, showing that it left no resources to pay for increases in A. He proposed a model in which A is financed from tax revenue collected by the government. Recent endogenous growth models have tended to follow Arrow (1962) and emphasize the private sector activities that contribute to technological advance rather than public sector funding for research. A subset of these models has tried to incorporate

both fact 4 (that technological advance comes from things people do) and fact 5 (the existence of monopoly rents). These are sometimes referred to as neo-Schumpeterian models because of Schumpeter's emphasis of the importance of temporary monopoly power as a motivating force in the innovative process.[5] In addition, there are two other distinct kinds of endogenous growth models. Spillover models have already been mentioned. Linear models will be described below.[6]

With the benefit of hindsight, it is obvious that growth theorists would eventually have to do what economists working at the industry and firm level have done: abandon the assumption of price-taking competition. Otherwise, there is no hope of capturing fact 5. Even at the time, the point received at least some attention. In his 1956 paper, Solow remarked in a footnote on the desirability of extending the model to allow for monopolistic competition. One of his students, William Nordhaus (1969), subsequently outlined a growth model that did have patents, monopoly power, and many firms. For technical reasons, this model still invoked exogenous technological change, so it is not strictly speaking a model of endogenous growth—but it could have been extended to become one. Because a general formal treatment of monopolistic competition was not available at the time, little progress in this direction took place for the next 20 years.

Even though it is obvious in retrospect that endogenous growth theory would have to introduce imperfect competition, this was not the direction that the first models of the 1980s pursued. Both my model (1986) and Robert Lucas's model (1988) included fact 4 without taking the final step and including step 5. In both of these models, the technology is endogenously provided as a side effect of private investment decisions. From the point of view of the users of technology, it is still treated as a pure public good, just as it is in the neoclassical model. As a result, firms can be treated as price takers and an equilibrium with many firms can exist.

This technique for introducing a form of aggregate increasing returns into a model with many firms was first proposed by Alfred Marshall (1890). To overturn the pessimistic predictions of Malthus and Ricardo, he wanted to introduce some form of aggregate increasing returns. To derive his downward sloping supply curve from an industry with many firms, Marshall introduced the new notion of increasing returns that were external to any individual firm. External effects therefore entered into economics to preserve the analytical

[5]Of course, Stigler's law applies in this case: The person that any result is named after was not the first person to derive or state the result. It just helps to have a label so that you can keep track of the players without a scorecard.

[6]Richard Nelson and Sidney Winter (1982) developed an alternative evolutionary model of growth. Their verbal, descriptive style of theory, which they label appreciative theory, was flexible enough to accommodate facts 1–5. This style of work can be thought of as a complement to formal theory, not a substitute for it. It leaves open the problem of constructing a formal theory that could accommodate these facts.

machinery of supply and demand curves and price taking in the presence of increasing returns. The analysis of other kinds of external effects—smoke, bees, and so on—came later.[7]

As noted in the previous discussion of spillover models, Arrow (1962) constructed a model along these lines. In a simplified form, output for firm j in his model can be written as $Y_j = A(K)F(K_j, L_j)$, where as before, K without a subscript denotes the aggregate stock of capital. For technical reasons, Arrow, like Nordhaus, did not emphasize the fact that his model could lead to sustained, endogenous growth. For the parameter values that he studies, if the size of the population is held constant, growth eventually comes to a halt.

Lucas's model has a very similar underlying structure. There, it is investments in human capital rather than physical capital that have spillover effects that increase the level of the technology. It is as if output for firm j takes the form $Y_j = A(H)F(K_j, H_j)$. Both of these models accommodated facts 1–4 but not fact 5.[8]

In my first paper on growth (Romer, 1986), I assumed in effect that aggregate output could be written as $Y = A(R)F(R_j, K_j, L_j)$ where R_j stands for the stock of results from expenditure on research and development by firm j.[9] I assumed that it is spillovers from private research efforts that lead to improvements in the public stock of knowledge A. This seemed appealing because it recognized that firms did research and development on purpose and that the relevant spillovers or incomplete property rights were associated with the results from research and development. (In the microeconomic analysis of research and development at the industry level, Zvi Griliches (1979) used this same kind of formulation.) But to make this model fit within the framework of price-taking with no monopoly power, I assumed that the function F was homogeneous of degree one in all of its inputs, including R. This, unfortunately, violates fact 2, that research is a nonrival good and fact 3, that only rival goods need to be replicated to double output. If I had admitted that R_j was nonrival, the replication argument would have implied that the firm faced increasing returns in the inputs R_j, K_j, and L_j that it controlled, because output would double merely by replicating K_j and L_j.

My sleight of hand in treating R_j as a rival good and making F homogeneous of degree 1 in all three of K, L, and R may seem like a trifling matter in an area of theory that depends on so many other short cuts. After all, if one is

[7]For an explicit treatment showing that Marshallian external increasing returns is ultimately an untenable way to model any process involving learning or knowledge, see Dasgupta and Stiglitz (1988).

[8]Lucas actually makes A depend on per capita H rather than total H. The difference between these two formulations is not relevant for the discussion here, but is important for some of the other implications of the model.

[9]For consistency with the rest of the discussion, I distinguish here between R and K. In the paper, I actually dropped physical capital from consideration so that I have only one state variable to deal with. This leads to a potential confusion because I also used the symbol K for knowledge instead of R.

going to do violence to the complexity of economic activity by assuming that there is an aggregate production function, how much more harm can it do to be sloppy about the difference between rival and nonrival goods? Unfortunately, quite a bit. The distinctions between rival and nonrival inputs, and the distinction between excludable and nonexcludable goods, are of absolutely fundamental importance in modeling and in policy formulation.

For years, the economic analysis of science and technology policy consisted of little more than a syllogism. The major premise was that the government should provide public goods and the private sector should provide private goods. The minor premise was that basic research is a public good and applied research is a private good. Once you think carefully about nonrivalry and excludability, it is clear that the major premise is misleading because it understates the possible role for collective action. Governments can usefully provide goods that are nonrival but are not true public goods, because they are potentially excludable. The minor premise is simply wrong. Applied research is not an ordinary private good. Discussion in policy circles is now taking place using new terms—critical technologies, generic research, and pre-competitive research—that are only vaguely defined but that take the discussion outside of the simple dichotomy between public goods and private goods. This is probably useful, but it would lend needed structure to this discussion if participants paid more attention to the distinction between the two different aspects of publicness (nonrivalry and nonexcludability) and looked more formally at the different kinds of policy challenges that nonrivalry and nonexcludability present.

The linear model branch of endogenous growth theory pursued even more aggressively the strategy I used.[10] If I could treat the part of knowledge that firms control as an ordinary input in production—that is, as an input that is rival and hence is not associated with increasing returns—why bother to allow for any nonrival inputs at all? In effect, these models assumed that output could be written as $Y = F(R, K, H)$ for a homogeneous of degree 1 production function F. These models assumed that research R, physical capital K, and human capital H were like ordinary inputs. If there are no nonrival goods, there are no increasing returns. It is then a relatively simple matter to build a perfectly competitive model of growth. To simplify still further, these models often aggregate R, K, and H into a single broad measure of capital. Suppose we call it X. Then we can write $F(X)$ as a linear function: $Y = F(X) = aX$, hence the name linear models. If we assume that a constant fraction of output Y is saved and used to produce more X, the model generates persistent, endogenous growth. Relative to the neoclassical model, these models capture fact 4—that technological change comes from investments that people make—at the cost of abandoning fact 2, that technology or knowledge is a nonrival good.

Proponents of the linear model and the neoclassical model have sometimes been drawn into pointless arguments about which model is worse. Proponents

[10]One of the early linear models was Uzawa (1965). Important recent papers in this line of work include Becker, Murphy, and Tamura (1990), Jones and Manuelli (1990), and Rebelo (1991).

of the linear growth models point out that the neoclassical model fails to capture fact 4. Proponents of the neoclassical model observe that the linear model cannot capture fact 2. This dispute is partly an outgrowth of the convergence controversy. Both sides specify that output takes the form $Y = K^{1-\beta}L^{\beta}$ and then argue about whether β is bigger than zero (as the proponents of the neoclassical model claim) or close to zero (as some versions of the linear growth model suggest).

This is not a very useful debate. There are circumstances in which each model can be a useful expositional device for highlighting different aspects of the growth process, but presumably the agenda for the profession ought to be to capture both facts 2 and 4 and pick up fact 5 to boot.

Neo-Schumpeterian Growth

Two steps were required for the neo-Schumpeterian models of growth to emerge. The first was that after struggling for years to preserve perfect competition, or at least price-taking in the presence of external effects, growth theorists had to decide to let go. It helped that economists working on industrial organization had given them something else to hang onto. By the late 1970s, there were aggregate models with many firms (fact 1), each of which could have market power (fact 5). The most convenient such model was developed by Avinash Dixit and Joseph Stiglitz (1977). William Ethier (1982) subsequently showed how their model of preferences over many goods could be interpreted as a production function that depended on a large number of inputs in production.

Once people who were interested in growth recognized that this approach offered the alternative to a competitive market structure, there was only one technical detail that remained to be resolved, the detail that had kept both Nordhaus and Arrow from producing models of endogenous growth. All models of growth need at least one equation which describes the evolution of something like $A(t)$.[11] This equation usually takes the form

$$\dot{A} = -A^{\phi}, \tag{2}$$

where A with a dot denotes the time derivative of A. Models that produce steady state growth fill in the blank with a constant and set the exponent ϕ equal to 1. For example, if we set ϕ equal to 1 and insert a constant g in the blank, we have the driving equation behind the neoclassical model with exogenous technological change.

Mathematically, this kind of formulation is not robust. If ϕ turns out to be even slightly greater than 1, the equation implies that the stock of technology will go to infinity in finite time. When we use this same kind of model to study population growth, this lack of robustness does not raise any particular

[11]Sometimes other variables like H or K are used in place of A, but the basic issues are the same.

difficulties. We understand that functional forms are always approximations, and that a linear differential equation leading to exponential growth is a particularly convenient approximation. But Nordhaus and Arrow both worked at a time when there was real concern about the knife-edge character of the assumptions about ϕ.[12] If it was less than one, growth eventually stopped. If it was even slightly greater than one, everything blows up. As a result, economists stayed well away from the edge and assumed that ϕ had to be strictly less than 1. In a model like Nordhaus's, growth can be kept going only by adding a second kind of knowledge A_2 that grows exogenously. (Formally, bringing in exogenous technological change amounts to bringing in a new equation in which the exponent corresponding to ϕ has already been set to 1, and it only takes one equation with this property to keep things going.)

I devoted a great deal of attention to this robustness problem in my analysis of the spillover models. I modified other functional forms elsewhere in the model to construct robust models of endogenous growth in which the level of output and its rate of growth stayed finite for all time for a range of values of ϕ that were strictly bigger than 1 (Romer, 1983; 1986). For values slightly less than 1, growth eventually stopped but could persist, nevertheless, for a very long time. The mathematical analysis in this more complicated robust model was much harder than the analysis that is possible when ϕ is equal to 1. The difference between the two models is the difference between studying the phase plane of a nonlinear differential equation system and solving a simple linear differential equation. Once it is clear that we could build a complicated model that is robust, there is every reason to work with the simple special case whenever possible.

By the late 1980s, economists like Kenneth Judd (1985) and Gene Grossman and Elhanan Helpman (1989) were working out models of growth with monopolistic competition. Like Nordhaus and Arrow, they stayed well away from the case where ϕ was equal to 1. Judd invoked exogenous technological change to keep his economy growing. Grossman and Helpman were investigating the connection between trade and growth, and settled for an analysis of transitional dynamics of the model as it converged to a steady state level of income where growth stopped. In each model, monopoly profits motivate discovery.

I took what I had learned about generating sustained growth from my analysis of spillover models and applied it to the monopolistic competition model. I constructed two very simple models of sustained growth that accommodated all five of the facts cited above. One of these did not invoke any spillover effects at all (Romer, 1987b). The other combined both monopoly power and spillovers—that is, incomplete intellectual property rights (Romer, 1990). In each of these models I set the analog of ϕ equal to 1. I knew that by

[12]See Stiglitz (1990) for a discussion of how people working on growth at the time perceived this problem.

repeating my analysis of the spillover model, it would be possible to construct more complicated robust models with the same qualitative implications.

Research on endogenous growth models in which monopoly profits motivate innovation has progressed rapidly since then and has uncovered a number of unexpected connections between market size, international trade, and growth, as the article by Grossman and Helpman in this symposium explains.

Conclusions

The economics profession is undergoing a substantial change in how we think about international trade, development, economic growth and economic geography.[13] In each of these areas, we have gone through a progression that starts with models based on perfect competition, moves to price-taking with external increasing returns, and finishes with explicit models of imperfect competition. It is likely that this pattern will repeat itself in other areas like the theory of macroeconomic fluctuations.

The effects of this general trend may be far-reaching. Ultimately, it may force economists to reconsider some of the most basic propositions in economics. For example, I am convinced that both markets and free trade are good, but the traditional answer that we give to students to explain why they are good, the one based on perfect competition and Pareto optimality, is becoming untenable. Something more interesting and more complicated is going on here.[14]

In each of the areas where our understanding has changed, evidence that challenged the models of perfect competition and supported the models with imperfect competition had been apparent all along. Everyone knew that there was lots of intra-industry trade between developed nations and little trade between the North and the South. Everyone knew that some developing countries grew spectacularly while others languished. Everyone knew that people do the things that lead to technological change. Everyone knew that the number of locally available goods was limited by the extent of the market in the city where someone lives and works.

In evaluating different models of growth, I have found that Lucas's (1988) observation, that people with human capital migrate from places where it is scarce to place where it is abundant, is as powerful a piece of evidence as all the cross-country growth regressions combined. But this kind of fact, like the fact about intra-industry trade or the fact that people make discoveries, does not come with an attached *t*-statistic. As a result, these kinds of facts tend to be

[13]Paul Krugman has made influential contributions in all of these areas. See Krugman (1990, 1991, 1993) for a discussion of the changes in these fields.

[14]Romer (forthcoming) offers a demonstration that, for example, the costs of trade restrictions in a developing country can be far greater in the context of a model with imperfect competition than they are in a model with perfect competition.

neglected in discussions that focus too narrowly on testing and rejecting models.

Economists often complain that we do not have enough data to differentiate between the available theories, but what constitutes relevant data is itself endogenous. If we set our standards for what constitutes relevant evidence too high and pose our tests too narrowly, we will indeed end up with too little data. We can thereby enshrine the economic orthodoxy and make it invulnerable to challenge.[15] If we do not have any models that can fit the data, the temptation will be to set very high standards for admissible evidence, because we would prefer not to reject the only models that we have.

When I look back on my work on growth, my greatest satisfaction comes from having rejected the first round of external effects models that I tried. I am glad that I was able to learn something about robustness and nonrivalry from struggling with these models, but was still able to let go when a better alternative became apparent. My greatest regret is the shift I made while working on these external effects models, a shift that took me away from the emphasis on research and knowledge that characterized my 1986 paper and toward the emphasis on physical capital that characterized the empirical work in the paper cited in the discussion of convergence (1987a). This paper contributed to the convergence controversy and to an emphasis on the exponents on capital and labor in aggregate production. I am now critical of this work, and I accept part of the blame. Looking back, I suspect that I made this shift toward capital and away from knowledge partly in an attempt to conform to the norms of what constituted convincing empirical work in macroeconomics. No international agency publishes data series on the local production of knowledge and inward flows of knowledge. If you want to run regressions, investment in physical capital is a variable that you can use, so use it I did. I wish I had stuck to my guns about the importance of evidence like that contained in facts 1 through 5.

If macroeconomists look only at the cross-country regressions deployed in the convergence controversy, it will be easy to be satisfied with neoclassical models in which market incentives and government policies have no effect on discovery, diffusion, and technological advance. But if we make use of all of the available evidence, economists can move beyond these models and begin once again to make progress toward a complete understanding of the determinants of long-run economic success. Ultimately, this will put us in position to offer policy-makers something more insightful than the standard neoclassical prescription—more saving and more schooling. We will be able to rejoin the ongoing policy debates about tax subsidies for private research, antitrust exemptions for research joint ventures, the activities of multinational firms, the

[15] In their discussion of real business cycle theories and the kind of evidence used to test them, Greg Mankiw (1989) and Robert Solow (1988) have both made a similar point about explicit statistical versus broader kinds of evidence.

effects of government procurement, the feedback between trade policy and innovation, the scope of protection for intellectual property rights, the links between private firms and universities, the mechanisms for selecting the research areas that receive public support, and the costs and benefits of an explicit government-led technology policy. We will be able to address the most important policy questions about growth: In a developing country like the Philippines, what are the best institutional arrangements for gaining access to the knowledge that already exists in the rest of the world? In a country like the United States, what are the best institutional arrangements for encouraging the production and use of new knowledge?

■ *I have benefitted from comments by Jeffrey Frankel, Alan Krueger, David Romer, Carl Shapiro, and Timothy Taylor on early drafts of this paper. This work was supported by NSF Grant #SES 9023469 and by the Canadian Institute for Advanced Research.*

References

Abramovitz, Moses, "Catching Up, Forging Ahead, and Falling Behind," *Journal of Economic History*, June 1986, *46*:2, 385–406.

Arrow, Kenneth J., "The Economic Implications of Learning by Doing," *Review of Economic Studies*, June 1962, *29*, 155–73.

Barro, Robert J., and Xavier Sala i Martin, "Convergence," *Journal of Political Economy*, April 1992, *100*:2, 223–51.

Barro, Robert J., and Xavier Sala i Martin, "Chapter 8: Diffusion of Technology." In Barro, R. J., and X. Sala-i-Martin, eds., *Economic Growth*. New York: McGraw Hill, forthcoming 1994.

Baumol, William J., "Productivity Growth, Convergence, and Welfare: What the Long-run Data Show," *American Economic Review*, December 1986, *76*:5, 1072–85.

Becker, G., K. Murphy, and R. Tamura, "Economic Growth, Human Capital, and Population Growth," *Journal of Political Economy*, October 1990, *98*:5 Part 2, S12–S137.

Dasgupta, P., and J. Stiglitz, "Learning-by-Doing, Market Structure, and Industrial and Trade Policies," *Oxford Economic Papers*, June 1988, *40*:2, 246–68.

De Long, J. Bradford, "Productivity Growth, Convergence and Welfare: Comment," *American Economic Review*, December 1988, *78*:5, 1138–54.

Dixit, A., and J. Stiglitz, "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, June 1977, *67*:3, 297–308.

Ethier, W. J., "National and International Returns to Scale in the Modern Theory of International Trade," *American Economic Review*, June 1982, *72*:3, 389–405.

Fagerberg, Jan, "A Technology Gap Approach to Why Growth Rates Differ," *Research Policy*, 1987, *16*, 87–99.

Griliches, Zvi, "Issues in Assessing the Contribution of Research and Development to Productivity Growth," *Bell Journal of Economics*, Spring 1979, *10*, 92–116.

Grossman, Gene, and Elhanan Helpman, "Product Development and International Trade," *Journal of Political Economy*, December 1989, *97*:6, 1261–83.

Heston, Alan, and Robert Summers, "The Penn World Trade (Mark 5): An Expanded Set of International Comparisons, 1950–1988," *Quarterly Journal of Economics*, May 1991, *106*, 327–68.

Jones, Lawrence, and Rodolfo Manuelli, "A Convex Model of Equilibrium Growth:

Theory and Policy Implications," *Journal of Political Economy*, October 1990, *98*:5 Part 1, 1008–38.

Jovanovic, Boyan, and Saul Lach, "Diffusion Lags and Aggregate Fluctuations," mimeo, New York University, August 1993.

Judd, K. L., "On the Performance of Patents," *Econometrica*, May 1985, *53*:3, 567–85.

King, Robert G., and Sergio Rebelo, "Transitional Dynamics and Economic Growth in the Neoclassical Model," *American Economic Review*, September 1993, *83*:4, 908–31.

Kremer, Michael, "Population Growth and Technological Change: One Million B.C. to 1990," *Quarterly Journal of Economics*, August 1993, *108*:3, 681–716.

Krugman, Paul, *Rethinking International Trade*. Cambridge: MIT Press, 1990.

Krugman, Paul, *Geography and Trade*. Cambridge: MIT Press, 1991.

Krugman, Paul, "Towards a Counter-Counter Revolution in Development Theory." *Proceedings of the Annual World Bank Conference on Development 1992*, Supplement, Washington, D.C., *World Bank Economic Review*, 1993, 15–38.

Lucas, Robert E., Jr., "On the Mechanics of Economic Development," *Journal of Monetary Economics*, July 1988, *22*:1, 3–42.

Maddison, A., *Phases of Capitalist Development*. Oxford: Oxford University Press, 1982.

Mankiw, N. Gregory, "Real Business Cycles: A New Keynesian Perspective," *Journal of Economic Perspectives*, Summer 1989, *3*:3, 79–90.

Mankiw, N. Gregory, David Romer, and David N. Weil, "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics*, May 1992, *107*, 407–37.

Marshall, Alfred, *Principles of Economics*. London: Macmillan, 1890.

Nelson, Richard R., and Edmund S. Phelps, "Investment in Humans, Technological Diffusion, and Economic Growth," *American Economic Review*, May 1966, *56*, 69–75.

Nelson, Richard R., and Sidney G. Winter, *An Evolutionary Theory of Economic Change*. Cambridge: The Belnap Press of Harvard University Press, 1982.

Nordhaus, William D., "An Economic Theory of Technological Change," *American Economic Review*, May 1969, *59*:2, 18–28.

Rebelo, Sergio, "Long Run Policy Analysis and Long Run Growth," *Journal of Political Economy*, June 1991, *99*:3, 500–21.

Romer, Paul M., "Dynamic Competitive Equilibria with Externalities, Increasing Returns and Unbounded Growth," Ph.D. dissertation, University of Chicago, 1983.

Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, October 1986, *94*:5, 1002–37.

Romer, Paul M., "Crazy Explanations for the Productivity Slowdown," In Fischer, S., ed., *NBER Macroeconomics Annual*. Cambridge: MIT Press, 1987a, 163–202.

Romer, Paul M., "Growth Based on Increasing Returns Due to Specialization," *American Economic Review*, May 1987b, *77*:2, 56–62.

Romer, Paul M., Endogenous Technological Change," *Journal of Political Economy*, 1990, *98*, S71–102.

Romer, Paul M., "Two Strategies for Economic Development: Using Ideas and Producing Ideas," *Proceedings of the Annual World Bank Conference on Development 1992*, Supplement, Washington, D.C., *World Bank Economic Review*, 1993,

Romer, Paul M., "New Goods, Old Theory and the Welfare Costs of Trade Restrictions," *Journal of Development Economics*, forthcoming February 1994, *43*:1.

Shell, Karl, "Toward a Theory of Inventive Activity and Capital Accumulation," *American Economic Review*, May 1966, *56*, 62–68.

Solow, Robert, "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics*, February 1956, *70*, 65–94.

Solow, Robert, "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, August 1957, *39*, 312–20.

Solow, Robert, "Growth Theory and After," *American Economic Review*, June 1988, *78*:3, 307–17.

Stiglitz, Joseph, "Comments: Some Retrospective Views on Growth Theory." In Diamond, Peter, ed., *Growth, Productivity, and Unemployment*. Cambridge: MIT Press, 1990, 50–68.

Summers, Robert, and Alan Heston, "A New Set of International Comparisons of Real Product and Price Levels Estimates for 130 Countries, 1950–1985," *Review of Income and Wealth*, March 1988, *34*:1, 1–25.

Uzawa, Hirofumi, "Optimum Technical Change in an Aggregative Model of Economic Growth," *International Economic Review*, January 1965, *6*, 18–31.