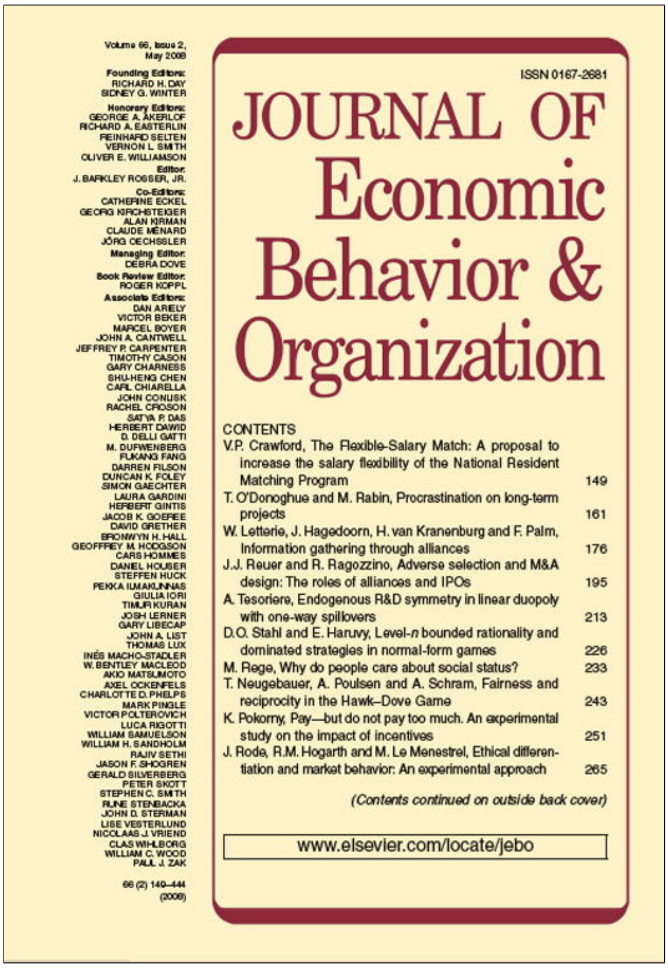

Volume 66, Issue 2,
May 2008

66 (2) 149–444
(2008)

ISSN 0167-2681

JOURNAL OF Economic Behavior & Organization

CONTENTS

(Contents continued on outside back cover)

www.elsevier.com/locate/jebo

ELSEVIER

Journal of Economic Behavior & Organization
Vol. 66 (2008) 403–419

JOURNAL OF
Economic Behavior
& Organization

www.elsevier.com/locate/econbase

Scale effects in growth: A role for institutions

Lewis S. Davis *,[1]

Department of Economics, Social Science Building 211D, 807 Union Street, Schenectady NY 12308, United States

Received 8 November 2004; accepted 25 May 2006
Available online 28 December 2006

Abstract

Reliance on a Dixit–Stiglitz production function leads the growth literature naïvely to associate economic scale with the size of a country's population. I develop an alternative approach in which market size is endogenous, reflecting a trade-off between the gains to exploiting non-rival skills and market transaction costs. Transaction costs reflect a country's institutional framework. The model supports scale effects in transitional growth rates and steady state income levels, suggesting scale effects may be important for developing countries. In this framework effective market size depends on a country's institutions. It is independent of population size and other macroeconomic variables.

JEL classification: O1; O4

Keywords: Institutions; Transaction costs; Specialization; Scale effects; Growth theory

1. Introduction

The presence of scale effects is frequently cited as a major shortcoming of the first generation endogenous innovation models of Romer (1990) and Grossman and Helpman (1991). Scale effects exist when per capita income growth is increasing in some measure of macroeconomic scale, usually taken to be population size. Empirically, scale effects have two troubling implications. First, large economies should grow faster than small ones, a prediction that is clearly hard to reconcile with the experience of developing countries (e.g. India and Singapore). Second, population growth should coincide with accelerating income growth. In what has become known as the

* Tel.: +1 518 388 8395; fax: +1 518 388 6988.
E-mail address: davisl@union.edu.
[1] I thank Roisin O'Sullivan, Bruce Elmslie, Tim Ford, participants at the meetings of the Southern Economic Association, and the editor of this journal for helpful comments on earlier drafts of this paper.

0167-2681/$ – see front matter
doi:10.1016/j.jebo.2006.05.007

"Jones critique," Charles Jones (1995b) has pointed out this is at odds with the historical record of the OECD countries, in which dramatic increases in population have coincided with relatively constant income growth.[2] Given these difficulties, most students of growth would probably agree with Lucas (1993, 263) that scale effects "are a nuisance implication that we want to dispose of."

Indeed, since the Jones critique considerable effort has gone into eliminating scale effects.[3] Scale effects arise because ideas are non-rival inputs in the production of intermediate goods, and the demand for each intermediate good is increasing in population size. An increase in population increases the return to innovation, drawing additional resources into research and raising the rate of growth. As Dinopoulos and Thompson (1999, 170) put it in a recent review article, the heart of the non-scale growth literature consists of various explanations for why "R&D becomes more difficult over time." Diminishing returns or declining horizontal spillovers serve to raise innovation costs, offsetting the impact of increases in the demand for ideas. While this line of inquiry has generated numerous insights regarding the economics of innovation, addressing *inter alia* the relationships between research inputs and patents and between horizontal and vertical innovation, it is unlikely to resolve the issue of scale effects fully.

As noted above scale effects result from the impact of population on the demand for new ideas. In treating scale effects as a knowledge production issue, the non-scale growth literature addresses the supply side of this market. Careful specification of the knowledge production technology may serve to neutralize scale effects, but it does not remove them at the source.[4] A more fundamental problem with the current approach is that scale effects are not unique to models of endogenous innovation. In general, scale effects may arise in any model with a reproducible non-rival input, such as Romer (1987).

With analytic attention focused on innovation costs, the demand for ideas has been formalized in a convenient but potentially misleading manner. The lockstep association of scale with a country's population, so prevalent in the existing literature, is a mathematical artifact, resulting from the use of labor and intermediate goods as complements in a Dixit–Stiglitz (DS) production function.[5] This means of formalizing demand is subject to two criticisms. First, the choice of "raw labor" is arbitrary. As shown in section two, if one uses effective labor input instead (labor force size augmented by average human capital) this framework generates *intensive scale effects*. That is, growth depends on not only on population size, a measure of the extent of production, but also on human capital per worker. This complicates the removal of scale effects because human capital accumulation, unlike population growth, is endogenous. Restricting attention to the size of the labor force may be an acceptable abstraction in some contexts, but the implications for the structure of scale effects should be treated lightly, not as an important insight into the effective demand for ideas.

[2] See also Backus et al. (1992) for evidence against scale effects. Kremer (1993) presents evidence of a positive relationship between population size and income growth in the pre-modern period.

[3] Prominent contributions to the non-scale growth literature include Jones (1995a), Kortum (1997), Segerstrom (1998), Dinopoulos and Thompson (1998), Peretto (1998), Young (1998), and Peretto and Smulder (2002). See Jones (1999) and Dinopoulos and Thompson (1999) for reviews.

[4] The newest set of non-scale growth models considers both horizontal and vertical innovation. In these models non-scale growth requires horizontal innovation or "entry" to become more difficult over time.

[5] Empirically, the use of a DS consumption function is less troubling. As the utility gains from increases in the variety of consumption goods are not reflected in GDP (Aghion and Howitt, 1998, Chapter 12), population size will not be related to growth. The question of whether population size is a good measure of variety demand, however, remains relevant whether a production or consumption function is used.

Second, it is unclear that scale ought to be associated with the size of the national economy. As Griliches (1990, 1968) has pointed out, "the 'reach' of any particular invention does not expand at the same rate as the growth of the overall economy, but only at the rate of its 'own' market." The DS production function provides the critical link between the demand for non-rival inputs and macroeconomic size, but other than analytic convenience there is little to recommend this function's use. The utility function developed by Dixit and Stiglitz (1977) was adopted into the theory of production as an *ad hoc* means of capturing gains to input specialization (Ethier, 1982; Romer, 1987). But there has been no attempt to derive this function from more primitive economic assumptions. The lack of clear microeconomic foundations for this reduced-form production function calls into question the validity of the relationships it embodies.

The third section of the paper draws on the microeconomics of production with specialized human capital to construct an alternative to the DS production function. Once acquired a skill may be used repeatedly in production, making it non-rivalrous and providing an incentive for workers to specialize (Rosen, 1983). Specialization allows workers to utilize their skills more intensively. Offsetting the gains to labor specialization are market transaction costs. These are incurred because a specialized worker must trade to acquire the complementary intermediates required to assemble the final good. The specialization model developed here is not entirely novel, being based on central elements of the existing literature on growth and labor specialization, but its application to the issue of scale effects is.[6]

The model supports scale effects in transitional growth rates and steady state income levels. What is new from the point of view of the literature on scale effects is that market size is endogenous, reflecting a trade-off between gains to specialization and market transaction costs. In particular, the model suggests that economic scale is determined by the degree of market integration, not by how many people live within a country's borders. This is the primary contribution of the paper.

The specialization model may help to explain the absence of evidence for scale effects as they are currently construed. Transitional scale effects, for example, are associated with domestic market integration and, thus, are unlikely to show up in data from developed countries. Moreover, population size is likely to be a poor proxy for economic scale: effective market size in Singapore and India may have less to do with their respective populations than with Singapore's integration into international markets and the fragmentation of India's domestic market. This broader understanding of scale effects is consistent both with a large literature relating openness to economic growth (Edwards, 1998; Alesina et al., 2000) and the growing body of evidence that the scale of local markets matters for growth and innovation (Sokoloff, 1988; Ciccone and Hall, 1996; Sedgley and Elmslie, 2000).

A second contribution regards the pivotal role of institutions, which following North (1990) and others I view as a critical determinant of market transaction costs. The dimension of institutional quality addressed by the model is different from that usually stressed in the growth literature, which focuses on factors that influence the cost of intertemporal exchange such as intellectual property rights. Here, the emphasis is on institutions that influence the cost of contemporaneous exchange, such as contract law, tariffs and trust, which impact growth through their effect on market size.

The idea that the rules governing contemporaneous exchange may influence growth goes back to Smith (1776) discussion of 18th century China, which stresses the growth retarding impact of

[6] See Yang and Borland (1991), Becker and Murphy (1992), Tamura (1992, 1996, 2002), and Davis (2003a, 2003b).

local tariffs. A similar point is made by Landes (1969, 46) in explaining why England industrialized prior to France: "By contrast [with England], a country like France, with more than three times as many people, was cut up by internal customs, obsolete tolls and charges ... into a mosaic of semi-autarkic cells."

The specialization model is an imperfect substitute for models based on endogenous innovation. Do market transaction costs, which constrain the exploitation of non-rival skills in the specialization model, similarly constrain the exploitation of ideas? This question is addressed in some detail in section four, but a fully satisfactory answer may await a model incorporating endogenous innovation and endogenous market size.

2. The diversity model

I call this the "diversity model" because the DS production function employed generates gains to the *number* of intermediate inputs, as distinct from the degree of input specificity as in the specialization model of the next section. The diversity model is intended to be canonical. For the most part it follows Jones (1998) but omits the research sector. By omitting innovation the model clearly locates the origin of scale effects in the demand for non-rival inputs. While this insight is not new (see Romer, 1987), it has been neglected in the recent debate over scale effects. In addition, this model serves as a point of departure for discussing the role of rising innovation costs in the non-scale growth literature and the modeling incentives that have led to the use of population size as a proxy for macroeconomic scale. As many readers may be familiar with models in this vein, details of the solution are relegated to the appendix.[7]

2.1. The model

The final good is produced by perfectly competitive firms employing labor and a range of intermediate goods:

$$Y = L^{1-\alpha} \int_0^N x_i^\alpha \, \mathrm{d}i, \tag{1}$$

where Y is the aggregate output, L the population size, x_i the input of the ith intermediate good, and $\alpha \in (0, 1)$. It is the manner in which intermediate goods enter production that was inspired by the Dixit–Stiglitz utility function.

Intermediate goods are produced using physical capital, with output of the ith good given by $x_i = k_i - f$, where K_i is physical capital employed by the ith firm. In the endogenous innovation literature, the constant f represents the rental rate on the patent for a new good (in units of capital), but it might just as easily include any fixed production cost. The presence of a fixed cost implies each intermediate good is produced by a single firm.

Using input demand functions derived from Eq. (1), profit maximization by intermediate good producers establishes the uniform equilibrium value for intermediate good output, $x = \alpha f/(1-\alpha)$. An adding-up condition for physical capital and free entry in the intermediate good sector allows us to solve for the number of intermediates, $N = (1-\alpha)K/f$. Substituting these results into (1) and

[7] The appendix is available on the Journal of Economic Behavior and Organization website.

dividing by population size gives the following expression for per capita output:

$$y = \frac{Y}{L} = Af^{\alpha-1}L^{1-\alpha}k, \tag{2}$$

where $A = \alpha^{\alpha}(1-\alpha)^{(1-\alpha)}$, $y = Y/L$ and $k = K/L$.

2.2. *Scale effects and innovation in diversity models*

It is clear why this model is so popular as a framework for considering innovation-driven growth. The market power enjoyed by intermediate good producers generates a stream of profits that provides the incentive to undertake costly research activities. By modeling innovation as a profit driven activity, this framework places the locus of knowledge creation within the economy. In this interpretation of the model, r is the rental price of capital and rf is the rental price of an intermediate good patent.

The presence of scale effects is readily apparent. Specifying a simple accumulation function, $\dot{K} = sY$, we have

$$g_{\mathrm{y}} = g_{\mathrm{k}} + (1-\alpha)(g_{\mathrm{L}} - g_{\mathrm{f}}) = sA\left(\frac{L}{f}\right)^{1-\alpha} - \alpha g_{\mathrm{L}} - (1-\alpha)g_{\mathrm{f}}, \tag{3}$$

where g_X denotes the growth rate of X. It is the first term in (3) that gives rise to scale effects: larger economies grow faster, and if population is growing, income growth accelerates. Eq. (3) also illustrates why non-scale growth theorists have been interested in motivating models with rising research costs: if f rises at the same rate as population, the first term is constant and scale effects are neutralized.

One gets similar results if saving is endogenized by appealing to dynamic optimization of a CIES utility function. Balanced growth requires a constant interest rate. With the return to capital given by

$$r = \alpha Af^{\alpha-1}L^{1-\alpha} \tag{4}$$

this requires research costs to grow in line with population so that L/f is constant. Otherwise, growth is explosive, and the objective function is unbounded.

2.3. *Extensive and intensive scale effects*

The scale effect in evidence here, and in most of the endogenous growth literature, is constrained in a somewhat artificial manner. In particular, the model above exhibits what might be called *extensive scale effects*. By this I mean that the return to capital depends on labor force size, a measure of the extent of production.

Suppose instead that skills matter and effective labor input is given by aggregate human capital, $H = hL$, where h is the uniform level of human capital in the work force. Substituting H for L in the model above, the returns to physical and human capital are $r_{\mathrm{k}} = \alpha Af^{(\alpha-1)}(hL)^{(1-\alpha)}$ and $r_{\mathrm{h}} = (1-\alpha)Af^{(\alpha-1)}L^{(1-\alpha)}h^{-\alpha}k$. Assuming agents act to equalize these returns, the uniform rate of return is given by $r = \alpha^{\alpha}(1-\alpha)^{(1-\alpha)}Af^{(\alpha-1)}L^{(1-\alpha)}k^{(1-\alpha)}$. In this case, in addition to the extensive scale effects indicated by the dependence of r on L, we have *intensive scale effects*: the return to capital depends on resources per worker or the intensity of production.

In general, this kind of scale effect is more difficult to eliminate. For a variety of accumulation functions, the feedback between human and physical capital accumulation will generate an explosive pattern of rising returns and accelerating output. To maintain the constant rate of return necessary for balanced growth, the fixed cost *f* must rise sufficiently quickly to offset both population increase and capital accumulation.

The discussion above is not, of course, an argument that aggregate human capital is the appropriate measure of labor input in this type of model (though a theory that gave some role to the intensity of production would seem to have a better chance in explaining innovation across, say, the US and India).[8] Rather, it is intended to point out that using raw labor instead, as is common, serves as something of an anchor for income growth rates. This has simplified (and unified) the effort to produce non-scale growth models. Population is biologically rather than economically reproducible, and the endogenous innovation literature regularly takes population growth to be independent of the return to labor. While this modeling strategy has had the intended effect of simplifying the task faced by non-scale growth theorists, it has had the unintended consequence of supporting a particularly narrow interpretation of scale effects, namely that growth rates are linear in population size.

3. The specialization model

3.1. Production

The specialization model employs a more disaggregate view of production. Each intermediate good corresponds to one of a large number of productive tasks undertaken by an individual worker, not the output of a specialized firm. Following Rosen gains to specialization arise due to non-rivalry in the use of task-specific human capital: having acquired the skills necessary to perform a given task, it benefits a worker to exploit those skills as intensively as possible. The structure of the model draws extensively from earlier division of labor growth models.

There is a continuum of intermediate goods arranged along the unit interval and indexed by $i \in [0,1]$. Intermediate goods are produced according to $x_i(\ell_i, h_i) = \ell_i^{\varphi} h_i^{\beta}$, where x_i is the output of intermediate good i, ℓ_i and h_i are time and human capital inputs into good i's production, and $\beta,\varphi \in (0,1)$.

Each of L workers is endowed with h units of human capital and one unit of time, which are allocated uniformly across v tasks, $v \le 1$. Labor specialization is the inverse of the number of tasks a worker performs, $s \equiv 1/v$, such that the time and human capital allocated to each task is increasing in labor specialization: $h_i = h/v = sh$ and $t_i = 1/v = s$. A worker's intermediate good output, z, is found by integrating x_i over the set of productive tasks:

$$z(s, h) = \int_0^1 x_i \, \mathrm{d}i = v x_i = s^{\varepsilon} h^{\beta}, \tag{5}$$

where $\varepsilon = \varphi + \beta - 1$. Gains to specialization, $\varepsilon > 0$, result from increasing returns in intermediate good production, which is assumed to hold.[9]

[8] In Romer (1990) aggregate human capital is the relevant measure of scale. Romer avoids explosive growth by holding level of human capital constant.

[9] As pointed out by Rosen, a sufficient condition for gains to specialization is the absence of task-specific fatigue, $\phi = 1$.

Note that the gains to specialization are increasing in the stock of human capital: $\delta z_s/\delta h > 0$. This relationship is the counterpart to Rosen's insight that specialization increases the return to capital: $\delta z_s/\delta h = \delta z_h/\delta s > 0$. These relationships provide the basis for a virtuous cycle of growth in which accumulation drives specialization, and specialization offsets diminishing returns to capital.

Workers produce the final good according to a Leontief technology with constant returns to scale. One unit of the final good requires one unit of each intermediate:

$$y = \min x_i, \quad \text{for } i \in [0, 1]. \tag{6}$$

The strong complementarities evident in (6) provide an incentive for specialized workers to engage in interpersonal trade as a means of acquiring the full-range of intermediate goods.

3.2. Exchange

Call a group of specialists who trade intermediate goods a *market*, and define market size, m, to be the number of market participants. Gains to specialization imply that market participants produce non-overlapping sets of intermediate goods, $v \leq 1/m$. Similarly, the final good technology implies each intermediate is produced by at least one market participant: $mv \geq 1$. Taken together, these restrictions imply that market size equals labor specialization: $m = 1/v = s$. In a market with 10 workers, each produces one-tenth of the intermediate goods.

Exchange is costly.[10] Transaction costs may be influenced by the distance between agents (Yang and Borland), monitoring costs (Becker and Murphy), and the number of relative prices (Coase, 1937). Following North, however, I assume transaction costs primarily reflect the nature of an economy's underlying institutions. As North (p. 34) persuasively argues, "Institutions provide the structure for exchange that . . . determines the cost of transacting." The quality of a country's institutions affects market transaction costs in a pervasive manner through its impact on uncertainty and opportunism in exchange (North; Williamson, 1979) and cost of defining and enforcing property rights (de Soto, 2000).

Transaction costs are assumed to rise with the number of market participants, partly due to increases in the complexity of contracting and partly due to the depersonalization of exchange, which reduces informal constraints on opportunism (North Chapter 5, Davis, 2006). The transaction costs incurred by a worker are given by

$$t(s) = as^{\gamma}, \tag{7}$$

where $a > 0$, $\gamma > 1$.

3.3. Optimization

Fig. 1 presents a schematic of the specialization model. Institutions and human capital determine market transaction costs and gains to specialization. These costs and benefits, in turn, affect optimal specialization and market size. The key benefit of this approach is that it allows us to treat market size as endogenous rather than equal to population size.

Workers choose labor specialization to maximize per capita income, equal to gross income less transaction costs, $y(s,h) = z(s,h) - t(s)$. Minimum specialization occurs when a worker pro-

[10] For specialization models in which firms avoid market transaction costs by organizing production internally, see Davis (2003b).

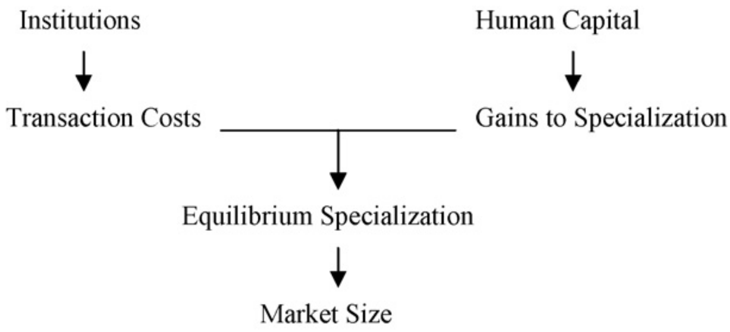

Fig. 1. Schematic of the specialization model.

duces the full range of intermediate goods, $s = 1$. Maximum specialization occurs when the whole population participates in a single market, $s = L$. To avoid complications involving the exercise of specialist monopoly power, it is assumed that interpersonal exchange proceeds according to contracts signed prior to specialization decisions. Since agents are *ex ante* identical, this ensures price taking behavior (Yang and Borland, 465). The worker's problem is thus

$$\max_{s \in [1,L]} s^{\varepsilon} h^{\beta} - a s^{\gamma}. \tag{8}$$

Kuhn-Tucker conditions imply equilibrium specialization is given by

$$s^{e}(h, L) = \begin{cases} 1, & h < h_1 \\ \left[\dfrac{\varepsilon}{\gamma a}\right]^{1/(\gamma-\varepsilon)} h^{\beta/(\gamma-\varepsilon)}, & h \in [h_1, h_2] \\ L, & h > h_2, \end{cases} \tag{9}$$

where $h_1 = (a\gamma/\varepsilon)^{1/\beta}$ and $h_2(L) = h_1 L^{(\gamma-\varepsilon)/\beta}$. Equilibrium specialization is weakly increasing in human capital, reflecting the impact of human capital on the gains to specialization in Eq. (5). Second-order conditions require $\gamma > \varepsilon$, which is assumed to hold.

The tripartite structure of (9) reflects different organizational equilibria that correspond in a stylized fashion to different stages of development. These are (1) an traditional economy with unspecialized, autarkic workers, (2) a developing economy in which the domestic market is fragmented into multiple independent markets, and (3) an organizationally mature economy with a single integrated market, the size of which is determined by its population.

Substituting (9) into the equation for per capita income, we get

$$y(h, L) = \begin{cases} h^{\beta} - a, & h < h_1 \\ \left[\dfrac{\gamma-\varepsilon}{\gamma}\right]\left[\dfrac{\varepsilon}{\gamma a}\right]^{\varepsilon/(\gamma-\varepsilon)} h^{\beta\gamma/(\gamma-\varepsilon)}, & h \in [h_1, h_2] \\ L^{\varepsilon} h^{\beta} - a L^{\gamma}, & h > h_2. \end{cases} \tag{10}$$

3.4. Dynamic optimization

The dynamic model uses a discrete-time framework in which the variables of the preceding subsection are functions of time. Population size grows at a constant exogenous rate n: $L(t+1) = (1+n)L(t)$. Dynastic agents maximize an infinite horizon utility function with logarithmic periodic utility, subject to an accumulation equation with full periodic depreciation. Complete depreciation of capital implies that specialization decisions depend only on the current level of human capital, allowing the use of the reduced-form function for per capita output in (11) and simplifying analysis of transitional dynamics.

The representative household's optimization problem is thus

$$\max U = \sum_{t=0}^{\infty} \rho^{-t} \ln(c(t)), \tag{11}$$

subject to $h(0) = h_0$, $L(t+1) = (1+n)L(t)$, and $(1+n)h(t+1) = y(h(t), L(t)) - c(t)$, where h_0 is the initial capital-labor ratio and $\rho > 1$ indicates a preference for current over future consumption. The first-order condition for the household's problem gives

$$g(t) = \frac{c(t+1)}{c(t)} - 1 = \frac{r(t)}{\rho(1+n)} - 1. \tag{12}$$

In the steady state the marginal product of capital is constant and per capita consumption, income and human capital grow at a common rate.

3.5. *Institutional flexibility and the return to capital*

Like output, the return to human capital depends on the organizational structure of the economy. Differentiating (10) with respect to human capital yields

$$r(h, L) = \begin{cases} \beta h^{\beta-1}, & h < h_1 \\ \beta \left[\dfrac{\varepsilon}{\gamma a}\right]^{\varepsilon/(\gamma-\varepsilon)} h^{(\varepsilon-\gamma(1-\beta))/(\gamma-\varepsilon)}, & h \in [h_1, h_2] \\ \beta L^{\varepsilon} h^{\beta-1}, & h > h_2. \end{cases} \tag{13}$$

The structure of (13) reflects the two forces acting on the return to capital. First, there are diminishing returns to capital in intermediate good production, captured by $\beta < 1$, implying that accumulation reduces the return to capital. I call this familiar negative effect the *production effect* of accumulation. Second, specialization allows workers to utilize specialized skills more intensively, increasing the return to capital. I call this the *specialization effect*.

The specialization effect is absent in traditional economies. Labor is unspecialized, and only diminishing marginal returns are evident in the first line of (13). In mature economies, labor specialization is constrained by population size. Thus, the return to capital is increasing in population size and decreasing in human capital, as seen in the third line of (13).

In developing economies the situation is more complex. Labor specialization is an endogenous function of the capital stock. Accumulation raises the return to labor specialization. Greater specialization, in turn, allows more intensive utilization of specialized skills, raising the return to capital. With both the specialization and accumulation effects linked to accumulation, the sign of the exponent in the second line of (13) is ambiguous. If the specialization effect is sufficiently strong, a developing economy will experience increasing marginal returns to capital. In particular, increasing marginal returns obtain if

$$\varepsilon > \gamma(1-\beta). \tag{14}$$

Intuitively, (14) holds if the gains to specialization, ε, outweigh the combined effect of rising transaction costs and diminishing returns to capital, $\gamma(1-\beta)$.

The condition in (14) indicates that transaction costs play a pivotal role in the return to capital. Recall that *gamma* is the elasticity of transaction costs with respect to market size. If *gamma* is large, transaction costs rise quickly with increases in specialization, reducing the impact of an increment of capital on equilibrium labor specialization. This reduces strength of the specialization effect.

Intuitively, a high value for gamma indicates that an economy has difficulty adopting the increasingly complex and impersonal patterns of exchange that accompany an advancing division of labor. In keeping with the institutional basis of transaction costs, I interpret this form of adaptability as a measure of institutional flexibility. The idea that the flexibility or adaptability, rather than static efficiency, of an economy's institutional framework that matters for growth appears repeatedly in economic historical work, such as Kuznets (1973), Abramovitz (1986). More specifically, define the flexibility of institutions as follows:

Definition 1. Let $\gamma^* = \varepsilon/(1-\beta)$, then

(1) an economy has *flexible institutions* provided $\gamma < \gamma^*$.
(2) an economy has *semi-flexible institutions* provided $\gamma = \gamma^*$, and
(3) an economy has *inflexible institutions* provided $\gamma > \gamma^*$.

This definition implies that flexible institutions correspond to increasing marginal returns to capital, semi-flexible institutions correspond to constant marginal returns, and inflexible institutions correspond to diminishing marginal returns.

3.6. Transitional dynamics and scale effects

In traditional economies workers are unspecialized and autarkic. Market size, equal to one, is independent of population size, and scale effects are notably absent. In this stage of development capital accumulation reduces the return to capital, causing growth rates to fall.

Transitional dynamics in developing economies exhibit intensive scale effects. The return to capital and rate of income growth depend on the intensity of production, as measured by human capital per worker. Because it reflects labor specialization decisions, market size also depends on the level of human capital, resulting in a systematic relationship between market size and growth rates. This relationship, however, is fundamentally different from that in the endogenous innovation literature: both market size and the rate of income growth are independent of population size, or indeed any measure of macroeconomic scale.

A second difference is that intensive scale effects may be positive, negative or neutral, depending on an economy's institutional flexibility. In an inflexible developing economy, rapidly rising transaction costs result in a small specialization effect and diminishing marginal returns to capital. Accumulation increases market size but decreases the rate of income growth. In a semi-flexible developing economy the return to capital is constant, the production and specialization effects just offsetting each other, and scale is neutral for growth. Strong specialization effects give rise to increasing marginal returns in flexible developing economies. With the return to capital rising, the rates of growth of income, market size expansion and human capital accumulation increase as growth progresses.

In sum, with market size endogenously determined, the specialization model predicts intensive scale effects (growth depends on the intensity of production, as measured by human capital per worker) and indicates a central role for institutions in determining whether endogenous market expansion is accompanied by rising or falling rates of growth.

3.7. Scale and growth in the steady state

The specialization model supports three steady state outcomes. Which steady state occurs depends on the rate of endogenous market expansion, a reflection of institutional flexibility, and the population growth rate. Since scale effects are absent in traditional economies, I rule out a low-level steady state by assuming $r(h_1) > \rho(1+n)$.

An inflexible economy converges to a stationary steady state in which income and market size are constant. Diminishing returns implies the rates of market expansion and human capital accumulation asymptotically approach zero. In the steady state, human capital, market size and income are constant at a level consistent with zero growth:

$$\begin{aligned} h_{\mathrm{I}}^{\mathrm{SS}} &= \left[\frac{\varepsilon}{\gamma a}\right]^{\varepsilon/(\gamma(1-\beta)-\varepsilon)} \left[\frac{\beta}{\rho(1+n)}\right]^{(\gamma-\varepsilon)/(\gamma(1-\beta)-\varepsilon)}, \\ s_{\mathrm{I}}^{\mathrm{SS}} &= \left[\frac{\varepsilon}{\gamma a}\right]^{1/(\gamma-\varepsilon)} [h_{\mathrm{I}}^{\mathrm{SS}}]^{\beta/(\gamma-\varepsilon)}, \\ y_{\mathrm{I}}^{\mathrm{SS}} &= \left[\frac{\gamma-\varepsilon}{\gamma}\right] \left[\frac{\varepsilon}{\gamma a}\right]^{\varepsilon/(\gamma-\varepsilon)} [h_{\mathrm{I}}^{\mathrm{SS}}]^{\beta\gamma/(\gamma-\varepsilon)}. \end{aligned} \tag{15}$$

The superscript and subscript in (15) indicate steady state values for an inflexible economy. For values of gamma consistent with inflexible institutions, the exponent on the second bracketed term of $h_{\mathrm{I}}^{\mathrm{SS}}$ is positive, implying that high discount rates and rapid population growth decrease the steady state levels of human capital, market size and per capita income.

The explosive trajectories of income, capital and market size in a flexible developing economy ceases once endogenous market size overtakes population, or equivalently once human capital exceeds the second threshold, $h_2(L(t)) = h_1 L(t)^{(\gamma-\varepsilon)/\beta}$. Beyond this point, the economy is organizationally mature, consisting of a single integrated market that expands exogenously at the rate of population growth. In the steady state, the return to capital given by the third line of (13) is constant, with capital growing at $g_{\mathrm{F}}^{\mathrm{SS}} = \varepsilon n/(1-\beta)$, a multiple of the population growth rate. Per capita income growth asymptotically approaches the same rate as transaction costs fall relative to output (see the third line of (10)).

In a semi-flexible developing economy the return to capital is constant. Gains to specialization exactly offset diminishing returns to capital and rising market transaction costs. The economy grows at a bounded, endogenous rate determined by its population growth rate, preferences and constant return to capital. This rate of growth is bounded in that if it is sufficiently high, market size eventually outstrips population size, and the economy transitions to the population-constrained steady state outlined above. Formally, this condition is $g_{\mathrm{SF}}^{\mathrm{SS}} > g_{h_2} = (\gamma-\varepsilon)n/\beta = \varepsilon n/(1-\beta)$, where the second equality holds due to the parameter constraints implied by the definition of semi-flexible institutions. Thus, the steady state growth rate for a semi-flexible economy is given by

$$g_{\mathrm{SF}}^{\mathrm{SS}} = \min\left\{\frac{\beta[\varepsilon/\gamma a]^{\varepsilon/(\gamma-\varepsilon)}}{\rho(1+n)} - 1, \frac{\varepsilon n}{1-\beta}\right\}, \tag{16}$$

where the first term is its endogenous growth rate as a developing economy and the second term relates to the population constrained steady state. From (9) and Definition 1, income growth is proportionate to the rate of market expansion in both equilibria, $g_{\mathrm{SF}}^{\mathrm{SS}} = [\varepsilon/(1-\beta)]g_{\mathrm{s}}$.

The central role of institutions is evident is these dramatically different steady state outcomes. Institutional flexibility determines whether an economy stagnates, grows at an endogenous rate, or grows at an exogenous rate proportional to the rate of population growth. In spite of these differences, all three steady states support scale effects in income levels: per capita income increases linearly with market size. As in much of the non-scale growth literature (e.g. Jones, 1995a), scale effects in income levels reflect the presence of some form of diminishing returns. In inflexible economies, the potential for a virtuous cycle of growth driven by mutually reinforcing accumulation and specialization is undermined by rapidly rising transaction costs. In a flexible economy, this virtuous cycle is realized, giving rise to a period of accelerating growth, but ultimately limited by the constraints of population size on endogenous market expansion.

3.8. Scale effects and economic integration

To address the effects of international trade on growth rates in the specialization model, I consider the simpler but related issue of integration among two identical economies in steady state equilibria. If the economies have inflexible or semi-flexible institutions, the number of markets doubles at the moment of integration, but market size itself is unaffected. Similarly, variables related to per capita economic performance such as income levels and growth rates are unchanged. If the economies have flexible institutions, however, then pre-integration market size is constrained by population, and integration allows greater specialization, increasing the return to human capital and the rate of growth during the transition to the new, higher level growth path along which income is proportional to combined population size.

This simple exercise suggests that the dynamic gains from trade may depend on the institutional flexibility of the economies involved. In particular, if market size is constrained by high transaction costs rather than population size, integration will not affect economic performance. This result suggests limits to the gains from South-South trade. It also suggests that trade among countries with flexible institutions will be of greater benefit to the country with the smaller population. With respect to the empirical literature on scale effects, it suggests that even in industrialized democracies one should not expect a strong relationship between income levels and population size, since international trade among such countries will tend to undermine the relationship between a nation's population and effective market size.

4. Discussion

4.1. Institutional structure and the diversity of dynamic behavior

The specialization model supports a rich diversity of dynamic behavior, including accelerating and decelerating growth along transitional trajectories and steady states in which growth is absent, endogenous or constrained by population growth. The combination of transitional scale effects in growth and steady state scale effects in income may help to resolve one dispute over the role of scale in growth. In particular, Romer (1996) agues that the failure to find scale effects in the cross-sectional data is at odds with the historical record. For example, Landes suggests that the relative integration of England's domestic market helped spur the industrial revolution, and Abramovitz and David (1992) ague that the large US market was a critical factor in explaining why the US overtook England in the late 19th and early 20th centuries.

The diversity of dynamic outcomes is also attractive when viewed more broadly against the diversity of growth experience. It is well-known that endogenous growth requires the presence of a "unit exponent" somewhere in production relationships of the model (Rebelo, 1991), a requirement that has been criticized for its "knife-edge" character (e.g. Solow, 2000, 100). Moreover, in most treatments the critical exponent reflects technical parameters that are assumed to be relatively uniform across countries, making it more difficult to account for the diverse growth experiences. In the specialization model the critical exponent depends on both technical and institutional parameters, providing some additional leeway. Associating this additional degree of freedom with institutions is appealing both because institutional structures vary dramatically across countries and because institutional heterogeneity is widely accepted to play a role in explaining growth differentials (Knack and Keefer, 1995; Acemoglu et al., 2001).

An important question the model does not address is what accounts for the flexibility of institutions, particularly the relationship between institutional structure and the ability to adopt increasingly complex patterns of production and exchange. One candidate is democratic political institutions. Rodrik (2000) has suggested that democracy serves as a "meta-institution" that lowers the costs of institutional change, but this is not a settled issue. For example, Olson (1982) argues that democracies are fertile ground for special interest politics that lead to an accumulation of restrictive rules, and some observers have praised the adaptability of authoritarian East Asian governments (World Bank, 1993).

4.2. Scale effects in the world economy

The lack of evidence of scale effects at the national level has led some economists to suggest that the relevant measure of economic scale is supranational, population size measured either globally or among the OECD countries (e.g. Jones, 2002). The specialization model offers a somewhat different perspective on growth and scale in the world economy.

First, consider the simpler but related issue of integration among two identical economies that are already in steady state equilibria. If the economies have inflexible or semi-flexible institutions, then market size is independent of population size. Integration doubles the number of markets but leaves market size unchanged. Consequently, the level and growth rate of per capita income is unaffected. If the economies have flexible institutions, however, then market size is population-constrained. In this case, integration allows greater labor specialization, increasing the return to human capital and the rate of growth during the transition to the new, higher level growth path along which income is proportional to combined population size.

This simple exercise suggests that the dynamic gains from trade may depend on the institutional flexibility of the economies involved. In particular, the impact of integration will be limited if market size is constrained by high transaction costs. With respect to the empirical literature on scale effects, it also suggests that international trade among countries with flexible institutions may undermine a strict relationship between income and population size (think Luxembourg or Singapore).

Alternately, the world economy may be seen as a single developing economy with incompletely integrated markets. Potentially global markets are fragmented due to high transaction costs, including transportation and information costs, language and cultural barriers, differing regulations and commercial law, various forms of domestic market protection, and the absence of international legal structures and third party contract enforcement.

From this perspective thought experiments involving frictionless integration probably overstate the impact of international trade. Thus, while international interactions reduce the relevance of national variables as a measure of economic scale (e.g. Ades and Glaeser, 1999), world or OECD population size probably overstates it. Effective scale depends less on world population size than on the level of international transaction costs and the resulting degree of global economic integration.[11] Support for this perspective is provided by evidence of significant border effects on inter-city trade (Engel and Rogers, 1996).

4.3. Scale effects in ideas and skills

A one-to-one mapping of the conceptual elements from the innovation literature to those of the specialization model fairs poorly: intermediate goods to productive tasks, non-rival ideas to non-rival skills, innovation costs to transaction costs. This raises an important question: if their theoretical foundations are largely orthogonal, then to what degree can the specialization model serve to inform the literature on innovation? In particular, is there any reason to believe that transaction costs, which limit the exploitation of specialized skills in the model above, also limit the utilization of a new invention?

Perhaps not. For example, while demand in US retail markets may be geographically fragmented, this would not prevent *Wal-Mart* from exploiting a new inventory management system nationally. It would simply introduce the new system into each of its stores. Despite the limited size of local retail markets, in this example it is the size of the national market that determines the return to the innovation.

Two considerations weigh against the argument in the paragraph above and suggest that transaction costs matter in determining effective scale for new inventions. First, using new inventions often requires workers with specialized skills. X-ray machines require X-ray technicians and radiologists, new software programs require trained users and technical support staff, electrification requires electricians. If the complementarity between invention and skills is sufficiently strong, say Leontief in structure, then the return to employing an innovation will depend on how intensively the complementary skills are utilized. In this case, transaction costs will play a role in determining the relevant measure of economic scale.

More generally, one might argue that transaction costs matter not only for coordinating production with specialized workers but also for coordinating production involving multiple intermediate goods. Indeed, this is one of the central themes of the new institutional literature on industrial organization (Williamson, 1979, 1985; Williamson and Masten, 1999), which stresses the difficulties of limiting opportunistic behavior in the presence of specialized investments. If so then the return to introducing a new invention would depend on the cost of integrating it into an increasingly complex structure of production. If production involving specialized intermediates is subject to non-trivial coordination costs, then institutions may limit the exploitation of non-rival ideas in much the same way that they limit the exploitation of non-rival skills. While it would be possible to use an ever increasing number of intermediate goods, it might not be profitable to do so.

5. Conclusion

This paper argues that in focusing on the microeconomics of knowledge production function, the non-scale growth literature has ignored the critical role played by the demand for non-rival

[11] This “microeconomic” understanding of market size is very much in line with that expressed by Mancur Olson (127): “I believe the greatest reduction of trade restrictions in history have come from reducing the mileage rather than the height of trade restrictions.”

goods. Furthermore, the excessively simple manner in which demand is formalized in this literature has led to the misleading association of scale with national population size.

In an attempt to address this theoretical lacuna, I develop two models. The diversity model suggests that the popular identification of scale with a nation's population size is the result of function forms adopted for analytical convenience, rather than a fundamental insight into the demand for non-rival goods. The second model is based on the exploitation of non-rival skills. This model suggests that the presence of market transaction costs and international trade may limit the relevance of *any macroeconomic variable* as a measure of the demand for non-rival goods. It is also argued that institutions play a central role in determining the structure of transaction costs and, consequently, that the presence of scale effects may be contingent on institutional structure. As argued in the preceding section, limits to the exploitation of non-rival skills may extend to non-rival ideas as well.

The paper has two primary implications for the non-scale growth literature. First, it stresses the importance of the overall theoretical project. With respect to the US, for example, if one includes both internal market integration and the expansion of international trade, then the potential for exploiting non-rival ideas has grown much faster than US population size. An account of why this has not resulted in accelerating income growth rates remains an important objective.

Second, it suggests that the empirical search for scale effects should be more nuanced than it currently is. Either poorly integrated domestic markets or international trade may result in national population size and other macroeconomic variables being poor proxies for the exploitation of non-rival inputs. This insight may be of particular importance in assessing the role of scale in less developed countries, in which underdeveloped institutions may lead to domestic market fragmentation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2006.05.007.

References

Abramovitz, M., 1986. Catching up, forging ahead, and falling behind. Journal of Economic History 46, 386–406.

Abramovitz, M., David, P.A., 1992. Convergence and deferred catch-up: productivity leadership and the waning of American exceptionalism. In: Landau, R., Taylor, T., Wright, G. (Eds.), The Mosaic of Economic Growth. Stanford University Press, Stanford, CA, pp. 21–62.

Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The colonial origins of comparative development: an empirical investigation. American Economic Review 91, 1369–1401.

Ades, A.F., Glaeser, E.L., 1999. Evidence on growth, increasing returns and the extent of the market. Quarterly Journal of Economics 114, 1025–1045.

Aghion, P., Howitt, P., 1998. Endogenous growth theory. MIT Press, Cambridge, MA.

Alesina, A., Enrico, S., Waciarg, R., 2000. Economic integration and political disintegration. American Economic Review 90, 1276–1296.

Backus, D.K., Kehoe, P.J., Kehoe, T., 1992. In search of scale effects in trade and growth. Journal of Economic Theory 58, 377–409.

Becker, G.S., Murphy, K.M., 1992. The division of labor, coordination costs, and knowledge. Quarterly Journal of Economics 57, 1137–1160.

Ciccone, A., Hall, A., 1996. Productivity and the density of economic activity. American Economic Review 86, 54–70.

Coase, R.H., 1937. The nature of the firm. Economica 4, 386–405.

Davis, L.S., 2003a. The division of labor and the growth of government. Journal of Economic Dynamics and Control 27, 1217–1235.

Davis, L.S., 2003b. Toward a unified transaction cost theory of economic organization. Journal of Institutional and Theoretical Economics 159, 571–593.

Davis, L.S., 2006. Growing apart: the division of labor and the breakdown of informal institutions. Journal of Comparative Economics 34, 75–91.

Dinopoulos, E., Thompson, P., 1998. Schumpeterian growth without scale effects. Journal of Economic Growth 3, 313–335.

Dinopoulos, E., Thompson, P., 1999. Scale effects in Schumpeterian models of economic growth. Journal of Evolutionary Economics 9, 157–185.

Dixit, A.K., Stiglitz, J., 1977. Monopolistic competition and optimum product diversity. American Economic Review 67, 297–308.

Edwards, S., 1998. Openness, productivity and growth: what do we really know? Economic Journal 108, 383–398.

Engel, C., Rogers, J.H., 1996. How wide is the border? American Economic Review 86, 1112–1125.

Ethier, W.J., 1982. National and international returns to scale in the modern theory of international trade. American Economic Review 72, 950–959.

Griliches, Z., 1990. Patent statistics as economic indicators: a survey. Journal of Economic Literature 28, 1661–1707.

Grossman, G.M., Helpman, E., 1991. Innovation and Growth in the Global Economy. MIT Press, Cambridge, MA.

Jones, C.I., 1995a. R&D-based models of economic growth. Journal of Political Economy 103, 759–784.

Jones, C.I., 1995b. Time-series tests of endogenous growth models. Quarterly Journal of Economics 110, 495–525.

Jones, C.I., 1998. Introduction to Economic Growth. W.W. Norton, New York.

Jones, C.I., 1999. Growth: with or without scale effects? American Economic Review 89, 139–144.

Jones, C.I., 2002. Sources of U.S. economic growth in a world of ideas. American Economic Review 92, 220–239.

Knack, S., Keefer, P., 1995. Institutions and economic performance: cross-country tests using alternative institutional measures. Economics and Politics 7, 207–227.

Kortum, S., 1997. Research, patenting, and technological change. Econometrica 65, 1389–1419.

Kremer, M., 1993. Population growth and technological change: one million B.C. to 1990. Quarterly Journal of Economics 108, 207–227.

Kuznets, S., 1973. Modern economic growth: findings and reflections. American Economic Review 63, 247–258.

Landes, D., 1969. The Unbound Prometheus. Cambridge University Press, London.

Lucas, R.E., 1993. Making a miracle. Econometrica 61, 251–272.

North, D.C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, New York.

Olson, M., 1982. The Rise and Decline of Nations. Yale University Press, New Haven.

Peretto, P., 1998. Technological change and population growth. Journal of Economic Growth 3, 283–311.

Peretto, P., Smulder, S., 2002. Technological distance, growth and scale effects. Economic Journal 112, 603–624.

Rebelo, S., 1991. Long-run policy analysis and long-run growth. Journal of Political Economy 99, 500–521.

Rodrik, D., 2000. Institutions for high-quality growth: what they are and how to acquire them. National Bureau of Economic Research Working Paper #7540.

Romer, P., 1987. Growth based on increasing returns due to specialization. American Economic Review 77, 56–62.

Romer, P., 1990. Endogenous technological change. Journal of Political Economy 98, S71–S102.

Romer, P., 1996. Why, indeed, in America? Theory, history, and the origins of modern economic growth. American Economic Review 86, 202–206.

Rosen, S., 1983. Specialization and human capital. Journal of Labor Economics 1, 43–49.

Sedgley, N., Elmslie, B., 2000. Agglomeration and congestion in the economics of ideas and technical change. American Journal of Economics and Sociology 60, 101–121.

Segerstrom, P., 1998. Endogenous growth without scale effects. American Economic Review 88, 1290–1310.

Smith, A. 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. Edwin Cannan (Ed.), Reprint 1976. University of Chicago Press, Chicago.

Sokoloff, K.L., 1988. Inventive activity in early industrial America: evidence from patent records. Journal of Economic History 48, 813–850.

Solow, R.M., 2000. Growth Theory: An Exposition, 2nd ed. Oxford University Press, New York.

de Soto, H., 2000. The Mystery of Capital. Basic Books, New York.

Tamura, R., 1992. Efficient equilibrium convergence: heterogeneity and growth. Journal of Economic Theory 58, 355–376.

Tamura, R., 1996. Regional economies and market integration. Journal of Economic Dynamics and Control 20, 825–845.

Tamura, R., 2002. Human capital and the switch from agriculture to industry. Journal of Economic Dynamics and Control 27, 207–242.

World Bank, 1993. The East Asian Miracle: Economic Growth and Public Policy. Oxford University Press, Oxford and New York.

Williamson, O.E., 1979. Transaction-cost economics: the governance of contractual relations. Journal of Law and Economics 22, 233–261.

Williamson, O.E., 1985. The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting. Free Press, New York.

Williamson, O.E., Masten, S.E. (Eds.), 1999. The Economics of Transaction Costs. Edward Elgar Publishing, Northampton, Massachusetts.

Yang, X., Borland, J., 1991. A microeconomic mechanism for economic growth. Journal of Political Economy 99, 460–482.

Young, A., 1998. Growth without scale effects. Journal of Political Economy 106, 41–63.