

*Journal of
Institutional
and Theoretical
Economics*

JITE

Vol. 166, No. 3

September 2010

Alexandre Sokic: Modelling the Transaction Role of Money and the Essentiality of Money in an Explosive Hyperinflation Context 387–396

Victor Nee, Jeong-han Kang, and Sonja Opper: A Theory of Innovation: Market Transition, Property Rights, and Innovative Activity 397–425

Shin-Hwan Chiang and Xiang Li: Market Correlation and Property Rights 426–438

Julia Angerhausen, Christian Bayer, and Burkhard Hehenkamp: Strategic Unemployment 439–461

Kazuhiro Ohnishi: Lifetime Employment Contract and Quantity Competition with Profit-Maximizing and Joint-Stock Firms 462–478

James H. Love: Opportunism, Hold-Up and the (Contractual) Theory of the Firm 479–501

Joachim Thøgersen: Unemployment, Public Pensions, and Capital Accumulation: Assessing Growth Effects of Alternative Funding Strategies 502–520

Juha-Pekka Niinimäki: Liquidity Creation without Bank Panics and Deposit Insurance 521–547

Hun-Chang Lee and Peter Temin: The Political Economy of Preindustrial Korean Trade 548–571



Mohr Siebeck

Modelling the Transaction Role of Money and the Essentiality of Money in an Explosive Hyperinflation Context

by

ALEXANDRE SOKIC*

This paper shows firstly that the cash-in-advance model presents exactly the same kind of limitations as the money-in-the-utility-function model for characterizing explosive hyperinflation. These limitations relate to sufficient money essentiality in the sense of SCHEINKMAN [1980]. Thereby this paper departs from GUTIERREZ AND VAZQUEZ [2004] and contributes to the understanding of failure of the Cagan inflationary finance models with perfect foresight. Secondly, it shows that the inclusion of the goods market equilibrium condition calls into question the validity of explosive hyperinflationary paths as equilibrium paths in the cash-in-advance model. Two solutions are proposed to save the validity of explosive hyperinflation paths. (JEL: E 31, E 41)

1 Introduction

Explosive hyperinflation¹ is not possible under perfect foresight in inflationary finance models based on CAGAN [1956].² Therefore, recent analytical studies abandon Cagan's money demand function and consider optimizing monetary models. GUTIERREZ AND VAZQUEZ [2004], henceforth called GV, follow this approach with the aim of characterizing agents' preferences compatibly with explosive hyperinflation "where an economy is just in a high-inflation scenario, there is perfect foresight, money demand depends only on inflation, and money is essential" (p. 314).

* Ecole Supérieure du Commerce Extérieur, Paris. I would like to thank Elmar Wolfstetter and two anonymous referees for their comments and suggestions, which significantly improved the paper. I am also very grateful to Jesús Vázquez for helpful discussions. Additionally, many thanks go to Meixing Dai, Michel Dévoluy, Gilbert Koenig, Gérard Lang, and Kirsten Ralf for useful comments and discussions.

¹ This paper is not about speculative hyperinflations, which are the focus of other works such as OBSTFELD AND ROGOFF [1983] or BARBOSA AND CUNHA [2003], for instance. Speculative hyperinflations, as defined by OBSTFELD AND ROGOFF [1983], are explosive price-level paths unrelated to monetary growth.

² EVANS [1995] and VAZQUEZ [1998] provide a survey of the literature concerning this failure.

GV consider two standard optimizing monetary models representing alternative ways of modelling the transaction role of money: a cash-in-advance model and a money-in-the-utility-function model. In both models they use specific agents' preferences represented by a constant-relative-risk-aversion utility function and show that standard maximizing models are consistent with explosive hyperinflationary dynamics. However, GV's analysis is that "the money-in-the-utility-function model presents more limitations than the simple cash-in-advance model for characterizing hyperinflation as an explosive process" (GUTIERREZ AND VAZQUEZ [2004, p. 324]). That leads those authors to point to the conclusion that the basic cash-in-advance model is a "natural framework" and a "sensible approach" to study hyperinflation (p. 323).

This paper extends GV's cash-in-advance model by considering agents' preferences represented by a general class of utility functions and taking into account the goods market equilibrium condition. The aim of the paper is twofold. Firstly, it shows that the cash-in-advance model presents exactly the same kind of limitations as the money-in-the-utility-function model for characterizing explosive hyperinflation. These limitations relate to sufficient money essentiality in the precise and formal sense of SCHEINKMAN [1980]. In this respect, this paper departs from GV but contributes to the understanding of the well-known failure of the Cagan inflationary finance models with perfect foresight by providing guidance about the choice of functional form of money demand for the analysis of explosive hyperinflation. Secondly, the paper shows that the inclusion of the goods market equilibrium condition, not taken into account in GV, calls into question the validity of explosive hyperinflationary paths as equilibrium paths in the cash-in-advance model. Two possible solutions are proposed to save the validity of explosive hyperinflation equilibrium paths in the cash-in-advance economy.

The paper is organized in the following way. Section 2 focuses on an extended setup of GV's cash-in-advance economy and provides a general characterization of agents' preferences compatible with explosive hyperinflation, relying on the formal concept of money essentiality. Section 3 deals with the consistency of explosive hyperinflation paths in the cash-in-advance model with the goods market equilibrium condition. Section 4 concludes and proposes further research tracks.

2 *Cash-in-Advance Economy, Hyperinflation, and Money Essentiality*

We focus on an extended setup of the cash-in-advance continuous-time model presented in GV where the economy consists of a large number of identical infinite-lived forward-looking households endowed with perfect foresight. The population is constant, and its size is normalized to unity for convenience. There is no uncertainty. Each household has a nonproduced endowment $y_t > 0$ of the nonstorable consumption good per unit of time.

The representative household's preferences are represented by a utility function depending only on per capita real consumption c_t . The household utility at time 0 is

$$(1) \quad \int_0^{\infty} e^{-rt} U(c_t) dt.$$

GV's setup is extended by considering a general class of utility functions U of which the constant-relative-risk-aversion (CRRA) utility function is a particular case. It is increasing and strictly concave in its single argument, real good consumption. Following GV, the rate r is the subjective discount rate, which is assumed to be equal to the real rate of interest. Financial wealth and the nominal interest rate are defined as

$$w_t = m_t + b_t,$$

$$i_t = r + \pi_t,$$

respectively, where $m_t = M_t/P_t$ is real per capita monetary balances (M is the per capita nominal stock of money; P is the price level), b_t denotes real per capita government debt, and π_t is the inflation rate. The household's budget constraint is

$$(2) \quad \dot{w} = y_t - \tau_t + rw_t - (c_t + i_t m_t),$$

where τ_t is a lump-sum tax assumed to be constant. In a cash-in-advance economy the role of money as a medium of exchange is captured by a cash-in-advance constraint: the assumption that money holding is strictly essential to buying the consumption good. In order to consume c units of the consumption good at time t , the household must hold a stock of real cash balances, m , greater than or equal to c :

$$m_t \geq c_t.$$

Assuming the existence of an interior solution for c , and that the nominal interest rate i is greater than zero (meaning that money is return-dominated by government bonds), it follows that the cash-in-advance constraint must hold with equality. Following GV, we have

$$(3) \quad m_t = c_t.$$

The representative household's optimization problem, consisting of maximizing (1) subject to the constraints given by (2) and (3), leads to the following first-order condition:

$$(4) \quad U'(m_t) = \lambda(1 + i_t),$$

which is the general expression of equation (9) in GV. The associated Lagrange multiplier λ is constant with respect to time because the agent's rate of time preference equals the real rate of interest, and real cash balances will indirectly enter the utility function according to (3). Equation (4) characterizes a demand for real money balances decreasing in the rate of inflation (or the cost of holding cash balances), because the utility function U is strictly concave. The optimum solution is completed by the transversality condition:

$$(5) \quad \lim_{t \rightarrow \infty} e^{-rt} \lambda w_t = 0.$$

Using the definition of the nominal interest rate, the first-order condition (4) can be rewritten as follows:

$$(6) \quad \pi_t = \frac{U'(m_t) - \lambda(1+r)}{\lambda},$$

which is the general expression of equation (11) in GV. Following GV, in usual inflationary finance models a constant per capita share of government's budget deficit, d , is financed by issuing high-powered money:

$$(7) \quad d = \frac{\dot{M}_t}{P_t} = \dot{m}_t + \pi_t m_t.$$

Substituting the value of π given by equation (6) in the latter expression leads to the inflationary finance model dynamics described by the following law of motion for real cash balances:

$$(8) \quad \dot{m}_t = d - \frac{1}{\lambda}(U'(m_t) - \lambda(1+r))m_t.$$

The differential equation (8), which is the general expression of the law of motion for real money balances represented in Figure 3 in GV, provides a complete characterization of real per capita money balance dynamics, which will be studied by using the technique of phase diagrams on $[0; +\infty[$. The main interesting point here is to examine whether this law of motion for real cash balances is able to produce hyperinflation paths. An explosive hyperinflation path will be observed if the law of motion presents a path leading to a zero level of real cash balances. Therefore, the conditions for this kind of paths should be identified. As the mathematical function representing the law of motion is continuous (which follows from standard assumptions on U), this kind of paths will be observed as long as (dropping the time index for convenience)

$$(9) \quad \lim_{m \rightarrow 0_+} \dot{m} < 0.$$

The calculation of $\lim_{m \rightarrow +\infty} \dot{m}$ will assess the existence of any steady state. Nevertheless, whatever the number of steady states, since we focus on possible explosive hyperinflation paths, we are only interested in the paths starting to the left of the first possible steady state when the condition $\lim_{m \rightarrow 0_+} \dot{m} < 0$ is met.

At this stage a second highly important point should be made clear. According to OBSTFELD AND ROGOFF [1983], in the context of speculative hyperinflations, any path leading to a zero value of real cash balances and crossing eventually the vertical axis at some finite point should be ruled out on grounds that such paths would not be feasible because the real stock of money would eventually become negative. However, we would rather follow the point made by BARBOSA AND CUNHA [2003], who contested OBSTFELD AND ROGOFF's [1983] approach by arguing that on such hyperinflationary paths "when the real quantity of money reaches zero hyperinflation has wiped out the value of money and the opportunity cost of holding money has become infinite" and "the economy is no longer a monetary economy" (BARBOSA AND CUNHA [2003, p. 192]). Therefore, we follow the point made by BARBOSA

AND CUNHA [2003] and consider the explosive hyperinflation paths corresponding to the condition $\lim_{m \rightarrow 0_+} \dot{m} < 0$ as relevant perfect-foresight paths.

Moreover, it is important to stress that the possible explosive hyperinflationary paths are explosive monetary hyperinflations, because along these paths the rate of growth of the money supply explodes. Rewriting the government budget constraint as

$$\frac{\dot{M}}{M} = \frac{d}{m},$$

we see that along the paths of continuously declining m , given that $d > 0$, the growth rate of the money supply increases continuously.

In this respect, according to the law of motion (8), the possibility of explosive hyperinflation in the cash-in-advance economy will depend on the condition

$$(10) \quad \lim_{m \rightarrow 0_+} [mU'(m)] > \lambda d.$$

The latter condition is basically a condition about a sufficient level of money essentiality. In the sense of SCHEINKMAN [1980], money is considered as essential if the inflation tax collected by the government does not tend to zero when the rate of inflation explodes. The interpretation is that “no matter how expensive it becomes to hold money people still hold a large quantity of it; that is money is very necessary to the system” (SCHEINKMAN [1980, p. 96]). From (7) we see that seigniorage obtained by printing money can be decomposed into two components: the change in the real stock of money, and the inflation tax πm , which can be written, according to equation (6),

$$\pi m = \left(\frac{U'(m) - \lambda(1+r)}{\lambda} \right) m.$$

Then, when the rate of inflation explodes, we consider the following limit:

$$\lim_{m \rightarrow 0_+} [\pi m] = \frac{1}{\lambda} \lim_{m \rightarrow 0_+} [mU'(m)].$$

Therefore, if $\lim_{m \rightarrow 0_+} [mU'(m)] > 0$ then $\lim_{m \rightarrow 0_+} \pi m > 0$ and money is essential. These findings point to the following proposition.

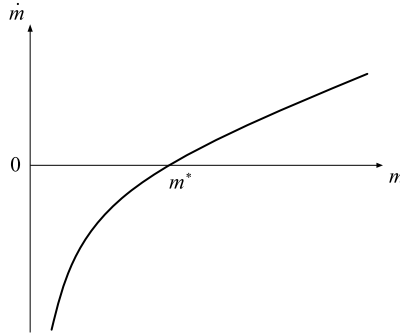
PROPOSITION (MONEY ESSENTIALITY PROPOSITION) *In a cash-in-advance economy with a general class of utility functions, explosive hyperinflations are possible only if money is sufficiently essential, that is, if $\lim_{m \rightarrow 0_+} [mU'(m)] > \lambda d$.*

PROOF The proof relies on the previous arguments and can be illustrated by the phase diagram depicted in Figure 1. The precise shape of the phase diagram depends on the first and second derivatives of \dot{m} with respect to m and thus depends on the precise choice of the utility function U . Other shapes than that depicted on Figure 1 could be chosen for the phase locus. However, as the important point for the analysis conducted here is the condition for $\lim_{m \rightarrow 0_+} \dot{m} < 0$, our analysis focuses only on the paths leading to a zero value of real cash balances. If $\lim_{m \rightarrow +\infty} \dot{m} > 0$, the locus \dot{m} will cross the horizontal axis at least once. We consider here a unique unstable steady state m^* , but the qualitative analysis for explosive hyperinflationary paths

does not change in the case of more steady states. All paths originating to the right of m^* are hyperdeflationary paths that can be ruled out because they violate the transversality condition (5). All paths starting to the left of m^* are explosive monetary hyperinflations paths.

Figure 1

Monetary Dynamics in a Cash-in-Advance Economy with Sufficient Money Essentiality



Q.E.D.

It should be noticed that the same kind of result can be obtained in the corresponding money-in-the-utility-function model with a general class of utility functions $U(c, m)$, where it can easily be shown, using the same methodology developed above, that the possibility of explosive hyperinflation paths depends on the condition

$$\lim_{m \rightarrow 0} \left[\frac{U'_m(c, m)}{U'_c(c, m)} m \right] > d,$$

which is the equivalent of the condition (10) in the cash-in-advance model. In the same way as previously done in this section, the latter condition can again be shown as corresponding to a sufficient level of money essentiality in SCHEINKMAN's [1980] formal sense.

Figure 1, which is equivalent to Figure 3 in GV, represents the phase diagram corresponding to GV's choice for a specific CRRA utility function with relative risk aversion parameter greater than one:

$$(11) \quad U(c) = \frac{c^{1-\alpha} - 1}{1 - \alpha} \quad \text{with } \alpha > 1.$$

This choice represents agent's preferences compatible with the possibility of explosive hyperinflation. It complies³ with the condition (10), as it leads to $\lim_{m \rightarrow 0+} [mU'(m)] = +\infty > \lambda d$. Therefore, the interesting results obtained in GV

³ The particular choice of (11) to represent agents' preferences does not need the argument made by BARBOSA AND CUNHA [2003] to consider explosive hyperinflation paths. But this argument remains necessary to deal with the more general case. GV cannot avoid the issue of dealing with a finite $\lim_{m \rightarrow 0} \dot{m} < 0$ for characterizing explo-

do not rely on the specificity of the cash-in-advance framework, but rather on a sufficient level of money essentiality, which is conveyed by the choice of the utility function given by (11).

3 *Cash-in-Advance Model Dynamics and the Goods Market Equilibrium Condition*

The GV framework is further completed by considering the equilibrium condition in the goods market. Following BARBOSA, CUNHA, AND SALLUM [2006] or VAZQUEZ [1998], “in the spirit of the traditional approach to the study of hyperinflationary phenomena, we assume that output and government expenditures are constant” (VAZQUEZ [1998, p. 438]). Therefore, the market for goods is in equilibrium when the constant supply of the good (y) equals the household consumption plus the constant per capita government expenditures (g):

$$(12) \quad y = c_t + g.$$

Explosive hyperinflation paths in the cash-in-advance economy raise an important issue not taken into account in GV. According to the cash-in-advance constraint (3), household real consumption will fall along explosive hyperinflation paths characterized by the declining value of real money balances. The fall of households’ real consumption will cause an increasing loss of welfare and represent the harmful effect of hyperinflation on the cash-in-advance economy. There is some evidence supporting this result. As pointed out by VAZQUEZ [1998], WEBB [1989] in his Table 5.4 shows evidence that consumption fell dramatically during the German hyperinflation. For instance, the consumption of butter, meat, and sugar fell to 5%, 39%, and 3% of the levels of consumption in 1913, respectively. Moreover, BRESCIANI-TURRONI [1937] describes how certain classes were hit by poverty during German hyperinflation.

Nevertheless, the goods market equilibrium condition (12) calls into question the validity of explosive hyperinflation paths as equilibrium paths in the cash-in-advance economy. According to the equilibrium condition (12), household real consumption c should be constant at the level $c = y - g$ because endowment in the nonstorable good is constant at level y and per capita government expenditures are constant at level g . Thus, any explosive hyperinflation path in the cash-in-advance economy does not comply with goods market equilibrium condition and should be therefore ruled out as not being an equilibrium path. That would seriously affect results obtained in GV.

Two solutions may be proposed to ensure the validity of explosive hyperinflation paths in the cash-in-advance economy. First, we could imagine that transactions not taking place in the monetary economy because of the declining value of real

sive hyperinflation in the particular money-in-the-utility-function model with a CRRA utility function. However, adopting the argument of OBSTFELD AND ROGOFF [1983], they impose an *ad hoc* restriction, admitting it as “not easy to motivate in the context of a money-in-the-utility-function model” (GUTIERREZ AND VAZQUEZ [2004, p. 323]), in order to prevent the \dot{m} schedule from eventually crossing the vertical axis.

cash balances might be offset by increasing resort to unofficial barter in the grey economy. Total per capita real consumption at time t could then be split into two components $c_t = c_{1t} + c_{2t}$, where c_{1t} would represent consumption constrained by holding of real cash balances, and c_{2t} would represent real consumption achieved through unofficial barter in the grey economy. In that case, the equilibrium condition on the goods market could be written as $y = c_{1t} + c_{2t} + g$. Then, along an explosive hyperinflation path, c_{1t} would decrease along with the real value of money balances and c_{2t} would increase consistently with goods market equilibrium, meaning that more and more transactions would be performed in the grey zone. Eventually, the cash-in-advance monetary economy would collapse and switch entirely to an unofficial barter grey economy. Second, we could imagine that the fall of real consumption might cause a fall of goods supply, leading eventually to the collapse of the economy. Then, at each time t , the goods supply y_t would adjust to the falling household real consumption c_t so that $y_t = c_t + g$. Recent evidence provided by the collapsing Zimbabwean economy may support this possibility (see Table 1). Zimbabwean real GDP has registered a drop of more than 40% since 1999.

Table 1
Economic Crisis and Inflation during Zimbabwean Hyperinflation

Year	2002	2003	2004	2005	2006	2007
GDP growth at constant prices (annual percent change)	-4.370	-10.363	-3.557 ^e	-3.953 ^e	-5.422 ^e	-6.092 ^e
Inflation, average consumer prices (annual percent change)	133.215	365.046	349.988	237.817	1 016.683	10 452.555 ^e

Source: IMF [2009].

Note: ^e IMF estimates.

4 Conclusion

The first result of this paper is that the possibility of explosive hyperinflation paths in an extended cash-in-advance model depends on a sufficient level of money essentiality defined in the formal sense of SCHEINKMAN [1980]. In that respect we depart from GV by showing that the cash-in-advance model presents exactly the same kind of limitations as the money-in-the-utility-function model for characterizing explosive hyperinflation paths. The cash-in-advance constraint does not convey by itself sufficient money essentiality, even if it makes money necessary for the transactions. The sufficient level of money essentiality is conveyed by the representative agent's

preferences.⁴ Therefore, this paper may contribute to the understanding of the well-known failure of Cagan inflationary finance models with perfect foresight. It may stimulate further research on the choice of an appropriate demand for real cash balances in hyperinflation contexts, for which microeconomic foundations should comply with the money essentiality requirement.

Another contribution of this paper is to complete GV in an important respect by considering the goods market equilibrium condition. That this major part is not taken into account by GV calls into question the validity of explosive hyperinflationary paths as equilibrium paths in the cash-in-advance model. Actually, considering the goods market equilibrium leads one to rule out explosive hyperinflation paths on the ground that they are not equilibrium paths. Two alternative solutions have been put forward to try to save the validity of explosive hyperinflation paths in the cash-in-advance economy. They imply either increasing resort to unofficial barter in the grey economy or the progressive collapse of the supply side of the monetary economy. Further research could be conducted to better integrate these possible solutions in the model.

References

- BARBOSA, F. H., AND A. B. CUNHA [2003], "Inflation Tax and Money Essentiality," *Economics Letters*, 78, 187–195.
- , —, AND E. M. SALLUM [2006], "Competitive Equilibrium Hyperinflation under Rational Expectations," *Economic Theory*, 29, 181–195.
- BRESCIANI-TURRONI, C. [1937], *The Economics of Inflation: A Study of Currency Depreciation in Post-War Germany 1914–1923*, George Allen & Unwin: London.
- CAGAN, P. [1956], "The Monetary Dynamics of Hyperinflation," pp. 25–117 in: M. Friedman (ed.), *Studies in the Quantity Theory of Money*, University of Chicago Press: Chicago.
- EVANS, J. L. [1995], "The Demand for Money: Hyperinflation or High Inflation Traps," *The Manchester School of Economic & Social Studies*, 63 (supplement), 49–56.
- GUTIERREZ, M. J., AND J. VAZQUEZ [2004], "Explosive Hyperinflation, Inflation Tax Laffer Curve and Modelling the Use of Money," *Journal of Institutional and Theoretical Economics*, 160, 311–326.
- IMF [2009], *World Economic Outlook Database*, April 2009, International Monetary Fund: Washington, D.C.
- OBSTFELD, M., AND K. ROGOFF [1983], "Speculative Hyperinflations in Maximizing Models: Can we Rule them Out?" *Journal of Political Economy*, 91, 675–687.
- SCHEINKMAN, J. [1980], "Discussion," pp. 91–96 in: J. Kareken and N. Wallace (eds.), *Models of Monetary Economies*, Federal Reserve Bank of Minneapolis: Minneapolis.

⁴ The requirement of sufficient money essentiality is relevant for the analysis of hyperinflationary paths beyond technical arguments. As pointed out by GV, money becomes more essential for purchasing goods during hyperinflation than during stable periods "because extreme inflation dramatically decreases credit transactions and in general the use of long term contracts" (GUTIERREZ AND VAZQUEZ [2004, p. 312]). Moreover, a sufficient level of money essentiality is crucial in inflationary finance models of hyperinflation, since the government needs the money to be essential to the system in order to get a sufficient inflation tax when inflation explodes.

- VAZQUEZ, J. [1998], "How High Can Inflation Get during Hyperinflation? A Transaction Cost Demand for Money Approach," *European Journal of Political Economy*, 14, 433–451.
- WEBB, S. B. [1989], *Hyperinflation and Stabilization in Weimar Germany*, Oxford University Press: New York.

Alexandre Sokic
Ecole Supérieure du Commerce Extérieur
92916 Paris La Défense
France
E-mail:
alexandre.sokic@esce.fr

A Theory of Innovation: Market Transition, Property Rights, and Innovative Activity

by

VICTOR NEE, JEONG-HAN KANG, AND SONJA OPPER*

The aim of this paper is to specify a theory to explain why transitions to a market economy cause a shift to a higher level of innovation. Marketization increases the power of economic actors relative to political actors, increases inter-firm competition, creates new opportunities for entrepreneurship, and subsequently motivates innovative activity. For our empirical application, we focus on China's transition economy, which offers a broad range of institutional environments to examine the relation between market transition and increasing innovative activity by entrepreneurs and firms. (JEL: O 31, P 31, P 3)

1 Introduction

The innovative process – SCHUMPETER's [1942] “perennial gale of creative destruction” – is the recognition of opportunities for profitable change and the pursuit of those opportunities all the way through until they are put into business practice. For Schumpeter, the entrepreneur – distinct from the capitalist and businessman – is the purveyor of innovations. For Marx, in contrast, innovation is a systemic feature of the underlying competitive dynamics of market capitalism. This view of innovation as an outgrowth of the ferocity of competitive pressures on capitalists has attracted new attention in the research on innovation. Insofar as innovation is a social process involving cooperation and competition within a larger institutional structure, incentives are matters not only of individual-level motives and decisions, but also of that institutional framework (SCHUMPETER [1912/1934], BAUMOL [2002]).

* Cornell University (corresponding author), Yonsei University – Seoul, and Lund University. This paper has been presented at the Annual Meeting of the International Society for New Institutional Economics at Reykjavik, Iceland, July 2007; the Macroeconomic Theory Seminar of the Department of Economics at Cornell University, August 2007; and the Conference on Capitalism and Entrepreneurship sponsored by the Center for the Study of Economy and Society at Cornell University, September 2007. We thank Henry Wan Jr., Kaushik Basu, Robert Braun, Brett de Bary, Michael Hannan, Patrick Park, Karl Shell, Henry Smith, Richard Swedberg, and Elmar Wolfstetter for helpful comments on an earlier draft.

This paper builds on the supposition that institutions matter in explaining innovative activity. Its core argument extends NEE's [1989] market transition theory to explain the rise of innovative activity (Propositions 1, 2, 4) and refines BAUMOL's [1990] supposition that the most effective way to stimulate productive entrepreneurial activity is to diminish relative rewards to unproductive or destructive rent-seeking and increase payoffs to productive entrepreneurial activity (Propositions 3, 5). By linking a theory of endogenous emergence of markets and entrepreneurial activities with Baumol's ideas on competition and innovation, our approach shifts the focus towards specifying the features of the institutional framework that enable, motivate, and guide innovative activity. We assert that the innovation literature neglects the role of real markets when it comes to the analysis of motivation and capability to innovate.

Our theory specifies the effect of marketization in transition economies on the relative payoffs to unproductive and productive entrepreneurial activity, and derives testable hypotheses. It is in transition economies where one finds a wide range in variability in the extent and scope of markets, which allows us to examine the effect of property rights and markets on entrepreneurial action as measured by innovative activities. Hypotheses derived from our model underscore the capacity of private enterprise to innovate, explain innovation as outgrowth of the competitive pressures on firms, and highlight the role of networks in regional clustering of innovation linking universities and research institutes with firms. Our empirical application focuses on China's transition economy as a strategic research site.

The remainder of the paper is organized as follows: The next section develops the theoretical framework explaining the shift of the reward structure as a consequence of economic transition. Section 3 derives our hypotheses predicting a close linkage between market transition, property rights, and innovation. Section 4 provides a summary account of market transition in China as our strategic research site and section 5 specifies data and method. Section 6 confirms that a principal cause of China's shift to innovation as a source of economic growth is the emergence of competitive markets. The final section concludes.

2 *Theoretical Framework*

We proffer propositions that specify elementary mechanisms embedded in markets as social institutions explaining the propensity to innovate (1–3), and then we lay out propositions (4–5) that explain increasing rate of innovations arising from market transition.

In market economies, the rules of the game of private property rights and decentralized markets provide powerful incentives for economic actors to innovate. Whether innovative activity is for the sake of the fruits of success, or for success itself, in price-making markets rewards are based on the competitive sorting and matching of quality and price. It is thus the restoration of consumer and producer sovereignty in transition economies, which activates incentives to innovate.

For convenience, we apply ROSEN's [1974] theory of hedonic prices to justify our first three propositions. Rosen's model of product differentiation – based on the hedonic assumption that goods are valued for their utility bearing attributes – illustrates how buyer and seller choices determine competitive equilibrium in a multi-dimensional plane. In fact, market pressure and subsequent innovation processes are the causal processes behind the illustrated product differentiation.

A class of goods is described by a vector of n measured characteristics $z = (z_1, z_2, \dots, z_n)$, where z_i measures the quantity of the good's product quality i . Products of a given class are thus described by distinct packages of z . Typically a spectrum of differentiated products will be available to choose from. Each product is associated with a market price $p(z) = p(z_1, z_2, \dots, z_n)$, which guides consumer and producer choices. $P(z)$ thus represents the minimum market price for a given package of product qualities.

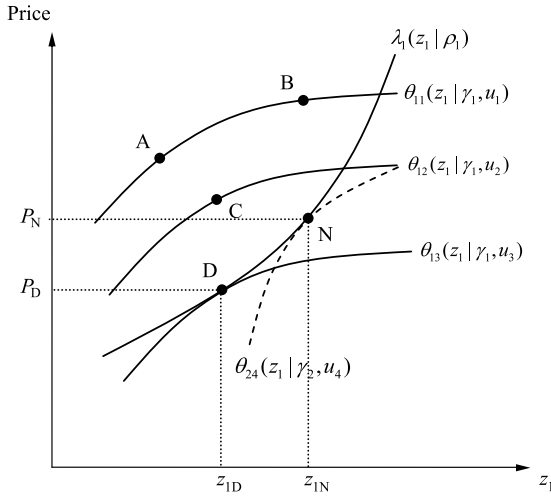
To identify the consumption decision, assume that the expenditures a consumer is willing to pay at a given consumer's taste and utility index is given by the value function $\theta(z; \gamma, u)$. Figure 1 depicts $\theta(z; \gamma, u)$ in z_1 given specific consumers' taste γ_i and utility level u_k (i.e., $\theta = \theta_{ik}(z_1|\gamma_i, u_k)$), holding constant other elements of vector z . θ_{11} , θ_{12} , and θ_{13} represent three utility levels ($u_1 < u_2 < u_3$) assuming a consumer type with taste γ_1 . Further, the producers' offer function given profit level ρ is $\lambda(z; \rho)$. Evidently utility is maximized in D, where $\theta(z; \gamma, u)$ and $\lambda(z, \rho)$ are tangent to each other. As customers prefer D's price- z_1 -offer over the three alternatives, firms A, B, and C have incentives to innovate by either lowering costs or adjusting the quantity of z_1 . Trivially, D will be in a short-term equilibrium at the tangential point of offer λ_1 and value function θ_{13} as long as consumer taste remains unchanged. In sum,

PROPOSITION 1 *Markets offer incentives to innovate insofar as rewards for performance depend on a match of quality and price or a match of cost and price.*

The entrepreneur-as-innovator is the person who pursues opportunities that others forgo. Suppose firms in a market sector, say cell-phones, face perfect competition so that the equilibrium prices of products provide only razor-thin margins. An entrepreneurial action in this setting would be to innovate by coming up with a new product based on the hunch that its novel features will break out of the standard mold and fetch a higher price. Our entrepreneur has accordingly purposively sought an opportunity that other firms in the industry have either implicitly forgone or have assessed as too costly to pursue. In this view the opportunity cost for forgoing investments in innovative activity is the *hidden cost* of firms pursuing the established business patterns and practices, which in our example locks them into a stable price structure. The market mechanism offers means to assess the potential costs and benefits from an innovation (HAYEK [1978]).

Consider the price- z_1 combination N in Figure 1. Given consumer taste γ_1 , the price-quality combination of good D would be preferred over N, but N in turn promises higher utility to a new latent customer class with taste $\gamma_2[\neq \gamma_1]$, here illustrated by a value curve $\theta_{24}(z_1|\gamma_2, u_4)$. A firm's ability to realize N rests on

Figure 1
Incentives and Opportunity to Innovate



the identification of new consumer preferences (i.e., consumer-differentiation). In our illustration, the new consumer class can be reached by offering a package of characteristics z , which includes a higher quantity of z_1 .

PROPOSITION 2 *The emergence of markets endogenously expands the opportunities for entrepreneurs and firms to identify new markets and prospects for profit-making.*

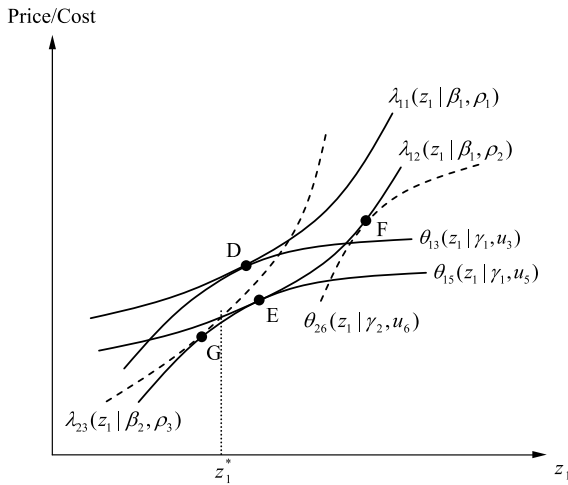
Market capitalism provides powerful incentives for innovation through the ferocity of competitive pressures (BAUMOL [2002]). The competition effect can be conveniently illustrated by incorporating multiple offer curves into the value map (Figure 2). A key consideration here is the differentiation of suppliers by introducing parameter β to offer curves (i.e., $\lambda = \lambda_{ji}(z_1 | \beta_j, \rho_i)$). β reflects inter-firm differences in factor prices and, more importantly, in “technology” and “R&D expenditure” (ROSEN [1974, p. 43]).

Given an offer curve $\lambda_{11}(z_1 | \beta_1, \rho_1)$ and a value curve $\theta_{13}(z_1 | \gamma_1, u_3)$ with an equilibrium offer D, market entry of new producers with the same price–quality offer will intensify competition. Subsequently, profit erodes and the offer curve shifts downward to $\lambda_{12}(z_1 | \beta_1, \rho_2)$ with $\rho_2 < \rho_1$ whereas consumer utility increases to $u_5 > u_3$. Eventually, competition will drive the offer curve down to the cost curve, which signals the technological frontier for producing a good with a given amount of product characteristic z_1 . To escape the competitive pricing situation, producers have to innovate: (1) By means of cost-saving innovations, firms at E may be able to shift the offer curve further down, and enjoy the higher profit margin until rival firms

discover a similar or even better technology. (2) Alternatively, firms can move to F to reach a new customer class which prefers a different quantity of product feature z_1 (as explained in Proposition 2). (3) Finally, firms can extend existing customers' choice options by reshaping the firms' cost structure (i.e., different β) without necessarily reducing overall costs. In Figure 2, a shift of the cost condition parameter β_1 to β_2 would shift the offer curve from $\lambda_{12}(z_1|\beta_1, \rho_2)$ to $\lambda_{23}(z_1|\beta_2, \rho_3)$. λ_{23} now has a cost advantage over λ_{12} as long as z_1 does not exceed z_1^* and the offer curve λ_{23} and value curve $[\theta_{15}]$ yield a new equilibrium G. The second and the third types of innovation respectively represent the process of consumer-differentiation and producer-differentiation. In sum,

PROPOSITION 3 *Irrespective of the distinct innovation type, the greater the market competition, the more firms are compelled to innovate.*

Figure 2
Innovation and New Combinations



The more both buyers and sellers differentiate and the more innovations made in accordance to the differentiation, the more equilibrium points are emerging crossing E, F, and G in Figure 2. Those points or an “envelope” signals as implicit prices or “hedonic” prices against which both buyers and sellers adjust their offer or value curves. It is the implicit prices or the “market” that intermediate between buyers and sellers (ROSEN [1974, p. 36]). In our terms, marketization is an effective mechanism to enhance innovation by realizing various Schumpeterian “new combinations” of γ , β , and $z = (z_1, z_2, \dots, z_n)$.

2.1 The Power Proposition

In state socialist economies, the political actors – party officials and bureaucrats – held monopoly power over the allocation of scarce resources. The emergence and growth of a decentralized market economy necessarily involves reducing the scope of state controls over resource allocation, hence diminishing the redistributive power of political actors, while economic actors – firms and entrepreneurs – gain power insofar as market transactions are based on voluntary agreement between buyers and sellers (NEE [1989]). Moreover, the shift to market allocation causes changes in relative rewards that reduce the payoffs for unproductive rent-seeking and offers incentives and opportunities for economic actors to engage in productivity-enhancing innovative activity.

PROPOSITION 4 *Market transition diminishes the relative power of political actors and empowers economic actors – firms and entrepreneurs.*

Assume that a firm can generate additional revenue through economic or political sources:

$$(1) \quad T_j = C\pi_j + P\phi_j,$$

where T_j is firm j 's total expected payoff, C expected revenue from competitive advantage through innovation (formalized in Figures 1 and 2), P expected rents from political sources, π_j (with $0 \leq \pi_j \leq 1$) firm j 's probability to realize additional revenue from innovation, and finally ϕ_j (with $0 \leq \phi_j \leq 1$) firm j 's probability to generate rents from political sources. In this model, a firm's expected pay-off T_j is determined as a linear combination between given structural parameters of the market (i.e., C and P) and firm-level parameters (i.e., π_j and ϕ_j). Note that we do not specify distinct market structures and resulting prices and costs, as our aim is not to model the market. For simplification we assume only one period and assume that failed efforts to pursue either innovation or political advantages yield no pay-off.

Further, the power proposition implies that expected innovation gains (C) and political rents (P) are functions of market transition with

$$(2) \quad C'(m) > 0 \text{ and } P'(m) < 0,$$

where m is the degree of marketization. In other words, the firm's income generation moves away from resources controlled by the state – political funds P – to income generated by innovative activities C as market transition changes the relative payoffs of unproductive and productive entrepreneurship.

We further assume that firm's probability to generate income from innovation (i.e., π_j) or political funds (i.e., ϕ_j) is a positive function of investment (including capital, time, skills, and efforts) in innovation I_{ji} (R&D activities) or politics I_{jp} (rent-seeking activities) by firm j :

$$(3) \quad \partial_{I_{ji}} \pi_j > 0 \text{ with } \pi_j|_{I_{ji}=0} = \pi_0 \quad \text{and} \quad \partial_{I_{jp}} \phi_j > 0 \text{ with } \phi_j|_{I_{jp}=0} = \phi_0,$$

where I_{ji} and I_{jp} are constrained by the investment budget B_j :

$$(4) \quad I_{ji} + I_{jp} = B_j.$$

Note in condition (3) that with no investment in innovation or politics (i.e., with $I_{ji} = 0$ or $I_{jp} = 0$), a firm's probability to generate income from innovation or political funds is assumed to reach a lower bound (i.e., π_0 or ϕ_0) which is independent of any firm characteristic j . Those lower bounds may be regarded zero for convenience. Also note that for convenience there are no financing costs.

2.2 The Politics Proposition

Whether through informal or formal arrangements, the reward structure for political actors is skewed to encourage the pursuit of innovative rent-seeking rather than productivity enhancing innovations (BAUMOL [1993]). In state socialist economies, the structure of incentives did not reward managers for innovating (SHLEIFER AND VISHNY [1994]). Given annual assignments of production quotas, managers bargained for more appropriations and lower production quotas. In other words, the payoff matrix rewarded managers with positional advantage and connections with politicians. Government bureaucracies lack the commitment to hard budget constraints, and hence the capacity for effective *ex post* screening required for divesting from innovation projects that are not viable (QIAN AND XU [1998]). For this reason, bureaucrats tend to rely on *ex ante* screening, which results in rejecting promising projects and funding fewer numbers of projects, especially those involving higher uncertainties and less research in the initial stages of development.

Our previous proposition assumed for convenience fixed firm-level probabilities of achieving innovation (π_j) or political advantages (ϕ_j), given an investment allocation between innovation (I_i) and politics (I_p). This simplification overlooks the critical role reward structures play in shaping economic activities and subsequent effectiveness for realizing innovation. Particularly the difference between economic actors, as profit maximizers, and political actors, who typically pursue multiple goals (such as employment generation and realization of social stability) may have a critical impact on the incentive to innovate. In line with earlier studies (HART, SHLEIFER, AND VISHNY [1997]), we claim that involvement of political actors may dilute both incentives and opportunities for productive entrepreneurship because it tends to skew the structure of rewards towards unproductive rent-seeking (WAN [2003]). We predict that firms respond to variability in the involvement of political actors by either strengthening competitive position through innovative activity or political advantage.

PROPOSITION 5 *When political actors are empowered to allocate resources in firms there are fewer innovations and more delays in bringing innovation projects to new products.*

The politics proposition implies that the marginal increase in the probability of successful innovation by a unit increase in investment (i.e., $\partial_{I_i}\pi$) depends on the

extent of involvement by political actors in the governance of the firm. While political control and involvement is usually hard to measure, state interference typically builds on the extent of government ownership of the firm. SHLEIFER AND VISHNY [1994], for instance, develop a formal model, where privatization limits the involvement of political actors in a firm's decisions. Based on the close connection between ownership form and involvement of politicians in the firm's governance, we introduce a as the proportion of private ownership in the politics proposition:

$$(5) \quad \partial_{a I_i} \pi > 0.$$

Condition (3) specifies a positive effect of investment in innovative activity on the probability of successful innovation (i.e., $\partial_{I_i} \pi > 0$). The politics proposition (condition (5)) further implies that this investment effect will be the stronger the larger the private ownership share a , and the less vulnerable the firm to political intervention. In addition, the politics proposition implies that the marginal increase in the probability of achieving political advantage by a unit increase in investment to politics (i.e., $\partial_{I_p} \phi [= \partial_{-I_i} \phi]$) will decrease with the extent of private ownership of the firm, a :

$$(5a) \quad \partial_{a I_p} \phi < 0 \text{ or } \partial_{a I_i} \phi > 0.$$

Essentially we hold that gains from innovative efforts are on average greater in private firms than in public enterprise.

Notwithstanding, this is not a claim that state-owned enterprises may not play a critical role in innovation. Given ready access to government funding for capital investment, for instance, state-owned enterprises enjoy advantages in sectors, where scale and scope effects lead to lower unit costs. In steel-production, for instance, the two state-owned enterprises Posco (Korea) and Shanghai Baosteel Group (China) rank among the top five steel producers globally. Similarly, China's state-owned enterprises outperform private companies in industries such as ship-building, aircraft, and mobile telephone.

3 Derived Hypotheses

Given the payoff structure in conditions (1) to (5a) and abstracting from financing costs, a firm will choose an optimal allocation of the budget B between two investment types I_i and I_p , so that it maximizes the expected payoff T :

$$(6) \quad \partial_{I_i} T = 0 \text{ and } \partial_{I_i I_i} T < 0.$$

Let I_i^* denote I_i satisfying condition (6). In other words, I_i^* is the optimal level of investment to innovation, for a given level of marketization (m) and a given proportion of private property rights (a). Then, we can deduce

$$(7) \quad \partial_m I_i^* > 0.$$

See Appendix, section A.1, for proof. Hence, for any firm, the optimal investment to innovation increases with market transition. Accordingly, the probability of successful innovation increases with market transition for any firm:

$$(8) \quad \partial_m \pi^* > 0.$$

See Appendix, section A.1, for details. Hence we specify:

HYPOTHESIS 1 *The greater the extent of market transition, the more dependent are firms on innovative activity for survival and profits.*

It is the growth of wealth-maximizing opportunities in competitive markets outside of the state-directed sectors of the transition economy that triggers a shift towards innovation, regardless of ownership form. In this process, the social structure of markets facilitates increasing reliance on regional technical and research cooperation. In specialty niches, entrepreneurs are embedded in multiplex networks of producers, suppliers, and distributors. Because many firms lack the technical capacity and resources to maintain in-house R&D departments, they seek informal and formal partnerships in innovation. In particular, the self-enforcing social structure of markets enables firms to develop close-knit inter-firm networks, in which cooperation helps to lower costs of innovation, facilitates learning, increases legitimacy of novel activities, and alleviates resource constraints (NEE AND OPPER [2010]). As an extension of Hypothesis 1 we specify:

HYPOTHESIS 2 *The greater the extent of market transition, the more developed markets for innovation, the more effective are R&D networks between firms and between firms and research universities and institutes.*

Hypotheses 1 and 2 are general properties independent of the firm's ownership structure. However, from our politics proposition (i.e., condition (5)) we derive that firms under tight political control will be less innovative than independent firms. Formally,

$$(9) \quad \partial_a \pi^* > 0.$$

Thus

HYPOTHESIS 3 *The higher the proportion of private ownership of a firm, the more likely a firm's chances of success in innovation projects (see Appendix, section A.2, for its proof).*

Condition (9) implies that state-owned enterprises will not invest in innovation at a level so that its probability of innovation π^* can keep up with that of private firms. Further examination of condition (6) reveals the underlying reason. First, in order to satisfy the requirement of $\partial_{l_i} T < 0$, we assume concavity of both π and ϕ , that is if investments in innovation projects/political efforts increase the same degree, we expect a *decreasing marginal improvement* in the probability of successful innovation and in that of allocation of political funds:

$$(10) \quad \partial_{l_i} \pi < 0 \text{ and } \partial_{l_p} \phi (= \partial_{l_i} \phi) < 0.$$

This concavity assumption implies an *increasing marginal degeneration* when investments are reduced. Second, $\partial_{I_i} T = 0$ in condition (6) is equivalent to

$$\frac{C(m)}{P(m)} = \frac{-\partial_{I_i} \phi}{\partial_{I_i} \pi}$$

Hence, as market transition proceeds income from innovative efforts $C(m)$ increases relative to income streams from political funds $P(m)$. A firm will rebalance its investment portfolio in favor of innovation projects, until $C(m)/P(m)$ equals the ratio between the marginal decrease in the probability of achieving political funds ($= -\partial_{I_i} \phi$) and the marginal increase in the probability of successful innovation ($= \partial_{I_i} \pi$). When $-\partial_{I_i} \phi / \partial_{I_i} \pi$ is smaller than $C(m)/P(m)$, the firm is under-investing in innovation and would benefit from a reduction of investments in political rents. If in contrast $-\partial_{I_i} \phi / \partial_{I_i} \pi$ is larger than $C(m)/P(m)$, the firm is over-investing in innovation as additional payoffs from innovation projects do not cover forgone income streams that could have been secured from political sources.

For a given investment I_i , the politics proposition (or conditions (5) and (5a)) implies that the marginal increase in the probability of successful innovation ($= \partial_{I_i} \pi$) for a government-owned firm is smaller than that of a private firm while its marginal decrease in the probability of achieving political advantages ($= -\partial_{I_i} \phi$) is larger than that of a private firm. As a result, $-\partial_{I_i} \phi / \partial_{I_i} \pi$ is larger for a government-owned firm than for a private firm. For government-owned firms, the probability of achieving political advantage decreases faster than that of private firms, relative to the probability that successful innovation increases. If the government-owned firm were to invest the same amount as the private firm at a certain level of marketization, the government firm would thus be over-investing in innovation. The lower levels of innovation for government-owned firms are therefore not only rooted in less effective innovative activity, but also attributable to smaller investments.

4 *The Transition to Market Economy in China*

China as a strategic research site provides an ideal case to test our theory due to its enormous inter-provincial variance in the extent of market transition. While many hinterland provinces remain locked in state-directed transition economies, most coastal provinces have completed the transition to a market economy. A review of the reform process confirms a close link between market transition and innovation. China embarked on a “dual-track” approach to economic reform, which emphasized a gradual approach to the diversification of allocation mechanisms and property forms over the shock-therapy undertaken in Eastern and Central Europe and the Soviet Union. Central planning was not immediately abolished in 1978, but complemented by a “market-track” which operated parallel to the “plan-track.” Under the dual-track system, producers enjoyed the right to sell their surplus production on free markets after fulfilling compulsory delivery obligations. By 1990 markets became the dominant allocation mechanism for most commodities. For industrial products, the share of economic transactions controlled by the state fell from

near 100% prior to economic reform to 45% in 1990; while market sales in the retail market approached 70% of total sales by 1990 (LAU, QIAN, AND ROLAND [2000]).

Once free markets operated alongside planned production, market niches, particularly in light industries notoriously neglected under central planning attracted entrepreneurial talents. Regulatory market entry barriers were gradually lowered and only few areas, such as finance, telecommunications, tobacco, selected heavy industries, and high-technology sectors, remained off-limits for private enterprise. Competition further intensified in the 1990s when, after a decade of organizational reforms, wide-ranging ownership reforms of state-owned enterprises (SOE) were initiated. Small and medium-size SOEs were privatized through auctions and management buy-outs, while key firms in strategic industries were corporatized, and as public corporations the largest were listed on the domestic stock markets. Formally the corporatization strategy intended to depoliticize enterprise decision making and to limit the state's interference in firm management. However, with the state as majority shareholder of two third of the listed firms and complete state ownership in many of the non-listed companies, political intervention persisted (NEE, OPPER, AND WONG [2007]).

In the early period of economic reform, new market entrants were mainly rural non-state firms (collective and privately run township and village enterprises) and foreign firms. But by the mid-1980s, the fledgling private enterprise sector grew rapidly in the expanding consumer and light industrial sector. Confronted with fierce competition from these start-ups, the contribution of state-owned enterprises decreased from 78% to only 35% of gross industrial production between 1979 and 2005 (NATIONAL BUREAU OF STATISTICS OF CHINA [2006]). Private firms spearheaded the development of China's technology-based industries in electronic and computer appliances. With an unprecedented founding rate of non-state firms, China developed into one of the most competitive market economies, with comparatively low market concentration ratios.¹

Between 1999 and 2003, national R&D-expenditures increased from 0.8% to 1.3% of GDP. The Ministry of Science and Technology projects that spending on R&D will increase to 2.5% of GDP by 2020 (CHONG [2006]). In parallel, the locus of research shifted from government institutions to the firm. With more than 60% of R&D funds provided by firms, the expenditure structure resembles that of advanced market capitalist economies. Also inter-firm technological collaboration and regional innovation clusters developed rapidly.

5 *Data and Method*

To analyze the relationship between market transition and innovativeness at the firm level, we use data from the World Bank Investment Climate Surveys. The 2002

¹ The five largest machinery builders in the U.S. have a combined market share of 69%, in Japan the top five hold 42%, whereas the top five manufacturers in China have only 20% of the market (OECD [2002, p. 403]).

survey includes firms in five middle-size and large Chinese cities ($N = 1,548$) and the 2003 survey includes firms in 18 middle-size and large cities ($N = 2,400$). Both surveys share a set of core questions on innovation activities and firm characteristics. Participating firms were randomly selected in each city. The industry mix comprises both labor-intensive and technology-intensive sectors across a broad spectrum of different production technologies and levels of competition.² Importantly, the World Bank data enables quantitative institutional analysis of a diverse sample of organizational and ownership forms – private, hybrid, and state-owned enterprises. A note of caution, however, should be added: there is currently no longitudinal data available that covers the type of in-depth organizational information ideally needed to test our theory. It is therefore technically not possible to fully rule out endogeneity concerns.

5.1 Model Specification

Our model tests for the impact of market forces and political influence on innovation. Formally, our model is

$$y_{ij} = X_{ij}\beta + v_i + \varepsilon_{ij},$$

where i denotes each city and j each firm. X_{ij} is a set of firm-level variables covering political control, research activities, competition and distinct firm characteristics and β is a vector of corresponding coefficients. v_i denotes regional level effects while ε_{ij} residuals.

5.2 Dependent Variable

To assess the broader concept of firm innovativeness, we employ three measures of innovation: (1) the introduction of new products, (2) the introduction of a new production process, and (3) the introduction of new quality-control measures. For all innovation measures the 2002-survey provides information for the years from 1998 to 2000, while the 2003-survey provides information for the years from 1999 to 2002.

In addition, we follow earlier work (SCHMOOKLER [1966]) and construct a dummy variable, which indicates whether a firm received a patent in the last available survey year (i.e., in 2002 and 2000). The use of patent applications as a measure of innovativeness, however, is not unproblematic. First, only patented inventions that are actually brought to the market are innovations (SCHUMPETER [1942]). In addition,

² The surveys used 16 different industry categories. The distribution of firms is: apparel & leather goods (14.61%), electronic equipment (9.73%), electronic parts (12.34%), household electronics (1.6%), vehicles & vehicle parts (14.61%), information technology (8.64%), accounting & related services (6.81%), marketing (6.18%), business logistics (9.75%), food processing (1.80%), chemical products & medicine (1.67%), biotech & Chinese medicine (0.91%), metallurgical products (4.00%), transport equipment (1.27%), communication services (1.85%), and consumer products (4.23%).

firms do not generally decide to patent their inventions. We therefore include patents mainly for the sake of complete coverage of standard innovation measures.

5.3 Independent Variables

Marketization. Measuring progress in market transition is not an easy endeavor. We employ a measure of local “private sector development”, as privatization is a close proxy of overall market transition and is also used to broadly classify capitalist systems (BRADA [1996]). The city level scores for private economic development are based on three components: (1) the proportion of private sector employees in total employment, (2) the proportion of private sector revenues to GDP, and (3) the contribution of private sector tax revenues to total revenues (CHINESE ACADEMY OF SOCIAL SCIENCES [2005]). The indicator is formulated as a relative measure, wherein 1 is assigned to the city with the most developed private sector in China. Due to missing information on five survey cities, our final sample is reduced to 18 cities, with private sector development ranging from 0.11 in Beijing to 0.56 in Shenzhen.³ To allow a closer analysis of the impact of market transition on innovation processes we also construct two sub-samples and divide the sample at the mean value of private sector development [= 0.277]. Note that the sub-mean sample does not indicate absolutely low levels of market transition.⁴ After more than 25 years of successful market reform, all survey cities have undergone extensive market reforms.

Research Activities. Whether a firm has invested in R&D over the last three years preceding the survey year is specified by a dummy variable (MAIRESSE AND MOHNEN [2002]). The average ratio of R&D expenditures to total sales over the last three years serves as an indicator of R&D intensity. Finally, we approximate the most recent stock of technological capital by noting whether a firm acquired patents over the preceding two years. This variable takes into account the path-dependent process of innovation wherein past experience and success has a positive impact on future innovation.

Research Networks. The emergence of innovation markets is measured through variables indicating the existence of contractual agreements for R&D cooperation in the last three years between the firm and (1) research institutes, (2) universities, and (3) other firms. Membership in business associations and location in industrial parks are proxies of the potential diffusion of information and knowledge through

³ The following cities are included in our sample. Private sector development scores are indicated in brackets: Beijing (0.12), Xian (0.16), Kunming (0.16), Chengdu (0.18), Harbin (0.20), Nanchang (0.21), Tianjin (0.27), Guangzhou (0.27), Wuhan (0.27), Zhengzhou (0.28), Chongqing (0.30), Changchun (0.34), Wenzhou (0.35), Hangzhou (0.35), Changsha (0.40), Dalian (0.41), Shanghai (0.43), Shenzhen (0.56).

⁴ The sub-mean sample includes the cities from Beijing to Guangzhou whereas the above-mean sample includes the cities from Wuhan to Shenzhen.

networks. This source of regional advantage does not rely on formal contractual research agreement, but on reduced information costs due to propinquity and inter-firm networks (POWELL, KOPUT, AND SMITH-DOERR [1996], BURT [2005], ARROW [2007]).

Political Control. We first note whether a firm is legally registered as a state-owned enterprise. State-owned enterprises in general operate under softer budget constraints and are subject to political involvement and rent-seeking. Public ownership is not limited to firms legally registered as state-owned enterprises. Many firms listed in China's two stock exchanges and joint-stock firms registered as private enterprises are partly or even majority state-owned. In order to capture such ownership effects, we differentiate four mutually exclusive levels of state ownership: (1) up to 25%, (2) between 25% and 50%, (3) between 50% and 99%, and (4) 100%. Fully privately held firms serve as benchmark category.

Table 1
Descriptive Statistics of Variables in Analysis

	Mean	Std. dev.	Min	Max
Product innovation	0.385	0.487	0	1
Process innovation	0.319	0.466	0	1
New quality control	0.490	0.500	0	1
Firm receives patent in 2002	0.116	0.320	0	1
Firm holds patents	0.114	0.318	0	1
Firm conducts R&D	0.353	0.478	0	1
Average R&D-to-sales ratio	0.020	0.604	0	32.370
Located in industrial park	0.240	0.427	0	1
Member of business association	0.568	0.495	0	1
R&D cooperation with firms	0.130	0.337	0	1
R&D cooperation with universities	0.140	0.347	0	1
R&D cooperation with research institutes	0.100	0.300	0	1
Legally registered as SOE	0.243	0.429	0	1
State holds up to 25% shares	0.020	0.138	0	1
State holds between 25% and 50%	0.020	0.138	0	1
State holds between 50% and 99%	0.023	0.150	0	1
State holds 100%	0.187	0.390	0	1
Market share >10%	0.265	0.441	0	1
Number of competitors in main business*	3.561	1.393	1	5
Firm exports	0.235	0.424	0	1
Firm is founded after 1978	0.810	0.393	0	1
Log of average firm assets	8.556	2.686	0	17.474
Log of average debt-asset ratio	1.013	0.869	0	7.291

Note: * 1: 1-3, 2: 4-6, 3: 7-15, 4: 16-100, 5: more than 100.

5.4 Control Variables

Competition. To separate effects of market transition from firm competition, we introduce five variables to measure competitive pressure: A binary variable indicates whether a firm controls more than 10% of the domestic market sales, in order to control for monopoly power (SCHUMPETER [1942], ARROW [1962]). Self-reported numbers of competitors in the relevant domestic market capture the perceived level of competition (we use a five-point scale with 1: 1–3, 2: 4–6, 3: 7–15, 4: 16–100, 5: more than 100). Because a certain threshold of competitive market pressure may be required to stimulate innovation, we allow for a non-linear relation (AGHION et al. [2005]) by specifying a square-term of the number of competitors. Whether firms participate in the export market is indicated by a dummy variable. Lastly, a set of dummy variables controls for 16 different industrial sectors, which serve as proxies of competitive pressure, technological opportunity conditions, and average innovativeness (MAIRESSE AND MOHNEN [2002]).

Additional Control Variables. Other firm characteristics – including age, size, financial leverage, and location – may correspond with a firm’s innovativeness. A firm’s age is believed to affect its adaptability and innovativeness. Older enterprises are encumbered by more structural inertia. A dummy variable for firms founded after the start of market reform in 1978 differentiates between new and older firms. To control for scale effects from firm size (SCHUMPETER [1942]), we include the natural logarithm of the average value of a firm’s net assets over the last three years. The natural logarithm of the average debt–asset ratio over the preceding two years indicates financial health. Finally we include a set of dummies for the northeast, coastal, central, southwest, and northwest region of China. Table 1 presents the descriptive statistics of variables in analysis.⁵

6 Results

To examine the impact of market forces on innovation, we provide estimates for the full sample of firms and for two subsamples representing firms in cities with below-mean and above-mean levels of market transition. We draw inferences on the impact of market transition on our predictor variables from changes of coefficient estimates across the two subsamples.

6.1 Hypothesis 1: Increasing Rates of Innovation with Marketization

To begin with, we test whether we can confirm a positive correlation between the extent of market transition as measured by the regional development of the private economy and the probability of success in innovative activity. For each of the four dependent variables, our benchmark model in Table 2 only includes

⁵ Correlation tables are available upon request from the authors.

Table 2
Innovativeness and Market Transition

	All	Firms founded before 1993	Sample 1 (sub-mean private economy development)	Sample 2 (above-mean private economy development)
<i>(1) Product Innovation</i>				
Development level of private economy	2.198*** (0.565)	1.469*** (0.502)	2.069*** (0.376)	0.311 (0.748)
Industry	yes	yes	yes	yes
Region	yes	yes	yes	yes
Pseudo R^2	0.104	0.1004	0.098	0.112
N	3247	1363	1799	1444
<i>(2) Process Innovation</i>				
Development level of private economy	1.034** (0.483)	0.874** (0.353)	5.450*** (0.824)	-0.260 (0.700)
Industry	yes	yes	yes	yes
Region	yes	yes	yes	yes
Pseudo R^2	0.112	0.109	0.106	0.140
N	3243	1362	1798	1445
<i>(3) Quality Control</i>				
Development level of private economy	1.006** (0.475)	0.821* (0.487)	2.138*** (0.524)	-0.153 (0.929)
Industry	yes	yes	yes	yes
Region	yes	yes	yes	yes
Pseudo R^2	0.058	0.0758	0.057	0.066
N	3239	1362	1796	1443
<i>(4) Patent Granted</i>				
Development level of private economy	-2.446*** (0.945)	0.233 (0.476)	-7.036 (4.754)	-0.260 (1.103)
Industry	yes	yes	yes	yes
Region	yes	yes	yes	yes
Pseudo R^2	0.154	0.087	0.173	0.155
N	2017	1314	783	1188

Note: In parentheses are robust standard errors clustered on city; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

measures of city-level development of private economy and control variables for industries and regions. The estimation results for the full sample show significant and positive coefficients for private sector development for all outcome variables except for patents granted, which confirms Hypothesis 1, indicating that incentives and opportunities arising from the emergence of a market economy trigger innovation processes at the firm level (Propositions 1 and 2).

The negative effect of the extent of market transition on patenting activities is most likely a reflection of China's weak legal protection of intellectual property. Anecdotal evidence from our firm interviews suggests that patenting activities are not only conducted to protect intellectual property rights but also to gain legitimacy within a government-controlled institutional environment. In this light it is not surprising that the results reported on patenting were driven by Beijing and Chengdu, both municipalities are tightly controlled by government bureaus.

Skeptics may raise the issue of reverse causality. One concern is that more innovative firms simply locate in more marketized cities, while the less innovation-prone firms choose cities with significantly lower marketization levels. To test selection effects at the firms' birth, we re-estimated the models in Table 2 for a subsample of firms, which were already founded before China's main liberalization drive in 1993. In all cases, the positive association between marketization and innovation is confirmed for the subsample of old firms. Even for patent acquisitions, the coefficient estimate is now positive, though insignificant. Reverse causality therefore seems not to be driving our results.

An intriguing pattern is revealed when we compare estimates between the two sub-samples sorted by marketization levels. Without exception, the strong positive effects of the extent of market transition in the sub-mean sample disappear in the above-mean sample to non-significant levels for the product, process, and quality control innovation. It implies that the system-level effect decreases with further marketization, which is consistent with the *decreasing marginal improvement* in the probability of successful innovation assumed in condition (10). Hence, after 25 years of market reform, China's most marketized cities no longer display direct system-level effects on innovation. Note that these findings are also confirmed under inclusion of the full set of control variables for process innovation (Table 4) and quality control innovation (Table 5). In contrast, significant system-level effects persist for product innovation (Table 3), which emphasizes the crucial role of markets for the placement of new products. While process innovation and quality control innovation may be essentially driven by cost competition, product innovation builds to a larger extent on the opportunity space that only markets with low entry barriers offer (see Table 3).

6.2 Hypothesis 2: Increasing Effectiveness of Research Activities and R&D Networks with Marketization

Progress in market transition (as measured by private sector development) has not only direct effects on firm innovativeness. Growing market incentives and opportunities to pursue productive rather than unproductive entrepreneurial activities also increase the effectiveness of ongoing research efforts (Propositions 1 and 2). Markets help entrepreneurs to distinguish between good and bad research initiatives, they increase incentives to monitor and guide research initiatives. These effects can be readily inferred from a comparative analysis of innovation effects of research activities on firm innovativeness (see Tables 3 to 6). To begin with, a firm's history

Table 3
 Probit Estimation: Product Innovation

	All	Sub-mean sample	Above-mean sample
<i>Marketization</i>			
Development level of private economy	1.357*** (0.289)	3.257*** (0.862)	0.656*** (0.233)
<i>Research Activity</i>			
Firm holds patent	0.331*** (0.097)	0.183** (0.077)	0.525*** (0.189)
Firm conducts R&D	0.527*** (0.067)	0.400*** (0.082)	0.649*** (0.081)
R&D-to-sales ratio	-1.436*** (0.512)	-1.445 (1.162)	-1.421** (0.650)
<i>Network/Cooperation</i>			
Located in industrial park	0.145** (0.061)	0.230** (0.102)	0.058 (0.082)
Member of business association	0.334*** (0.049)	0.308*** (0.081)	0.305*** (0.072)
R&D cooperation with firms	0.498*** (0.070)	0.312*** (0.066)	0.688*** (0.078)
R&D cooperation with universities	0.238*** (0.058)	0.349*** (0.083)	0.092 (0.070)
R&D cooperation with research institutes	0.372*** (0.097)	0.458*** (0.132)	0.325** (0.164)
<i>Political Control</i>			
Legally registered as SOE	0.007 (0.111)	0.157** (0.062)	-0.137 (0.194)
State holds up to 25% ownership	0.052 (0.224)	-0.187 (0.315)	0.432** (0.194)
State holds 25% to 50% ownership	-0.339*** (0.116)	-0.417*** (0.142)	-0.234 (0.209)
State holds 51% to 99% ownership	-0.163 (0.111)	-0.041 (0.140)	-0.316* (0.178)
State holds 100% ownership	-0.051 (0.106)	-0.307*** (0.089)	0.275 (0.172)
<i>Competition</i>			
Market share >10%	0.166* (0.090)	0.047 (0.094)	0.355* (0.186)
# of competitors in main business	0.220 (0.169)	0.322 (0.302)	0.111 (0.151)
# of competitors in main business (squared)	-0.044 (0.027)	-0.062 (0.049)	-0.025 (0.024)
Firm exports goods	0.127** (0.061)	0.075 (0.099)	0.237*** (0.074)
Industry	yes	yes	yes
<i>Firm Characteristics</i>			
Founded after reform	0.025 (0.081)	-0.026 (0.084)	0.042 (0.160)
Log value of assets	0.043** (0.018)	0.051** (0.025)	0.021 (0.029)
Log of average debt-to-asset ratio	0.051 (0.035)	0.068 (0.057)	0.014 (0.036)
Region	yes	yes	yes
Constant	-1.928*** (0.310)	-2.190*** (0.581)	-1.642*** (0.320)
Pseudo R ²	0.218	0.200	0.253
N	2635	1361	1270

Note: In parentheses are robust standard errors clustered on city; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4
 Probit Estimation: Process Innovation

	All	Sub-mean sample	Above-mean sample
<i>Marketization</i>			
Development level of private economy	0.134 (0.391)	7.745*** (0.624)	-0.354 (0.509)
<i>Research Activity</i>			
Firm holds patent	0.451*** (0.059)	0.393*** (0.065)	0.504*** (0.114)
Firm conducts R&D	0.328*** (0.057)	0.268** (0.112)	0.346*** (0.073)
R&D-to-sales ratio	-0.552 (0.383)	1.263 (1.795)	-0.723* (0.424)
<i>Network/Cooperation</i>			
Located in industrial park	0.066 (0.041)	0.090 (0.067)	0.028 (0.050)
Member of business association	0.175*** (0.060)	0.244*** (0.091)	0.102 (0.079)
R&D cooperation with firms	0.434*** (0.111)	0.263** (0.123)	0.523*** (0.168)
R&D cooperation with universities	0.224** (0.092)	0.252*** (0.077)	0.147 (0.164)
R&D cooperation with research institutes	0.403*** (0.075)	0.438*** (0.099)	0.383*** (0.115)
<i>Political Control</i>			
Legally registered as SOE	0.011 (0.086)	0.109 (0.113)	-0.055 (0.127)
State holds up to 25% ownership	0.215 (0.225)	0.445 (0.298)	-0.007 (0.322)
State holds 25% to 50% ownership	-0.090 (0.129)	-0.122 (0.139)	-0.010 (0.290)
State holds 51% to 99% ownership	-0.182 (0.175)	-0.331* (0.171)	-0.003 (0.232)
State holds 100% ownership	-0.010 (0.092)	-0.164 (0.145)	0.145* (0.081)
<i>Competition</i>			
Market share >10%	0.145* (0.080)	0.131 (0.121)	0.244* (0.129)
# of competitors in main business	0.264** (0.116)	0.280 (0.195)	0.306* (0.159)
# of competitors in main business (squared)	-0.044** (0.020)	-0.046 (0.035)	-0.053** (0.024)
Firm exports goods	0.130 (0.081)	0.053 (0.096)	0.202 (0.132)
Industry	yes	yes	yes
<i>Firm Characteristics</i>			
Founded after reform	0.096 (0.063)	0.048 (0.093)	0.139 (0.091)
Log value of assets	0.063*** (0.014)	0.083*** (0.023)	0.043* (0.022)
Log of average debt-to-asset ratio	0.024 (0.040)	0.049 (0.049)	-0.011 (0.067)
Region	yes	yes	yes
Constant	-1.632*** (0.317)	-2.911*** (0.357)	-1.409*** (0.360)
Pseudo R^2	0.209	0.205	0.241
N	2632	1360	1272

Note: In parentheses are robust standard errors clustered on city; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5
 Probit Estimation: New Quality Control

	All	Sub-mean sample	Above-mean sample
<i>Marketization</i>			
Development level of private economy	0.119 (0.348)	2.931*** (0.714)	-0.321 (0.718)
<i>Research Activity</i>			
Firm holds patent	0.148 (0.099)	0.128 (0.145)	0.219* (0.121)
Firm conducts R&D	0.230*** (0.069)	0.144* (0.075)	0.315*** (0.110)
R&D-to-sales ratio	1.602 (1.011)	2.638 (1.856)	1.096 (0.811)
<i>Network/Cooperation</i>			
Located in industrial park	0.158*** (0.058)	0.158** (0.075)	0.157 (0.097)
Member of business association	0.220*** (0.049)	0.219*** (0.040)	0.235** (0.096)
R&D cooperation with firms	0.312*** (0.081)	0.287** (0.119)	0.278** (0.114)
R&D cooperation with universities	0.224*** (0.077)	0.206* (0.119)	0.224** (0.098)
R&D cooperation with research institutes	0.440*** (0.099)	0.429*** (0.069)	0.472*** (0.171)
<i>Political Control</i>			
Legally registered as SOE	-0.132* (0.077)	-0.080 (0.101)	-0.205* (0.124)
State holds up to 25% ownership	-0.113 (0.162)	-0.038 (0.188)	-0.191 (0.272)
State holds 25% to 50% ownership	-0.193* (0.105)	-0.225 (0.145)	-0.202 (0.195)
State holds 51% to 99% ownership	0.102 (0.093)	0.166* (0.094)	0.060 (0.192)
State holds 100% ownership	-0.106 (0.087)	-0.245** (0.107)	0.078 (0.134)
<i>Competition</i>			
Market share >10%	0.115 (0.084)	0.148 (0.123)	0.116 (0.132)
# of competitors in main business	0.139 (0.109)	0.354*** (0.123)	-0.137 (0.144)
# of competitors in main business (squared)	-0.021 (0.017)	-0.053*** (0.019)	0.017 (0.025)
Firm exports goods	0.172** (0.081)	0.169 (0.130)	0.158 (0.116)
Industry	yes	yes	yes
<i>Firm Characteristics</i>			
Founded after reform	0.289*** (0.081)	0.228*** (0.076)	0.366** (0.160)
Log value of assets	0.085*** (0.018)	0.096*** (0.015)	0.069* (0.037)
Log of average debt to asset ratio	0.026 (0.016)	0.033** (0.017)	0.026 (0.036)
Region	yes	yes	yes
Constant	-1.478*** (0.312)	-2.273*** (0.411)	-0.826*** (0.301)
Pseudo R^2	0.144	0.138	0.167
N	2627	1359	1268

Note: In parentheses are robust standard errors clustered on city; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6
 Probit Estimation: Patent Granted in 2002

	All	Sub-mean sample	Above-mean sample
<i>Marketization</i>			
Development level of private economy	-2.706*** (0.687)	-6.190*** (1.876)	-0.740 (1.306)
<i>Research Activities</i>			
Patents were granted in both preceding years	2.209*** (0.381)	1.713*** (0.419)	2.999*** (0.501)
Conducts R&D	0.126 (0.099)	0.293** (0.147)	0.010 (0.165)
R&D-to-sales ratio	1.037 (0.693)	1.580 (1.182)	0.877 (0.807)
<i>Network/Cooperation</i>			
Located in industrial park	0.173* (0.105)	0.152 (0.151)	0.294* (0.158)
Member of business association	0.214** (0.090)	0.292 (0.211)	0.169 (0.110)
R&D cooperation with firms	0.075 (0.152)	0.176 (0.286)	0.069 (0.226)
R&D cooperation with universities	0.335* (0.187)	0.164 (0.228)	0.471 (0.329)
R&D cooperation with research institutes	0.024 (0.166)	0.014 (0.302)	0.144 (0.126)
<i>Political Control</i>			
Legally registered as SOE	-0.079 (0.151)	-0.034 (0.174)	-0.153 (0.232)
State holds up to 25% ownership	-0.252 (0.249)	-0.124 (0.461)	-0.443 (0.445)
State holds 25% to 50% ownership	0.179 (0.338)	0.620*** (0.196)	-0.946*** (0.113)
State holds 51% to 99% ownership	-0.232 (0.174)	0.004 (0.346)	-0.539** (0.231)
State holds 100% ownership	-0.241** (0.099)	-0.134 (0.157)	-0.445** (0.192)
<i>Competition</i>			
Market share >10%	0.169 (0.116)	0.179 (0.152)	0.105 (0.197)
# of competitors in main business	0.220* (0.131)	0.059 (0.208)	0.485** (0.212)
# of competitors in main business (squared)	-0.054*** (0.019)	-0.035 (0.034)	-0.087*** (0.027)
Firm exports goods	-0.069 (0.111)	-0.061 (0.189)	-0.134 (0.165)
Industry	yes	yes	yes
<i>Company Characteristics</i>			
Founded after reform	0.206 (0.136)	0.250 (0.158)	0.220 (0.284)
Log value of assets	0.082** (0.033)	0.081* (0.049)	0.102*** (0.035)
Log of average debt-to-asset ratio	0.134** (0.056)	0.164*** (0.057)	0.101 (0.104)
Region	yes	yes	yes
Constant	-2.145*** (0.597)	-1.346* (0.764)	-3.699*** (0.648)
Pseudo R ²	0.411	0.392	0.466
N	1804	664	1030

Note: In parentheses are robust standard errors clustered on city; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

of patenting activities (firm holds patents/patents were granted in both preceding years) is a strong and statistically significant predictor for most of the dependent variables.⁶ A cross-sample comparison consistently confirms stronger effects for the above-mean sample for all innovation measures. Similarly, R&D efforts (firm conducts R&D) increasingly determine innovation rates. Estimated effects are stronger in the above-mean sample than in sub-mean sample, except for patenting activities (Table 6) where the R&D effect in the above-mean sample seems weakened by the overwhelming effect of the lagged patent variable. Finally, the R&D-to-sales ratio has either statistically insignificant positive effects or significant negative effects. We suspect that this is due to a skewed distribution of the variable, as about 72% of the sample firms do not conduct R&D.

In sum, both innovation capacity (measured by patent-holding) and R&D efforts (measured by a dummy for R&D activity) determine innovation and increasingly so the greater the extent of market transition. The growing effectiveness of R&D is consistent with stronger incentive to invest in innovation projects in a market economy and our assumption that political involvement in the economic sphere declines with market transition (*power proposition*, Proposition 4).

Also the five dummy variables for R&D cooperation or networks, in general, show strong effects for product, process, and quality control innovation (see Tables 3 through 5) and moderate effects for patenting activities (see Table 6). A comparison of coefficient estimates across the sub-mean and above-mean samples indicates that network ties with other firms and research institutes show robust effects across different levels of market transition. Most notably for product innovation, R&D networks with other *firms* are exceptionally effective in the above mean sample (see Table 3). Its coefficient [= 0.688] is more than doubled compared to the sub-mean sample [= 0.312] and at least twice as large as those for other R&D network dummies in the above-mean sample. For process innovation, our results show a similar pattern with doubled coefficient estimates at advanced levels of market transition indicating a general and increasing superiority of firm-to-firm collaboration (see Table 4). This indicates that particularly inter-firm collaborations become more effective with the emergence and growth of free markets. The underlying causal effect may well be a closer incentive alignment with firms. Other research partners such as universities and research institutes are to a lesser extent motivated by market forces.

Another reverse causality concern refers to the impact of research activities and research networks on firm innovativeness. In order to rule out a sorting effect in the way that only innovative firms enter into research activities and network collaboration, we have focused on a subsample of firms, which have not yet proven their ability to innovate by formal patent-holdings. Our substantial results, however, remained unaffected further supporting Hypothesis 2.⁷

⁶ The only exceptions are for quality control, in the sub-mean sample and consequently in the total sample.

⁷ Regression results are available upon request from the authors.

6.3 Hypothesis 3: Positive Effects of Private Ownership on Innovation

Our final hypothesis argues that private ownership limits political involvement and unproductive entrepreneurship and thereby increases innovation rates (*power and politics propositions*, Propositions 4 and 5). The innovation advantage of private firms is widely confirmed as most of the coefficient values for state-owned enterprises and state ownership shares are negative though not all significant. In a straight forward interpretation, however, Hypothesis 3 calls for increasingly negative coefficient values the larger the representation of the state as a shareholder.

All three innovation types (Table 3 to Table 5) share two common patterns. First, our estimation results show stronger support for Hypothesis 3 in the sub-mean sample than in the above-mean sample. Across product, process, and quality control innovations, the state ownership share dummies tend to form negative slopes at large in the sub-mean sample. Second, there is a U-shaped pattern in the above-mean sample. Coefficient estimates even turn positive for the 100% state-share (0.275, 0.145, and 0.078 respectively in reference to zero state-share). We suspect that these estimates most likely reflect a selection effect. Local governments typically divest less competitive state-owned enterprises, while profitable firms in key sectors remain under government control and enjoy various forms of government protection. If the firm sample is to some extent affected by such a selection effect, which might have been particularly pronounced during the survey years (due to the ongoing privatization wave), this might help explain the unexpected result for wholly state-owned firms. As the concept of market transition itself also reflects progress in enterprise reforms, the most marketized regions may be characterized by stronger performance of state-owned firms simply because the less successful firms have already been divested. This is consistent with the fact that the proportion of 100% state-owned firms is 14.9% in the above-mean sample in comparison to 20.3% in the sub-mean sample.

For patenting activities, Hypothesis 3 is mainly confirmed for the above-mean sample. Coefficients for the three largest state ownership share dummies (i.e. larger than 25%) are negative and significant in the model (-0.946 , -0.539 , and -0.445 respectively). The above-mean sample, however, shows a slightly U-shaped pattern because the lowest innovation rate is found in the middle range of state ownership shares in 25 to 50% [= -0.946].

6.4 Other Control Variables

Most of our control variables for competition show the expected signs though coefficient estimates are not statistically significant in all models. Large market shares tend to increase innovativeness in the case of product and process innovations (Tables 3 and 4). Further, in line with AGHION et al. [2005] we identify an inverted U-shape relation between number of competitors and innovativeness, with statistically significant coefficients in the case of process innovation (Table 4) and patenting (Table 6). It is worth noting that competition (as measured by the number

of competitors) is not simply a proxy of progress in market transition. The correlation coefficients with marketization are -0.4 for the full sample and -0.17 for the above-mean sample, which confirms the distinctiveness of competition and the broader concept of market transition. Overall, firms in more marketized cities report almost the same number of competitors as firms in less marketized cities.

Among variables for firm characteristics, firm size measured by the natural logarithm of firm assets yields consistently strong effects on innovativeness. Larger firms seem to benefit from scale effects, which help them to succeed in a wide range of innovative activities. Also, new firms founded after the reform seem more successful in quality control. The debt-to-asset ratio reveals that more financially leveraged firms are more successful in patent acquisition and quality control.

Finally, we note some concerns: First, there is limited availability of survey data on firm innovativeness. Due to the survey design, our econometric tests are confined to non-linear estimation techniques. The survey for instance contains no information on the number of implemented innovation projects or their economic value. Given the reasonable assumption that firms in less marketized regions and state-owned firms perform fewer innovation projects, the survey design creates a statistical convergence which may not accurately reflect the current situation. It is likely that cross-ownership differences in firm innovativeness are downplayed in our firm sample.

Secondly, the lack of availability of city-level measures of market transition for Guiyang, Lanzhou, Nanning, Benxi, and Jiangmen has led to the exclusion of these five cities covered in the Investment Climate Survey. With the exception of Jiangmen, all of these cities belong to the less marketized cities with weak private sector development. We therefore assume that a broader sample covering the complete range of market transition would most likely show even stronger system level effects on firm behavior and innovativeness.

To sum up, we understand our empirical application as a first approximation. In spite of all limitations, we hope that these findings inspire future research exploring the distinct impact of the nature of the institutional environment on firm innovativeness.

7 Conclusion

We proposed a theory of innovation and applied it to explain why market transition caused a shift to a higher level of innovation in China's manufacturing economy. We proffered five propositions asserting that the endogenous emergence of markets increases the power of economic actors relative to political actors, increases inter-firm competition and creates new opportunities for entrepreneurship, and subsequently motivates endogenously innovative activity. The mechanisms explain that firms will become more inclined to invest in innovative activity and by doing so link marketization, economic power, and innovation into one theoretical framework. By linking Nee's ideas of endogenous emergence of decentralized markets and entrepreneurial

activities with Baumol's ideas on innovation, our theory highlights how market forces that shape motivation are embedded in institutions.

This paper also makes methodological and empirical contributions to understanding the workings of institutions: First, we use a quantitative approach to comparative institutional analysis explaining variations in innovativeness by the level of market orientation. Significant advances have been made in recent decades in understanding of institutions through historical case studies. For example, Avner GREIF's [2006] applications of game theory to explain institutional change in late medieval Europe utilizes context-bound models and analyses to shed explanatory light on the endogenous emergence and decline of economic institutions. General propositions from which predictive hypotheses can be derived and confirmed with quantitative measures of institutions drawn from a randomized sample can contribute to further advances in understanding why institutions matter in economic performance. The variability of regional transition economies in China, from the highly marketized southeastern coastal provinces to the less marketized hinterland provinces, makes for a large canvas to test hypotheses linking institutions to rates of innovative activity.

Second, our empirical results confirm that the individual rate of innovation increases with the level of local marketization as measured by private firm activities. We find evidence consistent with the view that markets do not just generate competitive pressure on individual firms, but sustain self-reinforcing institutional change that enable and motivate innovative activities. We also confirm an increasing effectiveness of research activities and R&D networks in the transition to a market economy. Further, we provide empirical evidence for the hypothesized negative effects of political involvement on the innovativeness of firms.

Our quantitative institutional analysis yields results in line with HAYEK's [2002, p. 19] contention that "lack in entrepreneurial spirit [...] is not an unchangeable attribute of individuals, but the consequence of limitations placed on individuals." The crucial role of the market economy as an institutional system driving innovation seems close to the Hayekian notion of "competition as a discovery procedure." While we emphasize the central role of competition as a mechanism to exploit undiscovered opportunities, our theory of innovation goes beyond the instrumental character of competition as a means for decentralized problem solving. Competition in the Hayekian sense is mainly a tool that responds to two observations: First, knowledge is dispersed and decentralized across society; secondly, rewards for knowledge generation are *ex ante* unpredictable as only consumer demand decides about success and failure. As a consequence, knowledge creation naturally relies on competition as a process of experimentation by a large number of entrepreneurs and the informational processing capability of free markets. In contrast, our concept capturing the market orientation of economic systems, attempts to reach down to the motivational foundations of human behavior.

While our empirical application focuses on the link between progress in market transition at the city-level and firm innovativeness in China's transition economy, we see scope for further development and applications: Our quantitative institutional approach can be applied to study cross-national variance of innovation. Another

field for future application is the cross-national comparison of innovativeness in distinct industrial sectors which underlie greatly varying regulatory regimes, which may balance firm interests either in the direction of active R&D or rent-seeking activities.

Appendix

A.1 Proof of $\partial_m I_i^* > 0$ and $\partial_m \pi^* > 0$

Condition (6) is equivalent to

$$(A1) \quad C \cdot \partial_{I_i} \pi + P \cdot \partial_{I_i} \phi = 0$$

and

$$(A2) \quad C \cdot \partial_{I_i I_i} \pi + P \cdot \partial_{I_i I_i} \phi < 0$$

because C and P are functions of marketization m while π and ϕ are functions of investment $I_i (= -I_p)$ and privatization a . We can solve (A1) for I_i as a function of m and a . In other words, the optimal investment level given a and $m [I_i^* = I_i(a, m)]$ is implicit in (A1).

Plugging $I_i^* = I_i(a, m)$ into (A1) and differentiating both sides with m ,

$$\partial_m(C(m) \cdot \partial_{I_i} \pi(a, I_i(a, m)) + P(m) \cdot \partial_{I_i} \phi(a, I_i(a, m))) = 0.$$

After some algebra we obtain

$$(A3) \quad \underbrace{C'(m) \cdot \partial_{I_i} \pi}_{(+)\times(+)} + \underbrace{P'(m) \cdot \partial_{I_i} \phi}_{(-)\times(-)} + \underbrace{(C \cdot \partial_{I_i I_i} \pi + P \cdot \partial_{I_i I_i} \phi)}_{(-) \text{ by (A2)}} \partial_m I_i(a, m) = 0.$$

In (A3), the first two terms are positive by (2) and (3). At the same time, $C \cdot \partial_{I_i I_i} \pi + P \cdot \partial_{I_i I_i} \phi$ is negative by (A2). In order to make the total sum zero, $\partial_m I_i(a, m)$ should be positive. Therefore, condition (7) is proved.

Let us denote innovation capacity at the optimal investment by π^* :

$$\pi^* = \pi(a, I_i^*) = \pi(a, I_i(a, m)).$$

Then,

$$\partial_m \pi^* = \partial_m \pi(a, I_i(a, m)) = \partial_{I_i} \pi \cdot \partial_m I_i^* > 0$$

by (3) and (7). Therefore, condition (8) is proved.

Q.E.D.

A.2 Proof of $\partial_a \pi^* > 0$

First, we show $\partial_a \pi > 0$.

From (5),

$$(A4) \quad \partial_{a I_i} \pi = \partial_{I_i a} \pi > 0$$

if $\pi(a, I_i)$ has continuous second partial derivatives. Let $\partial_a \pi(a, I_i) \equiv f(a, I_i)$. Then, (A4) can be re-written such that

$$(A4a) \quad \partial_{I_i} f(a, I_i) > 0.$$

Recall from condition (3) that π_0 is the lower bound of innovation capacity, when no investment into innovation is made at all. Namely,

$$(A5) \quad \pi(a, 0) = \pi_0$$

for all a . From (A5),

$$(A6) \quad \partial_a \pi(a, 0) = f(a, 0) = 0.$$

From (A4a), $f(a, I_i)$ is an increasing function with I_i with an initial value zero at $I_i = 0$ by (A6). It holds for any given a . Therefore,

$$(A7) \quad f(a, I_i) = \partial_a \pi(a, I_i) > 0 \quad \text{for any } a \text{ and } I_i$$

which completes the proof of $\partial_a \pi > 0$.

Second, we show:

$$(A8) \quad \partial_a I_i^* > 0.$$

By plugging $I_i^* = I_i(a, m)$ into (A1) and differentiating both sides with a ,

$$\partial_a(C(m) \cdot \partial_{I_i} \pi(a, I_i(a, m)) + P(m) \cdot \partial_{I_i} \phi(a, I_i(a, m))) = 0.$$

After some algebra we obtain

$$(A9) \quad C \underbrace{\partial_a I_i \pi}_{(+)\text{ by (5)}} + P \underbrace{\partial_a I_i \phi}_{(+)\text{ by (5a)}} + \underbrace{(C \cdot \partial_{I_i I_i} \pi + P \cdot \partial_{I_i I_i} \phi)}_{(-)\text{ by (A2)}} \partial_a I_i(a, m) = 0.$$

In order to make the total sum equal to zero on the left side of (A9), $\partial_a I_i(a, m) = \partial_a I_i^*$ should be positive and condition (A8) is proved.

Finally, we can prove $\partial_a \pi^* > 0$ by (A7) and (A8) because:

$$\partial_a \pi^* = \partial_a \pi(a, I_i(a, m)) = \underbrace{\partial_a \pi}_{(+)\text{ by (A7)}} + \underbrace{\partial_{I_i} \pi}_{(+)\text{ by (3)}} \cdot \underbrace{\partial_a I_i(a, m)}_{(+)\text{ by (A8)}} > 0.$$

References

- AGHION, P., N. BLOOM, R. BLUNDELL, R. GRIFFITH, AND P. HOWITT [2005], "Competition and Innovation: An Inverted-U Relationship," *The Quarterly Journal of Economics*, 120, 701–728.
- ARROW, K. J. [1962], "Economic Welfare and the Allocation of Resources for Invention," pp. 609–626 in: National Bureau of Economic Research (eds.), *The Rate and Direction of Inventive Activity: Economic and Social Factors*, Princeton University Press: Princeton, NJ.
- [2007], "The Macro-Context of the Microeconomics of Innovation," pp. 20–27 in: E. She-shinski., R. J. Strom, and W. J. Baumol (eds.), *Entrepreneurship, Innovation, and the Growth Mechanism of the Free-Enterprise Economies*, Princeton University Press: Princeton, NJ.
- BAUMOL, W. J. [1990], "Entrepreneurship: Productive, Unproductive and Destructive," *Journal of Political Economy*, 98, 893–921.
- [1993], *Entrepreneurship, Management and the Structure of Payoffs*, The MIT Press: Cambridge, MA.
- [2002], *The Free-Market Innovation Machine: Analyzing the Growth Miracle of Capitalism*, Princeton University Press: Princeton, NJ.

- BRADA, J. C. [1996], "Privatization Is Transition – or Is it?" *The Journal of Economic Perspectives*, 10, 67–86.
- BURT, R. S. [2005], *Brokerage & Closure: An Introduction to Social Capital*, Oxford University Press: Oxford.
- CHINESE ACADEMY OF SOCIAL SCIENCES [2005], *Annual Report of Urban Competitiveness*, Social Sciences Academic Press: Beijing.
- CHONG, W. [2006], "China Unveils Plans for Science-Based Development," *Science and Development Network*, February 10, available at <http://www.scidev.net/en/news/china-unveils-plans-for-sciencebased-development.html>.
- GREIF, A. [2006], *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*, Cambridge University Press: Cambridge, MA.
- HART, O., A. SHLEIFER, AND R. W. VISHNY [1997], "The Proper Scope of Government: Theory and an Application to Prisons," *The Quarterly Journal of Economics*, 112, 1127–1161.
- HAYEK, F. A. V. [1978], "Competition as a Discovery Procedure," pp. 179–190 in: F. A. v. Hayek, *New Studies in Philosophy, Politics, Economics and the History of Ideas*, The University of Chicago Press: Chicago, IL.
- [2002], "Competition as a Discovery Procedure," *The Quarterly Journal of Austrian Economics*, 5, 9–23.
- LAU, L. L., Y. QIAN, AND G. ROLAND [2000], "Reform without Losers: An Interpretation of China's Dual-Track Approach to Transition," *Journal of Political Economy*, 108, 120–143.
- MAIRESSE, J., AND P. MOHNEN [2002], "Accounting for Innovation and Measuring Innovativeness: An Illustrative Framework and an Application," *The American Economic Review: Papers and Proceedings*, 92, 226–230.
- NATIONAL BUREAU OF STATISTICS OF CHINA [2006], *China Statistical Yearbook*, China Statistics Press: Beijing.
- NEE, V. [1989], "A Theory of Market Transition: From Redistribution to Markets in State Socialism," *American Sociological Review*, 54, 663–681.
- AND S. OPPER [2010], "Endogenous Institutional Change and Dynamic Capitalism," *Sociologia del Lavoro*, forthcoming.
- , —, AND S. WONG [2007], "Developmental State and Corporate Governance in China," *Management and Organization Review*, 3, 19–53.
- OECD [2002], *China in the World Economy: The Domestic Policy Challenges*, OECD: Paris.
- POWELL, W., K. KOPUT, AND L. SMITH-DOERR [1996], "Interorganizational Collaboration and the Locus of Innovation: Networks of Learning in Biotechnology," *Administrative Science Quarterly*, 41, 116–145.
- QIAN, Y., AND C. XU [1998], "Innovation and Bureaucracy under Soft and Hard Budget Constraints," *The Review of Economic Studies*, 65, 151–164.
- ROSEN, S. [1974], "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82, 34–55.
- SCHMOOKLER, J. [1966], *Invention and Economic Growth*, Harvard University Press: Cambridge, MA.
- SCHUMPETER, J. A. [1912/1934], *The Theory of Economic Development*, Harvard University Press: Cambridge, MA.
- [1942], *Capitalism, Socialism and Democracy*, Harper & Row: New York.
- SHLEIFER, A., AND R. W. VISHNY [1994], "Politicians and Firms," *The Quarterly Journal of Economics*, 109, 995–1025.

WAN, H. Y. [2003], "Reform Unleashed Korean Growth," *German Economic Review*, 4, 19–34.

Victor Nee
Department of Sociology
Cornell University
Uris Hall 332
Ithaca, NY 14853
U.S.A.
E-mail:
vgn1@cornell.edu

Jeong-han Kang
Department of Sociology
Yonsei University
134 Shinchon-dong,
Seodaemun-gu
Seoul 120-749
South Korea
E-mail:
jhk55@yonsei.ac.kr

Sonja Opper
Department of Economy
Lund University
P.O. Box 7082
22007 Lund
Sweden
E-mail:
Sonja.Opper@nek.lu.se

Market Correlation and Property Rights

by

SHIN-HWAN CHIANG AND XIANG LI*

This paper examines the origins of property rights in the presence of production uncertainty. Since stealing others' possessions is permitted under anarchy, the winner is able to enjoy the lion's share of total outputs produced by all parties, and this generates a diversification effect, since the random outputs are polled together. Taking this effect into account, we characterize the subgame-perfect equilibrium for our two-stage game. Specifically, the emergence of property rights is shown to depend on players' incentives to fight, variances, and market correlations. The model predicts that property rights are more likely to emerge when market correlations increase. (JEL: D 23, D 82, D 81)

1 Introduction

In the world of Robinson Crusoe property rights play no role. As individuals (or regions or states for that matter) become more socially and economically interdependent, conflict of interest begins to develop. Under anarchy, the conflict is resolved through brutal means, viz., warfare, since there are no laws and orders that provide the necessary protection against bad behavior. In the absence of well-defined property rights, stealing and fighting are permitted. Thus, the incentive that drives good effort is lost. Consequently, resource allocations and investment decisions get distorted, thus adversely affecting individuals' well-being.

War and peace carry very different implications with respect to wealth redistribution. Needless to say, war produces a "winner" and a loser. Obviously, its immediate consequence is that wealth gets redistributed and the loser becomes the victim of the entire process. The winner seemingly benefits from it; but the victory may not always translate itself into positive net benefits when the costs associated with it are taken into consideration. One critical issue is whether it makes sense for individuals (or regions) to live peacefully together with well-defined property rights protecting them, or it is better for them to be in a community where there exist no rules and orders, so that stealing others' possessions is allowed if one chooses. The question of conflict has drawn considerable attention not only in economics

* Both authors are from York University, Toronto, Xiang Li being the corresponding author. We wish to thank Elmar Wolfstetter and an anonymous referee of this Journal for their helpful comments and suggestions. The usual disclaimer applies.

but also in political science. Interest in this area has heightened in recent years in view of the formation of the European Union – an example that replaces regional conflict by mutual cooperation and respect. Various models of conflict have been developed with the intention to explain why abstinence from fighting can be a self-enforcing equilibrium (for example, HAFER [2006] and MUTHOO [2004], among many others).

There is a large literature that examines the effect of property rights on economic outcomes. An early key contribution was made by Adam Smith, who argued that the expectation of profit from “improving one’s stock of capital” rests on private property rights. Property rights encourage the property holders to develop the property, generate wealth, and efficiently allocate resources. Later, COASE [1960] showed that in a zero-transaction-costs world, well-defined property rights can lead to economic efficiency. He also forcefully argued that the structure of property rights should have no effect on efficiency. Along the same line, CHEUNG [1963] realized the importance of transaction costs and showed that if there are none, there is no difference in using different institutional arrangements (e.g., market or government). Moreover, DEMSETZ [1967] proved that the allocation of property rights was a precondition for the efficient functioning of markets. As we deviate from the frictionless world, the distribution of property rights begins to have an effect on economic efficiency. For instance, GROSSMAN AND HART [1986] showed that with incomplete contracting, the distribution of property rights has a nonmarginal effect on economic efficiency.

Related to above studies, SKAPERDAS [1992] studied the origins of property rights. He showed that in the absence of property rights, cooperation can still be an equilibrium outcome, although warfare remains an option in resolving the conflict. ESTEBAN AND RAY [1999] proved that there exists a nonlinear relationship between social conflict and societywide distribution of individual characteristics. HIRSHLEIFER [1995] argued that anarchy may be a pattern of relationships that constitutes, at least potentially, a stable system. For anarchy to be an equilibrium outcome, the returns of fighting effort should be strongly diminishing and the income should be greater than a critical value. In a dynamic setting, HAFER [2006] analyzed the emergence of property rights in a model of conflict and production in the absence of institutions of enforcement. She showed that in the steady state of the game, the population sorts into two stable groups: “haves” and “have-nots.” As the system reaches the steady-state equilibrium, no one will have any interest in stealing others’ possession. In the same vein, MUTHOO [2004] also studied the origins of property rights, explaining why, when, and how such rights can emerge and be made secure in a repeated setting. He explores, in particular, the roles of the players’ fighting and productive skills in the emergence of property rights. His main contribution is that the property rights can be supported as an equilibrium outcome if the game is repeated and the players are patient enough.

The present paper is closely related to MUTHOO [2004]. It examines the origins of property rights when productions are subject to some random disturbances. In his two-period repeated framework, MUTHOO [2004] assumes that the players are endowed with fighting and productive skills, which are taken as exogenously

given. The chance of winning a fight is fixed by the relative physical strength of the players; no investment decisions in fighting can be chosen strategically by the players to influence the outcome. He also assumes that the production function is deterministic. As in MUTHOO [2004], we also have a two-stage model in which the players can strategically determine their time allocations in stage 1 and then choose whether to fight or not in stage 2. However, our model differs from his in two important aspects. First, we allow for parties to make their investment decisions regarding farming and fighting to enhancing their individual fighting skills, which in turn affect their individual probabilities of winning the fight should one occur. This important ingredient is missing in Muthoo's model. Second, we assume that the players' productions are subject to some idiosyncratic shocks, which in our view is a critical element in explaining the origins of property rights, and which is at the heart of our analysis. We believe that the introduction of stochastic elements in production adds a new dimension to players' investment decisions and allows us to gain some insights into their effects on the emergence of property rights. While we recognize that the reason underlying any social conflict is complex, there has been little discussion in the literature for the issue of property rights from a risk-sharing perspective. In effect, the intention to steal resources from others is reinforced when the combined outputs achieve some kind of diversification effect. This is because a decrease in correlations reduces overall risk, thus making a fight more attractive and strengthening the player's incentive to fight. However, by an increase in covariance (for example, making it sufficiently positive), the benefit of winning a fight is reduced, making the fighting less likely to occur. Property rights will therefore emerge as an equilibrium outcome. The result is self-enforcing, since no one is willing to deviate from it.

In Muthoo's model, the emergence of property rights is driven by a repeated procedure without which property rights will never emerge. Here, we wish to show that although our model is static in nature, property rights may still emerge as a self-enforcing equilibrium. The underlying reason behind this is the motive to diversify risk. Clearly, a battle produces only one winner who can enjoy the lion's share of total outputs produced by all parties involved, particularly for a winner-take-all battle. Thus, winning allows the winner to achieve the diversification effect because the random outputs are pooled together. Of course, the cost associated with it is that the player may end up losing the fruits of his effort if he is defeated. Higher covariance implies lower diversification effect, thus making fighting less attractive. Consequently, property rights emerge with higher probability. In short, there is a strong connection between property rights and market correlations. We take this to be a natural motive for property rights to emerge.

The paper is organized as follows. Section 2 presents a two-period, two-player model with uncertainty where players determine their time allocations in stage 1 and then choose whether or not to fight in stage 2. Section 3 characterizes the subgame-perfect equilibrium for our two-stage game and examines the effect of market correlations on the emergence of property rights. Finally, section 4 concludes the paper.

2 The Basic Model

Consider a two-stage model with two players in an environment without rules and orders. The society provides no protection, so that stealing is possible if it is profitable for the players. Each player can strategically allocate his time endowment between two different activities, farming and fight-related training, in stage 1, and then chooses whether or not to fight in stage 2. That is, the player can enhance his consumption by producing more himself or alternatively by stealing from the others. Farming and fight-related investment are subject to some random shocks. Each player is assumed to know his own output at the beginning of stage 2, but this piece of information is not known to his rival. Upon observing his output, player i ($i = 1, 2$) then decides whether to fight or not. Fighting produces a winner and a loser. The payoff of each player is determined after the realization of the random variables.

Players are symmetric with respect to their talents in farming and fighting. Player i 's output is assumed to take the following form:

$$O_i = l_i e_i, \quad i = 1, 2,$$

where l_i is the time spent in production and e_i is a random variable distributed according $h(e_i)$.¹ Assume that $E(e_i) = 1$ and $\text{var}(e_i) = \sigma^2$. The random shocks are potentially correlated, i.e., $\sigma_{ij} = \text{cov}(e_i, e_j) = E(e_i e_j) \neq 0$. That is, e_i and e_j are moving in the same (opposite) direction if $\text{cov}(e_i, e_j) > (<) 0$. Note that e_i is an idiosyncratic shock that affects players' output in a multiplicative manner and is assumed to be individual-specific.

Besides farming, the players can also invest time in improving their fighting skills:

$$(1) \quad S_i = f_i v_i, \quad i = 1, 2,$$

where f_i is the time investment in fight-related activities and v_i is a random variable with $E(v_i) = 1$ and $\text{var}(v_i) = 1$. For simplicity, further assume that v_1 and v_2 are independent of each other. Players decide whether or not to fight at the beginning of stage 2. A fight will occur as long as one of them decides to do so. The winner will take all the output from the loser. Given (1), the sufficient and necessary condition for player i to win the fight is

$$S_i = f_i v_i > S_j = f_j v_j.$$

Let $v = v_j/v_i$. The cumulative density function for v is $F(v)$, which is continuous and differentiable. Thus, the winning probability for player i ($i = 1, 2$) is

$$\begin{aligned} P_i &= P_i \left(v \leq \frac{f_i}{f_j} \right) \\ &= F \left(\frac{f_i}{f_j} \right), \quad \text{with } F' \left(\frac{f_i}{f_j} \right) > 0. \end{aligned}$$

Note that $P_i + P_j = 1$. Farming and fight-related training are costly. The consumption-enhancing activities through legitimate means such as farming may carry

¹ Normalization with unit properly chosen allows us to avoid output being negative.

a different meaning, at least at a personal level, from that of the illegitimate ones such as fighting. Thus, it is not unreasonable to assume that the cost function $C_i(l_i, f_i)$ displays imperfect substitutability in its arguments; accordingly, we set

$$C_i(l_i, f_i) = \frac{1}{2} l_i^2 + \frac{1}{2} f_i^2.$$

Player i 's preference is assumed to be separable and is given by

$$U_i(c, l_i, f_i) = u(c) - C_i(l_i, f_i),$$

where $u(c)$ is the utility derived from c units of consumption. Since outputs are subject to some random disturbances, c is therefore random. Player i 's expected utility function can be written as

$$\begin{aligned} EU_i(c, l_i, f_i) &= E[u(c) - C_i(l_i, f_i)] \\ &= E(c) - k \text{var}(c) - \frac{1}{2} l_i^2 - \frac{1}{2} f_i^2. \end{aligned}$$

Note that the expected utility depends only on mean and variance. We implicitly assume that either the utility is quadratic or the density function is normal.

3 Market Correlation and Property Rights

In his paper, MUTHOO [2004] shows that in order to sustain property rights as an equilibrium outcome, a repeated game must be introduced and players must be patient enough. Here, we show that even in a one-shot two-stage game, property rights can still emerge as an equilibrium outcome when productions are subject to some random disturbances.² In what follows, we will characterize the subgame-perfect equilibrium. The problem is solved by backward induction, beginning with the stage-2 game.

3.1 Stage 2

At the beginning of stage 2, player i observes his own output $l_i e_i$ but has no information regarding the output of his rival, $l_j e_j$. Player i determines whether to engage in a fight with his rival. Fighting will produce a winner and a loser. When a fight occurs, player i will receive $l_i e_i + l_j e_j$ with probability $F(f_i/f_j)$, and 0 with probability $1 - F(f_i/f_j)$. Thus, player i 's expected payoff is

$$\begin{aligned} EU_i^f &= Eu \left[F \left(\frac{f_i}{f_j} \right) (l_i e_i + l_j e_j) \right] \\ &= F \left(\frac{f_i}{f_j} \right) [l_i e_i + (l_j - k l_j^2 \sigma^2)]. \end{aligned}$$

² One can easily find examples that relate wars to market correlations. For example, the nomadic Xiongnu were a constant threat to China's northern frontier for many centuries. Fights between Xiongnu and Han Chinese were often driven by the negative correlation in food productions due to significant differences in their weather patterns.

With no fight, player i 's expected payoff is

$$\begin{aligned} EU_i^n &= Eu(l_i e_i) \\ &= l_i e_i . \end{aligned}$$

Let Ψ_i be player i 's incentive to fight:

$$(2) \quad \Psi_i = EU_i^f - EU_i^n ,$$

where $i, j = 1, 2$ and $i \neq j$. If $\Psi_i > 0$, the expected benefit from fighting exceeds the cost associated with it. In this case, player i will choose to fight. Conversely, if $\Psi_i < 0$, the benefit is too small to justify a fight for player i . His willingness to establish property rights is therefore characterized by the following condition:

$$(3) \quad \begin{aligned} \Psi_i &= P_i(l_j - kl_j^2\sigma^2) - P_j l_i e_i \\ &= F\left(\frac{f_i}{f_j}\right)(l_j - kl_j^2\sigma^2) - \left[1 - F\left(\frac{f_i}{f_j}\right)\right]l_i e_i < 0 \quad \text{for } i = 1, 2 . \end{aligned}$$

Under certainty, $\sigma^2 = 0$ and $e_i = 1$. The above condition is reduced to $P_i l_j - P_j l_i < 0$,³ confirming MUTHOO's [2004] result that fighting does not take place if $P_i l_j = P_j l_i$ but fighting does occur, either voluntarily or involuntarily, if $P_i l_j \neq P_j l_i$.⁴ In the presence of uncertainty, it is clear from (3) that there is a wide range of σ^2 such that $\Psi_i < 0$. Thus, even if $P_i l_j \neq P_j l_i e_i$, fighting may not always happen. For property rights to emerge, one requires $\Psi_i < 0$, which holds if

$$(4) \quad \sigma^2 > \frac{F\left(\frac{f_i}{f_j}\right)l_i e_i + F\left(\frac{f_i}{f_j}\right)l_j - l_i e_i}{F\left(\frac{f_i}{f_j}\right)kl_j^2} = \widehat{\sigma}_i^2$$

for $i = 1, 2$. Note that $\widehat{\sigma}_i^2$ is the cutoff for i below which i will choose to fight. This yields

PROPOSITION 1 *When players observe only their own outputs at the beginning of stage 2, property rights emerge in stage 2 if and only if $\sigma^2 > \widehat{\sigma}_i^2$ for $i = 1, 2$.*

The intuition behind Proposition 1 is simple. In our model, player i ($i = 1, 2$) makes the stage-2 decision of whether or not to fight after knowing what his output is. At this point, player j 's output is not unknown to player i . Higher volatility will adversely affect player i 's incentive to fight. While winning a fight is good for a player, it comes at a cost, particularly when the degree of uncertainty about j 's output is high. Higher σ^2 increases the overall income risk, thus making fighting less attractive. Player i is less likely to fight.

³ When $\Psi_i = 0$, the players are indifferent between fighting and no fighting. We assume that no fighting will occur when $\Psi_i = 0$.

⁴ Under certainty, $\Psi_i + \Psi_j = 0$ ($i, j = 1, 2$ and $i \neq j$). When $P_i l_j \neq P_j l_i$, $\Psi_i > 0$ implies $\Psi_j < 0$ and vice versa. Thus, fighting will always occur in this case.

As shown above, player i will get higher payoff with property rights than without when $\Psi_i < 0$. It is clear from (3) that $\Psi_i < 0$ if

$$e_i \geq \frac{F\left(\frac{f_i}{f_j}\right)l_j(1 - kl_j\sigma^2)}{\left[1 - F\left(\frac{f_i}{f_j}\right)\right]l_i} = e_i^0.$$

Given this, the probability that player i chooses not to fight in stage 2 is therefore

$$Q_i(e_i^0) = 1 - H(e_i^0),$$

where $H(e_i)$ is the cdf of e_i . Clearly, $H'(e_i^0) > 0$ and $Q'_i(e_i^0) < 0$. Given e_i^0 , it is easily verified that⁵

$$(5) \quad \begin{aligned} & \text{(i) } \frac{\partial e_i^0}{\partial f_i} > 0, \quad \text{(ii) } \frac{\partial e_i^0}{\partial f_j} < 0, \quad \text{(iii) } \frac{\partial e_i^0}{\partial l_i} < 0, \quad \text{(iv) } \frac{\partial e_i^0}{\partial l_j} > 0, \\ & \text{(v) } \frac{\partial e_i^0}{\partial \sigma^2} < 0, \quad \text{(vi) } \frac{\partial e_i^0}{\partial k} < 0. \end{aligned}$$

Given $Q'_i(e_i^0) < 0$, results (i) and (ii) say that player i will be more (less) likely to fight in stage 2 when he (his rival) invests more in fight-related activities in stage 1. Intuitively, investment in fighting skills will become sunk if fighting does not occur. Those who invest heavily in fighting will have a better chance of defeating their enemies and will therefore be more inclined to fight. Conversely, a heavy investment in training by j will consequently discourage any aggressive behavior by i .

Similarly, results (iii) and (iv) say that a player will be less likely to gamble the fruit of his effort when he (his rival) invests more (less) in farming. Clearly, more investment in farming generally implies a higher level of output. With more at stake, the risk-averse players are more likely to adopt a conservative strategy for the sake of risk avoidance. Needless to say, higher outputs will invite his rival to launch an attack against him, since the marginal benefit from an attack increases.

Result (v) illustrates the effect of market volatility on a risk-averse player. It indicates that higher volatility will adversely affect player's incentive to fight. Winning a fight is good for a player, but it comes at a cost. At the beginning of stage 2, player i knows his own output but has no information about j 's output. Higher uncertainty about j 's output increases overall income risk, thus making fighting less attractive. Therefore, player i is less likely to fight.

Result (vi) is obvious. It implies that more risk-averse individuals are reluctant to take risks, resulting in a lower probability of fighting.

To summarize, we obtain

PROPOSITION 2 *Player i 's willingness to fight increases with f_i and l_j but decreases with increasing f_j , l_i , k , and σ^2 .*

⁵ Details can be found in Appendix A.1.

3.2 Stage 1

If either player decides to fight, fighting will occur. Thus, property rights emerge only when both players choose not to fight. This occurs with probability Q :

$$\begin{aligned} Q &= Q_i Q_j \\ &= [1 - H(e_i^0)][1 - H(e_j^0)]. \end{aligned}$$

Given this, the players make their optimal choices so as to maximize their expected utility. As discussed earlier, players' payoffs are contingent on the actions taken by them and on the outcome of a fight. If both players choose not to fight, each player enjoys his own output, $l_i e_i$. But if either of them decides to fight, player i receives $l_i e_i + l_j e_j$ with probability $F(f_i/f_j)$ and 0 with probability $1 - F(f_i/f_j)$. Player i 's expected payoff at the end of stage 2 is therefore $Ql_i e_i + (1 - Q)F(f_i/f_j)(l_i e_i + l_j e_j)$. Let (f_i, l_i) solve player i 's stage-1 maximization problem:

$$(6) \quad \begin{aligned} EU_i &= Q[l_i - kl_i^2\sigma^2] + [1 - Q]F\left(\frac{f_i}{f_j}\right)[l_i + l_j - k(l_i^2\sigma^2 + l_j^2\sigma^2 + 2\sigma_{ij})] \\ &\quad - \frac{1}{2}l_i^2 - \frac{1}{2}f_i^2, \end{aligned}$$

where $i = 1, 2$. Symmetry implies that when a Nash equilibrium exists, $f_i = f_j = f = f(\sigma_{ij}, \sigma^2, k)$ and $l_i = l_j = l = l(\sigma_{ij}, \sigma^2, k)$. By substitution, we have

$$Q = [1 - H(e^0)]^2,$$

where $e_i^0 = e_j^0 = e^0 = 1 - kl\sigma^2$. To ease our calculation, assume that h is uniformly distributed. The effect of covariance on the probability of having peace is summarized in the following proposition:

PROPOSITION 3

- (1) *Property rights are more likely to emerge as an equilibrium outcome if players invest more in farming; but the probability is independent of time investment in training. Specifically, $\partial Q/\partial l > 0$ and $\partial Q/\partial f = 0$.*
- (2) *If σ^2 is sufficiently large that a Nash equilibrium in pure strategy exists, the probability that property rights emerge as an equilibrium outcome increases with the market correlation σ_{ij} .*

PROOF See Appendix A.2.

Proposition 3(1) can be understood as follows. More investment in farming implies higher expected output. With more output at stake, players are therefore reluctant to risk the fruits of their effort, since there is always a positive chance that they may lose the fight. This explains $\partial Q/\partial l > 0$. Aside from this, time spent in training is shown to produce no substantive effect on the fighting probability. This is because $f_i = f_j = f$ in equilibrium and no player has an advantage over the other. The outcome is purely random, yielding $\partial Q/\partial f = 0$. In short, players' stage-1 decisions produce nonmarginal effects on the emergence of property rights.

Proposition 3(2) says that when σ^2 is sufficiently large that a Nash equilibrium in pure strategy exists, property rights are more likely to emerge as an equilibrium outcome when the covariance increases. This result can be explained as follows. A battle will produce a winner and a loser. The obvious benefit of being a winner is that one is able to take over the entire outputs. Apart from that, the winner can enjoy the diversification effect due to pooling. The flip side is that he may end up losing a fight, in which case he loses the fruits of his effort. Clearly, higher covariance reduces the diversification effect, thus making a fight less attractive. Players will more likely choose to be self-sufficient, and consequently, property rights emerge with higher probability. The result is Pareto-improving, since in this case the players devote more time to farming rather than waste their time in preparing for a fight. Unlike Muthoo's result, our result is obtained without relying on a repeated procedure. More importantly, the equilibrium is self-enforcing, since any deviation from it can only make the players worse off.

A casual observation suggests that our result can be used to justify the formation of the European Union (EU). Germany and France are members of the EU, though they had been struggling for dominance in continental Europe for 80 years and had fought many wars. The political climate after the end of World War II favored western European unity, seen by many as an escape from the extreme forms of nationalism, which had devastated the continent. One successful proposal for European cooperation came in 1951 with the European Coal and Steel Community. This had the aim of bringing together control of the coal and steel industries of its member states (principally France and West Germany), so that war between them would no longer be possible. The origin of the EU was in the European Economic Community (EEC), formed in 1957. Since then, the EU has grown in size through the accession of new member states and has increased its powers by the addition of new policy areas to its remit. In 1993, the Maastricht Treaty established the current legal framework. Now, the EU is a political and economic community of twenty-seven member states with supranational and intergovernmental features, located primarily in Europe. The EU operates a single economic market across the territory of all its members and uses a single currency among the 13 members of the euro zone. Considered as a single economy, all the member countries including Germany and France get along peacefully. The emergence of the EU is not surprising, as the states have become closely connected to each other.

4 Conclusion

We have considered a two-player, two-stage noncooperative game in an environment where there are no rules or orders. Players determine their time allocations in stage 1 and then choose whether to fight or not in stage 2. Time allocations along with random shock jointly determine the level of production and players' chances of winning a fight if it indeed occurs. Productions are subject to some idiosyncratic shocks, which are potentially correlated. Our main results include:

- (1) Property rights may emerge as an equilibrium outcome with higher probability when the covariance increases. This result is driven by the fact that the diversification effect diminishes in its magnitude when market correlations increase. Put differently, when risks facing the players are positively correlated, perhaps due to similar weather patterns, the benefit from fighting is reduced and therefore a clash between two sides becomes less likely to occur.
- (2) Symmetry implies $f_i = f_j = f$ in equilibrium, so winning or losing is purely random. The likelihood for property rights to emerge (i.e., Q) depends only on the amount of investment in farming and is independent of time investment in training.
- (3) Player i 's willingness to fight increases with f_i and l_j but decreases with increasing f_j , l_i , k , and σ^2 .

Our results are contingent on the assumption that players can observe only their own output when making their fighting decision and this decision is not reversible. The problem arises if the realized outcomes differ substantially ex post. When this occurs, players may revise their decisions. Thus, ex ante decisions may be revoked in the face of ex post opportunity. This requires the model be extended to a three-stage framework. By limiting ourselves to a two-stage game, we are able to show that there is a strong connection between market correlations and property rights.

Our analysis also rests on the assumption that the random disturbances facing the two players are symmetrical. While this simplifies our analysis, it will be useful to see whether our results hold when players face different production risks. Intuitively, the player facing higher uncertainty is eager to achieve the diversification effect and is therefore more likely to behave aggressively, whereas the one facing lower uncertainty may adopt a more conservative strategy. Whether property rights are more or less likely to emerge remains an open question. We leave it for future inquiry.

Appendix

A.1 Proof of (5)

Given $e_i^0 = F(f_i/f_j) l_j(1 - kl_j\sigma^2)/[1 - F(f_i/f_j)]l_i$, it is straightforward to obtain

$$\frac{\partial e_i^0}{\partial f_i} = \frac{F' \left(\frac{f_i}{f_j} \right) l_j (1 - kl_j \sigma^2)}{\left[1 - F \left(\frac{f_i}{f_j} \right) \right]^2 l_i f_j} > 0, \quad \frac{\partial e_i^0}{\partial f_j} = \frac{F' \left(\frac{f_i}{f_j} \right) f_i l_j (-1 + kl_j \sigma^2)}{\left[1 - F \left(\frac{f_i}{f_j} \right) \right]^2 l_i f_j^2} < 0,$$

$$\frac{\partial e_i^0}{\partial l_i} = \frac{-F \left(\frac{f_i}{f_j} \right) l_j (1 - kl_j \sigma^2)}{\left[1 - F \left(\frac{f_i}{f_j} \right) \right] l_i^2} < 0, \quad \frac{\partial e_i^0}{\partial l_j} = \frac{F \left(\frac{f_i}{f_j} \right) (1 - 2kl_j \sigma^2)}{\left[1 - F \left(\frac{f_i}{f_j} \right) \right] l_i} > 0,$$

$$\frac{\partial e_i^0}{\partial \sigma^2} = \frac{-F\left(\frac{f_i}{f_j}\right)l_j^2k}{\left[1 - F\left(\frac{f_i}{f_j}\right)\right]l_i} < 0, \quad \frac{\partial e_i^0}{\partial k} = \frac{-F\left(\frac{f_i}{f_j}\right)l_j^2\sigma^2}{\left[1 - F\left(\frac{f_i}{f_j}\right)\right]l_i} < 0,$$

since $1 - kl_j\sigma_j^2 > 0^6$ and $F'(f_i/f_j) > 0$.

This gives (5).

Q.E.D.

A.2 Proof of Proposition 3

Maximization of (6) with respect to f_i and l_i yields the following first-order conditions:

(A1)

$$\frac{\partial EU_i}{\partial f_i} = \frac{\partial Q}{\partial f_i} [l_i - kl_i^2\sigma^2] - \frac{\partial Q}{\partial f_i} F[l_i + l_j - k(l_i^2\sigma^2 + l_j^2\sigma^2 + 2l_i l_j \sigma_{ij})] - \frac{1}{f_j} (1 - Q)F'[l_i + l_j - k(l_i^2\sigma^2 + l_j^2\sigma^2 + 2l_i l_j \sigma_{ij})] - f_i = 0,$$

(A2)

$$\frac{\partial EU_i}{\partial l_i} = \frac{\partial Q}{\partial l_i} [l_i - kl_i^2\sigma^2] + Q[1 - 2kl_i\sigma^2] - \frac{\partial Q}{\partial l_i} F[l_i + l_j - k(l_i^2\sigma^2 + l_j^2\sigma^2 + 2l_i l_j \sigma_{ij})] + (1 - Q)F[1 - k(2l_i\sigma^2 + 2l_j\sigma_{ij})] - l_i = 0,$$

where $i = 1, 2$. Symmetry implies that when a Nash equilibrium exists, $f_i = f_j = f$ and $l_i = l_j = l$. Thus, $e_i^0 = e_j^0 = e^0 = 1 - kl\sigma^2$. Moreover, these equations also imply that $Q = [1 - H(e^0)]^2$ and $F(f_i/f_j) = F(1) = 1/2$. Taking the derivatives of Q with respect to f_i and l_i and evaluating them at $(f_i, l_i, f_j, l_j) = (f, l, f, l)$, we obtain

(A3)

$$\left. \frac{\partial Q}{\partial f} \right|_{(f,l,f,l)} = 0,$$

(A4)

$$\left. \frac{\partial Q}{\partial l} \right|_{(f,l,f,l)} = 2H'(e^0)[1 - H(e^0)]k\sigma^2 > 0,$$

since $H'(e^0) = \partial H(e^0)/\partial e^0 > 0$ and $1 - H(e^0) > 0$. This proves Proposition 3(1).

Substituting (A3) and (A4) into (A1) and (A2), and then using symmetry, we have

(A5)

$$\frac{\partial EU_i}{\partial f} = \frac{2[1 - (1 - H(e^0))^2]F'(1)[l - k(l^2\sigma^2 + l^2\sigma_{ij})]}{f} - f = 0,$$

(A6)

$$\frac{\partial EU_i}{\partial l} = \frac{1}{2}[1 - H(e^0)]^2[1 - 2kl(\sigma^2 - \sigma_{ij})] + H'(e^0)[1 - H(e^0)]k^2l^2\sigma^2\sigma_{ij} + \frac{1}{2}[1 - 2kl(\sigma^2 + \sigma_{ij})] - l = 0.$$

Our primary interest is to examine the effect of market correlation on the emergence of property rights. Since f produces no effect on Q (see (A3)), we can simply

⁶ We assume that $Eu(l_j e_j) = l_j - kl_j^2\sigma_j^2 = l_j(1 - kl_j\sigma_j^2) \geq 0$. This implies that $1 - kl_j\sigma_j^2 \geq 0$.

focus our attention on the effect of σ_{ij} on Q through a change in l . Further note that (A6) is independent of f ; that is, $\partial^2 EU_i / \partial l \partial f = 0$. A simple differentiation of (A6) with respect to σ_{ij} and l yields

$$(A7) \quad \frac{\partial l}{\partial \sigma_{ij}} = - \frac{\partial^2 EU_i}{\partial l \partial \sigma_{ij}} \bigg/ \frac{\partial^2 EU_i}{\partial l^2},$$

where

$$\partial^2 EU_i / \partial l^2 < 0$$

(the second-order condition assumed to be satisfied) and

$$\partial^2 EU_i / \partial l \partial \sigma_{ij} = [kl\sigma^2(1 - H(e^0))H'(e^0) + (1 - H(e^0))^2 - 1]kl.$$

One can easily verify that $\partial^2 EU_i / \partial l \partial \sigma_{ij}$ is increasing in σ^2 when h is uniformly distributed. Thus, there exists a $\tilde{\sigma}^2$ such that $\partial^2 EU_i / \partial l \partial \sigma_{ij} = 0$. For $\sigma^2 > \tilde{\sigma}^2$, we have $\partial^2 EU_i / \partial l \partial \sigma_{ij} > 0$.⁷ This along with (A4) (i.e., $\partial Q / \partial l > 0$) implies that

$$\frac{\partial Q}{\partial \sigma_{ij}} = \frac{\partial Q}{\partial l} \frac{\partial l}{\partial \sigma_{ij}} > 0$$

if σ^2 is sufficiently large. This proves Proposition 3(2).

Q.E.D.

References

- CHEUNG, S. [1963], "Private Property Rights and Sharecropping," *Journal of Political Economy*, 76, 1107–1122.
- COASE, R. H. [1960], "The Problem of Social Cost," *The Journal of Law & Economics*, 3, 1–23.
- DEMSETZ, H. [1967], "Toward a Theory of Property Rights," *The American Economic Review*, 57, 347–359.
- ESTEBAN, J., AND D. RAY [1999], "Conflict and Distribution," *Journal of Economic Theory*, 82, 379–415.
- GROSSMAN, S., AND O. HART [1986], "The Costs and Benefits of Ownership: A Theory of Lateral and Vertical Integration," *Journal of Political Economy*, 94, 691–719.
- HAFER, C. [2006], "On the Origins of Property Rights: Conflict and Production in the State of Nature," *The Review of Economic Studies*, 73, 119–143.
- HIRSHLEIFER, J. [1995], "Anarchy and its Breakdown," *Journal of Political Economy*, 103, 26–52.
- LAZEAR, E. P., AND S. ROSEN [1981], "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 89, 841–864.
- MUTHOO, A. [2004], "A Model of the Origins of Basic Property Rights," *Games and Economic Behavior*, 49, 288–312.

⁷ For tournament models, LAZEAR AND ROSEN [1981, p. 845] note that "a pure strategy solution exists provided that σ^2 is sufficiently large: Contests are feasible only when chance is a significant factor. This result accords with intuition and is in the spirit of the old saying that a (sufficient) difference of opinion is necessary for a horse race." In light of this, σ^2 must be large enough to ensure the existence of a Nash equilibrium in pure strategies. To simplify, we assume that the required level of σ^2 exceeds $\tilde{\sigma}^2$.

SKAPERDAS, S. [1992], "Cooperation, Conflict, and Power in the Absence of Property Rights," *The American Economic Review*, 82, 720–739.

Shin-Hwan Chiang
Xiang Li
Department of Economics
York University
4700 Keele Street
Toronto, Ontario
M3J 1P3
Canada
E-mail:
schiang@econ.yorku.ca
lixiang@econ.yorku.ca

Strategic Unemployment

by

JULIA ANGERHAUSEN, CHRISTIAN BAYER, AND BURKHARD HEHENKAMP*

The empirical literature on happiness finds that employment significantly contributes to well-being. We propose a dynamic model that explains why individuals may nonetheless be reluctant to pick up low-paid work. Accepting low-paid work will put them in an adverse position in future wage bargaining, as employers could infer the individual's low reservation wage from his working history. Employers will exploit their knowledge by offering low wages to this individual in the future. Therefore, employees with low reservation wage *strategically* opt into unemployment to signal a high reservation wage. (JEL: D 82, J 30, J 64)

1 Introduction

A standard assumption in the economics literature is that there is a significant disutility from work. However, the empirical literature on happiness and unemployment suggests that this disutility of work does not exist for the majority of workers (see FREY AND STUTZER [2002] for an overview). A typical finding is that unemployment spells affect happiness in an adverse way that goes beyond the loss in income on average. For example, the estimates of FRIJTERS, HAIKEN-DENEW, AND SHIELDS [2004] imply a utility *gain* from employment that is roughly twice as large as the utility gain from the income earned.¹ Conversely, even when the income level is controlled for, unemployment correlates with substantial unhappiness (CLARK AND OSWALD [1994]; WINKELMANN AND WINKELMANN [1998]; FRIJTERS, HAIKEN-DENEW, AND SHIELDS [2004]; FRIJTERS et al. [2006]; CLARK, FRIJTERS, AND SHIELDS [2006]).

* Technische Universität Dortmund, Università Commerciale L. Bocconi, and Technische Universität Dortmund (corresponding author). We thank two anonymous referees, Patrick Herbst, Kornelius Kraft, Wolfgang Leininger, Johannes Münster, Wolfram Richter, Oz Shy, and seminar and conference participants at the SSRC Berlin, the University of Bergen, Universität Dortmund, and the GEABA 2006 for helpful comments and discussion.

¹ The estimates by WINKELMANN AND WINKELMANN [1998] imply even larger compensating differentials. CLARK AND OSWALD [1994] argue that the correlation should be viewed as a causality running from unemployment to unhappiness; see also WARR, JACKSON, AND BANKS [1988].

In contrast, the macro labor literature has estimated a substantial subjective disutility from work. In the standard search-and-matching setup a job searcher balances a wage offer against her outside alternative, which is made up of unemployment benefits, the subjective (dis)utility of work, and the continuation value of labor market search (see, e.g., CHRISTOFFEL AND KUESTER [2008]). In such a setting, COSTAIN AND REITER [2008] find that the instantaneous opportunity cost of work, i.e., the sum of unemployment benefits and subjective disutility of work, has to be roughly 75% of average productivity in order to make their business-cycle version of the search-and-matching model (MORTENSEN AND PISSARIDES [1994]; PISSARIDES [1985], [2000]) match the aggregate data. HAGEDORN AND MANOVSKII [2008] obtain even higher estimates (95%). Unemployment benefits imply replacement rates substantially below these numbers. Since the difference between replacement rates and the estimated instantaneous opportunity costs of work is the disutility from work, these findings imply the latter to be significant in size. In turn, the empirical findings from the microeconomic happiness literature are at odds with the findings from the macroeconomic literature – at least at first sight.

In other words, the discrepancy between micro and macro findings calls for a theory in which employees typically *behave* as if they suffer from a significant disutility of work, even if working in fact creates utility for most of them. In this paper, we provide such a theory, building on the asymmetry of information about reservation wages. This is how we suggest reconciling the seemingly contradictory findings described above.

We put forward a model in which a worker meets a monopsonistic employer and behaves as if his payoff from not working is large even if it is low in fact. This behavior stems from an asymmetry of information about reservation wages. Specifically, we analyze a two-period model with two principals and one agent. Each principal is incompletely informed about the agent's reservation wage. Principal 1 offers a wage contract in period 1, and the agent decides whether to accept it or not. Being informed about the agent's decision in period 1, principal 2 offers a wage contract in period 2 that the agent may again accept or reject.

This complicates the decision problem to the agent in period 1, as his first-period behavior sends a signal to future employers. By rejecting an employment offer, he can signal a high disutility of work or – in the context of a search-and-matching setup – a high continuation value of search. Conversely, accepting an employment offer, the agent reveals his reservation wage to be at most the offered wage. As a consequence, future employers will not be willing to make better offers. The associated signaling activity can result in unemployment when screening the agents' types is either ineffective or too costly to principal 1. The corresponding type of unemployment is what we call *strategic unemployment* in the following. Strategic unemployment is voluntary, but – as will turn out – second-best inefficient.²

² A related result of unemployment due to strategic reasons has been obtained by MA AND WEISS [1993]. They present a model in which unemployment serves as a device to burn money in order to signal productivity.

Our model is set up general enough to be applied to other decisions than the one to work or not to work. A particular example is the stay-or-go decision of CEOs and opera, sport, and other superstars. A superstar may have a high or a low *personal* inclination to switch employers. In the very beginning of his career this will be his private information. Later on, when headhunters have come into play with first bids to make the superstar switch employers, that is no longer the case. In this context, a superstar with a high switching inclination might profit from mimicking superstars with a low switching inclination and decline to switch employers for a small increase in income in order to earn better offers in the future.

The setup of our model resembles that of the seminal paper by HART AND TIROLE [1988]. They address the issue of contract renegotiation in a multiperiod buyer–seller model, where the seller is incompletely informed about the buyer’s reservation price and where all bargaining power goes with the seller. Consequently, revelation of information in early periods is very costly to the buyer in the later periods of play, so that extensive pooling takes place. VINCENT [1998] departs from the strongly asymmetric distribution of bargaining power, investigating linear-pricing contracts (as opposed to the nonlinear-pricing contracts that maximize a monopolistic seller’s profit). If the seller’s bargaining power is reduced in this way, there is comparatively more revelation of information in early periods.

Having the labor market in mind, it is natural to assume that asymmetric information about reservation wages goes with the buyer, which is the firm here. Unlike firms’ profits, workers’ disutility from work primarily represents a psychological construct, which is much harder to observe. HART [1983, section 5.C] and MOORE [1985] propose models in this spirit.

Both HART [1983] and MOORE [1985] consider the case of privately observed reservation wages. HART [1983] focuses on the productive inefficiency resulting from asymmetry in information. MOORE [1985] addresses the issue of involuntary layoffs and retentions. Each author examines a multiperiod model where firms propose long-term contracts to workers in period 1. At the time of contracting, the reservation wage is unknown to both parties. Subsequently, workers learn their reservation wage and may report it to the firm thereafter. Now, the firm decides whether to lay off the worker or to continue the relationship, paying a wage conditional on the reported reservation wage. The terms contracted on in the first period apply to both of these options, and renegotiation of the contract is assumed to be infeasible.

Our setup differs in two important aspects. First, we concentrate on reservation wages that are private information to the worker already at the instant of contracting. Second, we extend the model to two periods of contracting, which leads to the key element of our paper. A firm may learn an agent’s reservation wage from his employment history. Thus, agents with a low reservation wage are reluctant to pick up badly paid jobs, as this has an adverse effect on the prospects of future earnings. As a consequence, strategic unemployment results from contracts that are *not* signed, even though employment would be first-best efficient.

The remainder of this paper is organized as follows: In section 2, we set up the model and solve for strategies and beliefs in weak perfect Bayesian equilibrium. We proceed with computing strategic and nonstrategic unemployment. Section 3 discusses the robustness of our results and three possible extensions. First, it examines to what extent the two-period setup can be interpreted in the same way as a model with an infinite time horizon. Second, it analyzes in how far vertical integration of the principals, firing costs, or a legal minimum wage can improve welfare by reducing strategic unemployment. Third, it shows that our baseline results from a model with two worker types carry over to a more general setting with infinitely many types. Section 4 concludes.

2 The Basic Model

In this section, we set up a model where employers (principals) in the labor market have imperfect information on the reservation wages of potential employees (agents). Subsequently, we solve for weak perfect Bayesian equilibrium.

2.1 Principals

We consider a situation in which employers randomly draw projects that have a fixed value of revenues π . The employer needs an agent to implement the project and generate the revenues. He randomly meets agents and bargains about the amount of the wage payment (as in JOVANOVIĆ [1979]). For simplicity, we assume that all bargaining power is with the principal.³ Consequently, bargaining takes the form of take-it-or-leave-it offers. The principal does not know the reservation wage of the agent, but is aware of his employment history including past wages.⁴ To keep the model simple, we consider a two-period situation. In each period $t = 1, 2$, the profitability π_t of the project is probabilistic. Profitability is independently and identically distributed according to a distribution function G , which is assumed continuous and increasing on a compact support. Furthermore, the profitability of a project is strictly positive ($\Pr\{\pi_t > 0\} = 1$).

Having learned about the profitability of his project, an employer makes a take-it-or-leave-it wage offer of w_t to the agent. To exclude strategic behavior on behalf of the principals, an agent does not work for the same principal in both periods. In the first period (the present), the employment history is completely uninformative about the reservation wage of the agent. In the second period (the future), the employer draws a new project and meets another agent. Accordingly, from the perspective of the employer, there is no strategic interaction between the two periods.

³ This can be seen as a simplification of HAGEDORN AND MANOVSKII'S [2008] result that the correct calibration of a search-and-matching business-cycle model requires very low bargaining power of workers (bargaining power in the generalized Nash solution equal to 0.06).

⁴ In the Appendix, we discuss what happens if the principal can only observe accepted wage offers.

2.2 Agents

However, for employees there is such interaction. Principals learn about an employee's reservation wage through his working history. For this reason, we formulate the model from the point of view of an agent, who meets a different principal in each period of his working life.

Agents have a type $\theta \in \{\underline{\theta}, \bar{\theta}\}$, which reflects the agent's disutility from labor and his opportunity costs from working (loss of unemployment benefit, home production, etc.). In what follows we refer to θ as an agent's reservation wage or, slightly imprecisely, as his disutility from labor. The reservation wage θ can be either high or low. The difference in reservation wages may, for example, represent different effort costs or differences in the continuation value of search, if we interpret our model as embedded in a larger search-and-matching framework. For the ease of exposition, the low reservation wage is normalized to zero: $\underline{\theta} = 0$, which ensures that it is always profitable for the principal to employ a low-type agent at his reservation wage.⁵ The probability of an agent's being of the low type is p . The agent is aware that potential employers cannot observe his reservation wage. Hence, he has to take into account that the acceptance or rejection of an employment offer will shape the beliefs of potential future employers with respect to his reservation wage.

2.3 Chronology

This yields the following chronology of events within our model: First, nature draws the profitability π_1 of principal 1's project and the agent's reservation wage θ , where $p = \Pr(\theta = \underline{\theta}) \in (0, 1)$ is the probability of the agent's being of the low type.

Knowing the profitability of his project, principal 1 then makes a wage offer w_1 . In a next step, the agent accepts the offer (a) and receives $w_1 - \theta$, or rejects it ($\neg a$) and receives $\underline{U} = 0$.

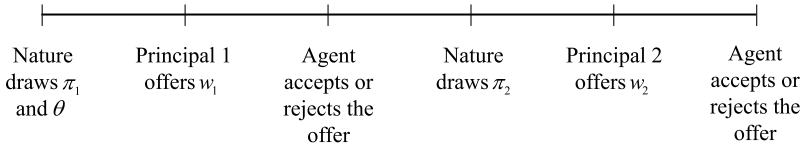
In the second period, nature draws the profitability of principal 2's project, π_2 , from the distribution G . Afterwards, principal 2 makes a wage offer w_2 , and again, the agent accepts it (a) and receives $w_2 - \theta$, or rejects it ($\neg a$) and receives \underline{U} . Figure 1 gives an overview of the succession of events.

2.4 Solution

The model is solved via backward induction. Consequently, we first characterize the agent's behavior facing possible wage offers in the second period. Anticipating

⁵ Suppose $\underline{\theta} \neq 0$. Then we can rewrite the model in terms of gains from employment instead of revenues, defining a new distribution function $\tilde{G} = G(\pi - \underline{\theta})$ as well as shifted reservation wages $\tilde{\underline{\theta}} = \underline{\theta} - \underline{\theta} = 0$ and $\tilde{\bar{\theta}} = \bar{\theta} - \underline{\theta}$. Even though some agent might experience a utility gain from working, it is unlikely that the sum of lost unemployment benefits and utility gain is positive, and hence we typically have that the sum of disutility from working, ω , and unemployment benefit, b – i.e., $b + \omega =: \underline{\theta}$ – is positive. Otherwise our model would imply that we should counterfactually observe negative wages from time to time.

Figure 1
Chronology of the Events



this behavior, the principal chooses an optimal wage offer that is based on his belief about the type of the agent. In the first period, the agent thus takes into account how accepting or rejecting the wage offer in period 1 will shape this belief. Finally, principal 1 offers a profit-maximizing wage to the agent in period 1.

2.4.1 Second Period

In the second period, all types of agents accept any wage offer that is larger than their reservation wage, $w_2 \geq \theta$. This implies that $\underline{\theta}$ -types will accept any positive wage offer, as their reservation wage equals zero. On the other hand, the high-type agents only accept wage offers that comply with $w_2 \geq \bar{\theta}$.

Principal 2 wants to pay a wage that is as low as possible, but needs to take into account that agents only accept wage offers that exceed their reservation wage. As the highest reservation wage is $\bar{\theta}$, a principal will never make a wage offer larger than this. The wage $w_2 = \bar{\theta}$ is sufficiently high to be accepted by all types of agents.

All the same, the principal may offer a low wage $w_2 = \underline{\theta} = 0$ that only meets with the reservation wage of the low-type agent. This strategy is risky. Agents with a high reservation wage will reject the offer $w_2 < \bar{\theta}$, and then the project cannot be implemented. The principal believes the agent to be of the low type $\underline{\theta}$ with probability $\mu(s_1, w_1)$ (and of the high-type agent with probability $1 - \mu(s_1, w_1)$, accordingly). This belief is based on the employment history (s_1, w_1) , where $s_1 \in \{a, \neg a\}$ refers to acceptance or rejection of the first-period wage offer w_1 .⁶

For principal 2 to make the low wage offer, the expected profit from choosing $w_2 = \underline{\theta}$ must be larger than the expected profit from a wage offer $w_2 = \bar{\theta}$. In the former case, this amounts to $\mu(s_1, w_1) \pi_2$, in the latter to $\pi_2 - \bar{\theta}$. Hence, he offers $w_2 = 0$ if

$$\mu(s_1, w_1) \pi_2 > \pi_2 - \bar{\theta},$$

and $w_2 = \bar{\theta}$ otherwise. The more likely the principal considers the low type to be, the higher is the probability of a low wage offer. On the other hand, the higher the revenue from the project, the more profitable is a high wage offer (unless

⁶ For expositional simplicity, we assume that principal 2 can observe rejected wage offers. In the Appendix, we discuss changes that would result from the alternative specification where rejected wage offers remain unobserved.

$\mu(s_1, w_1) = 1$, where $w_2 = \underline{\theta}$ is always optimal). This trade-off between the two possible wage offers characterizes a threshold for π_2 , which is

$$\bar{\pi}_2(\mu(s_1, w_1)) = \frac{\bar{\theta}}{1 - \mu(s_1, w_1)}.$$

For all revenue realizations below this threshold, the agent receives a low wage offer. Consequently, the low-type agent’s expected payoff in period 2 is

$$\begin{aligned} V(s_1, w_1|\underline{\theta}) &= \bar{\theta}[1 - G(\bar{\pi}_2(\mu(s_1, w_1)))] + \underbrace{\underline{\theta}G(\bar{\pi}_2(\mu(s_1, w_1)))}_{=0} \\ &= \bar{\theta}\underbrace{[1 - G(\bar{\pi}_2(\mu(s_1, w_1)))]}_{\Pr(\pi \geq \bar{\pi}_2(\mu(s_1, w_1)))}. \end{aligned}$$

Since $V(s_1, w_1|\underline{\theta})$ depends on the principal’s belief μ , there are two critical values of V . The value

$$\bar{V} = \bar{\theta}[1 - G(\bar{\theta})]$$

corresponds to the fully informative belief $\mu(s_1, w_1) = 0$, which assumes that all agents with history (s_1, w_1) are of the high type. Conversely,

$$\underline{V} = \bar{\theta} \left[1 - G\left(\frac{\bar{\theta}}{1 - p}\right) \right]$$

corresponds to the belief $\mu(s_1, w_1) = p$, i.e., the first-period ex ante probability of the low type. Put differently, history is completely uninformative about the agent’s type.

For a high-type agent, the expected second-period payoff is invariant to the principal’s belief, so that $V(s_1, w_1|\bar{\theta})$ is constant. Specifically, we have $V(s_1, w_1|\bar{\theta}) = 0$, as the high-type agent either is paid his reservation wage or will be unemployed.

2.4.2 First Period

Since the high type’s expected second-period payoff is invariant to the working history (s_1, w_1) , he cannot gain from acting strategically in the first period. So the high-type agent accepts a wage offer w_1 if

$$(1) \quad w_1 \geq \bar{\theta}$$

and rejects otherwise.

The decision-making of the agent with a low reservation wage is less straightforward. He has to take into account that working in period 1 may affect the beliefs of future employers. This, in turn, will have consequences for his second-period payoff. As a result, he agrees to work for a wage w_1 only if

$$(2) \quad w_1 - \underline{\theta} + \delta V(a, w_1|\underline{\theta}) \geq \delta V(\neg a, w_1|\underline{\theta}).$$

The inequality compares the discounted expected payoff over both periods for the two alternatives, acceptance and rejection. The discount factor is δ . Inserting $\underline{\theta} = 0$,

inequality (2) reduces to

$$(3) \quad w_1 \geq \delta[V(\neg a, w_1|\underline{\theta}) - V(a, w_1|\underline{\theta})] =: \delta\Delta(w_1).$$

In words, the wage offer in period 1 must compensate for the discounted differential $\delta\Delta$ in information rents $-V(s_1, w_1|\underline{\theta})$, $s_1 \in \{a, \neg a\}$ – that the low-type agent could realize in period 2.

2.4.3 Equilibrium

Considering a model with incomplete information, we solve for weak perfect Bayesian equilibrium. Accordingly, we have to determine both equilibrium strategies and equilibrium beliefs.

In equilibrium, any wage offer in period 1 larger than the high reservation wage, $w_1 \geq \bar{\theta}$, will be accepted by all agents. We have already argued that the high type always accepts this offer. Taking this into account, the low type would reveal his type if he rejected the offer, thus decreasing the value of future income. Therefore, he accepts the offer, expecting a future income of \underline{V} . Wage offers above $\bar{\theta}$ do not discriminate either type. In other words, acceptance of any wage offer $w_1 \geq \bar{\theta}$ does not create information, so that principal 2 continues to have the belief $\mu(a, w_1 \geq \bar{\theta}) = p$.⁷

In contrast to the above, the high-type agent will reject any offer $w_1 < \bar{\theta}$. This generates an incentive for the low type to mimic the behavior of the high type. To influence the principal’s belief, he may reject wage offers that are above his reservation wage but below the high type’s reservation wage. He *strategically* opts for unemployment.

In equilibrium, the principal’s belief $\mu(s_1, w_1)$ in facing a low-type agent has to be equal to the true probability of observing this type conditional on the employment history (s_1, w_1) . Let $q(w_1)$ be the probability of the agent’s accepting a wage offer w_1 . As p is the probability of a low type in the population, since all high types reject an offer $w_1 < \bar{\theta}$ and since $1 - q(w_1)$ is the probability of a low type’s to rejecting, we obtain for $w_1 < \bar{\theta}$

$$(4) \quad \mu(\neg a, w_1) = \frac{\overbrace{p(1 - q(w_1))}^{\text{Probability of low type AND rejection}}}{\underbrace{(1 - p) + p(1 - q(w_1))}_{\text{Probability of rejection}}} = p \frac{1 - q(w_1)}{1 - pq(w_1)},$$

$$\mu(a, w_1) = 1$$

as an equilibrium condition.

If $q \in (0, 1)$, then the agent mixes over the pure strategies “rejection” ($\neg a$) and “acceptance” (a). However, the agent will choose a mixed strategy only if the underlying belief of the principal makes him indifferent between accepting and

⁷ Since rejection is off the equilibrium path, we assume $\mu(\neg a, w_1 \geq \bar{\theta}) = p$ for simplicity.

rejecting the offer, i.e., if the inequality (3) is binding. Since $\mu(a, w_1) = 1$ implies $V(a, w_1|\underline{\theta}) = 0$, he is indifferent if

$$(5) \quad w_1 = \delta V(\neg a, w_1|\underline{\theta}) = \delta \bar{\theta} \left[1 - G \left(\frac{\bar{\theta}}{1 - \mu} \right) \right].$$

Together with (4), this implicitly defines an acceptance probability $q^+(w_1)$ that is consistent with the wage offer w_1 :

$$w_1 = \delta \bar{\theta} \left[1 - G \left(\frac{\bar{\theta} [1 - pq^+(w_1)]}{1 - p} \right) \right].$$

Since $G(\cdot)$ is increasing, the inverse $G^{-1}(\cdot)$ exists and we obtain

$$(6) \quad q^+(w_1) = \frac{1}{p} - G^{-1} \left(1 - \frac{w_1}{\delta \bar{\theta}} \right) \frac{1 - p}{\bar{\theta} p}.$$

Yet, this formula may well yield values $q^+ \notin [0, 1]$. Recall that \bar{V} and \underline{V} were defined as the expected second-period payoffs corresponding to $\mu = 0$ and $\mu = p$, respectively. Since q^+ is defined by (4) and (5), we obtain that $q^+(\delta \underline{V}) = 0$ and $q^+(\delta \bar{V}) = 1$. Additionally, the following lemma shows that $q^+(w_1)$ is monotonically increasing. We can thus infer that $q^+(w_1) \in [0, 1]$ only for $w_1 \in [\delta \underline{V}, \delta \bar{V}]$.

LEMMA 1 *If $G(\cdot)$ is continuous and increasing, then $q^+(w_1)$ is continuous and increasing on $W = [0, \delta \bar{\theta}]$.*

PROOF Since $G(\cdot)$ is continuous and increasing, the inverse $G^{-1}(\cdot)$ exists and is continuous and increasing as well. This implies that the acceptance probability q^+ given in (6) is continuous and increasing in w_1 as well. By continuity of $G(\cdot)$, the inverse $G^{-1}(\cdot)$ is defined on the whole interval W . *Q.E.D.*

Outside the interval $[\delta \underline{V}, \delta \bar{V}]$, $q^+(w_1)$ is no longer a probability measure, i.e., $q^+(w_1) \notin [0, 1]$. This means that wage offers outside $[\delta \underline{V}, \delta \bar{V}]$ induce an equilibrium in pure strategies. Wage offers above $\delta \bar{V}$ are always accepted by the agent; wage offers below $\delta \underline{V}$ are always rejected.

Combining this argument with the argument for wage offers above $\bar{\theta}$, the equilibrium probability that a low-type agent accepts an offer w_1 is given by

$$(7) \quad q^*(w_1) = \begin{cases} 0 & \text{if } w_1 < \min(\delta \underline{V}, \bar{\theta}), \\ 1 & \text{if } w_1 \geq \min(\delta \bar{V}, \bar{\theta}), \\ q^+(w_1) & \text{if } \min(\delta \underline{V}, \bar{\theta}) \leq w_1 < \min(\delta \bar{V}, \bar{\theta}). \end{cases}$$

This leads us to the following proposition.

PROPOSITION 1 *There exists a Bayesian Nash equilibrium*

$$\{s_1^*(w_1|\theta), s_2^*(w_2|\theta), w_2^*(\pi_2, s_1, w_1)\} \times \{\mu^*(s_1, w_1)\}$$

in the subgame after principal 1 has set wage w_1 , which is characterized as follows:

(1) The strategy of the high reservation type is

$$s_1^*(w_1|\bar{\theta}) = \begin{cases} a & \text{if } w_1 \geq \bar{\theta}, \\ -a & \text{if } w_1 < \bar{\theta}. \end{cases}$$

(2) The strategy of the low reservation type is

$$s_1^*(w_1|\underline{\theta}) = \begin{cases} a & \text{with probability } q^*(w_1), \\ -a & \text{with probability } 1 - q^*(w_1), \end{cases}$$

where $q^*(w_1)$ is given by equation (7).

(3) Principal 2 believes the probability of facing an agent with low reservation wage to be

$$\mu^*(s_1, w_1) = \begin{cases} p & \text{if } w_1 \geq \bar{\theta}, \\ p \frac{1-q^*(w_1)}{1-pq^*(w_1)} & \text{if } s_1 = -a \text{ and } w_1 < \bar{\theta}, \\ 1 & \text{if } s_1 = a \text{ and } w_1 < \bar{\theta}. \end{cases}$$

(4) Principal 2 offers a wage $w_2^*(\pi, s_1, w_1)$ to an agent with history (s_1, w_1) that is given by

$$w_2^*(\pi_2, s_1, w_1) = \begin{cases} \bar{\theta} & \text{if } \pi_2 \geq \frac{\bar{\theta}}{1-\mu^*(s_1, w_1)}, \\ \underline{\theta} = 0 & \text{if } \pi_2 < \frac{\bar{\theta}}{1-\mu^*(s_1, w_1)} \end{cases}$$

if he has a project of revenues π .

(5) Each type of agent θ accepts the wage offer w_2 if $w_2 \geq \theta$:

$$s_2^*(w_2|\theta) = \begin{cases} a & \text{if } w_2 \geq \theta, \\ -a & \text{if } w_2 < \theta. \end{cases}$$

PROOF As argued in the text.

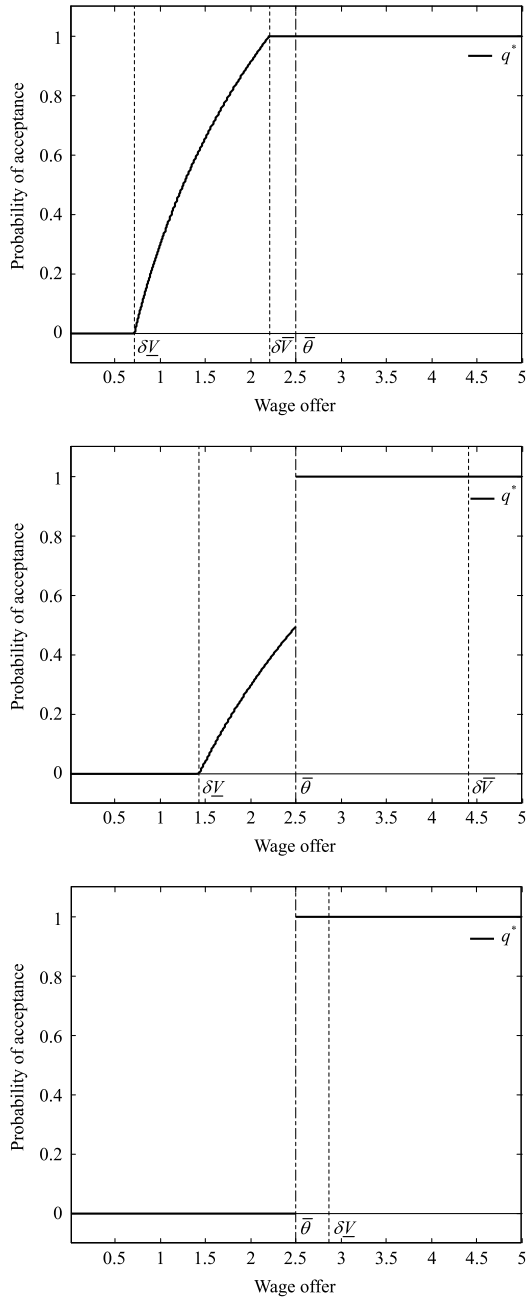
Q.E.D.

This characterization of the equilibrium in each w_1 -subgame embraces three situations with different implications for the ability of principal 2 to infer the type of the agent. For one set of model parameters, there exist some wage offers that induce a full revelation of the agent's type. For another set of model parameters, a wage offer can achieve partial revelation at best. Finally, there are model parameters for which no wage offer can achieve a revelation of the type of the agent.

The three panels in Figure 2 display these different situations. The *standard case* is $\delta\underline{V} < \delta\bar{V} < \bar{\theta}$, where all wage offers between $\delta\bar{V}$ and $\bar{\theta}$ achieve complete screening. The top panel in Figure 2 displays q^* for this situation. The next case is that of $\delta\underline{V} < \bar{\theta} < \delta\bar{V}$, illustrated by the second panel. Here, only *partial revelation* of the type can be achieved. Wage offers between $\delta\underline{V}$ and $\bar{\theta}$ reveal the type of the agent only partially, because some low-type agents remain unemployed for strategic reasons. Finally, if $\bar{\theta} \leq \delta\underline{V}$, *no revelation* of the type can be induced. All agents with a low reservation wage remain strategically unemployed if they receive an offer $w_1 < \bar{\theta}$. This is shown in the bottom panel of Figure 2.

The central parameter that discriminates the three cases is the discount factor δ : the more important the future, the more likely is strategic unemployment. This can be

Figure 2
Probability of Accepting a Wage Offer by the Low-Type Agent



illustrated by reformulating the critical values of the standard, the partial-revelation, and the no-revelation case:

$$\begin{aligned}
 \bar{\theta} > \delta \bar{V} &\Leftrightarrow \delta < [1 - G(\bar{\theta})]^{-1} && \text{(standard case),} \\
 \delta \bar{V} \leq \bar{\theta} < \delta \underline{V} &\Leftrightarrow [1 - G(\bar{\theta})]^{-1} \leq \delta < \left[1 - G\left(\frac{\bar{\theta}}{(1-p)}\right)\right]^{-1} && \text{(partial revelation),} \\
 \bar{\theta} < \delta \underline{V} &\Leftrightarrow \delta > \left[1 - G\left(\frac{\bar{\theta}}{(1-p)}\right)\right]^{-1} && \text{(no revelation).}
 \end{aligned}$$

Obviously, $\delta > 1$ is necessary for the partial- or no-revelation outcome. Accordingly, strategic unemployment becomes more significant when the future is relatively more important than the present. This is the case if period 2 represents a much longer period of time than period 1. For example, period 1 could be a period of temporary employment while in period 2 employment was permanent (with comparable job characteristics).⁸ An alternative interpretation would be that period 2 summarizes a sequence of many future employment periods. We discuss this latter interpretation in section 3.1.

Before we come back to this issue in the next section, we close the model by characterizing the wage-setting behavior of principal 1. As in the second period, the principal needs to compare the secure gain from offering $\bar{\theta}$ with the lottery from offering a wage $w_1 < \bar{\theta}$. Facing a low-type agent, it would be optimal for principal 1, who has a project of value π_1 , to offer

$$w_1^+(\pi_1) = \arg \max_{w_1 \geq 0} q^*(w_1)[\pi_1 - w_1].$$

He will meet a low type with probability p , and therefore offers $w_1^*(\pi_1) = w_1^+(\pi_1)$ if

$$\begin{aligned}
 pq^*(w_1^+(\pi_1))[\pi_1 - w_1^+(\pi_1)] &> \pi_1 - \bar{\theta} \\
 \Leftrightarrow \pi_1(1 - pq^*(w_1^+(\pi_1))) &< \bar{\theta} - w_1^+(\pi_1)pq^*.
 \end{aligned}$$

Otherwise, he will offer the high reservation wage $w_1^*(\pi_1) = \bar{\theta}$.

2.5 Unemployment in Equilibrium

The second period represents a standard monopsony situation. The equilibrium belief determines principal 2's degree of information about the agent's reservation wage. If he is fully informed, he can achieve perfect price discrimination and the monopsonistic market outcome is efficient. Some workers with a high reservation wage are unemployed,

$$U_2 = (1 - p)G(\bar{\theta}),$$

but this is efficient unemployment.

⁸ Clearly, if job characteristics change significantly between the two periods, also the disutility of work can be expected to change. This should limit the information revealed by the decision in the first period.

If the principal remains uninformed ($\mu = p$), he will offer a wage of $\underline{\theta}$ more often, leading to inefficiently high unemployment of high types:

$$(8) \quad U_2 = (1 - p)G(\bar{\pi}_2).$$

In the first period, also strategic unemployment of low types adds to the inefficient unemployment for monopsony reasons. Let $\bar{\pi}^*$ be the threshold value of revenues at which the firm opts for $w_1^*(\pi) = \bar{\theta}$, i.e., $\bar{\pi}_1^* := \inf\{\pi | w_1^*(\pi) = \bar{\theta}\}$. Then we obtain

$$(9) \quad U_1 = p \int_0^{\bar{\pi}_1^*} [1 - q^*(w_1^*(\pi))] dG(\pi) + (1 - p)G(\bar{\pi}_1^*).$$

The first part of this sum reflects *strategic unemployment*; the second, monopsony-induced unemployment. The monopsonistic firm can hire all agents at wage $\bar{\theta}$, or alternatively hire a low-type employee at $w_1^*(\pi)$. Since the latter option has lower value to the principal in the presence of strategic unemployment, we obtain

$$\bar{\pi}_1^* < \bar{\pi} = \frac{\bar{\theta}}{1 - p}$$

for the threshold value. Accordingly, strategic unemployment *reduces* the ability of the principal to exert monopsony power. Consequently, nonstrategic unemployment is reduced by the strategic behavior of the agent. However, the additional strategic unemployment may well outweigh this reduction. A particularly interesting case is the one of no revelation. In this case $\bar{\pi}_1^* = \bar{\theta}$ follows and equation (9) reduces to

$$U_1 = p \int_0^{\bar{\theta}} [1 - 0] dG(\pi) + (1 - p)G(\bar{\theta}) = G(\bar{\theta}).$$

We can easily compare this expression with the expression (8) for monopsony unemployment. It is clear that unemployment in period 2 coincides with that in the case of an uninformed monopsonist, since no information is revealed in period 1.⁹

PROPOSITION 2 *If*

$$G(\bar{\theta}) > (1 - p)G\left(\frac{\bar{\theta}}{1 - p}\right),$$

then aggregate unemployment is larger in the no-revelation case than in the case of an uninformed monopsonist.

LEMMA 2 *If G is (strictly) concave, i.e., its density is (strictly) decreasing, then*

$$G(\bar{\theta}) \geq (>) (1 - p)G\left(\frac{\bar{\theta}}{1 - p}\right)$$

for any p . If G is (strictly) convex, the reverse inequality applies.

⁹ One should note, however, that the relative increase in unemployment due to strategic reasons is limited, as period 2 has to be longer than period 1 ($\delta > 1$) as a necessary condition for no revelation.

PROOF For a (strictly) concave function G , we obtain

$$\begin{aligned} (1-p)G\left(\frac{\bar{\theta}}{1-p}\right) &= (1-p)G\left(\frac{\bar{\theta}}{1-p}\right) + pG(0) \\ &\leq (<) G\left((1-p)\frac{\bar{\theta}}{1-p} + p \cdot 0\right) = G(\bar{\theta}). \end{aligned}$$

Q.E.D.

Compared to the unemployment in a perfectly competitive labor market,

$$U_{\text{comp}} = \min[(1-p), G(\bar{\theta})],$$

the unemployment rate is lower under perfect price discrimination and may be higher under strategic unemployment (if there is no revelation). This is the case because in a perfectly competitive market, high-profit firms are matched with low-reservation-wage workers and low-profit firms with high-reservation-wage workers. Therefore, in any random matching situation, such as we assume for our model, additional contracts are formed between low-reservation-wage workers and low-profit firms as well as some between high-reservation-wage workers and high-profit firms.

Correspondingly, looking just at the differences in unemployment is misleading for welfare judgements. For example, the increase in unemployment from $(1-p)G[\bar{\theta}/(1-p)]$ to $G(\bar{\theta})$ when there is no revelation of reservation wages will understate the welfare loss due to strategic unemployment. It is the most efficient matches that are destroyed by the strategic reasoning. Compared to the (standard) uninformed monopsony case, only fewer *and less efficient* matches are formed between high-reservation-wage workers and employers with projects of a value between $\bar{\pi}$ and $\bar{\theta}$. Thus, strategic unemployment may be substantially welfare-harming and will be most prevalent in the no-revelation case.

3 Extensions and Discussion

3.1 Infinite Time Horizon

Yet, why should the no-revelation case be particularly relevant? We have seen that $\delta > 1$ is necessary to establish this case. So far, we have just argued informally that this assumption may reflect the future being more important than the present. One way to incorporate this would be a model with more than two periods. However, an exhaustive analysis quickly becomes much less tractable as the number of periods grows. Moreover and more importantly, it does not provide us with many additional general insights. Therefore, we abstain from presenting the model with an infinite horizon. Instead, we concentrate on showing that the case of strategic unemployment without revelation of types requires a much weaker assumption on the discount factor within the extension to an infinite time horizon.

Let $\beta < 1$ be the discount factor for each period. Suppose unrevealed low-type agents never accept an offer below $\bar{\theta}$. Then principals will offer all agents $w = \bar{\theta}$

as long as revenues suffice for doing so. We will show that this constitutes an equilibrium. In this situation, a low-type agent who never reveals his type has an expected payoff of $\bar{V} = \bar{\theta}(1 - G(\bar{\theta}))$ in each period. This gives him a discounted expected future payoff of

$$\Phi := \frac{\beta}{1 - \beta} [\bar{\theta}(1 - G(\bar{\theta}))].$$

On the other hand, once he has revealed his type, his future payoff is zero. This implies that not accepting, and hence not revealing the type, is the best response to any wage offer that fulfills $w_1 < \Phi$. Consequently, revelation can only be achieved if there exist wages between Φ and $\bar{\theta}$, i.e., the interval $[\Phi, \bar{\theta}]$ is nonempty. This, in turn, implies the following proposition.

PROPOSITION 3 *In the infinite-horizon case, there is an equilibrium with no revelation of types if*

$$\beta > \frac{1}{2 - G(\bar{\theta})}.$$

PROOF As argued, there is no revelation of types if the interval $[\Phi, \bar{\theta}]$ is empty. This means

$$\frac{\beta}{1 - \beta} [\bar{\theta}(1 - G(\bar{\theta}))] > \bar{\theta} \quad \Leftrightarrow \quad \beta(1 - G(\bar{\theta})) > 1 - \beta \quad \Leftrightarrow \quad \beta > \frac{1}{2 - G(\bar{\theta})}.$$

If β is close to 1, then $\beta > 1/(2 - G(\bar{\theta}))$ is only a very loose restriction, and strategic unemployment becomes a significant and constant equilibrium phenomenon in a model with an infinite horizon.

3.2 Countermeasures: Vertical Integration, Firing Costs, and Minimum Wages

As we have argued in section 2.5, strategic unemployment imposes a welfare loss. This motivates us to discuss labor market institutions that can mitigate this welfare loss. In particular, we discuss vertical integration of employers, firing costs, and minimum wages as such countermeasures against the welfare loss due to strategic unemployment.

If there is partial revelation at least, integration of both principals is a possibility to soften the strategic unemployment problem. Principal 1 then takes into account the effect his wage offer has on the knowledge of principal 2 about the type of the agent. Then, the principal can strategically choose a wage that enables him to screen the agents.

However, in the no-revelation case even integration of the principals does not solve the screening problem. If $\delta \bar{V} > \bar{\theta}$, then no wage offer by principal 1 will screen the agents' types. The low-type agent loses too much if he reveals his type. Revelation forces the principal to pay zero wage in the second period. This is the key to the no-revelation result. The principal cannot credibly commit to pay a wage

above the low reservation wage in the second period, although he might wish to do so in the first period.

This lack of commitment can be healed by firing costs or a minimum wage (see HART AND TIROLE [1988]). Suppose both principals vertically integrate and offer a two-period contract in period 1. In period 2, the principal may change the contract with the agent, but in case he does, he will have to pay a firing cost of c . Therefore, any offer $w_2 \leq c$ for period 2 is credible in period 1.

Suppose the principal offers the same wage w for both periods. If $w < \bar{\theta}$, then the agent compares the expected income from working, $(1 + \delta)w$, with the information rent $\delta\Delta$ from not revealing his type. Consequently, the two-period wage offer changes the inequality (3) to

$$(1 + \delta)w > \delta\Delta \quad \Leftrightarrow \quad w > \frac{\delta}{1 + \delta}\Delta.$$

As long as $w < c$, this offer is credible. The upper bound to the information rent is $\bar{\theta}$. This means that there are wage offers that fulfill

$$\bar{\theta} > w > \frac{\delta}{1 + \delta}\bar{\theta} > \frac{\delta}{1 + \delta}\Delta.$$

Consequently, if the firing costs c exceed $\delta\bar{\theta}/(1 + \delta)$, screening is a possible option for the principal.

The extent to which the principal uses screening depends on two factors. One is the efficiency gain from perfect discrimination in period 2. The other is the loss of monopsony power in period 1 by offering rents to the low-type agent. Say the principal finds it optimal to fully screen the agents. This means that no high-type agents are inefficiently unemployed in period 2 (there is perfect price discrimination). Compared to the situation with one-period contracts, strategic unemployment is reduced in period 1. A smaller wage offer is required to induce the agent to reveal his type and accept to work. Hence, firing costs may lower unemployment overall.

Minimum wages have a similar effect, but additionally they reduce the monopsony power in the first period. Therefore, they also reduce the inefficient unemployment of high types due to monopsony power. However, unlike firing costs, the optimal minimum wage needs to be determined by a central authority.

3.3 More than Two Types of Workers

If we suppose there are more than just the two types of workers, the analysis becomes more involved and the problem less tractable. This subsection shows by example that the basic intuition remains: low-reservation-wage types reject low offers even if they are above their reservation wage, to obtain higher wage offers in the future. If the future is sufficiently important, this can lead to no revelation of types.

We have to generalize our previous notation slightly. Let agents' types be independently and identically distributed according to a distribution function F with continuous density function f and a compact support. Moreover, we denote by $q(w_1|\theta)$ the probability that an agent of type θ will accept an offer of wage w_1 in period 1.

It is clear that $q(w_1|\theta) = 0$ for $w_1 \leq \theta$. The belief of the principal that an agent is of type $\theta' \leq \theta$ is denoted by $\mu(\theta|w_1, s_1)$. Weak perfect Bayesian equilibrium requires

$$\mu(\theta|w_1, s_1) = \frac{\Pr\{\theta' \leq \theta \cap s_1|w_1\}}{\Pr\{s_1|w_1\}}$$

if $\Pr\{s_1|w_1\} > 0$. For an accepted wage offer $s_1 = a$, this means

$$\begin{aligned} \mu(\theta|w_1, s_1 = a) &= \frac{\Pr\{\theta' \leq \theta \cap a|w_1\}}{\Pr\{a|w_1\}} = \frac{\int_0^\theta q(w_1|\theta') f(\theta') d\theta'}{\int_0^{\bar{\theta}} q(w_1|\theta') f(\theta') d\theta'} \\ (10) \qquad \qquad \qquad &= \frac{\int_0^\theta q(w_1|\theta') f(\theta') d\theta'}{\int_0^{w_1} q(w_1|\theta') f(\theta') d\theta'} \end{aligned}$$

where the last equality follows from $q(w_1|\theta) = 0$ for $w_1 \leq \theta$. Conversely, we obtain

$$(11) \qquad \mu(\theta|w_1, s_1 = \neg a) = \frac{\int_0^\theta (1 - q(w_1|\theta')) f(\theta') d\theta'}{\int_0^{\bar{\theta}} (1 - q(w_1|\theta')) f(\theta') d\theta'}$$

Now consider the principal's wage offer of period 2. She will offer a wage that maximizes

$$w_2^*(\pi_2, w_1, s_1) = \arg \max_{w_2 \geq 0} \mu(w_2|w_1, s_1)[\pi_2 - w_2]$$

From the first-order condition to this problem we obtain

$$(12) \qquad w_2^+ = \pi_2 - \frac{\mu(w_2^+|w_1, s_1)}{\mu'(w_2^+|w_1, s_1)}$$

where μ' is the first derivative of μ with respect to w_2 . If the expression exists and if $w_2^+ \geq 0$, then $w_2^* = w_2^+$. This expression shows that agents facing a wage offer $w_1 > \theta$ in period 1 may have an incentive to manipulate the principal's belief by rejecting such an offer, because this will alter wage offers in period 2.

We show that this incentive exists and leads to strategic unemployment for an example with uniformly distributed types and profits, for which we obtain the following proposition.

PROPOSITION 4 *Suppose the types of agents are uniformly distributed according to $F(\theta) = \theta/\bar{\theta}$ with $\theta \in [0, \bar{\theta}]$ and the distribution of profits is uniform on $[0, 2\bar{\theta}]$, $G(\pi) = \pi/(2\bar{\theta})$. Then:*

(a) *For all $\delta > 0$ there is some strategic unemployment, i.e.,*

$$q^*(w_1|\theta) \neq \begin{cases} 0 & \text{if } w_1 \leq \theta, \\ 1 & \text{if } w_1 > \theta. \end{cases}$$

In other words, there is no equilibrium in which agents accept all offers in period 1 that are above their reservation wage.

(b) *If $\delta \geq 2$, there is a no-revelation equilibrium. All agents reject offers $w_1 < \bar{\theta}$ in period 1, and principals in period 2 have a belief system given by $\mu(\theta|w_1, \neg a) = \min(1, \theta/\bar{\theta})$, and $\mu(\theta|w_1, a) = 1$.*

PROOF See Appendix.

4 Conclusion

We have proposed a model in which workers choose to be unemployed in order to signal a high reservation wage if they value the future sufficiently more than the present (e.g., because it is a longer period of time). This may lead to persistently high unemployment for strategic reasons. In each period, agents with a low reservation wage reject low wage offers so as to not reveal their type and to avoid being exploited in the future. A key feature of such an equilibrium with strategic unemployment is that low-reservation-wage workers *behave as if* their reservation wage is high. From a positive perspective, this result can serve as a justification for assuming a significant disutility from work when modeling employee behavior. However, one should be careful when drawing normative conclusions from such models. We have seen that strategic unemployment is inefficient, although it is voluntary.

Crucial for our result is a lack of commitment power on behalf of the principals. They can neither commit to make once-and-for-all low wage offers, nor commit not to exploit the knowledge about agents' reservation wages in the future. Legal institutions, such as multiperiod contracts combined with firing costs or a legal minimum wage, may help to mitigate this problem. They can lower the unemployment induced by the strategic interaction, but will not make it disappear completely.

Against the backdrop of strategic unemployment, there is no longer a contradiction between the standard presumption of a positive reservation wage and the finding in the happiness literature that work significantly contributes to well-being.

Appendix

A.1 Unobserved Rejected Wage Offers

So far, we have assumed that principal 2 can observe the wage that principal 1 offered to the agent in period 1. Because we have the lower segment of the labor market in mind, this is most plausible in the case where the agent has accepted the offer. In this segment wages may typically be inferred from the naming of the job, e.g., because a wage table has been negotiated between employers and unions. As to the case of rejected offers, however, our assumption would require an intermediating institution keeping record of rejected wage offers by the agent (e.g., a state-run employment agency). In the following, we discuss the changes that would result from an alternative setup where the specific wage that has been rejected cannot be observed by principal 2 (while rejection itself remains observable).

Following backward induction, we see that the acceptance decision of each type of agent to the wage offer by principal 2 remains unaltered. By contrast, the formation of beliefs can no longer condition on rejected wages. From the perspective of principal 2, the wage in period 1, w_1 , now represents a censored variable. Let \hat{w}_1 denote the observed wage (where $\hat{w}_1 = 0$ indicates no observation). Principal 2 now forms his belief on the basis of \hat{w}_1 . It is clear that all agents accept any wage offer above $\bar{\theta}$. Therefore, an observation of $\hat{w}_1 \geq \bar{\theta}$ reveals no information, so that

$\mu^*(\theta|\hat{w}_1) = p$. On the other hand, an accepted wage offer $\hat{w}_1 < \bar{\theta}$ identifies the agent as being of low type, $\mu^*(\theta|\hat{w}_1) = 1$. If the wage offer is rejected, inference becomes more complicated. Let $\Gamma(w_1)$ be the equilibrium distribution function of wage offers by principal 1 conditional on $w_1 < \bar{\theta}$. The posterior belief of principal 2 upon observing a rejection is the probability of observing a low type and a rejection divided by the overall probability of rejection. The former evaluates as $p \int (1 - q^*(w))\Gamma'(w)dw$, the latter as $\int (1 - pq^*(w))\Gamma'(w)dw$. Principal 2's belief thus reads

$$\mu^*(\theta|\hat{w}_1) = \begin{cases} p & \text{if } \hat{w}_1 \geq \bar{\theta}, \\ \frac{p \int (1 - q^*(w))\Gamma'(w)dw}{\int (1 - pq^*(w))\Gamma'(w)dw} & \text{if } \hat{w}_1 = 0, \\ 1 & \text{if } 0 < \hat{w}_1 < \bar{\theta}. \end{cases}$$

Observe that for all rejections the belief is a fixed number not depending on w_1 . Denote this number by

$$\bar{\mu} = \frac{p \int (1 - q^*(w))\Gamma'(w)dw}{\int (1 - pq^*(w))\Gamma'(w)dw}.$$

The indifference condition (5) simplifies to

$$(A1) \quad w_1 = \delta \bar{\theta} \left[1 - G \left(\frac{\bar{\theta}}{1 - \bar{\mu}} \right) \right].$$

Implicitly, this defines a threshold value \bar{w}_1 above which low-type agents always accept and below which they always reject. This means that

$$q^*(w_1) \begin{cases} = 0 & \text{if } w_1 < \bar{w}_1, \\ \in [0, 1] & \text{if } w_1 = \bar{w}_1, \\ = 1 & \text{if } w_1 > \bar{w}_1. \end{cases}$$

Consequently, the integral defining $\bar{\mu}$ reduces to

$$(A2) \quad \begin{aligned} \bar{\mu} &= \frac{p\Gamma(\bar{w})}{\Gamma(\bar{w}) + (1 - p)(1 - \Gamma(\bar{w}))} \\ &= \frac{p\Gamma(\bar{w})}{1 - p(1 - \Gamma(\bar{w}))}. \end{aligned}$$

Combining (A1) and (A2), we obtain

$$\begin{aligned} \bar{w} &\leq \delta \bar{\theta} \left[1 - G \left(\frac{\bar{\theta}}{1 - \frac{p\Gamma(\bar{w})}{1 - p(1 - \Gamma(\bar{w}))}} \right) \right] \\ &= \delta \bar{\theta} \left[1 - G \left([1 - p - p\Gamma(\bar{w})] \frac{\bar{\theta}}{1 - p} \right) \right] \end{aligned}$$

as a combined equilibrium condition. In case there exists a \bar{w} satisfying the above condition with equality, this pins down \bar{w} .

For the no-revelation case of the original model,

$$\delta > \left[1 - G \left(\frac{\bar{\theta}}{1 - p} \right) \right]^{-1},$$

there is no $\bar{w} < \bar{\theta}$ that fulfills the combined equilibrium condition. Hence, $\bar{w} = \bar{\theta}$. The partial-revelation and the standard case,

$$\delta < \left[1 - G \left(\frac{\bar{\theta}}{(1-p)} \right) \right]^{-1},$$

imply that a $\bar{w} < \bar{\theta}$ exists meeting the equilibrium condition with equality.

Thus, for the model with unobserved rejected wage offers, strategic unemployment still is an equilibrium phenomenon. However, the solution of the model becomes much more complicated, as we need to solve also for Γ in equilibrium. Put differently, we cannot determine an explicit equilibrium of the subgame conditional on the wage of principal 1 without solving his problem of optimal wage offers.

A.2 Proof of Proposition 4

The following lemma prepares the proof of part (a).

LEMMA A1 *Suppose the types of agents are uniformly distributed according to $F(\theta) = \theta/\bar{\theta}$ with $\theta \in [0, \bar{\theta}]$. Further assume that*

$$(A3) \quad q(w_1|\theta) = \begin{cases} 0 & \text{if } w_1 \leq \theta, \\ 1 & \text{if } w_1 > \theta. \end{cases}$$

Then the difference in the expected second-period wage offer $Ew_2^+(\pi_2, w_1, \neg a) - Ew_2^+(\pi_2, w_1, a)$ is strictly positive.

PROOF By assumption, the density function reads $F'(\theta) = f(\theta) = 1/\bar{\theta}$. Inserting this expression in equations (10) and (11), we obtain

$$\mu(\theta|w_1, s_1 = a) = \frac{\int_0^\theta q(w_1|\theta') f(\theta') d\theta'}{\int_0^{w_1} q(w_1|\theta') f(\theta') d\theta'} = \frac{\int_0^\theta q(w_1|\theta')/\bar{\theta} d\theta'}{\int_0^{w_1} q(w_1|\theta')/\bar{\theta} d\theta'} = \frac{\int_0^\theta q(w_1|\theta') d\theta'}{\int_0^{w_1} q(w_1|\theta') d\theta'}$$

and

$$\mu(\theta|w_1, s_1 = \neg a) = \frac{\int_0^\theta (1 - q(w_1|\theta')) f(\theta') d\theta'}{\int_0^{\bar{\theta}} (1 - q(w_1|\theta')) f(\theta') d\theta'} = \frac{\int_0^\theta (1 - q(w_1|\theta')) d\theta'}{\int_0^{\bar{\theta}} (1 - q(w_1|\theta')) d\theta'}$$

respectively. Under the assumption (A3), the two beliefs reduce to

$$\begin{aligned} \mu(\theta|w_1, s_1 = a) &= \frac{\int_0^{\min(w_1, \theta)} q(w_1|\theta') d\theta' + \int_{\min(w_1, \theta)}^\theta q(w_1|\theta') d\theta'}{\int_0^{w_1} q(w_1|\theta') d\theta' + \int_{w_1}^{\bar{\theta}} q(w_1|\theta') d\theta'} \\ &= \begin{cases} 1 & \text{if } w_1 \leq \theta, \\ \frac{\theta}{w_1} & \text{if } w_1 > \theta \end{cases} \end{aligned}$$

and

$$\begin{aligned} \mu(\theta|w_1, s_1 = \neg a) &= \frac{\int_0^{\min(w_1, \theta)} (1 - q(w_1|\theta')) d\theta' + \int_{\min(w_1, \theta)}^{\theta} (1 - q(w_1|\theta')) d\theta'}{\int_0^{w_1} (1 - q(w_1|\theta')) d\theta' + \int_{w_1}^{\bar{\theta}} (1 - q(w_1|\theta')) d\theta'} \\ &= \begin{cases} \frac{\theta - w_1}{\bar{\theta} - w_1} & \text{if } w_1 \leq \theta, \\ 0 & \text{if } w_1 > \theta, \end{cases} \end{aligned}$$

respectively. We hence obtain

$$w_2^+(\pi_2, w_1, \neg a) = \begin{cases} \frac{\pi_2 + w_1}{2} & \text{if } \pi_2 \geq w_1, \\ \pi_2 & \text{if } \pi_2 < w_1 \end{cases}$$

for $s_1 = \neg a$, and

$$w_2^+(\pi_2, w_1, a) = \begin{cases} w_1 & \text{if } \pi_2 \geq 2w_1, \\ \frac{1}{2}\pi_2 & \text{if } \pi_2 < 2w_1 \end{cases}$$

for $s_1 = a$. The expected wage in period 2 thus reads

$$\begin{aligned} Ew_2^+(\pi_2, w_1, \neg a) &= \frac{1}{2\bar{\theta}} \left(\int_0^{w_1} \pi_2 d\pi_2 + \int_{w_1}^{2\bar{\theta}} \frac{\pi_2 + w_1}{2} d\pi_2 \right) \\ &= \frac{1}{2\bar{\theta}} \left(\frac{1}{2}w_1^2 + \bar{\theta}^2 + \bar{\theta}w_1 - \frac{1}{4}w_1^2 - \frac{1}{2}w_1^2 \right) \\ &= \frac{1}{2\bar{\theta}} \left(\bar{\theta}^2 + \bar{\theta}w_1 - \frac{1}{4}w_1^2 \right) \end{aligned}$$

and

$$\begin{aligned} Ew_2^+(\pi_2, w_1, a) &= \frac{1}{2\bar{\theta}} \left(\int_0^{2w_1} \frac{1}{2}\pi_2 d\pi_2 + \int_{2w_1}^{2\bar{\theta}} w_1 d\pi_2 \right) \\ &= \frac{1}{2\bar{\theta}} (w_1^2 - 2w_1^2 + 2\bar{\theta}w_1) = w_1 \frac{1}{2\bar{\theta}} (2\bar{\theta} - w_1). \end{aligned}$$

The difference between the two is

$$\begin{aligned} Ew_2^+(\pi_2, w_1, \neg a) - Ew_2^+(\pi_2, w_1, a) &= \frac{1}{2\bar{\theta}} \left(\bar{\theta}^2 + \bar{\theta}w_1 - \frac{1}{4}w_1^2 \right) - w_1 \frac{1}{2\bar{\theta}} (2\bar{\theta} - w_1) \\ &= \frac{1}{2\bar{\theta}} \left(\bar{\theta}^2 - \bar{\theta}w_1 + \frac{3}{4}w_1^2 \right) \\ &= \frac{1}{2\bar{\theta}} \left(\left(\bar{\theta} - \frac{1}{2}w_1 \right)^2 + \frac{1}{2}w_1^2 \right) > 0. \end{aligned}$$

Q.E.D.

PROOF OF PROPOSITION 4 (A) Suppose to the contrary that (A3) holds true. By Lemma A1, the expected income difference is strictly positive for all w_1 . Therefore,

agents with $\theta = 0$ have an incentive to reject offers sufficiently close to zero, and hence there must be some strategic unemployment in equilibrium, i.e., for all $\delta > 0$,

$$q(w_1|\theta) \neq \begin{cases} 0 & \text{if } w_1 \leq \theta, \\ 1 & \text{if } w_1 > \theta. \end{cases}$$

PROOF OF PROPOSITION 4 (B) First, observe that the proposed belief system is consistent with the decision rule. When all agents reject any offer below $\bar{\theta}$, prior and posterior upon observing $\neg a$ must coincide. Off the equilibrium path, i.e., for acceptance a , we can choose μ freely.

Second, we need to show that nonacceptance of any wage below $\bar{\theta}$ is a best response in period 1. Under the belief system, the optimal wage set in period 2 is $w_2^*(\pi_2, w_1, \neg a) = \pi_2/2$ according to (12) if the agent has rejected a wage offer in period 1. If he accepted a wage offer in period 1 (off the equilibrium path), the wage offer in period 2 will be zero (he is assumed to be of type $\theta = 0$).

This implies that the expected utility differential in period 2 between nonacceptance and acceptance of a wage offer w_1 in period 1 is

$$\begin{aligned} E(w_2 - \theta | w_2 > \theta) &= \frac{\int_{2\theta}^{2\bar{\theta}} \frac{(\frac{1}{2}\pi_2 - \theta)}{2\bar{\theta}} d\pi_2}{\int_{2\theta}^{2\bar{\theta}} \frac{1}{2\bar{\theta}} d\pi_2} \\ &= \frac{\bar{\theta}^2 - \theta^2 - 2\bar{\theta}\theta + 2\theta^2}{2(\bar{\theta} - \theta)} = \frac{\bar{\theta} - \theta}{2}. \end{aligned}$$

The largest wage that will be offered in period 1 is $\bar{\theta}$. Hence, we only need to check whether for all $\theta \in [0, \bar{\theta}]$

$$\delta \left(\frac{\bar{\theta} - \theta}{2} \right) \geq \bar{\theta} - \theta,$$

which holds for all $\bar{\theta} \geq \theta \geq 0$ if $\delta \geq 2$.

References

- CHRISTOFFEL, K., AND K. KUESTER [2008], "Resuscitating the Wage Channel in Models with Unemployment Fluctuations," *Journal of Monetary Economics*, 55, 865–887.
- CLARK, A. E., P. FRIJTERS, AND M. A. SHIELDS [2006], "Income and Happiness: Evidence, Explanations and Economic Implications," PSE Working Papers 2006-24, PSE (Ecole normale supérieure).
- AND A. J. OSWALD [1994], "Unhappiness and Unemployment," *The Economic Journal*, 104, 648–659.
- COSTAIN, J. S., AND M. REITER [2008], "Business Cycles, Unemployment Insurance, and the Calibration of Matching Models," *Journal of Economic Dynamics & Control*, 32, 1120–1155.
- FREY, B. S., AND A. STUTZER [2002], "What Can Economists Learn from Happiness Research?" *Journal of Economic Literature*, 60, 402–435.

- FRIJTERS, P., I. GEISHECKER, J. P. HAIKEN-DENEW, AND M. A. SHIELDS [2006], "Can the Large Swings in Russian Life Satisfaction be Explained by Ups and Downs in Real Incomes?" *Scandinavian Journal of Economics*, 108, 433-458.
- , J. P. HAIKEN-DENEW, AND M. A. SHIELDS [2004], "Money does Matter! Evidence from Increasing Real Income and Life Satisfaction in East Germany Following Reunification," *The American Economic Review*, 94, 730-740.
- HAGEDORN, M., AND I. MANOVSKII [2008], "The Cyclical Behavior of Equilibrium, Unemployment and Vacancies Revisited," *The American Economic Review*, 98, 1692-1706.
- HART, O. D. [1983], "Optimal Labour Contracts under Asymmetric Information: An Introduction," *The Review of Economic Studies*, 50, 3-35.
- AND J. TIROLE [1988], "Contract Renegotiation and Coasian Dynamics," *The Review of Economic Studies*, 55, 509-540.
- JOVANOVIĆ, B. [1979], "Job Matching and the Theory of Turnover," *Journal of Political Economy*, 87, 972-990.
- MA, C.-T. A., AND A. M. WEISS [1993], "A Signaling Theory of Unemployment," *European Economic Review*, 37, 135-157.
- MOORE, J. [1985], "Optimal Labour Contracts when Workers have a Variety of Privately Observed Reservation Wages," *The Review of Economic Studies*, 52, 37-67.
- MORTENSEN, D., AND C. PISSARIDES [1994], "Job Creation and Job Destruction in the Theory of Unemployment," *The Review of Economic Studies*, 61, 397-415.
- PISSARIDES, C. [1985], "Short-Run Equilibrium Dynamics of Unemployment, Vacancies, and Real Wages," *The American Economic Review*, 75, 676-690.
- [2000], *Equilibrium Unemployment Theory*, 2nd ed., The MIT Press: Cambridge, MA.
- VINCENT, D. R. [1998], "Repeated Signalling Games and Dynamic Trading Relationships," *International Economic Review*, 39, 402-435.
- WARR, P. B., P. R. JACKSON, AND M. BANKS [1988], "Unemployment and Mental Health: Some British Studies," *Journal of Social Issues*, 44, 47-68.
- WINKELMANN, L., AND R. WINKELMANN [1998], "Why Are the Unemployed so Unhappy? Evidence from Panel Data," *Economica*, 65, 1-15.

Julia Angerhausen
 Burkhard Hehenkamp
 Department of Economics
 Technische Universität Dortmund
 Vogelpothsweg 87
 44221 Dortmund
 Germany
 E-mail:
 Julia.Angerhausen@uni-dortmund.de
 Burkhard.Hehenkamp@udo.edu

Christian Bayer
 IGIER
 Università Commerciale L. Bocconi
 Via Röntgen 1
 20136 Milano
 Italy
 E-mail:
 Christian.Bayer@unibocconi.it

Lifetime Employment Contract and Quantity Competition with Profit-Maximizing and Joint-Stock Firms

by

KAZUHIRO OHNISHI*

This paper studies two-stage Cournot duopoly competition with a profit-maximizing firm and a joint-stock income-per-unit-of-capital-maximizing firm. In the first stage, each firm noncooperatively decides whether to offer lifetime employment as a strategic commitment. In the second stage, both firms noncooperatively choose actual outputs. The paper shows the equilibrium outcome of the mixed model and finds that the introduction of lifetime employment into the analysis of Cournot mixed competition is profitable only for the joint-stock firm. Furthermore, the paper examines interactions between an incumbent and a potential entrant as well as a pair of established firms. (JEL: C 72, D 21, L 20)

1 Introduction

Many economic theory models assume that firms maximize profits. Therefore, the behavior of profit-maximizing (PM) firms has been most frequently encountered in the literature on economic theory. However, in the real world, not all firms adopt PM behavior.¹ Certain firms have other agendas to consider besides profit maximization. For example, economic market models that incorporate labor-managed (LM) income-per-worker-maximizing firms are sometimes analyzed by economists. The pioneering work on a theoretical model of an LM firm was done by WARD [1958]. Since then, economists have analyzed strategic choice models that incorporate LM firms, such as capacity investment (STEWART [1991]), managerial delegation (STEWART [1992]), and capital investment (FUTAGAMI AND OKAMURA [1996], NEARY AND ULPH [1997], LAMBERTINI AND ROSSINI [1998]).

The behavior of joint-stock (JS) income-per-unit-of-capital-maximizing firms is hardly ever encountered in the literature on economic theory. MEADE [1972] shows

* Institute for Basic Economic Science, Osaka. The author would like to thank Yvan Lengwiler and an anonymous referee for helpful comments and suggestions on earlier drafts of this paper.

¹ See, for example, JAMES AND ROSE-ACKERMAN [1986], BONIN AND PUTTERMAN [1987], HAY AND MORRIS [1991], MILGROM AND ROBERTS [1992], MONTIAS, BEN-NER, AND NEUBERGER [1994], FARAZMAND (ed.) [2001], and DOW [2003].

the differences in incentives, short-run adjustment, and so forth among PM, LM, and JS firms. HEY [1981] restricts attention to the case of a perfectly competitive firm producing a single output with two inputs, labor and capital, and examines the behavior of PM, LM, and JS firms. However, strategic choice competition with PM and JS firms has not been studied.

This paper focuses on JS firms and considers mixed market competition in which a PM firm and a JS firm can offer lifetime employment as a strategic commitment. MEADE [1972] called any firm that maximizes income per unit of capital a JS firm. The empirical evidence by KAPLAN, DIRLAM, AND LANZILLOTTI [1958] suggested that an assumed objective of earning a rate of return on invested capital is in fact a common phenomenon. Kaplan, Dirlam, and Lanzillotti conducted interviews with officials of 20 large U.S. firms. Each of these firms was among the 200 largest industrial corporations, and over one-half were among the 100 largest industrials, in terms of assets. The research by KAPLAN, DIRLAM, AND LANZILLOTTI [1958] and the summary by LANZILLOTTI [1958] suggested that corporate giants such as Alcoa, Du Pont, Esso (Standard Oil of New Jersey), General Electric, General Motors, International Harvester, Johns-Manville, Union Carbide, and U.S. Steel set prices to achieve a target return rate on invested capital. Lanzillotti reports that the average of targets mentioned was 14% (after taxes); only one was below 10%; and the highest was 20%. Furthermore, Lanzillotti states that no single motivation hypothesis such as profit maximization is likely to impose an unambiguous course of action on the firm for any given action. WALDMAN AND JENSEN [2007] survey the findings of Kaplan, Dirlam, and Lanzillotti, and state that because firms faced widely differing elasticities of demand, they targeted widely differing returns on investment. Furthermore, Waldman and Jensen conclude that there is little doubt that instead of the PM method, many firms claim to use others, such as a target return policy on invested capital.

The practice of lifetime employment is mainly found in Japan and is one of the main features that characterize the Japanese labor market.² The elements of the Japanese employment system include lifetime employment, a seniority system of compensation, a seniority system of promotion and appraisal, generalist training, enterprise unionism, and consensus decision-making. Many large Japanese firms focus their hiring on new male graduates from schools or universities, and these firms offer lifetime employment to the employees they recruit. The employees are recruited at the outset of their career without any particular concern for specific acquired skills. These firms expect to keep the employees they recruit until the age of compulsory retirement, which generally occurs at between 55 and 65 years of age. HASHIMOTO AND RAISIAN [1985] show that the numbers of cumulative new jobs held by males of various ages in the United States for 1978 and Japan for 1977 are 4.40 and 2.06 at age 20–24, 7.40 and 3.11 at age 30–34, 10.25 and 4.21

² A great many works dealing with the Japanese lifetime employment system have been published. See, for example, PETERSON AND SULLIVAN [1990], ITO [1992], NOMURA [1996], BROWN et al. [1997], DALY [1998], and KNELLER [2003] for recent surveys.

at age 40–54, and 10.95 and 4.91 at age 55–64, respectively. That is, the numbers of cumulative new jobs held by males are much lower in Japan than in the United States.

Although Japan is a small island society that possesses few natural resources, it achieved rapid economic growth from the end of World War II to the great oil shock of 1973 and became the world's second largest economy. Japan also produced a bubble economy in the second half of the 1980s. The Japanese economy received the world's attention during most of the 1970s and 1980s, and the Japanese lifetime employment system was considered an indispensable ingredient of the successful Japanese economy.³ OECD [1986] reported that employment arrangements in Japan tended to be the most durable among all OECD countries.

However, the Japanese economy faced a serious recession with the collapse of the bubble economy in the 1990s. The economic slowdown has allegedly been eroding the environments favorable to the lifetime employment practice. Therefore, KATO [2001] analyzed whether the practice of lifetime employment had survived in Japan since the burst of the bubble economy, and showed that contrary to the popular rhetoric of its demise, evidence points to the enduring nature of this practice in Japan. Specifically, he found little evidence for any major decline in the job retention rates of Japanese employees from the period prior to the burst of the bubble economy in the late 1980s to the postbubble period. Furthermore, ONO [2007] examines whether lifetime employment in Japan is changing and shows that the incentives among workers, managers, and executives are aligned to preserve the lifetime employment system.

This paper studies the strategic behavior of a PM firm and a JS firm in two-stage Cournot duopoly competition. In the first stage, each firm noncooperatively decides whether to adopt lifetime employment as a strategic commitment.⁴ If a firm offers lifetime employment, then it chooses an output level and enters into a lifetime employment contract with the number of employees necessary to achieve the output level. This irreversible behavior causes changes to the quantity-setting competing environment of the second stage. We examine the equilibrium outcome of the Cournot mixed market model. We then find that the introduction of lifetime employment into the analysis of Cournot mixed competition is profitable for the JS firm but is not profitable for the PM firm. This result depends upon the organizational forms of the firms, namely, the slopes of the reaction functions.

In addition, the paper examines interactions between an incumbent and a potential entrant rather than a pair of established firms. We then find that there are equilibria in which the PM incumbent and the JS incumbent each deter entry by offering lifetime employment as a strategic commitment. Since the offer of lifetime employment by an incumbent increases its aggressiveness regardless of the slope of the reaction function, the introduction of lifetime employment into the analysis of

³ See, for example, CHRISTAINSEN AND HOGENDORN [1983], HASHIMOTO AND RAISIAN [1985], LEIBENSTEIN [1987], and PETERSON AND SULLIVAN [1990].

⁴ For details see OHNISHI [2001], [2002].

entry deterrence is profitable for both the JS incumbent and the PM incumbent in entry deterrence.

This paper is organized as follows. In section 2, we formulate the model. Section 3 gives supplementary explanations of the model. Section 4 discusses the equilibrium of the model. Section 5 considers interactions between an incumbent and a potential entrant rather than a pair of established firms. Section 6 concludes the paper. Finally, the Appendix provides formal proofs.

2 The Model

Let us consider a mixed duopoly model with one PM firm and one JS firm, producing perfectly substitutable goods. The market price is determined by the inverse demand function $p(Q)$, where $Q = q_P + q_J$. Subscripts P and J denote the PM firm and the JS firm, respectively. We assume that $p' < 0$ and $p'' \leq 0$. The two stages of the game run as follows. In the first stage, both firms simultaneously and independently decide whether to adopt lifetime employment. If a firm offers lifetime employment, then it chooses an output level $q_i^* > 0$, employs the necessary number of employees to produce q_i^* , and enters into a lifetime employment contract with all of the employees ($i = P, J$). At the end of the first stage, each firm observes the behavior of the other firm. In the second stage, both firms simultaneously and independently choose actual outputs $q_P \geq 0$ and $q_J \geq 0$, and both the PM firm's profit and the JS firm's income per unit of capital are decided.

Therefore, the PM firm's profit is given by

$$(1) \quad \pi_P = \begin{cases} p(Q)q_P - rk_P(q_P) - wq_P - f & \text{if } q_P > q_P^*, \\ p(Q)q_P - rk_P(q_P) - wq_P^* - f & \text{if } q_P \leq q_P^*, \end{cases}$$

where $r > 0$ denotes the unit cost of capital (capacity), $k_P(q_P)$ the PM firm's capital input function, $w > 0$ the labor cost per output, and $f > 0$ the fixed cost.⁵ If the PM firm produces output q_P within the limit of the output level it has chosen in the first stage (i.e., $q_P \leq q_P^*$), then its marginal cost becomes rk'_P , because its labor cost is sunk as a fixed cost. On the other hand, if the PM firm wishes to produce $q_P > q_P^*$ in the second stage, then it must employ additional labor to match its output in the second stage, and its marginal cost rises to $rk'_P + w$.

Furthermore, the JS firm's income per unit of capital is given by

$$(2) \quad \psi_J = \begin{cases} \frac{p(Q)q_J - wq_J - f}{k_J(q_J)} & \text{if } q_J > q_J^*, \\ \frac{p(Q)q_J - wq_J^* - f}{k_J(q_J)} & \text{if } q_J \leq q_J^*, \end{cases}$$

where $k_J(q_J)$ denotes the JS firm's capital input function.⁶ If the JS firm produces output q_J within the limit of the output level it has chosen in the first stage (i.e.,

⁵ See DIXIT [1980] and STEWART [1991] for PM firms' profits with capacity precommitments.

⁶ See MEADE [1972] and HEY [1981] for analyses without precommitments. STEWART [1991] gives an LM firm's income per worker with capacity installed.

$q_J \leq q_J^*$), then its marginal labor cost becomes zero because its labor cost is sunk as a fixed cost. On the other hand, if the JS firm wishes to produce $q_J > q_J^*$ in the second stage, then it must employ additional labor to match its output in the second stage, and its marginal labor cost rises to w . That is, if labor is expanded as a flow simultaneously with production, then its cost is not sunk. Thus, each firm's marginal cost exhibits a discontinuity at $q_i = q_i^*$ ($i = P, J$).⁷

We assume that $k'_i > 0$ and $k''_i > 0$. This assumption means that the marginal capital input is increasing. We use subgame perfection as an equilibrium concept. The fact that inverse demand is defined only for nonnegative outputs ensures that all outputs obtained in equilibrium are nonnegative.

3 Supplementary Explanations

In this section, we give supplementary explanations of the model formulated in the previous section. First, we derive the PM firm's best reaction function from (1). If the PM firm produces output q_P within the limit of the output level it has chosen in the first stage, then its reaction function is defined by

$$R_P^W(q_J) = \arg \max_{q_P} [p(Q)q_P - rk_P(q_P) - wq_P^* - f],$$

and if the PM firm wishes to produce $q_P > q_P^*$ in the second stage, then its reaction function is defined by

$$R_P(q_J) = \arg \max_{q_P} [p(Q)q_P - rk_P(q_P) - wq_P - f].$$

Therefore, if the PM firm selects q_P^* and offers lifetime employment, then its best reaction function is as follows:

$$(3) \quad R_P^L(q_J) = \begin{cases} R_P(q_J) & \text{if } q_P > q_P^*, \\ q_P^* & \text{if } q_P = q_P^*, \\ R_P^W(q_J) & \text{if } q_P < q_P^*. \end{cases}$$

The equilibrium occurs where each firm maximizes its objective with respect to its own output level, given the output level of its rival. That is, the PM firm aims to maximize its profit with respect to its own output level, given the output level of the JS firm. Therefore, we obtain

$$R'_P(q_J) = R_P^{W'}(q_J) = -\frac{p''q_P + p'}{p''q_P + 2p' - rk''_P},$$

where the denominator is the second-order condition for maximization and is negative. Hence, both $R_P(q_J)$ and $R_P^W(q_J)$ slope downward. This means that the PM firm treats quantities as strategic substitutes.⁸

⁷ From (2), we see that the JS firm's marginal cost does not exhibit a discontinuity by capacity investment à la DIXIT [1980] and STEWART [1991].

⁸ The concepts of strategic substitutes and complements were introduced by BULOW, GEANAKOPOLOS, AND KLEMPERER [1985].

Second, we derive the JS firm's best reaction function from (2). If the JS firm produces output q_J within the limit of the output level it has chosen in the first stage, then its reaction function is defined by

$$R_J^W(q_P) = \arg \max_{q_J} \left[\frac{p(Q)q_J - wq_J^* - f}{k_J(q_J)} \right],$$

and if the JS firm wishes to produce $q_J > q_J^*$ in the second stage, then its reaction function is defined by

$$R_J(q_P) = \arg \max_{q_J} \left[\frac{p(Q)q_J - wq_J - f}{k_J(q_J)} \right].$$

Therefore, if the JS firm selects q_J^* and offers lifetime employment, then its best response is shown as follows:

$$(4) \quad R_J^L(q_P) = \begin{cases} R_J(q_P) & \text{if } q_J > q_J^*, \\ q_J^* & \text{if } q_J = q_J^*, \\ R_J^W(q_P) & \text{if } q_J < q_J^*. \end{cases}$$

The JS firm aims to maximize its income per unit of capital with respect to its own output level, given the output level of the PM firm. Therefore, we obtain

$$(5) \quad R_J'(q_P) = - \frac{p''q_Jk_J + p'(k - q_Jk_J')}{(p''q_J + 2p')k_J - (pq_J - wq_J - f)k_J'}$$

and

$$(6) \quad R_J^{W'}(q_P) = - \frac{p''q_Jk_J + p'(k - q_Jk_J')}{(p''q_J + 2p')k_J - (pq_J - wq_J^* - f)k_J'}.$$

The denominators of (5) and (6) are second-order conditions for maximization and are negative. Furthermore, since $k_J'' > 0$, we have $k_J - q_Jk_J' < 0$, so that $p''q_Jk_J + p'(k_J - q_Jk_J')$ is positive; that is, both $R_J(q_P)$ and $R_J^W(q_P)$ slope upward. This means that the JS firm treats quantities as strategic complements.

Third, we consider the following two lemmas.

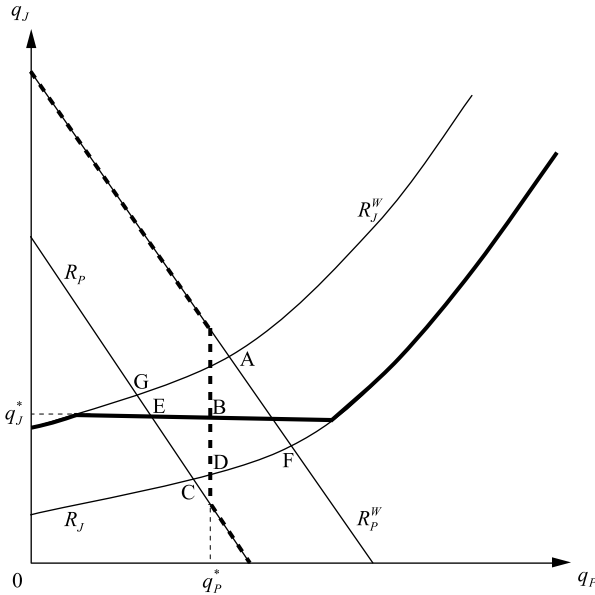
LEMMA 1 *If the PM firm (the JS firm) offers lifetime employment and a duopoly equilibrium is achieved, then in equilibrium $q_P = q_P^*$ ($q_J = q_J^*$).*

LEMMA 2 *Each firm's optimal output is higher when it offers lifetime employment than when it does not.*

Lemmas 1 and 2 provide characterizations of lifetime employment as a strategic commitment. Lemma 1 means that in a duopoly equilibrium neither firm employs extra employees. Lemma 2 means that the offer of lifetime employment by a firm increases its aggressiveness.

Fourth, we consider each firm's Stackelberg leader output. That is, the PM firm maximizes its profit $\pi_P(q_P, R_J(q_P))$ with respect to q_P , and the JS firm maximizes its income per unit of capital $\psi_J(q_J, R_P(q_J))$ with respect to q_J . We present the following two lemmas.

Figure 1
Reaction Curves in the Quantity Space



LEMMA 3 *Suppose the quantity-setting mixed game with no lifetime employment. Then the PM firm's Stackelberg leader output q_P^S is lower than its Cournot output q_P^C .*

LEMMA 4 *Suppose the quantity-setting mixed game with no lifetime employment. Then the JS firm's Stackelberg leader output q_J^S is higher than its Cournot output q_J^C .*

Fifth, we illustrate both firms' reaction curves, which are drawn in Figure 1. R_P and R_J are reaction curves without lifetime employment, and R_P^W and R_J^W are reaction curves with zero marginal labor costs. For intuitive explanations, this figure is drawn very simply. R_P and R_P^W are downward sloping, whereas R_J and R_J^W are upward sloping. Suppose that the PM firm offers lifetime employment in the first stage. By strategic choice of lifetime employment, the PM firm's best response becomes (3). The offer of lifetime employment by the PM firm thus creates kinks in the reaction curve at the level of q_P^* . That is, if the PM firm chooses q_P^* and offers lifetime employment, then its reaction curve becomes the kinked bold broken lines drawn in this figure. Furthermore, if the JS firm chooses q_J^* and offers lifetime employment, then from (4), its reaction curve becomes the kinked bold lines.⁹

⁹ As illustrated in STEWART [1992], FUTAGAMI AND OKAMURA [1996], NEARY AND ULPH [1997], and LAMBERTINI AND ROSSINI [1998], strategic choices of managerial delegation and capital investment do not create kinks in the reaction curves of firms.

If both firms offer lifetime employment, then their reaction curves intersect at a point like B. The reaction curve of each firm will have a flat segment at q_i^* . The PM firm can increase its profit by reducing q_p^* , and the JS firm can increase its income per unit of capital by reducing q_j^* . That is, each firm wants to deviate from B. Hence, we see that there is no equilibrium in which both firms offer lifetime employment.

4 Equilibrium

In this section, we discuss the equilibrium of the mixed model formulated in section 2. First, consider the case in which only the PM firm can offer lifetime employment. Since the JS firm does not offer lifetime employment in the first stage, its reaction curve is R_J , drawn in Figure 1. If the PM firm does not offer lifetime employment, then its reaction curve is R_P , drawn in Figure 1. The offer of lifetime employment by the PM firm reduces its marginal cost and increases its optimal output (Lemma 2). If the PM firm chooses q_p^* and offers lifetime employment, then its reaction curve shifts for $q_p \leq q_p^*$ and becomes the kinked bold broken lines. The equilibrium is decided in a Cournot fashion, i.e., the intersection of the PM firm's and the JS firm's reaction curves gives us a unique equilibrium. The PM firm's unilateral offer solution can occur at the appropriate point of the segment CF. The PM firm's Stackelberg leader point is to the left of C on R_J (Lemma 3). In R_J , the PM firm's profit is the highest at its Stackelberg leader point, and the further the point on R_J deviates from its Stackelberg leader point, the more its profit decreases. Hence, the PM firm's profit is the highest at C in CF.

We can now state the following proposition:

PROPOSITION 1 *Suppose that only the PM firm can offer lifetime employment. Then the equilibrium coincides with the Cournot solution with no lifetime employment.*

Second, consider the case in which only the JS firm can offer lifetime employment. Since the PM firm does not offer lifetime employment in the first stage, its reaction curve is R_P , drawn in Figure 1. The offer of lifetime employment by the JS firm reduces its marginal cost and increases its optimal output (Lemma 2). If the JS firm chooses q_j^* and offers lifetime employment, then its reaction curve shifts for $q_j \leq q_j^*$. The intersection of the PM firm's and the JS firm's reaction curves gives us a unique equilibrium. The JS firm's unilateral offer solution can occur at the appropriate point of the segment CG. The JS firm's Stackelberg leader point is to the left of C on R_P (Lemma 4). If E on CG is the JS firm's Stackelberg leader point, then the JS firm's income per unit of capital is the highest at E on R_P . Therefore, the JS firm chooses q_j^* corresponding to E in the first stage, and its reaction curve becomes the kinked bold lines drawn in Figure 1. Hence, the JS firm's unilateral offer equilibrium occurs at E.

If the JS firm's Stackelberg leader point is to the left of G on R_P , then the equilibrium cannot occur at that point. In R_P , the JS firm's income per unit of capital is the highest at its Stackelberg leader point, and the further the point on R_P

deviates from its Stackelberg leader point, the more its income per unit of capital decreases. Hence, the JS firm's income per unit of capital is the highest at G. Therefore, the JS firm chooses q_j^* corresponding to G in the first stage. That is, the JS firm's unilateral offer equilibrium occurs at G.

On the other hand, if neither firm offers lifetime employment in the first stage, then the equilibrium occurs at C. Hence, we can see easily that the JS firm's unilateral offer solution increases its income per unit of capital.

We can now present the following proposition.

PROPOSITION 2 *Suppose that only the JS firm can offer lifetime employment. Then in equilibrium, the JS firm's income per unit of capital is higher than in the Cournot equilibrium with no lifetime employment, whereas the PM firm's profit is lower than in the Cournot equilibrium with no lifetime employment.*

Third, consider the case in which both firms can offer lifetime employment. If both firms offer lifetime employment, then the intersection of their reaction curves becomes a point like B in Figure 1. The reaction curve of each firm will have a flat segment at q_i^* . The PM firm can increase its profit by reducing q_p^* , and the JS firm can increase its income per unit of capital by reducing q_j^* . Each firm wants to deviate from B. Hence, B is not an equilibrium. Furthermore, D is not an equilibrium, because the PM firm can increase its profit by reducing q_p^* .

The main result of this study is described by the following proposition.

PROPOSITION 3 *In the quantity-setting mixed duopoly regime, there exist two asymmetric equilibria in which only one firm introduces lifetime employment: (i) the PM firm's unilateral offer equilibrium and (ii) the JS firm's unilateral offer equilibrium. In (i), the PM firm gets the Cournot profit with no lifetime employment, and the JS firm gets the Cournot income per unit of capital with no lifetime employment. In (ii), the JS firm's income per unit of capital is higher than in the Cournot equilibrium with no lifetime employment, whereas the PM firm's profit is lower than in the Cournot equilibrium with no lifetime employment.*

Proposition 3 (i) means that the introduction of lifetime employment into the analysis of Cournot mixed competition is unprofitable for both firms. On the other hand, Proposition 3 (ii) means that the introduction of lifetime employment into the analysis of Cournot mixed competition is profitable for the JS firm but not profitable for the PM firm. First, we explain the intuition behind Proposition 3 (i). We consider the possibility that the PM firm offers lifetime employment. If the PM firm offers lifetime employment and increases its output, then its profit decreases (see Figure 1). Hence, the PM firm has no incentive to increase its output by offering lifetime employment. Furthermore, if the JS firm offers lifetime employment and increases its output, then the PM firm's profit decreases (see Figure 1). Therefore, the PM firm chooses q_p^{*C} corresponding to its Cournot output and offers lifetime employment. The reaction curve of the PM firm will have a flat segment at q_p^{*C} . If the JS firm offers lifetime employment and increases its output, then its income per unit of capital decreases, and thus it has no incentive to do so. The equilibrium coincides with the Cournot solution with no lifetime employment.

Next, we explain the intuition behind Proposition 3 (ii). We consider the possibility that the JS firm offers lifetime employment. If the JS firm offers lifetime employment and increases its output, then the PM firm's profit decreases. Since the PM firm decreases its output because of strategic substitutes, the JS firm's income per unit of capital increases (see Figure 1). If the PM firm offers lifetime employment and increases its output, then its profit decreases further, and hence it has no incentive to do so. The unilateral offer of lifetime employment by the JS firm increases its own income per unit of capital, whereas it decreases the PM firm's profit.

5 Entry

In this section, the rival is taken to be a potential entrant rather than an established firm. That is, the following situation is considered. In the first stage, an established firm decides whether to offer lifetime employment. At the end of the first stage, a potential entrant observes the behavior of the established firm and decides whether to enter the market. In the second stage, if the potential entrant enters, a Cournot duopoly equilibrium is achieved, whereas if the potential entrant does not enter, the established firm prevails as a monopolist.

Figure 2 illustrates a single PM incumbent facing a single JS potential entrant. The JS potential entrant's best response if it were to enter the market is given by R_J . The PM incumbent can offer lifetime employment in order to deter entry. R_P denotes the PM incumbent's best response without lifetime employment, and R_P^W the PM incumbent's best response with zero marginal labor cost. The JS potential entrant enters the market if and only if its postentry income per unit of capital is positive. In Figure 2, R_J meets R_P at N and R_P^W at W. Here, q_P^N denotes the PM incumbent's output corresponding to N, and q_P^W its output corresponding to W. Rightward movements along R_J involve successively lower levels of income per unit of capital in the JS entrant. The JS entrant would then choose not to enter, so at point Y its reaction curve exhibits a discontinuity, dropping to $q_J = 0$. In the figure, Z_P denotes this point on the horizontal axis.

Furthermore, Figure 3 illustrates a single JS incumbent facing a single PM potential entrant. R_P denotes the PM potential entrant's best response should it enter the market, R_J the JS incumbent's best response without lifetime employment, and R_J^W the JS incumbent's best response with zero marginal labor cost. In this figure, R_J^W meets the horizontal axis at q_J^X . The PM potential entrant enters the market if and only if its postentry profit is positive.

In each of Figures 2 and 3, there are three cases to consider.

Case 1: $Z_i \leq q_i^N$. The potential entrant does not try to enter the market at all. Entry being irrelevant, the incumbent will enjoy a pure monopoly.

Case 2: $q_i^W < Z_i$. The incumbent cannot hope to prevent entry, so it can only seek the best available duopoly position. Depending on the incumbent's choice of q_i^* ,

Figure 2
The PM Incumbent

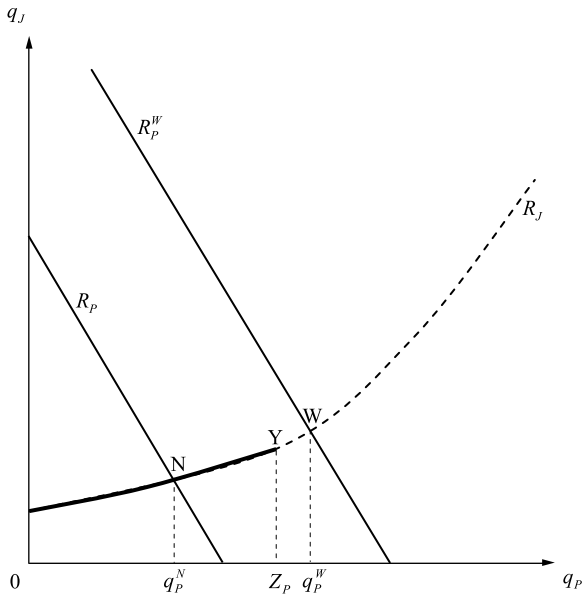
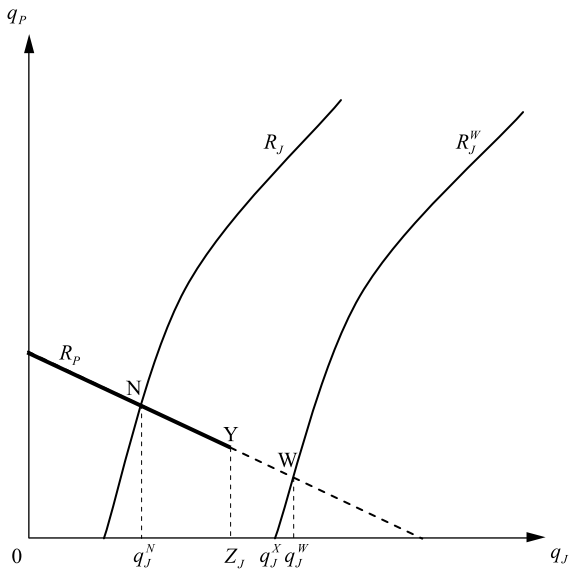


Figure 3
The JS Incumbent



the postentry equilibrium can be at any point between N and W on the entrant's reaction curve. The point N corresponds to the incumbent's smallest output that can be sustained as a Cournot equilibrium. The point W is the incumbent's largest output that can be sustained as a Cournot equilibrium. That is, the postentry equilibrium occurs at the most appropriate point for the incumbent between N and W.

Case 3: $q_i^N < Z_i \leq q_i^W$. The potential entrant is either accommodated into the market or deterred. If accommodated, the postentry equilibrium occurs at the most appropriate point for the incumbent between N and Y on the entrant's reaction curve. If deterred, the incumbent chooses q_i^* corresponding to Z_i and offers lifetime employment. Hence, if $q_p^N < Z_p \leq q_p^W$ and $q_j^N < Z_j \leq q_j^X$, then the entry-deterrence equilibrium occurs at Z_i , while if $q_j^X < Z_j \leq q_j^W$, then there is idle labor and the entry-deterrence equilibrium becomes q_j^X .

The PM incumbent firm and the JS incumbent do not always deter entry by offering lifetime employment as a strategic commitment, but we have established that it is a possibility for each.

6 Conclusion

We have first examined Cournot mixed competition, where a PM firm and a JS firm can offer lifetime employment as a strategic commitment. We have established that there exist two asymmetric equilibria in which only one firm introduces lifetime employment: one is the PM firm's unilateral offer equilibrium, and the other is the JS firm's unilateral offer equilibrium. The former equilibrium coincides with the Cournot equilibrium with no lifetime employment. In the latter equilibrium, the JS firm's income per unit of capital is higher than in the Cournot equilibrium with no lifetime employment, whereas the PM firm's profit is lower than in the Cournot equilibrium with no lifetime employment. Therefore, we have found that the introduction of lifetime employment into the analysis of Cournot mixed competition is profitable for the JS firm but not profitable for the PM firm.

Next, we have considered interactions between an incumbent and a potential entrant. We have established that the PM incumbent and the JS incumbent both may deter entry by offering lifetime employment as a strategic commitment. As a result, we see that the introduction of lifetime employment into the analysis of entry deterrence is profitable for the PM incumbent as well as for the JS incumbent.

Appendix

PROOF OF LEMMA 1 We prove that if the PM firm offers lifetime employment and a duopoly equilibrium is achieved, then in equilibrium $q_p = q_p^*$. First, consider the possibility that $q_p < q_p^*$ in equilibrium. From (1), if $q_p < q_p^*$, the PM firm must

employ the extra employees necessary to produce $q_P^* - q_P$. That is, the PM firm can increase its profit by reducing q_P^* , and the equilibrium point does not change in $q_P \leq q_P^*$. Hence, $q_P < q_P^*$ does not result in an equilibrium.

Next, consider the possibility that $q_P > q_P^*$ in equilibrium. From (1), the PM firm has to incur the full marginal costs of producing any given quantity. It is impossible for the PM firm to change its output in equilibrium, because such a strategy is not credible. That is, if $q_P > q_P^*$, lifetime employment does not function as a strategic commitment.

The proof of offer by the JS firm is omitted, since it is the same as the proof of offer by the PM firm. *Q.E.D.*

PROOF OF LEMMA 2 We prove that the JS firm's optimal output is higher when it offers lifetime employment than when it does not. From (2), we see that lifetime employment will never increase the marginal cost of the JS firm. If the JS firm does not offer lifetime employment, then the first-order condition is

$$(p'q_J + p - w)k_J - (pq_J - wq_J - f)k'_J = 0,$$

and if the JS firm produces output q_J within the limit of the output level it has chosen in the first stage, then the first-order condition is

$$(A1) \quad (p'q_J + p)k_J - (pq_J - wq_J^* - f)k'_J = 0.$$

Here, w is positive. Furthermore, Lemma 1 shows that if the JS firm offers lifetime employment and maximizes its income per unit of capital, then $q_J = q_J^*$. To satisfy (A1), $(p'q_J + p - w)k_J - (pq_J - wq_J - f)k'_J$ must be negative. Thus, the JS firm's optimal output is higher when it offers lifetime employment than when it does not.

The proof of offer by the PM firm is omitted, since it is the same as the proof of offer by the JS firm. *Q.E.D.*

PROOF OF LEMMA 3 The PM firm selects q_P , and the JS firm selects q_J after observing q_P . When the PM firm is the Stackelberg leader, it maximizes its profit $\pi_P(q_P, R_J(q_P))$ with respect to q_P . Therefore, the PM firm's Stackelberg leader output satisfies the first-order condition

$$(A2) \quad p'q_P + p - rk'_P - w + p'q_P R'_J = 0.$$

Because $p' < 0$ and $R'_J > 0$, to satisfy (A2), $p'q_P + p - rk'_P - w$ must be positive. Thus, Lemma 3 follows. *Q.E.D.*

PROOF OF LEMMA 4 The JS firm selects q_J , and the PM firm selects q_P after observing q_J . When the JS firm is the Stackelberg leader, it maximizes its income per unit of capital, $\psi_J(q_J, R_P(q_J))$, with respect to q_J . Therefore, the JS firm's Stackelberg leader output satisfies the first-order condition

$$(A3) \quad (p'q_J + p - w)k_J - (pq_J - wq_J - f)k'_J + p'q_J R'_P k_J = 0.$$

From $p' < 0$ and $R'_p < 0$, to satisfy (A3), $(p'q_J + p - w)k_J - (pq_J - wq_J - f)k'_J$ must be negative. Thus, Lemma 4 follows. Q.E.D.

PROOF OF PROPOSITION 1 Lemma 2 shows that the PM firm's profit-maximizing output is higher when the PM firm offers lifetime employment than when it does not. On the other hand, Lemma 3 shows that $q_p^S < q_p^C$. Furthermore, $\pi_p = p(Q)q_p - rk_p(q_p) - wq_p - f$ is continuous and concave. In R_J , π_p is highest at the PM firm's Stackelberg leader point, and the further the point on R_J gets from the PM firm's Stackelberg leader point, the more π_p decreases. The PM firm has no incentive to increase q_p by offering lifetime employment, and thus Proposition 1 follows. Q.E.D.

PROOF OF PROPOSITION 2 Lemma 2 shows that the JS firm's optimal output is higher when the JS firm offers lifetime employment than when it does not. Lemma 4 shows that $q_J^S > q_J^C$. Furthermore, $\psi_J = [p(Q)q_J - wq_J - f]/k_J(q_J)$ is continuous and concave. In R_p , ψ_J is highest at the JS firm's Stackelberg leader point, and the further the point on R_p gets from the JS firm's Stackelberg leader point, the more ψ_J decreases. Lemma 1 shows that in equilibrium $q_J = q_J^*$. Thus, the JS firm increases q_J by offering lifetime employment, and ψ_J is higher in the Cournot mixed game with lifetime employment than in the Cournot mixed game with no lifetime employment.

The JS firm increases q_J by offering lifetime employment. Furthermore, $\pi_p = p(Q)q_p - rk_p(q_p) - wq_p - f$ is continuous and concave. Since $\partial\pi_p/\partial q_J = p'q_p < 0$, increasing q_J decreases π_p given q_p , and thus Proposition 2 follows. Q.E.D.

PROOF OF PROPOSITION 3 (i) Suppose that the PM firm unilaterally offers lifetime employment. Lemma 2 states that the PM firm's profit-maximizing output is higher when the PM firm offers lifetime employment than when it does not. That is, the PM firm's optimal output does not decrease by offering lifetime employment. However, Lemma 3 states that $q_p^S < q_p^C$. Furthermore, $\pi_p = p(Q)q_p - rk_p(q_p) - wq_p - f$ is continuous and concave. In R_J , π_p is highest at the PM firm's Stackelberg leader point, and the further the point on R_J gets from the PM firm's Stackelberg leader point, the more π_p decreases. Hence, the PM firm chooses q_p^{*C} corresponding to the Cournot point with no lifetime employment. The PM firm's marginal cost exhibits a discontinuity at $q_p = q_p^{*C}$. The PM firm's reaction curve is kinked at the level equal to $q_p = q_p^{*C}$ and becomes the kinked bold broken lines, while the JS firm's reaction curve is R_J (see Figure A1). Here, if only the PM firm (only the JS firm) changes q_i^* and/or q_i , then π_p (then ψ_J) decreases. That is, neither firm has an incentive to deviate from the Cournot point given that the other firm does not deviate. Lemma 1 states that in equilibrium $q_p = q_p^*$. Thus, Proposition 3 (i) follows.

(ii) Suppose that the JS firm unilaterally offers lifetime employment. Lemma 2 states that the JS firm's optimal output is higher when the JS firm offers lifetime

Figure A1
The PM firm's Unilateral Offer Equilibrium

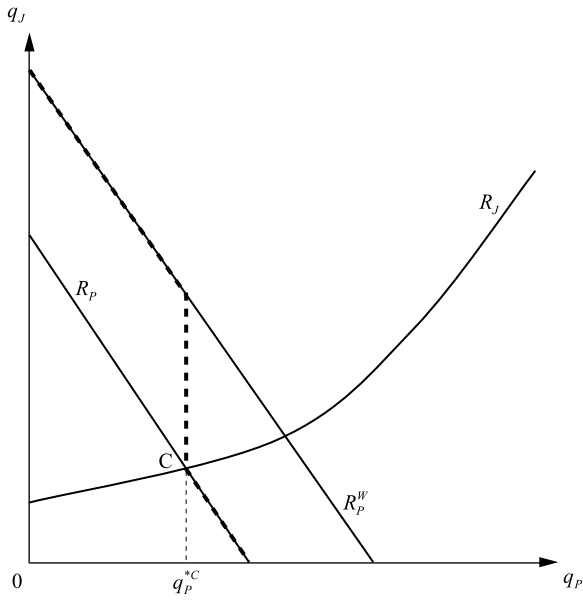
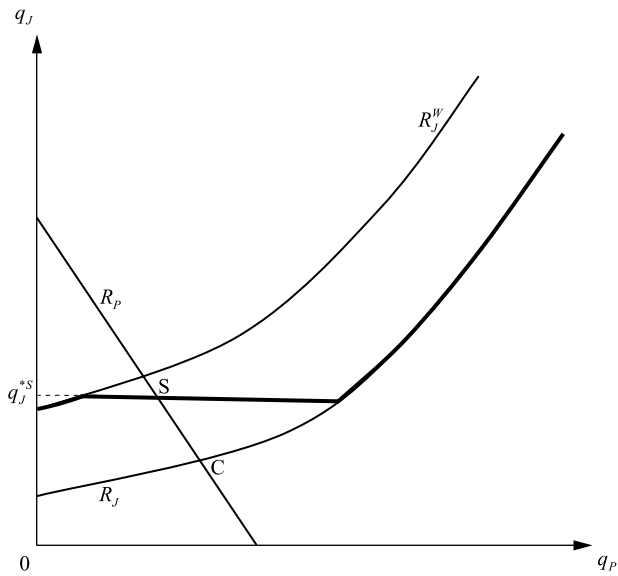


Figure A2
The JS firm's Unilateral Offer Equilibrium



employment than when it does not. Lemma 4 states that $q_J^S > q_J^C$. Furthermore, $\psi_J = [p(Q)q_J - wq_J - f]/k_J(q_J)$ is continuous and concave. In R_P , ψ_J is highest at the JS firm's Stackelberg leader point S, and the further the point on R_P gets from S, the more ψ_J decreases. If S is the highest possible income per unit of capital for the JS firm, then the JS firm chooses q_J^{*S} corresponding to S. The JS firm's marginal cost exhibits a discontinuity at $q_J = q_J^{*S}$. The JS firm's reaction curve is kinked at the level equal to $q_J = q_J^{*S}$ and becomes the kinked bold lines, while the PM firm's reaction curve is R_P (see Figure A2). Here, if only the JS firm (the PM firm) changes q_i^* and/or q_i , then ψ_J (π_P) decreases. That is, neither firm has an incentive to deviate from S given that the other firm does not deviate. Lemma 1 shows that in equilibrium $q_J = q_J^*$. Thus, the JS firm unilaterally offers lifetime employment, and ψ_J is higher in the Cournot mixed game with lifetime employment than in the Cournot mixed game with no lifetime employment.

The JS firm increases q_J by offering lifetime employment. Furthermore, $\pi_P = p(Q)q_P - rk_P(q_P) - wq_P - f$ is continuous and concave. Since $\partial\pi_P/\partial q_J = p'q_P < 0$, increasing q_J decreases π_P given q_P , and thus Proposition 3 (ii) follows. *Q.E.D.*

References

- BONIN, J. P., AND L. PUTTERMAN [1987], *Economics of Cooperation and the Labor-Managed Economy*, Harwood Academic Publishers: Chur.
- BROWN, C., Y. NAKATA, M. REICH, AND L. ULMAN [1997], *Work and Pay in the United States and Japan*, Oxford University Press: New York.
- BULOW, J., J. GEANAKOPOLOS, AND P. KLEMPERER [1985], "Multimarket Oligopoly: Strategic Substitutes and Complements," *Journal of Political Economy*, 93, 488–511.
- CHRISTAINSEN, G. B., AND J. S. HOGENDORN [1983], "Japanese Productivity: Adapting to Changing Comparative Advantage in the Face of Lifetime Employment Commitments," *The Quarterly Review of Economics and Business*, 23, 23–39.
- DALY, G. G. [1998], "Entrepreneurship and Business Culture in Japan and the U.S.," *Japan and the World Economy*, 10, 487–494.
- DIXIT, A. K. [1980], "The Role of Investment in Entry-Deterrence," *The Economic Journal*, 90, 95–106.
- DOW, G. K. [2003], *Governing in the Firm: Workers' Control in Theory and Practice*, Cambridge University Press: Cambridge.
- FARAZMAND, A. (ed.) [2001], *Privatization or Public Enterprise Reform?* Greenwood Press: Westport, CT.
- FUTAGAMI, K., AND M. OKAMURA [1996], "Strategic Investment: The Labor-Managed Firm and the Profit-Maximizing Firm," *Journal of Comparative Economics*, 23, 73–91.
- HASHIMOTO, M., AND J. RAISIAN [1985], "Employment Tenure and Earnings Profiles in Japan and the United States," *The American Economic Review*, 75, 721–735.
- HAY, D. A., AND D. J. MORRIS [1991], *Industrial Economics and Organization: Theory and Evidence*, Oxford University Press: Oxford.
- HEY, J. D. [1981], "A Unified Theory of the Behaviour of Profit-Maximising, Labour-Managed and Joint-Stock Firms Operating under Uncertainty," *The Economic Journal*, 91, 364–374.
- ITO, T. [1992], *The Japanese Economy*, The MIT Press: Cambridge, MA.
- JAMES, E., AND S. ROSE-ACKERMAN [1986], *The Nonprofit Enterprise in Market Economies*, Harwood Academic Publishers: Chur.

- KAPLAN, A. D. H., J. B. DIRLAM, AND R. F. LANZILLOTTI [1958], *Pricing in Big Business: A Case Approach*, The Brookings Institution: Washington, D.C.
- KATO, T. [2001], "The End of Lifetime Employment in Japan? Evidence from National Surveys and Field Research," *Journal of the Japanese and International Economies*, 15, 489–514.
- KNELLER, R. [2003], "Autarkic Drug Discovery in Japanese Pharmaceutical Companies: Insights into National Differences in Industrial Innovation," *Research Policy*, 32, 1805–1827.
- LAMBERTINI, L., AND G. ROSSINI [1998], "Capital Commitment and Cournot Competition with Labour-Managed and Profit-Maximising Firms," *Australian Economic Papers*, 37, 14–21.
- LANZILLOTTI, R. F. [1958], "Pricing Objectives in Large Companies," *The American Economic Review*, 48, 921–940.
- LEIBENSTEIN, H. [1987], *Inside the Firm: The Inefficiencies of Hierarchy*, Harvard University Press: Cambridge, MA.
- MEADE, J. E. [1972], "The Theory of Labour-Managed Firms and of Profit Sharing," *The Economic Journal*, 82, 402–428.
- MILGROM, P., AND J. ROBERTS [1992], *Economics, Organization and Management*, Prentice Hall: Englewood Cliffs, NJ.
- MONTIAS, J. M., A. BEN-NER, AND E. NEUBERGER [1994], *Comparative Economics*, Harwood Academic Publishers: Chur.
- NEARY, H. M., AND D. ULPH [1997], "Strategic Investment and the Co-Existence of Labour-Managed and Profit-Maximising Firms," *Canadian Journal of Economics*, 30, 308–328.
- NOMURA, M. [1996], "The Concept of Lifetime Employment: A Historical Critique," *Japanese Economic Studies*, 24, 39–57.
- OECD [1986], *Flexibility in the Labour Market: The Current Debate: A Technical Report*, OECD: Paris.
- OHNISHI, K. [2001], "Lifetime Employment Contract and Strategic Entry Deterrence: Cournot and Bertrand," *Australian Economic Papers*, 40, 30–43.
- [2002], "On the Effectiveness of the Lifetime-Employment-Contract Policy," *The Manchester School*, 70, 812–821.
- ONO, H. [2007], "Lifetime Employment in Japan: Concepts and Measurements," SSE/EFI Working Paper Series in Economics and Finance No. 624, Stockholm School of Economics.
- PETERSON, R. B., AND J. SULLIVAN [1990], "The Japanese Lifetime Employment System: Whither it Goes?" pp. 169–194 in: S. B. Prasad (ed.), *Advances in International Comparative Management*, Vol. 5, JAI Press: Greenwich, CT.
- STEWART, G. [1991], "Strategic Entry Interactions Involving Profit-Maximising and Labour-Managed Firms," *Oxford Economic Papers*, 43, 570–583.
- [1992], "Management Objectives and Strategic Interactions among Capitalist and Labour-Managed Firms," *Journal of Economic Behavior & Organization*, 17, 423–431.
- WALDMAN, D. E., AND E. J. JENSEN [2007], *Industrial Organization: Theory and Practice*, Pearson/Addison Wesley: Boston, MA.
- WARD, B. [1958], "The Firm in Illyria: Market Syndicalism," *The American Economic Review*, 48, 566–589.

Kazuhiro Ohnishi
 Institute for Basic Economic Science
 Tsugaryo 102
 Hanjo 2-15-12
 Minoo, Osaka 562-0044
 Japan
 E-mail: ohnishi@e.people.or.jp

Opportunism, Hold-Up and the (Contractual) Theory of the Firm

by

JAMES H. LOVE*

This paper considers the role of opportunism in three contractual theories of the firm: rent-seeking theory, property rights theory, and agency theory. In each case I examine whether it is possible to have a functioning contractual theory of the firm without recourse to opportunism. Without opportunism firms may still exist as a result of issues arising from (incomplete) contracting. Far from posing a problem for the theory of the firm, questioning the role of opportunism and the ubiquity of the hold-up problem helps us understand more about the purpose and functions of contracts which go beyond mere incentive alignment. (JEL: L 14, L 22)

1 Introduction

It would be hard to think of a more fundamental question in economics than ‘why do firms exist?’ Yet this has proved to be a theoretical – and empirical – puzzle for economists for decades. This is not for a want of theories; indeed, there is a plethora of theories which attempt, with varying degrees of success, to answer the questions of why firms exist, what determines their scale and scope, what advantages they have over market contracting, and the economic forces that give rise to their internal organization.

My concern is with one class of theories – contractual theories of the firm – and with one key issue within them: the role of opportunism and the associated ‘hold-up problem’. The reason for restricting the analysis to contractual theories of the firm is that it places the role of opportunism in sharp focus. There is a body of literature, deriving mainly from the resource-based view, which suggests that opportunism has no role to play when the firm is conceived of in terms of the development of resources giving rise to sustainable economic rents.¹ I do not deal with these approaches, not because they are uninteresting or do not tell us anything

* Aston University, Birmingham. This paper was written while I was an academic visitor at CIBAM, Judge Business School, University of Cambridge, and Visiting Fellow of Wolfson College, Cambridge. I am grateful to Neil Kay and two anonymous referees for valuable comments on earlier drafts.

¹ For a debate on this issue see CONNER [1991]; CONNER AND PRAHALAD [1996]; FOSS [1996a], [1996b]; MAHONEY [2001].

about the nature of the firm, but because they are principally concerned with issues other than contracting, the area in which the use of opportunism is most likely to be profitable and in which measures to guard against it are most likely to be required. Developing an opportunism-free theory of the firm where hold-up is unlikely to prosper seems a rather hollow victory.² I therefore deal specifically with the role of opportunism in three theories which are predicated on issues of contracting. Two of these deal explicitly with costs arising from hold-up, mainly in the make-or-buy decision: rent-seeking theory (particularly the Williamsonian transaction cost version), and property rights theory. The third, principal-agent analysis, deals with issues of contracting and is a theory of the firm by default. If opportunism has a necessary place in the contractual theory of the firm it is likely to be found in these three theories.

In each case I lay out an informal description of the basic theory, examine the type and role of opportunism which is embedded within it, and then determine whether it is possible to have a functioning contractual theory of the firm without recourse to opportunism and hold-up. This is done by considering the two basic questions which a theory of the firm ought to be able to answer: why firms exist, and what determines their boundaries³ (GARROUSTE AND SAUSSIER [2005]). Crucially, a workable theory ought to show clearly the trade off between integration and non-integration; thus the advantage of firm over market has to be made explicit (GIBBONS [2005]). This issue is also considered in turn for each of the three contractual theories.

If we remove the threat of opportunistic hold-up from the theory of the firm it transpires that we still have something that is contractual in nature: firms may exist at least in part as a result of issues arising from (incomplete) contracting even in the absence of opportunism. In addition, far from posing a problem for the theory of the firm, questioning the role of opportunism actually helps us understand more about the purpose and functions of contracts which go beyond mere incentive alignment or the prevention of wrongdoing, and so enhances our understanding of the nature of the firm.

2 Definitions

Taxonomy is not the principal focus of this paper, but some basic definitions are clearly in order to allow the analysis to proceed. Here I define what I take to mean by ‘the firm’, ‘contractual theories’, ‘opportunism’ and ‘hold-up’.

2.1 The Firm

Interestingly, none of the three contractual theories considered below explicitly defines a firm, and in fact this is surprisingly rarely done. A reasonable starting

² However, MAHONEY [2001] argues that opportunism is necessary even for a functioning resource-based theory of the firm.

³ I do not consider the related question of how firms are internally organized.

point is Hodgson's definition based on the legal as well as economic basis of the firm:

"A firm is defined as an integrated and durable organization involving two or more people, acting openly or tacitly as a 'legal person', capable of owning assets, set up for the purpose of producing goods or services, with the capacity to hire or sell these goods or services to customers." (HODGSON [2002, p. 56])

This has some desirable features from the current perspective. First, it clearly differentiates firms from markets, which are a different form of institutional structure. Second, it stresses the role of asset ownership, central to property rights theory and highly relevant to rent-seeking theory. However, Hodgson's definition also has omissions for current purposes. Although it usefully excludes single-person 'firms', the employment relationship is not part of the definition above:⁴ the inclusion of this relationship is necessary, especially for principal-agent analysis as it specifically compares independent contracting with the employment contract. Therefore to Hodgson's definition I add the key element that the 'integrated and durable organization' has employees.

In adopting this definition it must be acknowledged that, in reality, a simple dichotomous relationship between firms and markets is not always apparent. There is, of course, as much heterogeneity within different types of firms as there is between firms and the market: a modern multinational enterprise with its decentralized structure and shifting set of alliances bears little obvious resemblance to the corner grocery store, and there are many forms of quasi-firm/quasi-market organizations. Nevertheless, in theoretical terms we want to explain at the margin why one form of organization gives way to another, and so the conceptual dichotomy is useful for our purpose.

2.2 *Contractual Theories*

By a contractual theory I mean one which explains the existence and boundaries of a firm of the type defined above principally in terms of contracting issues which cannot be satisfactorily resolved within the institutional form known as the market. This includes, but is not restricted to, those theories which see the firm simply as the replacement of one set of contracts by another, either in terms of agency relationships (ALCHIAN AND DEMSETZ [1972]) or property rights analysis (GROSSMAN AND HART [1986], HART [1989], HART AND MOORE [1990]). This definition also includes the class of theories viewing the firm as a hierarchy which, although created by contract, is not itself a contractual relationship, but rather "a mechanism of post-contractual governance, to handle issue that can neither be adequately defined in complex contracts which are necessarily incomplete, nor resolved by allocating the rights to make crucial decisions to the owners of critical assets" (LOASBY [1999, p. 83]).

⁴ Partnerships without employees could be included in Hodgson's definition, as could worker cooperatives or even slave enterprises (HODGSON [2002, p. 53]).

I expressly exclude from consideration resource-based, capabilities or evolutionary theories of the firm. As indicated earlier, this is not because these theories have nothing to say, but because they generally deal with issues other than the existence and boundaries of the firm, such as how firms differ in growth rates and levels of return and how these differences persist through time. These are important issues, but not generally those which are central to contractual theories of the firm.

2.3 *Opportunism*

It is difficult for economists to hear the word opportunism without immediately thinking of the work of Oliver Williamson and his version of transaction cost analysis. To his credit, Williamson has expressed more clearly than most the assumptions underlying his theory and has attempted to define and explain its crucial components. Williamson's famous definition is "[...] self-interest seeking with guile. This includes but is scarcely limited to more blatant forms, such as lying, stealing, and cheating. Opportunism more often involves more subtle forms of deceit" (WILLIAMSON [1985, p. 47]). Williamson does not assume that all economic agents are always opportunistic, or even that opportunism always pays. The difficulty arises in determining *ex ante* which possible contracting partners are likely to behave in this way, and discriminating between those circumstances under which opportunism will pay and those when it will not. Transaction costs can then arise from the need to protect against the likelihood of, and potential loss that may arise from, opportunistic behaviour by another party.

It is perfectly reasonable to measure Williamson's theory against his own measuring rod of opportunism, (e.g. HODGSON [2004], LOVE [2005]). However, if we are to examine the role of opportunism in contractual theories of the firm more generally it is unfair to allow Williamson's rather doleful definition to dominate, with its emphasis on "calculated efforts to mislead, distort, disguise, obfuscate, or otherwise confuse" (WILLIAMSON [1985, p. 47]). This is because Williamson's view of opportunism is rather out of step with the normal meaning of the word, and with the views of other writers. Consider the following definitions, both from the *Concise Oxford English Dictionary* (2002):

Opportunistic: 'exploiting immediate opportunities, especially in an unplanned or selfish way'.

Opportunist: 'person who takes advantage of opportunities as and when they arise, regardless of planning or principle'.

Rather than the deliberate deceit which is required by Williamson, these definitions stress the somewhat unplanned nature of opportunism: it may involve selfish and unprincipled behaviour, but is scarcely restricted to the actions of the guileful or deceitful. Some writers have adopted visions of opportunism which encompass a much wider range of human behaviour than is implied by Williamson's definition, and are more in accord with the dictionary definition, notably that opportunism

can simply be the act of following self-interest at the expense of another party (GOLDBERG [1984]). As STEPHEN [1996] points out, this milder version of opportunism may still give rise to transaction costs because of the need for parties investing in transaction-specific capital to guard against *ex post* hold-up. The threat of virtually any kind of self-interested behaviour short of actual deceit may preclude the use of a self-enforcing 'general clause' to get round the problem of contractual incompleteness: it does not depend on consciously guileful behaviour. Since problems arising from incomplete contracts form the essence of the contractual theory of the firm, this is an important point.

This suggests that in assessing the role of opportunism in the contractual theory of the firm, we must consider two degrees of opportunism: the strong (Williamson) form which requires guileful behaviour, and the weaker form which suggests a willingness to follow self-interest at another's expense without the necessity for systematically deceitful behaviour.

2.4 *Hold-Up*

Opportunism, in whatever guise, is unlikely to be an issue unless it gives rise to the potential for hold-up. As HOLMSTRÖM AND ROBERTS [1998] note, 'the hold-up problem' has become the central issue of much of the theoretical work on the existence and boundaries of the firm since the 1970s. At its simplest, "[h]oldup occurs when one contracting party threatens another with economic harm unless concessions are granted by the threatened party" (SMITH AND KING [2009, p. 18]). Clearly, hold-up must be a credible threat to impel economic actors to prefer one form of economic organization over another, and the precise nature of the hold-up threat will be considered in turn in each theory.

3 *Hold-Up Arising from Rent-Seeking Behaviour*

Since opportunism is synonymous with the work of Oliver Williamson it is logical to start with the theory of the firm most closely associated with him. Although Williamson's approach is often seen as following naturally from COASE [1937], in many respects it has more in common with the analysis of KLEIN, CRAWFORD, AND ALCHIAN [1978] and their analysis of the potential for haggling over appropriable quasi-rents arising from the ownership and use of a contract-specific investment.

Where one or other party to a transaction invests in an asset which has lower value in the absence of the contract, then there is the opportunity for hold-up, i.e. the extraction of appropriable quasi-rents, through 'opportunistic recontracting'. Note that the hold-up could be by either party. Take the example of an assembly firm contracting out the production of a key component to a supplier which invests in specialised machinery that has lower – possibly zero – profitable alternative uses. The assembly firm may exercise hold-up by recognising that the supplier has limited alternative use for the investment, and so bid down the terms of the

contract for supply of the relevant component. But the supplier may also be able to exercise hold-up because it may be difficult or impossible for the assembler to find an alternative source of supply within a reasonable timescale, especially where the market is thin and the investment highly specialised. In one case investing in a contract-specific investment exposes the investor to the risk of hold-up, in the other it permits the exercise of hold-up by the investor. Of course, the hold-up may not be couched in such blatant terms: renegotiation may occur because one or other party claims, for example that the quality of the contracted for good or service has turned out to be a problem. The problem is then deciding when it is a genuine (unforeseen) problem or when it is just a subtle form of opportunism.

3.1 *The Advantage of Firm over Market*

In the rent-seeking approach, the benefit of the firm (or in Williamsonian terms, 'hierarchy') over the market is that the costs of haggling over appropriable quasi-rents are reduced or eliminated by vesting ownership of the assets within a single organization. Lest this should lead to the conclusion that all assets should be vested in one massive vertically integrated firm, a trade-off is introduced, but not one that depends on opportunism. The trade-off is between transaction costs and production costs: internal procurement of a component may imply foregoing economies of scale or scope in production which are available to an external supplier. This is based on WILLIAMSON's [1985], [1989] heuristic model, which envisages a relationship between asset specificity and the potential gains from economies of scale arising from market procurement. It is argued that the production cost advantage of the market declines as assets become more specific because it becomes increasingly difficult to aggregate the different demands of a number of buyers.

There is a second trade-off in the Williamson model, between the high-powered incentives of the market and the relatively low-powered incentives of the hierarchy. It is less clear, however, exactly how hierarchy gets over the problem of opportunism, and so makes the low-powered incentives a desirable option. Having made so much of the issue, Williamson is never entirely clear on how opportunism is mitigated in the firm, a point picked up by early commentators on his work (MCKEAN [1971]) and by property rights theorists (see below). Therefore the advantage of the firm over the market is less manifest than one might wish. In addition, because it treats the market as the default position, transaction cost analysis never fully considers the mechanisms and advantages which market contracting may possess in overcoming the threat of hold-up: "In transaction cost economics, the functioning market is as much a black box as the firm in neoclassical microeconomic theory" (HOLMSTRÖM AND ROBERTS [1998, p 77]).

3.2 *Is Hold-Up Possible in the Absence of Opportunism?*

(a) *The Self-Reinforcing Range of Contracts.* For Williamson, opportunism and hold-up are inextricably linked. The contract-specific nature of many investments and the need to guard against the hazards of opportunism are both essential for

hold-up problems to occur. If assets are non-specific there will always be a market for them in an alternative use, and so there are limited or zero quasi-rents over which to haggle. Equally, if opportunism is not a problem we can rely on transactors sticking to the terms of the contract and not using guile or deceitfulness to twist events to their advantage.

KLEIN [1996], [2000a], [2000b] offers an alternative view of why hold-up occurs which puts a different slant on opportunism and moral hazard issues. Klein points out that contracts involving specific assets often occur between relatively sophisticated contracting parties such as large corporations: yet instances of hold-up still occur. He cites the example of the Fisher Body–GM case as one in which hold-up occurred despite its being “a transaction between two large, sophisticated business firms with no evidence of any pre-contract deception on either transactors’ part” (KLEIN [1996, p. 445]). The reason for this, Klein argues, is because contracts are generally set up to self-reinforce within some range of contractual performance. There is some range of hold-up gain or loss which will be tolerated by the contracting parties because it is more costly to attempt renegotiation within these limits than to accept the gain or loss. But if conditions change sufficiently, for example through a rapid shift in demand as in the case of Fisher Body–GM or a quadrupling in the price of aluminium as in the case of another of his examples, Alcoa–Essex, the contract may move outside its self-reinforcing range and permit the possibility of hold-up by one or other party. The parties are aware of this possibility *ex ante*, but regard the likelihood as being so remote as to be outside the terms of the contracts self-reinforcing range, and so not worth considering in contractual terms. This probabilistic view of hold-up results in contractual non-performance, not through opportunism but through the onset of unanticipated events (i.e. events given *ex ante* a very low probability of occurring).

This can also explain why relatively incomplete contracts are often used in preference to ‘complete’ contracts: in the latter case there is a problem of contractual rigidity which itself can be a source of hold-up if conditions change dramatically. So solving one source of hold-up (i.e. tying down the behaviour of the parties *ex ante*) may give rise to another (arising from contractual rigidity) if conditions change in an unanticipated way.

The Fisher Body–GM case has been the subject of much debate and reinterpretation, with considerable disagreement on whether it involved issues of hold-up at all (CASADESUS-MASANELL AND SPULBER [2000]; COASE [2000], [2006]; FREELAND [2000]; GOLDBERG [2008]; KLEIN [2000b], [2007], [2008]; PAGANO [2000]; ROIDER [2004], [2006]). In a recent debate over the Fisher Body–GM merger, COASE [2006] argues that the exercise of opportunism is checked by the need to consider future loss of business. This, he argues, is similar to the need for a firm committing fraud to have to consider the future loss of custom which this would engender. In this sense, “opportunism is analogous to fraud (and may indeed involve fraud)” (COASE [2006, p. 260]).

KLEIN [2007] sees Coase’s point as indicating the need for deceit for the existence of hold-up, a view with which Klein disagrees:

“The major problem here is a semantic one. In contrast to Coase’s assertion that ‘Opportunism is analogous to fraud’ (2006: 260), the existence of a holdup does not mean that one transactor has deceived its transacting partner. All that is necessary for a holdup to occur is that the contract governing a relationship does not cover some unanticipated change in market conditions and that reputational capital is insufficient to prevent one transactor from taking advantage of these circumstances to shift rents in their favour. *Perhaps ‘post-contractual opportunism’ is a less charged description of what is taking place than ‘holdup’.*” (KLEIN [2007, p. 17], emphasis added)

For Klein, therefore, opportunism and hold-up are essentially one and the same, but neither depends on deceit or guile in the way that Williamson describes. They do, however, depend on one or other party being willing to take advantage of a situation in which an unexpected situation arises that places events outside the self-reinforcing range of a contract. Intriguingly, in more recent work Williamson appears to have adopted a less strong definition of opportunism which echoes KLEIN’s [1996] analysis of the self-reinforcing range within contracts: “The self-interestedness assumption is that of opportunism, on which account parties to a long-term contract will contemplate defection from the spirit of a contract and revert to self-interested bargaining when a contract is pushed out of alignment by significant disturbances” (WILLIAMSON [2003, p. 922]).

(b) The Small Numbers Condition. There is, of course, an extensive body of empirical support for the transaction cost/rent-seeking approach, and especially for the link between asset specificity and vertical integration.⁵ However, even taken on its own terms, it can be shown that opportunism and hold-up arising from it are not necessary for hierarchy to replace the market in the transaction cost framework. The insight that investment by either party in transaction-specific assets fundamentally transforms the nature of the transaction has been extremely powerful. Fear of exposure to opportunistic hold-up leads, it is postulated, to vertical integration. Subsequently, asset specificity, in conjunction with opportunism, has been seen as the driving force behind vertical integration. However, neither asset specificity nor opportunism directly gives rise to hold-up. It is the small numbers situation which gives hold-up its credibility: there is no alternative party with whom to contract or at least no alternative party with whom the transaction can be completed without a cost penalty being incurred that is greater than that imposed by acceding to the opportunistic contractor’s demands. Asset specificity coupled with the possibility of opportunism is one possible means by which a small numbers situation may exist, but it is not the only one. It is the small numbers situation itself which is the necessary condition for the threat of hold-up to be credible and in turn vertical integration to be seen as the appropriate mode of governance.

⁵ In a review of the empirical literature on transaction-cost economics, DAVID AND HAN [2004] find that the role of asset specificity is the most widely supported element of Williamson’s approach. However, other reviews dispute whether there is conclusive evidence even on the role of asset specificity (e.g. CARTER AND HODGSON [2006]).

The collapsing of the small numbers condition into that of asset specificity has had the unfortunate consequence of distracting attention from other sources of hold-up and indeed from the nature of hold-up itself. Another obvious source of the small numbers condition is if there are few potential parties with whom to contract. The extreme source of hold-up is monopoly (and monopsony): supply (or a market) is denied unless the monopolist's (monopsonist's) terms are accepted. Clearly, barriers to entry must be present for the monopolist to extract monopoly rents. However, unless they arise from proprietary knowledge, resource monopoly or significant economies of scale, self-supply could protect against hold-up. Time-critical production has been posited as a further source of vertical integration. Here delay of a supplier in delivering a component or providing a service at the time specified in the contract engenders a production delay which is costly to the customer and, perhaps, triggers a penalty clause in a downstream contract with the customer's own client. This has been characterised in the empirical literature as 'temporal specificity', one variant of asset specificity⁶ (MASTEN, MEEHAN, AND SNYDER [1991]).

However, the requirement for supply of a time-critical component, delay with which gives rise to large losses for the company concerned, may justify a 'make' decision regardless either of the number of potential suppliers (there may simply be no time to switch to an alternative supplier), or of the likelihood of opportunistic hold-up being exercised by a supplier (delay may be engendered for 'genuine' i.e. non-opportunistic reasons). Thus a time-critical component may be produced in-house because the company simply cannot afford the risk of failure to supply for *any* reason. This need not simply be an example of the 'fundamental transformation' from *ex ante* competitive conditions to *ex post* bilateral monopoly, because the component need not depend on investment in specific assets of any kind nor on the exercise of (or fear of) opportunism, yet its time criticality may be great. Note, therefore, that this arises without the need for either 'post-contractual opportunism' or hold-up as defined by KLEIN [2007].

STEPHEN AND LOVE [2000] find some empirical support for this (non-opportunistic) element of time criticality in a study of the make-or-buy decision in naval shipbuilding. Their survey obtained information on 70 components and tasks (31 'make', 39 'buy') performed during the building of a class of vessel which the shipbuilder had designed for, and for which it had obtained several construction orders from, the UK's Ministry of Defence. They found that 'make' components were heavily biased towards high values for 'time criticality' and 'cost of delay' regardless of the number of potential suppliers of the components. In detailed empirical estimation, the costs of 'hold-up'⁷ were related to physical asset specificity, the severity of delay in the production programme from non-supply, and complexity: the number of potential suppliers had no effect on hold-up costs. The decision to

⁶ Along with physical asset specificity, human asset specificity, site specificity and dedicated assets.

⁷ Note that in this research the term 'hold-up' was used to indicate the costs of any form of delay in supply, regardless of the reason for this.

'make' rather than 'buy' was in turn found to be positively related to the size of hold-up costs. Thus for some crucial components, the 'make' option may be used because the cost of delay is too great even if there is no danger of opportunistic hold-up by potential suppliers.

So the choice between governance structures (in Williamsonian terms) may occur because the cost of delay or non-supply of a time-critical component is so high that self-supply is the only cost-efficient option under *any* circumstances. And on closer inspection it becomes clear that this is related to Klein's analysis of conditions changing to push contractual performance outside the self-reinforcing range. Stephen and Love's analysis of self-supply in response to the costs of delay from a time-critical component is an extreme case of Klein's probabilistic view of contract failure: a contract with no self-reinforcing range or one so limited that it effectively precludes market contracting, even in the absence of opportunism (i.e. self-interested behaviour) or the potential for hold-up.

Here, the boundary of the firm exists at the edge of the self-reinforcing range of relevant contracts. Outside these limits either self-supply is used or a potentially profitable trade does not occur at all. The benefit of the firm is that it permits a class of profitable 'transactions' to occur which would not otherwise take place because they are outside this self-reinforcing range. The trade-off between transaction costs and production costs still exists, as it will in any example of vertical integration. But where the self-reinforcing range is zero or very limited, as in the case of delay in a time-critical component, there is no need to tackle the issue of how opportunism is mitigated within the firm; the firm does not arise for reasons of opportunism, but because conditions change beyond the anticipated feasible range of the (incomplete) contract to the benefit of one party whether or not that party has acted in a consciously self-interested manner.

4 Property Rights Theory

The 'modern' property rights theory⁸ (GROSSMAN AND HART [1986], HART [1989], HART AND MOORE [1990]) puts the emphasis on ownership of assets in resolving hold-up problems under conditions of incomplete contracting. Indeed, so powerful has the property rights approach become that it is often regarded synonymously with incomplete contracts, although there are various classes of incomplete contract theory which do not depend on the Grossman–Hart–Moore (GHM) approach (SCHMITZ [2001]).

As with the version of transaction cost economics which deals with rent-seeking behaviour, the issue here is one of hold-up. Specifically, the issue is the potential for hold-up where contract-specific assets are used for efficient specialization. The GHM approach accepts the transaction cost arguments that there will be contracting costs where contracts are incomplete, e.g. haggling over contract revisions; not

⁸ See KIM AND MAHONEY [2005] for a brief overview of the differences between 'modern' and 'classical' property rights literature.

only may *ex post* bargaining be costly, but the parties may, because of asymmetric information, simply fail to reach an efficient agreement. This is not a problem if parties can easily switch to different suppliers, but becomes a problem if they are bound together by some forms of relationship-specific investment, which makes it costly to switch supplier at the (inevitable) renegotiation stage of an incomplete contract. This is seen as one of the hazards of incomplete contracting.

However, while the GHM approach tackles the same issue as Williamsonian transaction cost analysis, it sees a different underlying cause of the problem and proposes another solution. Indeed, HART [1995, p. 27f.] argues that the main weakness of transaction cost analysis is that it fails to explain exactly how haggling and hold-up is reduced when a market transaction is replaced by hierarchy. It is unsatisfactory to assume that parties simply become less opportunistic – why is this the case? As it stands, Hart argues, transaction cost analysis cannot provide an answer.

Because contracts are frequently incomplete, as a result of the costs or sheer impossibility of anticipating every conceivable contingency, issues arise in the sharing of the joint output between the parties to a contract, and – as with transaction cost analysis – problems may arise in giving suitable incentives for investing in relation-specific assets. The GHM approach sees the problem – and its solution – as one of asset ownership, or more strictly, the residual control rights which ownership gives. These control rights are defined as “the rights to decide all usages of the asset in any way not inconsistent with a prior contract, custom, or law” (HART [1995, p. 30]). This is not the usual idea of ownership which is about possession of residual *income* from an asset. By vesting residual control rights with the party which has most to gain over control of the asset’s use, incentive problems are minimised and hold-up problems attenuated. Opportunism need not disappear or be reduced by common ownership, but the appropriate vesting of control rights aligns incentives and therefore reduces the incentive for the use of opportunism.

The boundaries of the firm are therefore determined by the efficient allocation of residual control rights, the rights to make decisions about production which are not made explicit in any contract. Where shared or complementary assets previously held by contracting parties become owned by one of the parties (e.g. via merger) a firm is created. This in turn not only defines the firm clearly as a bundle of assets under common ownership, but also tells us something about the nature of the relationship between employers and employees. Employers have the residual rights with respect to the physical assets which determine the employees’ productivity; this is not the case when an independent contractor is hired. The GHM property rights approach therefore gives us a contractual theory of the firm.

Unlike Williamson’s analysis, opportunism is never explicitly defined in the GHM model, although it is made clear that dealing with issues of opportunism arising from incomplete contracting is the essence of the approach.⁹ However, the

⁹ “[T]o develop a theory of the firm, one must analyse a situation where [...] reputational forces are not strong enough to eliminate all problems of opportunism” (HART [1995, p. 67]).

form of opportunistic behaviour with which the theory deals can be inferred. In drawing attention to the general aspects of contracts, HART [1995] stresses two features:

“The first is that contracts are incomplete. The second is that, because of this, the *ex post* allocation of power (or control) matter. Here power refers roughly to the position of each party if the other party does not perform (e.g. if the other party behaves opportunistically).” (p. 4)

And earlier:

“We are all looking for a contract that will ensure that, whatever happens, each side has some protection against opportunistic behaviour by the other party and against bad luck.” (p. 2)

Opportunism, therefore, appears to involve consciously self-interested behaviour motivated by the prospect of gain at the expense of another contracting party, but excludes ‘windfall’ gains and losses induced by wholly unforeseen events. This is a rather wider interpretation of opportunism than Williamson’s ‘self-interest seeking with guile’ (WILLIAMSON [1985, p. 47]), as it need not involve any form of deceit, blatant or otherwise. It merely requires that one party recognise that it is in their interests not to perform to the terms of the contract and therefore seems closer to the simpler notion of following self-interest at the expense of others. This will still give rise to the types of incomplete contracting problems envisaged by Hart, however. As pointed out earlier, the risk of virtually any kind of self-interested behaviour short of actual deceit may preclude the use of a self-enforcing ‘general clause’ to get round the problem of contractual incompleteness where there is transaction-specific capital (STEPHEN [1996]).

4.1 Authority and Residual Rights to Non-Human Assets

In the GHM model incentive problems are resolved by the appropriate allocation of residual control rights to assets. From the property rights approach, the firm is defined by asset ownership and any employment relation follows as a result. However, by considering the firm as an employment relationship for which asset ownership follows, the issue of authority becomes central, and this is crucial in considering the prospect of a property rights approach which is not conditional on opportunism and its associated hold-up problems. It permits a theory of the firm based on asset ownership but not dependent on hold-up.

HOLMSTRÖM [1999] argues that the property rights approach does not explain the very issue which HART [1995] raises as fundamental to the theory of the firm: why the firm is superior to market transacting under some circumstances:

“The strength of the property rights view is that it articulates so clearly the role of market incentives and how they can be altered by shifts in asset ownership. But it says nothing about the incentives that can be created within firms.” (HOLMSTRÖM [1999, p. 76])

In its pure form, the property rights approach does not say why *firms* own assets, which Holmström sees as “one of the most significant and robust empirical regularities to be explained by any theory of the firm” (p. 75). Indeed, the property rights approach says very little about firms at all: rather it is a theory of asset ownership by individuals.¹⁰ The GHM model, Holmström argues, is subject to precisely the same critique levelled at ALCHIAN AND DEMSETZ [1972], “that organizational affiliations did not matter for transactions” (HOLMSTRÖM [1999, p. 87]). This is important, because the individual ownership of assets does not provide a theory of the firm unless individuals are firms: so the GHM approach is not a theory of firm boundaries, but an entrepreneurial theory of the firm.

Holmström then goes on to suggest alternative reasons as to why firms rather than workers generally own nonhuman productive assets. By having assets under a single authority, firms can assign workers in a manner which is much more varied than under separate ownership. The firm can promote, dismiss or move workers in any way it deems fit, but only within the set of assets which it owns: it is a form of internal capital market. Crucially, “[b]y focusing on holdups alone, the property rights approach overlooks the great variety of instruments that can be used to influence employee incentives” (HOLMSTRÖM [1999, p. 89]), and these incentives go far beyond the simple ones of bonus payments and so on, but include much more subtle arrangements such as the control of information channels, the assignment of tasks, and even espousing a particular corporate culture. Thus the leverage which the firm has over its human assets via the holding of contracting rights over non-human assets arises not merely from dealing with the hold-up power of employees, but from the range of instruments which the firm has at its disposal to deal with these issues. This also helps explain why the low-powered incentives of the firm are preferred in some circumstances, the problem with which Williamsonian transaction cost analysis struggles, according to HART [1995]. The relatively low-powered incentives provided by hierarchy encourages employee cooperation in situations where performance is imperfectly measured and such cooperation might be compromised by strong market incentives.

Another view of the authority relationship also considers a theory of the firm which does not depend on the resolution of incentive conflicts. WERNERFELT [1997] explicitly equates the firm with an employment contract, and tackles the issue of why employees submit to the authority of their employers. This was initially raised by ALCHIAN AND DEMSETZ [1972] in their suggestion that there was no difference between authority in employment and (contractual) ‘authority’ in a market relationship, and is resolved in the property rights approach by the firm’s ownership of the residual rights to productive assets. However, Wernerfelt points out that authority exists even in circumstances where there is little room for the kind of incentive conflicts which give rise to the firm in the property rights analysis, such

¹⁰ “[T]his is a theory of solo entrepreneurs (single actors who own entire asset combinations) and drone employees (who own nothing and hence, in this model, face no incentives and so do nothing)” (GIBBONS [2005, p. 206]).

as members of volunteer organizations. If incentive conflicts can be separated from authority *per se*, this suggests that submission to authority can be an efficient communication pattern where there are no incentive conflicts, and even where such conflicts exist as long as the parties can settle possible conflicts up front. Wernerfelt develops a model which explains when each of three possible types of contracting for human asset services will be most efficient, based on the communication costs of adjustment involved in each. The three types of contract are the hierarchy form (i.e. an employment relationship), the negotiation-as-needed form, and the price list form. Wernerfelt shows theoretically that the employment relation is most efficient where diverse adaptations are needed, and it would be overwhelmingly expensive to negotiate separately for each adaptation. In a relatively small study of manufacturers Wernerfelt finds empirical support for the view that there is a shift from price-list to hierarchical relationship as the need for diverse adaptation increases among salesforces.¹¹

The importance of Wernerfelt's analysis lies in showing that there is no necessary link between asset ownership and residual rights on one hand, and authority relations¹² on the other. This work has the advantage that the benefits of hierarchy over the market are made explicit: parties can coordinate (i.e. avoid haggling) over a wider range of actions in the hierarchy. Where coordination costs are non-zero and diverse adaptations are needed, this can give a substantial advantage, one which depends on the willingness to accept an authority relationship, but does not depend on incentive conflicts or the threat of hold-up.

There are three key points from this analysis. First, the possession of residual rights to non-human productive assets fails to explain the existence of authority in many circumstances (WERNERFELT [1997]), and therefore cannot be the explanation for some aspects of the contractual properties of the firm, specifically the employment relationship. Second, as Holmström indicates, in the property rights model the leverage which ownership of assets provides over 'human assets' is entirely related to forms of hold-up and the resulting payoffs, rather than relating to the much wider variety of instruments which asset ownership provides. In addition, the property rights model is really a theory of individual asset ownership: it fails as a theory of the firm. Third, and crucially, for neither Wernerfelt nor Holmström is the control of opportunistic behaviour necessary for hierarchy to have advantages over the market. For Holmström the firm is about asset ownership (like GHM), but not *just* about controlling hold-up. For Wernerfelt, the employment relationship is the essence of the firm, and this organizational form is preferred where there is a need for diverse adaptations in contracted-for human services. Both are contractual theories, both offer explanations for why the firm may be preferred to the market, but neither depends crucially on opportunism or hold-up.

¹¹ Wernerfelt was unable to test satisfactorily that frequency (rather than diversity) of adaptation leads to a shift from negotiation-as-needed to hierarchy.

¹² Or 'direction' as DEMSETZ [1997] calls the authority relationship.

5 *Principal–Agent Analysis*

The essence of the principal–agent problem is how to design a contract which encourages the agent to act in the best interests of the principal when there is both information asymmetry between principal and agent, and where the two parties have a different objective function. Under these conditions, two agency problems arise: adverse selection (the principal cannot ascertain if the agent accurately represents her ability to perform the task for which she is being paid) and moral hazard (because of measurement problems or asymmetric information, the principal cannot be sure if the agent has put maximum effort into the task).

These problems mean that fixed wage contracts are not always the optimal form of relationship between principals and agents (JENSEN AND MECKLING [1976]). A fixed wage might create an incentive for the agent to shirk (i.e. act opportunistically) since the agent's compensation remains the same regardless of the quality of work or effort level. When agents have incentive to shirk, it may be more efficient to replace fixed wages with compensation based on residual claims on the profits of the firm (ALCHIAN AND DEMSETZ [1972]). Carefully applied, such ownership rights reduce the incentive for agents' adverse selection and moral hazard since it makes their compensation dependent on their performance. According to JENSEN AND MECKLING [1976] the agency problem can be costly, because it includes monitoring costs incurred by the principal, bonding costs incurred by the agent, and 'residual loss' which must be borne by the principal.

More modern approaches to this issue (HOLMSTRÖM AND MILGROM [1991], [1994]) deal specifically with the incentive systems between principals and agents. This approach is not directly concerned with the vertical integration decision which characterises both the transaction cost and property rights theories. But because the Holmström–Milgrom theory deals both with internal incentives and with how asset ownership affects the payoffs to, and incentives of, agents, it provides a theory of the firm by default: it is an 'accidental' theory of the firm (GIBBONS [1998], [2005]).

As with property rights theory, the nature of opportunism assumed by agency theorists is rarely made explicit. However, adverse selection, moral hazard and shirking clearly depend to some extent on opportunistic behaviour by agents. Given the nature of limited and asymmetric behaviour, this need not be Williamsonian guileful behaviour, but certainly requires consciously self-interested behaviour on the part of agents in order to give rise to the agency costs outlined by Jensen and Meckling. Therefore principal–agent theory requires at least 'weak' opportunism.

5.1 *Agency Problems without Opportunism*

Issues of incentive alignment and shirking may not be the greatest problem facing an agency relationship, and are certainly not the only ones. LANGLOIS AND FOSS [1999] point out that exclusive emphasis on issues of incentive alignment can lead to other important areas being ignored, such as how to link together one

person's productive knowledge with that of another, and this has been shown to be of particular importance in terms of principal-agent analysis.

HENDRY [2002], for example, examines the problems inherent in principal-agent contracting, but considers different aspects of the agency problem. Instead of the usual problems of moral hazard, adverse selection and shirking which may arise in the agency relationship, he deals with two other problems which are not induced by self-interested opportunistic behaviour: honest but limited 'incompetence' and the difficulty of specifying objectives. The former issue has no pejorative connotation, but arises from the limits on human knowledge and understanding of what another person wants. An agent may genuinely be unable to understand precisely what the principal expects. In addition, even if the agent believes he understands the role expected of him, he may simply be mistaken; people frequently make genuine mistakes, especially where relatively complex tasks such as running a business require the exercise of (boundedly rational) judgement. The second problem is that of specifiability. Principals may not know precisely what they want at a point in time as it may be contingent on future events, and even if they did know "it may be so contingently complex as to resist accurate specification" (HENDRY [2002, p. 101]). So there are limits in the ability of principals to formulate and communicate to an agent what they want, and problems in the agent (honestly) understanding and implementing this, even where goodwill is evident on both sides.¹³

While Hendry concludes that his version of the agency problem leads to very similar predictions to that of standard agency theory (with one crucial difference, discussed in section 6 below), this does have implications for the design of contracts, and possibly of firms. If the problems he highlights are present and substantial there are implications for the way that agency relationships develop and their attendant contracts are designed: less attention should be paid to the monitoring of opportunism and prevention of malfeasance, and more paid to "combinations of guidance, training and monitoring" (HENDRY [2002, p. 111]),¹⁴ designed to help attenuate the problems of honest incompetence and the limited specifiability of objectives.

HODGSON [2004] outlines a similar argument. His intention is to show that opportunism cannot play the central role within transaction cost economics ascribed to it by Oliver Williamson, and that contract default may occur for reasons other than those of opportunism. However, his argument is entirely couched not in terms of the make-or-buy decision, but in terms of agency problems. Like Hendry, Hodgson's analysis hinges on the problems of communication and interpretation: certainly, wil-

¹³ Hendry's approach is somewhat different to the problems of 'honest disagreements' (ALCHIAN AND WOODWARD [1988]), or 'future uncertainty' (LANGLOIS [1997]). These both involve situations where, even though the parties to a contractual agreement trust each other in virtually every respect, they may have such genuine and irreconcilably different views of what their contractual obligations are under conditions of great uncertainty that joint ownership is the only method of "reconciling divergent visions of the uncertain future" (LANGLOIS [1997, p. 16]).

¹⁴ The issue of monitoring without a clear division of property rights is also discussed by HELPER, MACDUFFIE, AND SABEL [2000].

ful misrepresentation can give rise to contractual default, but this can also arise from “different cognitive frameworks or because of differences in knowledge” (p. 410). Instructions transmitted from one party to another in a contractual relationship may be misunderstood because the parties have different cognitive frameworks or differences in knowledge or cultural background which may render it difficult to seamlessly transfer even codified knowledge without misinterpretation, even where there is no deliberate will to be dishonest or even consciously self-interested. Like Hendry, Hodgson outlines a class of agency problem that is dependent neither on Williamson’s strong opportunism nor on Goldberg’s weaker version. Hodgson also concludes that this may have implications for organizational design: the tasks of managers are more geared towards enabling learning and innovation rather than communicating orders.

Clearly, agency-type contractual problems do not arise solely because of opportunism.¹⁵ But for a theory of the firm we have to be able to show how integration can successfully overcome the kinds of communication, specification and interpretation problems outlined above, and the circumstances under which this will occur. Neither Hendry nor Hodgson explicitly attempts to do so, but this is unsurprising as their aim is to explore issues of contractual design rather than to develop a theory of the firm *per se*. However, I show below that the analysis of agency problems without opportunism and its attendant problems is consistent with Klein’s probabilistic view of the self-reinforcing range of contracts and with the advantages of authority considered by Wernerfelt and Holmström, which together do provide a contractual theory of the firm.

6 Authority, Adaptability and the Firm

It is not difficult to show that non-opportunistic contractual problems may arise within property rights analysis, agency theory, and consideration of the hold-up problem. The discussion above shows that the implicit or explicit assumption of opportunism underlying each of the three main contractual theories of the firm can be questioned on a variety of grounds. The more interesting issue is to show why, and under what circumstances, the firm is more likely to attenuate these problems than the market. Despite their apparent differences, there is a common thread underlying each of the alternative approaches discussed above which helps answer this: the adaptability or flexibility of the firm and the role of authority.

KLEIN [1996], [2000a] shows that, far from being a source of problems where transaction-specific assets are involved, incomplete contracts perform a useful job of permitting flexibility and adaptability up to some limit (the self-reinforcing

¹⁵ The communication and coordination problems highlighted by Hendry and by Hodgson can be seen as elements of LANGLOIS’s [1997] ‘dynamic transaction costs’, albeit embedded in the specific context of the agency relationship.

range).¹⁶ If conditions (unexpectedly) move outside this range, contracts no longer perform the function of protecting against hold-up. Where the self-reinforcing range is very limited, as in costly non-supply of time-critical components (STEPHEN AND LOVE [2000]), contracts may fail as an efficient institutional form even where the parties have no inclination to act in a knowingly self-interested way, and another organizational form – the firm – is preferred. Thus contract terms, and the choice between contractual and hierarchical relationships, goes beyond the role simply of aligning incentives (KLEIN [1996, p. 462]).

The advantage of the firm lies in its authority relations, but this is not merely because authority permits the attenuation of opportunism within the firm. Authority structures exist even where there are no incentive conflicts arising from the exercise of opportunism (WERNERFELT [1997]). Authority provides a usefully flexible form of organization which is efficient where diverse adaptation is needed, and provides a variety of methods of influencing employee incentives and behaviour which go far beyond those dealing with issues of hold-up alone (HOLMSTRÖM [1999]).

Therefore the suggestion that the firm is superior to market contracting because authority controls opportunism within the firm (MCFETRIDGE [1995], MAHONEY [2001]) is only part of the story. Authority within the firm may well play this role on some occasions, and some firms may arise for this reason: but not always, and not everywhere. Authority is not *just* about the suppression of opportunism or the alignment of incentives, and therefore opportunism and hold-up are not *necessary* to explain its existence, or that of the firm, within a contractual framework. Authority has other attributes, which depend “not only on the inducements which are offered but also on the perceived convenience of a ready-made framework for defining problems and seeking appropriate solutions” (LOASBY [1999, p. 101]). Outside the self-reinforcing range of a contract, authority provides the adaptability which an incomplete contract provides within it.

In his more recent writing Oliver Williamson’s version of transaction cost economics – or ‘the lens of contract’ – has evolved in a way which fits more closely with this view of the adaptive properties of the firm vis-à-vis the market.

“More attention to the choice of governance structures which have good adaptive properties (and less to concentrating all of the action in the *ex ante* incentive alignment stage) is one of the central lessons of viewing economic organization through the lens of incomplete contracting.” (WILLIAMSON [2003, p. 925])

Williamson goes on to stress the benefits of hierarchy in terms of ‘cooperative adaptation’. This is not just gathering information through prices as in the market:

“Instead, a wider range of information is gathered and shared, otherwise divergent expectations are supplanted by a common information base from which common projections are made [...], coordinated investments and operating responses are also reached, contingent plans for revisiting the issues are worked up, and disputes are settled not in the courts but

¹⁶ See FINCH [2002] for an insightful example of flexibility and adaptability in contracts in the oil sector which go far beyond the role of incentive alignment.

internally. The adaptive properties of markets and hierarchies differ for each of these reasons.” (p. 925)¹⁷

This view is entirely consistent with that of Wernerfelt’s analysis of the benefits of the employment relationship (i.e. hierarchy) where ‘diverse adaptations’ are needed, and with Holmström’s consideration of the variety of instruments which firms have for influencing behaviour, and how the low-powered incentives of the firm encourage cooperation and adaptability. But, as explained above, neither of these approaches is predicated on the need to guard against opportunism.

Where issues other than protecting against opportunism and hold-up come to the fore, the design of agency/employment contracts, and organizational design more generally, also assumes a wider and more flexible remit. Authority and the agency relationship become less concerned about giving orders, and more concerned with issues such as enabling guidance and training in combination with the more traditional concerns of monitoring and providing appropriate incentives (HENDRY [2002], HODGSON [2004]). Again, flexibility and adaptability are key issues, this time in terms of contractual and organizational design, and again the firm may be preferred where adaptability is at a premium. The clearest example of this can be seen in the one aspect of HENDRY [2002] which differs from standard agency theory. This is in the link between asymmetric information and contract design, with standard theory suggesting that information asymmetry hinders monitoring, increases the scope for agent opportunism, and so increases the likelihood of the principal favouring outcome-based rewards. By contrast, in the non-opportunistic case, information asymmetry encourages the principal to rely on the (dutiful) discretion of the agent, and avoid the use of outcome-based rewards. This is because of the very real costs to the principal of specifying objectives, and the concomitant dangers of mis-specification. Therefore, where adaptability and flexibility are key issues, a simple fixed-wage employment contract may be preferred to a self-employment contract: the firm may be preferred to the market.

7 Conclusions

I began by arguing that a workable theory of the firm ought to show clearly the advantage of firm over market, both for three standard contractual theories and their opportunism-free variants. However, it is worth pointing out that not everyone would even regard this as a useful question to ask. If, like CHEUNG [1983], you believe that firms and markets are just contracts of a different form (one form of contract gives way to another) then the question is of little consequence, unless you regard the employment contract as being *sui generis* and worthy of attention in its own right. From this perspective, what we are interested in is good contract design, and whether

¹⁷ Williamson approvingly cites MALMGREN [1961] at this point, which is ironic as FOSS [1996c] shows that Malmgren developed a theory of the firm without resort to opportunism.

we call one of these contractual forms a firm is of no consequence: the search for the firm is futile. Similarly, if you believe that firms and markets are fundamentally different, that they are complements rather than substitutes, and that firms should not be regarded as markets *manqué*, then you will be equally uninterested in the question but for quite different reasons: not because it is uninteresting but because it is – literally – meaningless. Firms produce things or make decisions (KAY [2000]) while markets provide price signals, so deciding whether one is ‘better’ than the other is like asking whether a bicycle is better than a chicken: it depends on whether you want to ride it or eat it.

I take an intermediate position between these two extremes. Granted, firms do not do everything markets do and vice versa, but contractual theories concern themselves with the aspects on which they overlap, and with precisely those areas in which a choice can be made about whether firm or market should best be adopted. Nor is there any doubt that opportunism and the threat of hold-up exist, that measure to guard against them are important, and that these measures can influence both contractual design and the existence and boundaries of firms under some circumstances. But it is possible to have explanations of the existence and boundaries of the firm which do not depend on opportunism and hold-up under all circumstances, while still operating within the framework of the three archetypical examples of contractual theories. There is therefore some purpose in considering a theory of the firm which is induced by (non-opportunistic) contracting problems, and asking what sort of issues of organizational design this will throw up.

By considering contractual issues which do not arise only because of opportunism we emphasise aspects of organizational and contractual design which are not all about monitoring and protecting against shirking or moral hazard. If we concentrate too heavily on hold-up, incentives and the prevention of contractual non-compliance, we ignore important issues such as the wider role of authority in providing adaptability and flexibility, and how authority can influence employee behaviour in ways that are not concerned only with hold-up. But by considering these wider issues, far from obscuring the analysis of firm versus market, we return to a rich analysis of the firm – and of incomplete contracting – that is more in keeping with COASE [1937], [1993] and his emphasis on the superior ability of firms in terms of directing factors of production and providing less complicated contractual arrangements than the market.¹⁸ Opportunism, in either its strong or weak forms, is a useful restrictive assumption in permitting the development of more formal statements of theories of the firm, because it allows us to concentrate on issues such as hold-up, incentive alignment, moral hazard and shirking. But the assumption of opportunism is not necessary to have a contractual theory of the firm, and concentrating exclusively on issues of incentive alignment entails costs: it distracts us from giving consideration to some of the more interesting aspects of contractual design and functions, and ultimately narrows our understanding of the nature of the firm.

¹⁸ See LOVE [2005] for a more detailed discussion of both Coase’s and Demsetz’s analyses of authority, direction and adaptability within the firm.

References

- ALCHIAN, A., AND H. DEMSETZ [1972], "Production, Information Costs, and Economic Organization," *The American Economic Review*, 62, 777–795.
- ALCHIAN, A. A., AND S. WOODWARD [1988] "The Firm is Dead; Long Live the Firm: A Review of Oliver E. Williamson's *The Economic Institutions of Capitalism*," *Journal of Economic Literature*, 26, 65–79.
- CARTER, R., AND G. M. HODGSON [2006], "The Impact of Empirical Tests of Transaction Cost Economics on the Debate on the Nature of the Firm," *Strategic Management Journal*, 27, 461–476.
- CASADESUS-MASANELL, R., AND D. F. SPULBER [2000], "The Fable of Fisher Body," *The Journal of Law & Economics*, 43, 67–104.
- CHEUNG, S. N. S. [1983], "The Contractual Nature of the Firm," *The Journal of Law & Economics*, 26, 1–21.
- COASE, R. H. [1937], "The Nature of the Firm," *Economica*, 4, 386–405.
- [1993], "The Nature of the Firm: Influence," pp. 61–74 in: O. E. Williamson and S. G. Winter (eds.), *The Nature of the Firm: Origins, Evolution and Development*, Oxford University Press: Oxford.
- [2000], "The Acquisition of Fisher Body by General Motors," *The Journal of Law & Economics*, 43, 15–31.
- [2006], "The Conduct of Economics: The Example of Fisher Body and General Motors," *Journal of Economics & Management Strategy*, 15, 255–278.
- CONNER, K. R. [1991], "A Historical Comparison of Resource-Based Theory and Five Schools of Thought within Industrial Organization Economics: Do we Have a New Theory of the Firm?" *Journal of Management*, 17, 121–154.
- AND C. K. PRAHALAD [1996], "A Resource-Based Theory of the Firm: Knowledge versus Opportunism," *Organization Science*, 7, 477–501.
- DAVID, R. J., AND S.-K. HAN [2004], "A Systematic Assessment of the Empirical Support for Transaction Cost Economics," *Strategic Management Journal*, 25, 39–58.
- DEMSETZ, H. [1997], *The Economics of the Business Firm: Seven Critical Commentaries*, Cambridge University Press: Cambridge.
- FINCH, J. H. [2002], "Transferring Exploration and Production Activities within the UK's Upstream Oil and Gas Industry: A Capabilities Perspective," *Journal of Evolutionary Economics*, 12, 55–81.
- FOSS, N. J. [1996a], "Knowledge-Based Approaches to the Theory of the Firm: Some Critical Comments," *Organization Science*, 7, 470–476.
- [1996b], "More Critical Comments on Knowledge-Based Theories of the Firm," *Organization Science*, 7, 519–523.
- [1996c], "Harald B Malmgren's Analysis of the Firm: Lessons for Modern Theorists?" *Review of Political Economy*, 8, 349–366.
- FREELAND, R. H. [2000], "Creating Holdup through Vertical Integration: Fisher Body Revisited," *The Journal of Law & Economics*, 43, 33–66.
- GARROUSTE, P., AND S. SAUSSIER [2005], "Looking for a Theory of the Firm: Future Challenges," *Journal of Economic Behavior & Organization*, 58, 178–199.
- GIBBONS, R. [1998], "Incentives in Organizations," *The Journal of Economic Perspectives*, 12, 115–132.
- [2005], "Four Formal(izable) Theories of the Firm?" *Journal of Economic Behavior & Organization*, 58, 200–245.
- GOLDBERG, V. P. [1984], "A Relational Exchange Perspective on the Employment Relationship," pp. 127–145 in: F. H. Stephen (ed.), *Firms, Organizations and Labour*, Macmillan: London.
- [2008], "Lawyers Asleep at the Wheel? The GM–Fisher Body Contract," *Industrial and Corporate Change*, 17, 1071–1084.

- GROSSMAN, S., AND O. HART [1986], "The Costs and Benefits of Ownership: A Theory of Lateral and Vertical Integration," *Journal of Political Economy*, 94, 691–719.
- HART, O. [1989], "An Economist's Perspective on the Theory of the Firm," *Columbia Law Review*, 89, 1757–1774.
- [1995], *Firms, Contracts, and Financial Structure*, Clarendon Press: Oxford.
- AND J. MOORE [1990], "Property Rights and the Nature of the Firm," *Journal of Political Economy*, 98, 1119–1158.
- HELPER, S., J. MACDUFFIE, AND C. SABEL [2000], "Pragmatic Collaborations: Advancing Knowledge while Controlling Opportunism," *Industrial and Corporate Change*, 9, 443–488.
- HENDRY, J. [2002], "The Principal's Other Problems: Honest Incompetence and the Specification of Objectives," *The Academy of Management Review*, 27, 98–113.
- HODGSON, G. M. [2002], "The Legal Nature of the Firm and the Myth of the Firm-Market Hybrid," *International Journal of the Economics of Business*, 9, 37–60.
- [2004], "Opportunism is Not the Only Reason why Firms Exist: Why an Explanatory Emphasis on Opportunism May Mislead Management Strategy," *Industrial and Corporate Change*, 13, 401–418.
- HOLMSTRÖM, B. [1999], "The Firm as a Subeconomy," *The Journal of Law, Economics, & Organization*, 15, 74–102.
- AND P. MILGROM [1991], "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *The Journal of Law, Economics, & Organization*, 7, 201–228.
- AND — [1994], "The Firm as an Incentive System," *The American Economic Review*, 84, 972–991.
- AND J. ROBERTS [1998], "The Boundaries of the Firm Revisited," *Journal of Economic Perspectives*, 12, 73–94.
- JENSEN, M., AND W. MECKLING [1976], "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership," *Journal of Financial Economics*, 3, 305–360.
- KAY, N. [2000], "Searching for the Firm: The Role of Decision in the Economics of Organization," *Industrial and Corporate Change*, 9, 683–707.
- KIM, J., AND J. T. MAHONEY [2005], "Property Rights Theory, Transaction Cost Theory, and Agency Theory: An Organizational Economics Approach to Strategic Management," *Managerial and Decision Economics*, 26, 223–242.
- KLEIN, B. [1996], "Why Hold-Ups Occur: The Self-Reinforcing Range of Contractual Relationships," *Economic Inquiry*, 34, 444–463.
- [2000a], "The Role of Incomplete Contracts in Self-Reinforcing Relationships," *Revue d'économie industrielle*, issue 92, 67–80.
- [2000b], "Fisher-General Motors and the Nature of the Firm," *The Journal of Law & Economics*, 43, 105–141.
- [2007], "The Economic Lessons of Fisher Body-General Motors," *International Journal of the Economics of Business*, 14, 1–36.
- [2008], "The Enforceability of the GM-Fisher Body Contract: Comment on Goldberg," *Industrial and Corporate Change*, 17, 1085–1096.
- , R. G. CRAWFORD, AND A. A. ALCHIAN [1978], "Vertical Integration, Appropriable Rents and Competitive Contracting Process," *The Journal of Law & Economics*, 21, 297–326.
- LANGLOIS, R. N. [1997], "Transaction Costs, Production Costs, and the Passage of Time," pp. 1–21 in: S. G. Medema (ed.), *Coasean Economics: Law and Economics and the New Institutional Economics*, Kluwer Academic Publishers: Dordrecht.
- AND N. J. FOSS [1999], "Capabilities and Governance: The Rebirth of Production in the Theory of Economic Organization," *Kyklos*, 52, 201–218.
- LOASBY, B. [1999], *Knowledge, Institutions and Evolution in Economics*, Routledge: London.
- LOVE, J. H. [2005], "On the Opportunism-Independent Theory of the Firm," *Cambridge Journal of Economics*, 29, 381–397.

- MAHONEY, J. T. [2001], "A Resource-Based Theory of Sustainable Rents," *Journal of Management*, 27, 651–660.
- MALMGREN, H. B. [1961], "Information, Expectations and the Theory of the Firm," *The Quarterly Journal of Economics*, 75, 399–421.
- MASTEN, S. E., J. W. MEEHAN, AND E. A. SNYDER [1991], "The Costs of Organization," *The Journal of Law, Economics, & Organization*, 7, 1–25.
- McFETRIDGE, D. G. [1995], "Knowledge, Market Failure and the Multinational Enterprise: A Comment," *Journal of International Business Studies*, 26, 409–415.
- McKEAN, R. [1971], "Discussion," *The American Economic Review*, 61, 124–125.
- PAGANO, U. [2000], "Public Markets, Private Orderings and Corporate Governance," *International Review of Law and Economics*, 20, 453–477.
- ROIDER, A. [2004], "Asset Ownership and Contractibility of Interaction," *The RAND Journal of Economics*, 35, 787–802.
- [2006], "Fisher Body Revisited: Supply Contracts and Vertical Integration," *European Journal of Law and Economics*, 22, 181–196.
- SCHMITZ, P. W. [2001], "The Hold-Up Problem and Incomplete Contracts: A Survey of Recent Topics in Contract Theory," *Bulletin of Economic Research*, 53, 1–17.
- SMITH, D. G., AND B. G. KING [2008], "Contracts as Organizations," *Arizona Law Review*, 51, 1–45.
- STEPHEN, F. H. [1996], "Lawyers, Transaction Costs and Opportunism," *Journal of Institutional and Theoretical Economics*, 152, 146–153.
- AND J. H. LOVE [2000], "Hold-Up Costs, Economies of Scale and the Make-or-Buy Decision," Research Paper RP0020, Aston Business School, Birmingham.
- WERNERFELT, B. [1997], "On the Nature and Scope of the Firm: An Adjustment-Cost Theory," *The Journal of Business*, 70, 489–514.
- WILLIAMSON, O. E. [1985], *The Economic Institutions of Capitalism*, Free Press: New York.
- [1989], "Transaction Cost Economics," Chapter 3 in: R. Schmalensee and R. Willig (eds.), *Handbook of Industrial Organisation*, North-Holland: Amsterdam.
- [2003], "Examining Economic Organization through the Lens of Contract," *Industrial and Corporate Change*, 12, 917–942.

James H. Love
Economics and Strategy Group
Aston Business School
Aston University
Birmingham
B4 7ET
United Kingdom
E-mail:
j.h.love@aston.ac.uk

Unemployment, Public Pensions, and Capital Accumulation: Assessing Growth Effects of Alternative Funding Strategies

by

JOACHIM THØGERSEN*

The paper develops an overlapping-generations model that interacts with a labor market characterized by equilibrium unemployment. This structure implies that young individuals can be in two different states, employed or unemployed. Hence, the social security system contains both old-age benefits and unemployment insurance. Including these features, the model seeks to assess growth effects of three different pension systems: one unfunded and two funded, where the distinction is made between actuarial and nonactuarial funding strategies. It is shown that both funded systems generate higher growth than an unfunded system. Moreover, the actuarial system fosters higher growth than the nonactuarial. (JEL: E 24, H 55, J 51, O 40)

1 Introduction

Several OECD countries experience population aging due to low fertility rates and higher life expectancies. At the same time many of these countries are plagued with high unemployment rates due to structural problems. A third observation is that many of these economies also suffer from a decline in productivity and slow economic growth. All of these problems are closely related to how the social security program is formed, and consequently social security reforms are highly prioritized on the policy agenda.

Since most of these countries have unfunded pension systems, population aging will trigger increased tax burdens and/or reduced benefits in the future. This can affect savings. These problems have caused several countries to reform their so-

* Oslo University College. I am indebted to Steinar Holden and Øystein Thøgersen for highly appreciated discussions, comments, and suggestions. I have also benefited from comments provided by Trond-Arne Borgersen; Kåre Bævre; Vidar Christiansen; Erling Steigum; participants at the SUERF Colloquium: Money, Finance and Demography – the Consequences of Ageing, Lisbon 2006; participants at Forskermøtet, Bergen 2009; and an anonymous referee. I gratefully acknowledge financial support from the University of Agder.

cial security systems. Typical changes have involved more funding strategies and changes in retirement incentives, in order to motivate workers to stay longer in the workforce. In most European economies these issues are accompanied by unemployment problems. High unemployment rates have been observed for a long time. Although there does not exist a consensus theory on this long-term unemployment, most economists agree that the problems have some structural flavor and can be associated with institutional arrangements, such as the wage formation or the social security system and welfare programs (LJUNGQVIST AND SARGENT [1998], LAYARD, NICKELL, AND JACKMAN [1991]). As social security has an influence on both employment issues and saving decisions, social security must also be linked to economic growth.

In this paper I will analyze the relationship between different social security systems, unemployment, and economic growth. To do this an overlapping-generations model in discrete time is applied, where the young generation can be in either of two states, employed or unemployed. The extent of unemployment is related to the wage bargaining. The old generation is assumed to be nonworking. As the model economy contains two groups that are nonworking, the social security system contains both unemployment insurance and old-age pension benefits. These characteristics and institutional features are relevant for European welfare states and labor markets.

Several papers have treated these issues separately, and the focus has often been on different aspects of the features mentioned above, analyzed in isolation. The relation between pensions and growth is typically analyzed in intertemporal models as treated in BREYER [1989], SAINT-PAUL [1992], and LAMBRECHT, MICHEL, AND VIDAL [2005]. Motivated by the pension reform debate, several authors have studied the transition from a nonfunding pension scheme to a more or completely funded system (VERBON [1989], PETERS [1991], THØGERSEN [2001]). The shift is known to have a short-run cost, as the transition generation is constrained to pay twice, both for its own retirement and for the old part of the population through a pay-as-you-go scheme. From a social welfare point of view the shift is therefore problematic. In BELAN, MICHEL, AND PESTIEAU [1998], however, the transition is studied in an endogenous growth model, and it is shown that a Pareto-improving social security reform is possible.

The link between social security and unemployment is, however, ignored in this literature, and usually studied in a separate class of models. AGHION AND HOWITT [1999] and PISSARIDES [2000] study unemployment and growth when unemployment is due to search frictions and mismatch, while BRÄUNINGER [2000] and LINGENS [2003] study the same variables, but on the assumption that unemployment is caused by the wage bargaining.

CORNEO AND MARQUARDT [2000] take a first step in integrating social security, unemployment, and growth, where social security refers to the combination of public pensions and unemployment insurance programs. The labor market in their model is characterized by union wage setting, where the wage is set by a monopoly union. They find that unemployment does not affect growth and that the Pareto-improving

pension reform studied in BELAN, MICHEL, AND PESTIEAU [1998] is maintained even when allowing for equilibrium unemployment. BRÄUNINGER [2005] studies the same relations as Corneo and Marquardt, but he assumes that the wage is determined through wage bargaining. Bräuninger concludes that both of the insurance components in the social security program have a negative effect on growth. First of all, the pension system has a direct negative effect, since pensions crowd out savings and therefore reduce capital accumulation and growth. Secondly, unemployment insurance has an indirect negative effect on economic growth by affecting unemployment. The unemployment benefits influence equilibrium unemployment through the wage bargaining.

To expand on this path of literature I model and compare how different pension systems affect capital accumulation and growth. To make the comparisons it is necessary to find analytical solutions for the growth factor of capital. I have therefore made some substantial changes to the setup by Bräuninger. In addition, not only a fully funded and a pay-as-you-go (PAYGO) scheme are considered, but also a third alternative. This alternative is assumed to be similar to a PAYGO system with respect to the governmental intervention, but dissimilar with respect to the financing. Pension payments are here assumed to be financed by a pension fund governed by the government, and where the young individuals pay taxes to finance their own generation's old-age need. But the system is also nonindividualized and nonactuarial; hence there is no link between tax contribution and pension benefit. Thus it is not equivalent to an individual and fully funded system, which is perfectly actuarial.¹ I will elaborate on the different pension schemes in section 5.3 and in the conclusion. Addressing policy issues in a setting like this is often avoided in the theoretical literature, though it is a highly important issue for policymakers. The question I ask in this paper is therefore how different social security systems will influence savings, capital formation, and growth in an economy where wage bargaining leads to long-term unemployment and the social security system accordingly comprises old-age pensions and unemployment insurance.

The rest of the paper is organized in the following way: Section 2 describes population in an overlapping-generation setting. Special attention is given to the two states possible for a young worker. In section 3, households, life-cycle consumption, and the individual savings function are modeled. Section 4 presents the production structure of the economy. Firms are assumed to act under monopolistic competition. This section also studies firms' interaction with trade unions and the wage bargaining. In section 5 I describe the government and the different social security systems. Then, section 6 assesses how capital accumulation and economic growth are affected by different funding strategies when capital is endogenous. Section 7 offers some concluding remarks.

¹ See FEHR AND THØGERSEN [2009] for a survey of different pension schemes and their effects on future generations.

2 Population and States

I consider a model with two cohorts. Each individual lives in two periods, and the ones who are young in period t are old in period $t + 1$. The number of the young in period t is N_t . Population grows at a constant rate n , and it follows that $N_{t+1} = (1 + n)N_t \Leftrightarrow n = (N_t/N_{t-1}) - 1$, where N_{t-1} is the number of old in period t . Young individuals all supply one unit of labor, though they can be in either of two different states: employed or unemployed. The proportion $u \in (0, 1)$ is unemployed, so the total numbers of unemployed and employed are, respectively, uN_t and $(1 - u)N_t$. Each of the working individuals earns the wage w_t . The wage rate, however, is taxed at a proportional rate τ , in order to finance unemployment insurance (b) and pensions. To which generation the pension is distributed depends on the pension system.

3 Households and Life-Cycle Consumption

A representative young agent in generation t will choose a consumption path to maximize the following lifetime utility function: $U_t = U(c_{1,t}, c_{2,t+1})$, where U is strictly concave and increasing in each of its two arguments, $c_{1,t}$ is consumption when young in period t , and $c_{2,t+1}$ is consumption when the same individual is old in the next period, $t + 1$. For analytical purposes I assume a logarithmic Cobb–Douglas description of the preference structure:

$$(1) \quad U_t = (1 - \delta) \log c_{1,t} + \delta \log c_{2,t+1},$$

where $\delta \in (0, 1)$ is the weight on consumption in the two periods and reflects the discount factor. When individuals are young, they can either work and earn $w_t(1 - \tau)$, or be unemployed and receive unemployment benefits. The working part of the population will allocate income between consumption and saving (s_t). All savings are allocated to investments, which yield a positive rate of return. In the second period of life, individuals are retired and receive a pension benefit along with a payoff from investments made when young. Pensioners consume all of their wealth, i.e., the model does not include bequests.

A crucial assumption in the current model is that taxes have distortionary effects. This assumption is necessary in order to distinguish between real effects of the two funded pension schemes. Due to the importance of this assumption, it will be further discussed in the last section. The modeling of the distortion can be done in several ways. One approach is to let labor supply be endogenous and include leisure in the utility function. However, such an approach, in an overlapping-generations model like the one applied here, would significantly complicate the capital market equilibrium. The dynamics of capital would then be characterized by a second-order difference equation, which might involve multiple equilibrium paths and nonuniqueness. Accordingly, I have adopted the approach by BARRO [1979] and BOHN [1992]. They argue that taxation involves collection costs and/or misallocation costs that are imposed on the private economy. Hence, the tax on labor income will

impose an excess burden on workers. This burden will not influence the unemployed, since they do not pay taxes. The excess burden is denoted $h(\tau)$; we have $h'(\tau) > 0$ and $h(0) = 0$. The net labor income of a worker in period t is therefore $w_t(1 - \tau - h(\tau))$. This simplified modeling of the distortionary effect is not crucial for the qualitative results of the analysis.

In order to assess effects on economic growth, it is convenient to derive an expression for aggregate savings. Aggregate savings are the sum of the savings made by the employed and the unemployed. The unemployed workers receive unemployment benefits while young, and pension benefits while retired. However, the unemployed workers also divide their income (transfers) between consumption and saving, and in the aggregate, their intertemporal choice is included. When analyzing growth implications, it is necessary to express both aggregate savings and potential wealth accumulation by the government. Whether the government can accumulate financial capital depends on the pension system. However, I start the exposition by setting up the intertemporal consumption decision of the working part of the population and thereby derive an expression for their individual saving. As the young and working individuals in period t will divide their net income between consumption and saving, consumption is given by

$$(2) \quad c_{1,t} = w_t(1 - \tau - h(\tau)) - s_t.$$

Consumption of the old is given by their benefits and accumulated saving:

$$(3) \quad c_{2,t+1} = \theta w_t + s_t R_{t+1},$$

where R_t denotes the interest factor, and $\theta < 1$ denotes the constant pension ratio. From (2) and (3) one can derive the intertemporal budget constraint:

$$(4) \quad c_{1,t} + \frac{1}{R_{t+1}} c_{2,t+1} = w_t(1 - \tau - h(\tau)) + \frac{1}{R_{t+1}} \theta w_t =: \Lambda_t,$$

where Λ_t denotes net life-cycle income. Equation (4) expresses the net life income received in the first period, plus the discounted value of pensions received in the second period of life. The decision problem for young workers born in period t is to maximize the lifetime utility (1) subject to the consolidated budget constraint in (4). To derive optimal individual savings, it is convenient to define the following savings function:

$$s_t := \arg \max_{c_{1,t}, c_{2,t+1}} \{U(c_{1,t}, c_{2,t+1}) \mid c_{1,t} + (R_{t+1})^{-1} c_{2,t+1} = \Lambda_t\}.$$

Using the Cobb–Douglas specification of the utility function along with the budget constraints in period t gives optimal individual savings as

$$(5) \quad s_t = \left[\delta(1 - \tau - h(\tau)) - (1 - \delta) \frac{\theta}{R_{t+1}} \right] w_t.$$

Equation (5) states that individual savings depend on individual income during the working period, the tax distortion, and the level of pensions. In section 6 I will include savings made by the unemployed in order to expand and derive an expression for total savings for all individuals.

4 Firms and Wage Bargaining

4.1 Technology and Market Structure

The firms act in a market characterized by monopolistic competition. There are therefore assumed to be a large number of firms, but all goods are imperfect substitutes such that there exists some profit to be shared between workers and firms. The sharing of these profits is a part of the bargaining. Each firm i makes use of two input factors, capital ($K_{i,t}$) and labor ($L_{i,t}$), in order to produce a variety of goods. The demand for these goods produced by firm i is given by $Y_{i,t} = \pi_{i,t}^{-\eta} Y_t$, where $\pi_{i,t}$ is the relative price of good i , η is the price elasticity of demand, and Y_t is an index of aggregate demand. Technology is given by a Cobb–Douglas production function with positive and diminishing marginal products of each input, constant returns to scale, and labor-augmenting technology for firm i :

$$Y_{i,t} = K_{i,t}^\alpha (A_t L_{i,t})^\beta, \quad \text{where } \alpha + \beta = 1, \alpha > 0, \text{ and } \beta > 0.$$

Effective labor is thus given by $A_t L_{i,t}$, where A_t measures labor efficiency, which each single firm takes as given. Firms maximize profits $\Pi_{i,t} = I_{i,t} - w_{i,t} L_{i,t} - R_{i,t} K_{i,t}$, where $I_{i,t}$ is revenue and given by $\pi_{i,t} Y_{i,t}$. It is assumed that capital fully depreciates each period. Insert the inverse demand function to obtain $I_{i,t} = Y_t^{1/\eta} Y_{i,t}^\kappa$, where $\kappa := 1 - 1/\eta$. Standard profit maximization implies that the marginal revenues of capital and labor equal their input prices, respectively,

$$\frac{\partial I_{i,t}}{\partial L_{i,t}} = \frac{\beta \kappa I_{i,t}}{L_{i,t}} = w_{i,t} \quad \text{and} \quad \frac{\partial I_{i,t}}{\partial K_{i,t}} = \frac{\alpha \kappa I_{i,t}}{K_{i,t}} = R_{i,t}.$$

4.2 Aggregate Production and Technological Spillovers

In order to assess growth effects in the aggregate economy, it is necessary to expand the profit-maximizing framework above to include all the agents, employed and unemployed. I assume that firms are symmetric; hence, in the aggregate, $K_{i,t} = K_t$ and $Y_{i,t} = Y_t$. It is also assumed that the real wage is determined in the wage bargaining process as presented in section 4.3. The symmetry of firms implies that all prices are equal; hence the relative price is $\pi_{i,t} = \pi = 1$, for all t . The revenue of each firm will then be equal to its output, i.e., $I_{i,t} = Y_{i,t}$. Furthermore, as all firms are assumed to be symmetric, the production function is equal for all firms, i.e., $Y_t = K_t^\alpha (A_t L_t)^\beta$. Since young people in generation t are either employed or unemployed, $L_t = (1 - u)N_t$, and aggregate production is thus given by

$$(6) \quad Y_t = K_t^\alpha [A_t (1 - u)N_t]^\beta.$$

The profit-maximizing first-order conditions in section 4.1 are at the aggregate level given by $\alpha \kappa Y_t / K_t = R_t$ and $\beta \kappa Y_t / L_t = w_t$. Inserting the production function in (6) gives the first-order conditions as

$$(7) \quad \alpha \kappa \hat{k}_t^{-\beta} (1 - u)^\beta = R_t \quad \text{and} \quad \frac{A_t \beta \kappa \hat{k}_t^\alpha}{(1 - u)^\alpha} = w_t,$$

where $\hat{k}_t := K_t/A_t N_t$ denotes capital per unit of efficient labor. To endogenize the labor productivity index A_t , I follow the setup originally formulated by ARROW [1962] and the extensions by ROMER [1986]. Following this approach implies a technological spillover from the size of the aggregate capital stock on labor productivity in individual firms. This positive externality permits an endogenous growth process. In order to ensure the existence of a steady-state equilibrium, the technological spillover is assumed to be linear in the aggregate capital stock per young individual:

$$(8) \quad A_t = \frac{1}{a} \frac{K_t}{N_t},$$

where a is a scaling productivity parameter reflecting the influence of capital intensity on labor productivity.² Moreover, the productivity of labor is decreasing in a . With respect to the production function, inserting (8) into (6) gives

$$Y_t = K_t(1-u)^\beta a^{-\beta},$$

i.e., the production function is of the AK type and linear in capital. Notice that unemployment reduces output. Such a relation is motivated by the empirical evidence of a negative effect of unemployment on growth.³ In BRÄUNINGER [2005], it is shown that a convenient way to model this aspect is to let the technological spillover depend on young individuals in general, and not only on the employed individuals. This feature is captured in (8), and it follows that unemployment has a negative effect on output and economic growth. From (8) and the definition of \hat{k}_t , it follows that capital per effective unit of labor is constant:

$$(9) \quad \hat{k}_t := \frac{K_t}{A_t N_t} = a.$$

It is straightforward from (7) to see that the interest factor (R_t) and the wage per efficient unit (w_t/A_t) are functions of \hat{k}_t . Substituting (9) into the first-order conditions in (7) yields

$$(10) \quad R_t = \alpha \kappa a^{-\beta} (1-u)^\beta = R \quad \text{for all } t,$$

$$(11) \quad w_t = \frac{A_t \beta \kappa a^\alpha}{(1-u)^\alpha}.$$

Thus, the interest factor is constant over time, and the wage rate is proportional to the level of labor productivity, i.e., grows at the rate of growth of the technological spillover.

4.3 Unions, Bargaining, and Equilibrium Unemployment

In this subsection I present the wage setting and the equilibrium unemployment rate. This is done in order to analyze how unemployment will affect economic

² See GROSSMAN AND YANAGAWA [1993], WIGGER [2002], and HOLLER [2007] for similar approaches.

³ See DAVERI AND TABELLINI [2000] and BRÄUNINGER AND PANNENBERG [2002] for empirical studies.

growth under different social security systems. Unemployment is the result of wage bargaining at the firm level. This exposition of the labor market is conventional and can be found in BOOTH [1995], LAYARD, NICKELL, AND JACKMAN [1991], or BRÄUNINGER [2005] among others.⁴

All workers, both employed and unemployed, are members of a trade union, and the total number of members is therefore N_i . The trade union is assumed to maximize the utility of a worker, given by net income. The workers can either be employed in firm i and earn $(1 - \tau)w_i$, or have an alternative income m_i , which comes either from employment in another firm or as unemployment benefit.⁵ Each individual is sometimes unemployed, depending on the outcome of the bargaining and the exogenous fluctuations in the labor market. It is assumed that the outcome of the bargaining problem is given by maximization of the Nash product $\Phi = (V_i - \bar{V}_i)^\gamma (\Pi_i - \bar{\Pi}_i)^{1-\gamma}$, where γ is the relative bargaining power of the union and $V_i = N_i v_i$, where v_i is the utility of a member. \bar{V}_i is the threat point of the union, and is given by the alternative income of a member in case of disagreement, $m_i N_i$. Likewise $\bar{\Pi}_i$ is the threat point of the firm, and is given by the firm's payoff in case of disagreement, $-RK_i$. Using these expressions, one obtains the following Nash product: $\Phi = (L_i((1 - \tau)w_i - m_i))^\gamma (I_i - w_i L_i)^{1-\gamma}$, and maximization implies

$$(12) \quad (1 - \tau)w_i = \mu m_i, \quad \text{where } \mu := 1 - \gamma + \frac{\gamma}{\beta\kappa},$$

i.e., the wage is equal to the alternative income multiplied by a fixed markup $\mu \geq 1$. The higher the union power, the higher is the markup. If the union has no power, $\gamma = 0 \Rightarrow \mu = 1$, i.e., the firm offers a wage equal to the alternative income.

Union members not employed in firm i face a probability $1 - \phi u$ of being employed in another firm and a probability ϕu of staying unemployed. The parameter ϕ is exogenous and describes fluctuations in the labor market. These fluctuations take place at much higher frequency than the periods in the overlapping-generations model, so they are not included formally in the model. By assuming that being employed in another firm gives the worker a net income of $(1 - \tau)w$, one can formulate the following relation for the alternative income: $m_i = \phi u b w + (1 - \phi u)(1 - \tau)w$. Inserting this equation into (12), one obtains

$$(13) \quad (1 - \tau)w_i = \mu (\phi u b w + (1 - \phi u)(1 - \tau)w) .$$

The assumption of identical firms and homogeneous workers implies $w_i = w$. Inserting this into (13) and solving for the equilibrium unemployment rate yields

$$u = \frac{(\mu - 1)(1 - \tau)}{\mu\phi(1 - \tau - b)},$$

which is equivalent to BRÄUNINGER [2005]. The expression shows that equilibrium unemployment depends on taxes.

⁴ The modeling of the wage setting follows LAYARD, NICKELL, AND JACKMAN [1991] and BRÄUNINGER [2005], and the reader is encouraged to consult these references for further discussion.

⁵ I drop time indices in this part of the exposition.

5 Government and Social Security

The government runs a social security system that consists of unemployment benefits and old-age pension benefits. In order to finance unemployment insurance, the government imposes a tax on the working individuals. This leads to an intragenerational transfer. Pension benefits are also financed by taxing the workers, but whether the transfer is intergenerational or intragenerational depends on the pension system. In the following, three different stylized systems are considered, one unfunded and two funded. While the funded schemes are intragenerational, the unfunded is intergenerational. The two funded systems differ with respect to the tax–benefit link, and whether they are individual or not. One system is assumed to be nonindividual and nonactuarial, i.e., a weak tax–benefit link. The other funded system assumes individual and fully actuarial funding of the pension benefits. Under a perfect capital market this system is equivalent to a system without governmental interventions, with respect to total savings and capital accumulation (BLANCHARD AND FISCHER [1989]). Accordingly, it can be analyzed by assuming that pension payments and their earmarked taxes are zero. Hence, the social security system will only consist of unemployment insurance. As elaborated on in section 5.3, these distinctions are crucial to the formulation of the governmental budget restrictions.

In the nonactuarial funding system the government accumulates financial wealth. This is due to the time lag between income from taxes and the payments to the old. This implies that the government under certain assumptions will contribute to the accumulation of national wealth. As will become clear in section 5.3, this is not the case with the other pension systems.

5.1 National Wealth

It is assumed that the economy is closed, which implies that a country's national wealth (Ω_t^n) consists of the capital in the economy. Since capital can be accumulated by the government in the case of a nonactuarial funding system, as well as by the private sector, we have $\Omega_t^n := K_t = \Omega_t^P + \Omega_t^G$, where Ω_t^P and Ω_t^G denote the wealth accumulated by the households and by the government, respectively. The employed workers in the economy pay taxes to finance the two components of the social security system. The government distributes these revenues among the entitled ones. For expositional reasons we distinguish between taxes paid to finance unemployment insurance (τ_u) and taxes paid to finance old-age pensions (τ_p). The government's wealth in the beginning of period $t + 1$ is accordingly

$$\Omega_{t+1}^G = R_{t+1}\Omega_t^G + \tau_p w_t N_t (1 - u) - PN_{t-1}.$$

Note that intragenerational transfers from the employed to the unemployed are excluded, due to their purely intragenerational distributional characterization. This part of the tax is not invested and therefore cannot contribute to wealth accumulation. The government's wealth in per-worker form is given by

$$(14) \quad (1 + n)\omega_{t+1}^G = R_{t+1}\omega_t^G + \tau_p w_t (1 - u) - P(1 + n)^{-1},$$

where $\omega := \Omega/N$ and $P = \theta w_t$. The pension P received by the old part of the population is proportional to the current wage.

5.2 Unemployment Insurance

The young and working individuals in period t , $(1-u)N_t$, finance unemployment benefits to the unemployed. These benefits (B) are fixed in relation to wages: $B = bw_t$, where $b < 1$ is the replacement ratio. Since the total number of unemployed is uN_t , the total expenditures to the unemployed in period t are buw_tN_t . Since expenditures must be equal to total taxes earmarked for unemployment benefits, the following budget restriction with respect to unemployment insurance must hold:

$$\tau_u w_t (1-u)N_t = buw_t N_t \iff \tau_u = \frac{bu}{(1-u)},$$

which implies that the tax rate with respect to unemployment insurance is independent of the pension system.

5.3 Pension Systems

As aforementioned, I distinguish between three different stylized pension schemes: pay-as-you-go (PAYGO), nonactuarial funding (NAF), and actuarial funding (AF).

PAYGO Financing. Within this regime it is assumed that taxes on labor income to the young part of the population are used to finance old-age pensions in the same period. This gives the following budget restriction with respect to pensions:

$$(15) \quad \tau_u w_t N_t (1-u) = \theta w_t N_{t-1}.$$

As long as taxes to finance pensioners are paid out in the same period as they are received, the government cannot accumulate wealth. This implies that $\Omega_t^G = 0$ for all t . One can then rewrite the restriction and solve for τ_p :

$$\tau_p = \frac{\theta N_{t-1}}{(1-u)N_t} = \frac{\theta}{(1-u)(1+n)}.$$

The total tax levied on the worker under a PAYGO system can then be expressed as

$$(16) \quad \tau^{\text{PAYGO}} = \tau_u + \tau_p = \frac{\theta + bu(1+n)}{(1-u)(1+n)}.$$

The tax is negatively related to n and positively related to the replacement ratio b , the pension ratio θ , and the unemployment rate u .

Nonactuarial Funding. This regime is also assumed to be nonindividualized, and the representative individual in each generation will pay taxes that are contributions to his own generation's pension fund. This means that the representative individual in the second period of life will receive a pension that equals his generations' earlier contributions plus accumulated interests. As the system is characterized by a nonactuarial relationship between each individual's contribution when young and

his benefit received when old, the pension received by each individual does not necessarily reflect his contribution in particular. The government therefore distributes income on a generation basis, and not on an individual one.⁶ In this system the government can contribute to the accumulation of national wealth. The government's financial wealth, Ω_t^G , is pension taxes that the government receives in period $t - 1$, and at the beginning of period t , the government has $R_{t+1}\Omega_t^G$ at its disposal. This implies that $\theta w_t N_{t-1} = R_{t+1}\Omega_t^G$, and that $R_{t+1}\omega_t^G = \theta w_t / (1 + n)$ in per-worker form. By inserting this into (14), it follows that

$$(17) \quad (1 + n)\omega_{t+1}^G = \tau_p w_t (1 - u).$$

Equation (17) reveals that the taxes paid by the employed in period t give the governmental wealth in period $t + 1$. The budget restriction that characterizes the NAF strategy is accordingly

$$(18) \quad \theta w_t N_t = R_{t+1}\tau_p w_t N_t (1 - u) \iff \tau_p = \frac{\theta}{R_{t+1}(1 - u)}.$$

The total tax under a NAF social security system is then given by

$$(19) \quad \tau^{\text{NAF}} = \tau_p + \tau_u = \frac{\theta + buR_{t+1}}{(1 - u)R_{t+1}}.$$

Equation (19) shows that the tax is now not directly affected by n . The growth in population will not, under a NAF regime, have a direct effect on the tax paid by workers. From equation (7), however, one can see that population growth has an effect on the interest rate and therefore an indirect effect on taxes.

Actuarial Funding. This regime is also assumed to be individualized, so the ordinary payments of the workers to social security are replaced with contributions to their own individual accounts. Workers save a mandated fraction of their labor income and invest it by themselves for their own old-age need. Provided that capital markets are perfect, an individualized and actuarial pension system is equivalent to the absence of a pension system. Thus, the social security system consists only of an unemployment insurance scheme. Hence, $\tau_u > 0$ and $\tau_p = \theta = 0$. Moreover, the government cannot accumulate wealth, i.e., $\omega_t^G = 0$ for all t . The total taxes on labor income under an AF strategy are then

$$(20) \quad \tau^{\text{AF}} = \tau_p + \tau_u = \frac{bu}{1 - u}.$$

Hence, an actuarial funding strategy implies that taxes are independent of population growth and the real interest factor. Notice also that taxes are lower in the AF system than in the NAF system. Therefore, the relative excess burden of taxes is higher in the NAF system. This feature illustrates a difference between AF and NAF that is important in the subsequent analysis.

The above analysis shows that the pension system affects the tax levied on workers.

⁶ A similar approach can be found in THØGERSEN [2001].

6 Capital Accumulation and Alternative Social Security Systems

This section studies how capital accumulation and growth depend on the choice of the pension system. To incorporate different funding strategies the model uses the different governmental budget restrictions from section 5.3. The growth factor is determined by aggregate savings and capital accumulation. I will therefore derive an expression for aggregate savings where the contributions from both the employed and the unemployed are included. Moreover, the analysis of different pension systems makes it necessary to derive explicit solutions for the growth factor of capital in the different settings. The following exposition follows BRÄUNINGER [2005] and expands his model by including tax distortions, different pension systems, and comparative studies.

6.1 Aggregate Savings

Total savings in the economy are defined as the sum of the savings by the employed workers and by the unemployed individuals. However, as the government can accumulate financial wealth if there is a time lag between its tax income and social security payments, its contribution must also be included in the analysis. Thus, considering the equilibrium in the capital market involves both total savings and governmental savings. Total savings are given by expanding equation (5) to include all the employed and unemployed workers. The proportion of the employed is $1 - u$, and their individual income is $(1 - \tau)w_t$. The proportion of the unemployed is u , and their income is bw_t . Aggregate savings are accordingly

$$S_t = \delta(1 - \tau)(1 - u)w_tN_t - \delta h(\tau)(1 - u)w_tN_t \\ - \frac{(1 - \delta)\theta w_t}{R_{t+1}}(1 - u)N_t + \delta buw_tN_t - \frac{(1 - \delta)\theta w_t}{R_{t+1}}uN_t.$$

Notice that the contributions from the employed must correspond to the benefits received by the unemployed in equilibrium. This means that $\tau_u(1 - u)w_tN_t = buw_tN_t$, and S_t can be simplified to

$$(21) \quad S_t = \delta(1 - u)w_tN_t[1 - \tau_p - h(\tau)] - \frac{(1 - \delta)\theta w_t}{R_{t+1}}N_t.$$

Equation (21) shows that savings depend on the wage rate, the unemployment rate, the tax distortion, and the pension ratio. Savings are however independent of the replacement ratio. The distortionary effects in this expression are related to the taxes paid to pensions. Due to this tax distortion, aggregate savings are lower than they would have been without it.

The pension ratio is important. How changes in the pension ratio will affect savings depends on the pension system under consideration.

6.2 Equilibrium Conditions and Capital Accumulation

In period t , the equilibrium in the economy as a whole is defined by equilibrium in three markets: the labor market, the capital market, and the final good market. In the labor market, equilibrium is given by $L_t = (1 - u)N_t$, which follows because a positive part of the population is at any time unemployed. The final good market equilibrium displays the resource constraint for the economy as a whole, and states that output can be used either for aggregate consumption C_t or for gross investment in period t :

$$Y_t = C_t + K_{t+1},$$

where

$$C_t := N_t c_{1,t} + N_{t-1} c_{2,t},$$

i.e., aggregate consumption is the sum of consumption by the young and the old individuals in period t .

As the depreciation rate of capital is assumed to be unity, capital evolves according to $K_{t+1} = S_t$. In the capital market the supply of capital comes from both private and governmental savings. The government's financial wealth is therefore essential in the consideration of the equilibrium condition for the capital market. The next period's capital is therefore given as

$$K_{t+1} = S_t + \Omega_{t+1}^G,$$

where aggregate savings by the private sector are given by equation (21). By dividing aggregate savings by N_t , one obtains savings per young individual, including both employed and unemployed. The dynamic behavior of capital per young individual is accordingly $(1 + n)k_{t+1} = S_t/N_t + (1 + n)\omega_{t+1}^G$, where $k_t := K_t/N_t$ denotes capital per young individual. Thus, inserting (21) and (14) gives the dynamics of capital in the economy as

$$(22) \quad (1 + n)k_{t+1} = \delta(1 - u)w_t[1 - \tau_p - h(\tau)] - \frac{(1 - \delta)\theta w_t}{R_{t+1}} + R_{t+1}\omega_t^G + \tau_p w_t(1 - u) - \frac{\theta w_t}{1 + n}.$$

The dynamic equilibrium in (22) is fundamental in the subsequent analysis of how social security and different public pension regimes affect the growth of capital in the economy. The long-run growth factor of capital in the economy is defined by

$$(23) \quad g := \frac{k_{t+1}}{k_t}.$$

To obtain analytical expressions for the growth factor the following equations are necessary: the intertemporal equilibrium in (22) and the first-order conditions in (10) and (11), and to distinguish between different pension systems one needs the governmental budget restrictions in (15), (18), and (20). It is also necessary to specify whether the government's wealth is equal to zero or not between two periods. Moreover, the following result turns out to be quite useful:

LEMMA 1 *The structure of the technological spillover applied in the production sector implies that $w_t/k_t = \text{const}$ for all t .*

PROOF Inserting (8) into (11) yields

$$\frac{w_t}{k_t} = \frac{A_t \beta \kappa a^\alpha}{k_t (1-u)^\alpha} = \frac{\beta \kappa}{(1-u)^\alpha a^\beta},$$

which is constant.

Q.E.D.

6.3 Comparing Funding Strategies

6.3.1 PAYGO Financing

In a PAYGO system an increase in the pension ratio will affect savings in two ways. First, it increases the contributions of the young, so that their net income declines. Secondly, it affects the young generation's motivation to save, since part of their consumption when old is financed by the next generation's tax payments. Moreover, with a PAYGO pension scheme the government cannot contribute to national wealth, as all governmental transfers are intergenerational, and tax income obtained in one period is transferred to pensioners within the same period. Hence, $\Omega_t^G = \omega_t^G = 0$ for all t . To implement the governmental budget restriction into the growth factor defined in (23), it is convenient to solve equation (15) for the pension ratio θ . This implies

$$(24) \quad \theta = \tau_p(1-u)(1+n),$$

where $(1+n)(1-u)$ is the ratio of the number of employed workers in period t , to that of all workers in period $t-1$. If population growth exceeds the ratio between the unemployment rate and the employment rate $(1-u)$, then the pension ratio exceeds the pension tax.

Inserting (24) and $\omega_{t+1}^G = 0$ into (22) gives

$$(25) \quad (1+n)k_{t+1} = \left\{ \delta[1 - \tau_p - h(\tau)] - \frac{(1-\delta)\tau_p(1+n)}{R_{t+1}} \right\} (1-u)w_t,$$

which displays the dynamic behavior of capital with a PAYGO pension scheme. Inserting (10), (25), and Lemma 1 into (23) yields the following growth factor:

$$(26) \quad g^{\text{PAYGO}} = \left\{ \frac{\delta[1 - \tau_p - h(\tau)]}{1+n} - \frac{(1-\delta)\tau_p}{R} \right\} \frac{\beta \kappa (1-u)^\beta}{a^\beta}.$$

Note that the growth factor is time-invariant. It is straightforward to see that unemployment reduces the growth in capital. The reason for this is that unemployment leads to reduced output, lower aggregate income, and therefore lower savings. Moreover, an increase in the pension tax decreases savings by young individuals and consequently the growth factor. Formally,

$$\frac{\partial g^{\text{PAYGO}}}{\partial \tau_p} = - \left\{ \frac{[1 + h'(\tau)]\delta}{1+n} + \frac{(1-\delta)}{R} \right\} \frac{\beta \kappa (1-u)^\beta}{a^\beta} < 0.$$

6.3.2 Nonactuarial Funding

As aforementioned, the NAF funding strategy implies that the young working generation finances its own pension, but on a generation basis, and not on an individual one. The intragenerational characterization of the NAF system also implies that the pension ratio is independent of population growth. In section 5.3, it was shown that the government can contribute to the national wealth, due to the time lag between tax income and pension payments. Hence, the government accumulates wealth according to (17). To derive the growth factor in an economy with a NAF pension scheme, it is necessary to solve the governmental budget restriction given in (18) for the pension ratio. This implies

$$(27) \quad \theta = \tau_p(1 - u)R_{t+1}.$$

Using (17) and (27), the dynamic behavior of capital per young individual within this pension regime becomes

$$(28) \quad (1 + n)k_{t+1} = [1 - h(\tau)]\delta(1 - u)w_t.$$

Consequently, the growth factor is derived by inserting (28) and Lemma 1 into (23):

$$(29) \quad g^{\text{NAF}} = \frac{\beta\delta\kappa[1 - h(\tau)](1 - u)^\beta}{(1 + n)a^\beta}.$$

Hence, the growth factor is time-invariant. As in the PAYGO program, unemployment and pension taxes reduce growth. The negative effect of the tax is due to the distortionary effect that remains as long as the working generation pays taxes.

The following proposition compares the PAYGO pension scheme with the NAF scheme, with respect to their effect on the growth factor of capital.

PROPOSITION 1 *Capital accumulation is higher in an economy with a NAF pension system than in an economy with a PAYGO pension system, i.e., $g^{\text{NAF}} > g^{\text{PAYGO}}$.*

PROOF The proposition is proved by contradiction. On assuming that $g^{\text{PAYGO}} \geq g^{\text{NAF}}$, and utilizing the expressions in (26) and (29), it follows that

$$\begin{aligned} g^{\text{PAYGO}} &\geq g^{\text{NAF}} \\ \Rightarrow \left\{ \frac{\delta[1 - \tau_p - h(\tau)]}{1 + n} - \frac{(1 - \delta)\tau_p}{R} \right\} \frac{\beta\kappa(1 - u)^\beta}{a^\beta} &\geq \frac{\beta\delta\kappa[1 - h(\tau)](1 - u)^\beta}{(1 + n)a^\beta} \\ \Leftrightarrow \delta[1 - \tau_p - h(\tau)] - \frac{(1 + n)(1 - \delta)\tau_p}{R} &\geq \delta[1 - h(\tau)], \end{aligned}$$

which leads to a contradiction, as $\tau_p > 0$, and the second term on the left-hand side is greater than zero. *Q.E.D.*

The reason for this result lies in the government’s contribution to the economy’s total savings through a pension fund. Establishing a public social security fund and thereby accumulating financial wealth, the government indirectly stimulates capital accumulation. In this system the government accumulates financial wealth because of the time lag between when tax money is received by the government and when it is transferred to the older generation. Hence, an important difference between this

system and the PAYGO system is that the government now contributes to capital accumulation.

According to neoclassical growth theory, capital accumulation will temporarily increase the growth rate. In this model, however, we also get a permanent increase in growth, since accumulation of capital increases the stock of knowledge and positive spillovers stimulate perpetual growth.

6.3.3 Actuarial Funding

The main assumption within this system is that individuals pay contributions to their own individual accounts. As noted in section 5.3, this implies that $\tau_p = \theta = 0$. But the employed workers still pay taxes in order to finance unemployment benefits. This implies that $\tau_u > 0$. Under a PAYGO and a NAF strategy, the tax consists of both τ_u and τ_p . Consequently, since $\tau = \tau_u + \tau_p > \tau_u$ and $h'(\tau) > 0$, the following inequality applies: $h(\tau) > h(\tau_u)$, i.e., the tax distortion is lower than with the other social security programs.

The government has no opportunity to accumulate wealth in the AF system, as the only tax received is intragenerational, i.e., $\omega_t^G = 0$ for all t . By inserting these assumptions and corollaries into equation (22) one obtains the equilibrium dynamics in the capital market as

$$(30) \quad (1+n)k_{t+1} = [1 - h(\tau_u)]\delta(1-u)w_t.$$

Comparing capital dynamics in the NAF system given by (28) and in the AF system shows that the only feature that distinguishes the social security programs is the size of the tax distortion $h(\cdot)$. Due to the lower tax in the AF program, the excess burden is higher in the NAF program.

Accordingly, the growth factor in the economy with an AF social security system is found by inserting (30) and Lemma 1 into (23):

$$(31) \quad g^{\text{AF}} = \frac{\beta\delta\kappa [1 - h(\tau_u)](1-u)^\beta}{(1+n)a^\beta}.$$

Hence, the growth factor is time-invariant. Inspection of (31) shows that the tax distortion still exists, but is now lower than in the other pension systems. This follows because the total tax rate on wages is now lower.

The following proposition compares capital accumulation in an economy with a NAF pension program and an economy with an AF pension program.

PROPOSITION 2 *Capital accumulation is higher in an economy with an AF pension system than in an economy with a NAF pension system, i.e., $g^{\text{AF}} > g^{\text{NAF}}$.*

PROOF The proposition is showed by contradiction. On assuming that $g^{\text{NAF}} \geq g^{\text{AF}}$, and utilizing the expressions in (26) and (31), it follows that

$$\begin{aligned} g^{\text{NAF}} \geq g^{\text{AF}} &\Rightarrow \frac{\beta\delta\kappa[1 - h(\tau)](1-u)^\beta}{(1+n)a^\beta} \geq \frac{\beta\delta\kappa[1 - h(\tau_u)](1-u)^\beta}{(1+n)a^\beta} \\ &\Leftrightarrow h(\tau) \leq h(\tau_u). \end{aligned}$$

Since $h(\tau) > h(\tau_u)$, the inequality is not fulfilled and the proof is complete. *Q.E.D.*

Individual net income is higher under an AF pension system than under a NAF system. This is due to both smaller taxes and smaller tax distortions. The government cannot contribute to national wealth in an AF program, as the unemployment tax is intergenerational and distributed within the cohort. However, the individuals in the economy compensate for this element by saving a larger proportion of their income. And as the tax and the collection costs are lower, the net income is higher. Therefore, total savings and capital accumulation are greater in an economy with an AF social security scheme.

Another feature of individual and actuarial pension schemes is that the motivation to save is higher under them than under a nonindividualized and nonactuarial pension system. This is due to the one-to-one connection between an individual's payments and received pensions (SØRENSEN, HANSEN, AND BOVENBERG [2006]).

7 Conclusion

In several European countries, population aging, unemployment, and economic growth are much-debated issues among professional economists and politicians. These issues are highly related through the social security system.

CORNEO AND MARQUARDT [2000] consider a model where individuals live for two periods. In the first period, individuals can be either employed or unemployed, and in the second period they are all retired. Due to this structure, the social security system refers to a combination of public pensions and unemployment insurance programs. A main point in Corneo and Marquardt is that unemployment is caused by the union wage setting. In their model the labor market is characterized by a monopoly union that determines the wage.

BRÄUNINGER [2005] expands the model of Corneo and Marquardt to include wage bargaining at an intermediate level. The labor market is therefore characterized by a Nash bargaining solution, rather than a monopoly trade union.

The current paper contributes to this theoretical literature by expanding Bräuninger's model to include three different pension schemes. Moreover, the setup of the model makes it possible to compare the different pension schemes with respect to growth implications. The PAYGO and the AF system are fairly standard in the literature, except for the inclusion of long-term unemployment. However, the NAF system is rarely considered, and it represents the counterpart to a fully funded system, regarding the tax–benefit link and the degree of individualism.

To distinguish real effects of the two funded schemes, it is necessary to assume some sort of distortion in the economy. In the current setup, this is done via the excess burden on workers due to tax payments. If $h(\tau) = 0$, the NAF and the AF schemes would be equivalent with respect to capital accumulation and output.

The second novel part of the paper lies in the combination of the modeling of the pension systems and the endogenous growth framework, which permits an analytical solution of the growth factor of capital. To be able to express the growth factor explicitly, the technological spillover is assumed to be linear in capital

intensity. Moreover, to simplify the growth model compared with some of the earlier literature, I model the growth factor by expressing it per young individual. These simplifications make it possible to do a comparative analysis of the different pension systems.

It is shown that growth is higher in an economy with a NAF system than with a PAYGO system. The result depends on the governmental contribution to national wealth, and thereby capital accumulation, within the NAF scheme. Comparing the two funding strategies reveals that an individual and actuarial pension scheme fosters higher growth than the nonindividual and nonactuarial pension scheme.

References

- AGHION, P., AND P. HOWITT [1999], *Endogenous Growth Theory*, The MIT Press: Cambridge, MA.
- ARROW, K. J. [1962], "The Economic Implications of Learning by Doing," *The Review of Economic Studies*, 29, 155–173.
- BARRO, R. J. [1979], "On the Determination of the Public Debt," *Journal of Political Economy*, 87, 940–971.
- BELAN, P., P. MICHEL, AND P. PESTIEAU [1998], "Pareto-Improving Social Security Reform with Endogenous Growth," *The Geneva Papers on Risk and Insurance Theory*, 23, 119–125.
- BLANCHARD, O. J., AND S. FISCHER [1989], *Lectures on Macroeconomics*, The MIT Press: Cambridge, MA.
- BOHN, H. [1992], "Endogenous Government Spending and Ricardian Equivalence," *The Economic Journal*, 102, 588–597.
- BOOTH, A. L. [1995], *The Economics of the Trade Union*, Cambridge University Press: Cambridge.
- BRÄUNINGER, M. [2000], "Wage Bargaining, Unemployment and Growth," *Journal of Institutional and Theoretical Economics*, 156, 646–660.
- [2005], "Social Security, Unemployment, and Growth," *International Tax and Public Finance*, 12, 423–434.
- AND M. PANNENBERG [2002], "Unemployment and Productivity Growth: An Empirical Analysis within the Augmented Solow Model," *Economic Modelling*, 19, 105–120.
- BREYER, F. [1989], "On the Intergenerational Pareto Efficiency of Pay-as-you-Go Financed Pension Systems," *Journal of Institutional and Theoretical Economics*, 145, 643–658.
- CORNEO, G., AND M. MARQUARDT [2000], "Public Pensions, Unemployment Insurance, and Growth," *Journal of Public Economics*, 75, 293–311.
- DAVERI, F., AND G. TABELLINI [2000], "Unemployment, Growth and Taxation in Industrial Countries," *Economic Policy*, 30, 47–104.
- FEHR, H., AND Ø. THØGGERSEN [2009], "Social Security and Future Generations," pp. 417–454 in: R. J. Brent (ed.), *Handbook on Research in Cost–Benefit Analysis*, Edward Elgar Publishing: Cheltenham.
- GROSSMAN, G. M., AND N. YANAGAWA [1993], "Asset Bubbles and Endogenous Growth," *Journal of Monetary Economics*, 31, 3–19.
- HOLLER, J. [2007], "Pension Systems and their Influence on Fertility and Growth," Working Paper 0704, Department of Economics, University of Vienna.
- LAMBRECHT, S., P. MICHEL, AND J.-P. VIDAL [2005], "Public Pensions and Growth," *European Economic Review*, 49, 1261–1281.
- LAYARD, R., S. NICKELL, AND R. JACKMAN [1991], *Unemployment: Macroeconomic Performance and the Labour Market*, Oxford University Press: Oxford.

- LINGENS, J. [2003], "The Impact of a Unionised Labour Market in a Schumpeterian Growth Model," *Labour Economics*, 10, 91–104.
- LJUNGQVIST, L., AND T. J. SARGENT [1998], "The European Unemployment Dilemma," *Journal of Political Economy*, 106, 514–550.
- PETERS, W. [1991], "Public Pensions in Transition: An Optimal Policy Path," *Journal of Population Economics*, 4, 155–175.
- PISSARIDES, C. A. [2000], *Equilibrium Unemployment Theory*, The MIT Press: Cambridge, MA.
- ROMER, P. [1986], "Increasing Returns and Long-Run Growth," *Journal of Political Economy*, 94, 1002–1035.
- SAINT-PAUL, G. [1992], "Fiscal Policy in an Endogenous Growth Model," *The Quarterly Journal of Economics*, 107, 1243–1259.
- SØRENSEN, P. B., M. I. HANSEN, AND A. L. BOVENBERG [2006], "Individual Savings Accounts and the Life-Cycle Approach to Social Insurance," EPRU Working Paper Series: Copenhagen.
- THØGERSEN, Ø. [2001], "Reforming Social Security: Assessing the Effects of Alternative Funding Strategies in a Small Open Economy," *Applied Economics*, 33, 1531–1540.
- VERBON, H. A. [1989], "Conversion Policies for Public Pensions Plans in a Small Open Economy," pp. 83–95 in: B. Gustafsson and N. A. Klevmarken (eds.), *The Political Economy of Social Security*, North-Holland: Amsterdam.
- WIGGER, B. U. [2002], *Public Pensions and Economic Growth*, Springer: Berlin.

Joachim Thøgersen
Faculty of Social Sciences
Oslo University College
Pilestredet 35
0130 Oslo
Norway
E-mail:
joachim.thogersen@sam.hio.no

Liquidity Creation without Bank Panics and Deposit Insurance

by

JUHA-PEKKA NIINIMÄKI*

This paper develops a panic-free bank system in an OLG model. A bank issues both demand deposits and time deposits (or bank stocks) so that the maturity-matching constraint is satisfied. The agents who cannot participate in capital markets put their savings in demand deposits; others favour marketable time deposits. Everyone receives a liquid saving asset, and the bank boosts the liquidity of the economy, even though it operates under maturity matching. The costs of stabilization are high if the bank's operating costs are substantial or if there are only a few agents who will participate in the capital markets without subsidies. (JEL: G 21, G 22, G 28)

1 Introduction

The recent subprime crisis has triggered a surge of bank panics. In Britain, for example, panic-stricken depositors in three days withdrew over £3 billion from Northern Rock Bank in September 2007. The collapse of Lehman Brothers, the fifth-largest U.S. investment bank, in September 2008 unleashed a further devastating wave of financial panics, generating a legion of bank failures.

This paper offers a method for eliminating bank panics. The economy involves two types of agents: active and passive. Both types are risk-averse and encounter privately observed shocks to intertemporal preferences for consumption. Active agents have access to capital markets, invest their funds directly in firm stocks, and attain liquidity by trading them. Thus, active agents do not need banks. Passive agents do not have access to capital markets and obtain liquidity only through demand deposits. Unfortunately, the bank is vulnerable to runs. If depositors panic and try to withdraw their deposits simultaneously, the liquidation value of assets does not cover the payments on deposits and the bank fails. This paper demonstrates that it is possible to construct a panic-free bank system utilizing maturity matching: A bank issues both liquid demand deposits and time deposits (alternatively, bank stocks) so that the total liquidation value of deposits is equal to the liquidation

* Bank of Finland. I thank Dominique Demougin and two anonymous referees for useful comments and suggestions.

value of the bank assets. Depositors know that the bank is always able to pay off deposits and so have no reason to panic. Time deposits are more productive than demand deposits and cannot be interrupted before maturity, but they can be resold in secondary markets. Since time deposits are more productive than demand deposits, active agents invest in them instead of investing directly in firm stocks. Passive agents favour demand deposits. Therefore, both agent types obtain a liquid saving asset.

If the costs of market access are large, there may initially be no active agents. The bank must pay very high interest rates on time deposits in order to encourage some agents to become active, invest in time deposits, and trade them. Since the payments on time deposits exceed the bank's income from its long-term assets, the bank can pay less on demand deposits. The stabilization effect of time deposits (or bank stocks) may be so expensive that it makes banking unprofitable. To avoid this, a bank regulator may decide to offer deposit insurance. Panics can be prevented without the excessive burden of time deposits, and banks can pay a moderate return on demand deposits.

Given maturity matching, it may appear that the bank does not increase the liquidity of the economy. But this is not the case. The bank provides a fundamental service to the economy by transforming the constant liquidity of its assets to deposits with different liquidity. Demand deposits are more liquid than bank assets. Time deposits are nominally less liquid than bank assets, since they cannot be interrupted before maturity. They can, however, be resold in secondary markets and thus are effectively liquid for the agents who can participate in the markets. Consequently, both demand deposits and time deposits are more liquid than the assets of the bank, and thus the bank increases the liquidity of the economy even under maturity matching.¹

The paper is related to previous analysis on bank panics, e.g., DIAMOND AND DYBVG [1983].² It is linked most closely to alternative suggestions for eliminating panics: deposit insurance or central-bank intervention (DIAMOND AND DYBVG [1983], ALLEN AND GALE [1998]), suspension of convertibility (WALLACE [1988]), narrow banking (WALLACE [1996]), and time deposits that cannot be traded (NIINIMÄKI [2003]). GREEN AND LIN [2003] suggest that panics could be prevented by replacing simple demand deposit contracts with more advanced contracts. This paper is most closely related to NIINIMÄKI [2003]. He shows how panics can be prevented if a bank attracts not only liquid demand deposits but also time deposits with low interruption value. When depositors face the different risk of a preference shock, high-risk agents will hold their savings as demand deposits, whereas low-risk

¹ Large-denomination time deposits (CDs), which can be resold in secondary markets, are a crucial source of funds for banks. CDs represent approximately 16% of the liabilities of commercial banks in the U.S. (MISHKIN [2007, p. 221]).

² Models that use DIAMOND AND DYBVG's [1983] framework are still extensively employed: VON THADDEN [1997], [1999], ALLEN AND GALE [1998], [2004], FREIXAS, PARIGI, AND ROCHET [2000], GREEN AND LIN [2000], [2003], QI [2003], ROCHET AND VIVES [2004], and CHEN AND HASAN [2006], [2008].

agents will prefer time deposits. Since the interruption value of a time deposit is lower than the initial deposit, time deposits are very risky. An agent is eager to resell his time deposits before maturity instead of interrupting them, but reselling is not allowed in the model. The analysis is extended in this paper so that it is possible to resell time deposits.

The paper uses a dynamic OLG model. In the one-generation economy of DIAMOND AND DYBVIK [1983], a bank can create liquidity for agents only if they live in isolation and are unable to contact each other or participate in markets. We cannot use the one-generation model, because time deposits are now marketable. Therefore, we use the OLG version, in which depositors can meet each other and participate in markets. Previous OLG models (QI [1994], BHATTACHARYA AND PADILLA [1996], FULGHIERI AND ROVELLI [1998], BHATTACHARYA, FULGHIERI, AND ROVELLI [1998]) study the possibility of sharing liquidity risk among generations, whereas this paper aims to design a panic-free bank system.

The assumption that only a fraction of agents can participate in capital markets is borrowed from DIAMOND [1997]. Diamond concentrates on liquidity insurance, whereas this paper also examines panics. In DIAMOND [1997], increasing participation in capital markets moderately reduces the liquidity that banks can create. In this paper, increasing participation in capital markets moderately *increases* the amount of liquidity that banks can create. The difference is based on model frameworks. DIAMOND [1997] uses the one-generation model, whereas this paper adopts the OLG framework.

Nowadays the threat of panics has widened from deposit banks to other types of firms, most of all investment banks. In 2008 Bear Stearns and Lehman Brothers, for example, relied greatly on extremely fragile short-term funding, e.g., 24-hour repo contracts. The drying up of short-term funding drove these investment banks to liquidity crisis and failure. Many nonbank firms that had funded their operations with short-term commercial paper also faced severe liquidity problems. Our findings on the stabilizing effects of capital stock and long-term funding in deposit banks can be generalized to other types of firms.

The rest of the paper is organized as follows. Section 2 sets out the economy, and section 3 describes a bank that is vulnerable to panic. A panic-free bank is set up with time deposits in section 4, and with bank stocks in section 5. In both cases, the economy has sufficiently many active agents. The opposite case is set out in section 6, and section 7 concludes.

2 Economy

Consider an infinite-horizon economy with an infinite sequence of overlapping generations of agents. A new generation is born at every time $t \in \{0, 1, 2, \dots\}$ and consists of a continuum of agents of measure 1. Each new-born agent is endowed with 1 unit of a homogeneous good, which can be spent on either consumption or production. An agent born at time t lives with certainty at $t + 1$ and possibly also at

$t + 2$. The agents of the first type thus live for one period only, consume at time $t + 1$, and are labelled *early consumers*. Those of the second type live for two periods, consume at $t + 2$, and are identified as *late consumers*. An agent born at t learns his type (early or later consumer) privately at time $t + 1$. Agents become early or late consumers with constant probabilities ε and $1 - \varepsilon$ respectively. The population is so large that there is no uncertainty regarding the aggregate distribution between early and late consumers. A constant share ε of agents born at t consumes at $t + 1$, whereas the rest consume at $t + 2$. A new-born's expected utility is

$$(1) \quad W = \varepsilon U(c_1) + (1 - \varepsilon)U(c_2).$$

In (1), c_1 denotes the level of consumption of an early consumer, and c_2 the consumption of a later consumer. The utility function satisfies DIAMOND AND DYBVIK'S [1983] assumptions. It is strictly increasing and concave: $U'(\cdot) > 0$, $U''(\cdot) < 0$, $U'(0) = \infty$, $U'(\infty) = 0$. It is also assumed that $U(0) = -\infty$. The last assumption plays an important role in the analysis. In a bank panic, a few agents lose their savings and achieve utility $U(0) = -\infty$. Given the risk of losing a deposit, no agent will initially deposit his endowment in the bank. Only a panic-free banking system can improve depositors' expected utility.

At each time t , $t \geq 2$, there are four groups of agents: (1) new-borns of generation t (agents born at t), whose total measure is 1; (2) early consumers of generation $t - 1$ (measure ε); (3) late consumers of generation $t - 1$ (measure $1 - \varepsilon$); (4) later consumers of generation $t - 2$ (measure $1 - \varepsilon$). At each time point there are $3 - \varepsilon$ agents, and groups (2) and (4) consume. Since there is no uncertainty regarding the aggregate distribution between early and late consumers, the demographic structure of the population is constant through time.

Two production technologies are available. The first is risk-free long-term *production* with constant returns to scale. It requires at least one unit of investment at time t and produces $R > 1$ units at $t + 2$. The liquidation value of production is $0 < l < 1$ units. The second technology is *storage*. Consumption can be transferred from one period to the next without any loss or depreciation. Two assumptions are made.

ASSUMPTION 1 $\varepsilon l + (1 - \varepsilon)R < 1$.

This assumption states that in autarky (each agent produces his own consumption) an agent prefers storage to long-term production. Given the risk in liquidating early and the low liquidation value, the expected returns from production are lower than the returns from storage. Therefore, there is no production in autarky. The assumption is not critical, but it makes it easy to analyze the benefits of bank formation, because without banks the agents do not invest at all. If a bank can offer a larger consumption bundle to agents than $(1, 1) =$ (consumption of an early consumer, consumption of a late consumer), it is optimal to create a bank. If Assumption 1 is relaxed, it is possible that a bank can pay positive income to agents, but the income is lower than in autarky. Assumption 2 ensures that the liquidation value of production is so low that banks are vulnerable to panics.

ASSUMPTION 2 $l < 1 - \varepsilon$.

If Assumption 2 is not satisfied, a bank is panic-free even without time deposits or capital stock.

New production can be started at each time point. This generates a sequence of overlapping technologies in different stages of the production process: production in the start-up stage, production in the intermediate stage, and production that materializes.

As in DIAMOND [1997], it is assumed that a fraction $0 < \alpha < 1$ of agents are *active* and can contact each other and participate in capital markets. The rest of the agents, $1 - \alpha$, are *passive*, live in isolation (WALLACE [1988]), cannot contact each other, and so cannot participate in the capital markets. They can, however, contact a bank. The fractions of early and late consumers are assumed to be independent of the agents' ability to participate in capital markets.

As is standard for OLG models, the analysis focuses on steady-state allocations that yield identical ex ante returns for all current and future generations. For the dynamic transition to the steady state, see BHATTACHARYA, FULGHIERI, AND ROVELLI [1998] and QI [1994].

3 Bank with Panics

This section highlights liquidity provision through a bank. Since a fraction $1 - \alpha$ of agents are passive and unable to participate in capital markets, the bank system is their only option for investment and liquidity. Unfortunately, this service renders the bank vulnerable to panic. This section utilizes the analysis of QI [1994], BHATTACHARYA AND PADILLA [1996], and FULGHIERI AND ROVELLI [1998], but their model frameworks are reconstructed by assuming that banking entails operating costs. This enriches the analysis in two ways. First, it is possible to compare levels of liquidity provision in banks and capital markets. Second, it is possible to examine the costs of bank stabilization via time deposits and equity capital (section 4).³ Obviously, the assumption of no operating costs is a specific case of our model. To begin, the occurrence of a panic is clarified.

DEFINITION 1 As in DIAMOND AND DYBVIK [1983], a sunspot generates a panic. That is, mechanisms that cause agents' beliefs to change are not modelled explicitly. The sunspot appears at each time point. Then, an agent anticipates a bank panic and joins the panic if it is rational to do so. The anticipated panic then actually occurs. But if it is not rational for him to join the panic, he does not do so, and, because identical agents behave identically, the anticipated panic is avoided. During a panic, the existing depositors rush to the bank to withdraw their deposits, and the new-borns join the panic by staying away from the bank.

³ The positive costs of operation are realistic. MISHKIN [2007, p. 221] documents: "In recent years, interest paid on deposits (checkable and time) has accounted for around 25% of total bank operating expenses, while costs involved in servicing accounts have been approximately 50% of operating costs."

The definition accords with those of DIAMOND AND DYBVIK [1983] and QI [1994].⁴ The panics are unrelated to changes in the real economy, and they self-fulfil prophecies. Importantly, even when a panic is anticipated, it can be avoided if an agent (a late consumer or a new-born) refuses to join it. Such an agent will not join the panic if it is profitable not to. Nobody then joins the panic, and the anticipated panic is avoided. Suppose, for example, that a sunspot appears when deposits are protected by deposit insurance. A rational agent knows that he has no reason to join the panic even if everyone else does. The anticipated panic is then avoided.

DEFINITION 2 *A bank operates under maturity mismatch when the liquidation value of its deposits exceeds the liquidation value of its assets. Under maturity matching, the liquidation value of deposits is equal to (or less than) the liquidation value of assets.*

Under maturity mismatch, the bank is vulnerable to panic. Since the liquidation value of production does not cover the liquidation value of deposits, the bank cannot pay off deposits, and it fails. Given the assumption of first come, first served, the last withdrawers lose their deposits. It is clear that new-borns will not deposit their endowments in a bank during a panic. A new-born knows that if he puts his endowment in the bank, the bank will use it to pay off the deposits of withdrawing agents. But, in spite of the endowment, the bank will fail and the new-born will lose his endowment. Thus the panic is rational; all agents have an incentive to join in. Hence, under maturity mismatch, a panic occurs, the bank fails, and a few depositors obtain utility $u(0) = -\infty$. Therefore, no bank will be established under maturity mismatch. On the contrary, under maturity matching, the bank can always pay on its deposits.

Let us first determine the optimal consumption allocation with maturity mismatch. A bank is established by passive agents who save their endowment in it. It attracts only demand deposits, and it pays D_1 for withdrawals made after one period and D_2 for withdrawals made after two periods. The payments are chosen so that the expected utility of a depositor is maximized:⁵

$$\begin{aligned}
 (2) \quad & (i) \quad \varepsilon u(D_1) + (1 - \varepsilon)u(D_2) \\
 & \text{s.t.} \\
 & (ii) \quad \varepsilon D_1 + (1 - \varepsilon)D_2 = 1 - I + L + R(I - L) - (2 - \varepsilon)B, \\
 & (iii) \quad (D_1)^2 \leq D_2, \\
 & (iv) \quad 0 \leq I \leq 1, \\
 & (v) \quad L \leq II, \\
 & (vi) \quad \varepsilon \left(R - \sqrt{R} \right) < B < (R - 1)/(2 - \varepsilon).
 \end{aligned}$$

⁴ The one and only difference is that in this paper a sunspot is assumed to appear with certainty. The exact probability (100%) is needed only in section 6.2. Otherwise, one can assume that a sunspot appears with some positive probability.

⁵ Here the bank is modelled as if its size were 1 instead of $1 - \alpha$. That is, the multiplications by $1 - \alpha$ are dropped.

In (2) the expected utility (i) is maximized subject to the resource constraint (ii). The left-hand side of (ii) represents consumption at t . Both the early consumers of generation $t - 1$ and the late consumers of generation $t - 2$ consume. The right-hand side gives the resources. The first term is the endowment, 1, from which the production investment, I , is subtracted. The resources also include returns from the liquidated production, L , production output, $R(I - L)$, and the costs of banking per size unit, B . In addition, (iii) is needed to eliminate arbitrage by depositors; a late consumer will not mimic an early consumer by withdrawing D_1 at $t + 1$ and redepositing it in another bank for a period. This strategy would yield $(D_1)^2$, and it cannot be more profitable than retaining the savings in the original bank. Since the agents are risk-averse, (iii) is binding. As regards (iv), $I \leq 1$ means that the production investment cannot exceed the endowment. Additionally, in (v), the returns from liquidation, L , cannot exceed the liquidation value of total production, II . It is easy to see that the right-hand side of (ii) is maximized when $L = 0$, so that no production is liquidated. Constraint (vi) determines the operating costs of banking. Since $B < (R - 1)/(2 - \varepsilon)$, the operating costs are so low that banking is profitable. The second part, $B > \varepsilon(R - \sqrt{R})$, sets a lower limit on the operating costs. The optimal allocation is solved in Appendix A.1.⁶

PROPOSITION 1 *The bank's optimal allocation satisfies $1 < D_1^* < D_2^* < R$, $D_1^* = \sqrt{D_2^*}$, and $I = 1$.*

Therefore, the bank can offer agents both a positive return and liquidity.

Unfortunately, the bank is under threat of panic. The late consumers of generation $t - 2$ withdraw $(1 - \varepsilon)D_2$, and the early consumers of the next generation withdraw εD_1 , even without a panic, since their true consumption time point is the present. During a panic, the late consumers of generation $t - 1$ mimic early consumers and withdraw $(1 - \varepsilon)D_1$. The new-borns also panic and do not save their endowments in the bank. Thus, the resources consist of materializing production, R , and, from the liquidated intermediate production, $II = l$. The resources cover the withdrawals if

$$R - B(2 - \varepsilon) - (1 - \varepsilon)D_2 - \varepsilon D_1 - (1 - \varepsilon)D_1 + l \geq 0.$$

The sum of the first four terms is zero owing to the resource constraint. Given Assumption 2, $1 - \varepsilon > l$. The liquidation value of the intermediate production is so low that it does not cover payments on deposits, and the bank fails due to maturity mismatch, $-(1 - \varepsilon)D_1 + l < 0$. Therefore, if an agent anticipates a panic,

⁶ It is easy to see from (ii) that the supply of liquidity is decreasing in operating costs. Without operating costs, the bank could supply $D_1 = \sqrt{R}$, $D_2 = R$, and the maximization problem would have to be supplemented with a constraint, $D_2 \leq R$, so that the bank would have no incentive to invest in other banks (see BHATTACHARYA AND PADILLA [1996, p. 1010]). Hence, this constraint is unnecessary when the operating costs are sufficiently high. Obviously, if the operating costs are too high, banking is unprofitable (see (vi)).

he rationally joins it and the panic actually occurs. Since the last withdrawers do not get anything, their utility is $u(0) = -\infty$. As a result, a rational agent will not immediately deposit his endowment in the bank, but will instead store it. No bank is established, although a panic-free bank system could raise the expected returns of the agents and provide the desired liquidity. The role of the banking system is theoretical if panics cannot be prevented.

4 *Panic-Free Banking System with Time Deposits*

This section discusses how panic can be prevented if the bank attracts not only liquid demand deposits but also time deposits, so that maturity mismatch is eliminated. Passive agents favour demand deposits, whereas active agents save via marketable time deposits. It is assumed that there are sufficiently many active agents to save in time deposits. The assumption is dropped in section 6.

To proceed, we investigate liquidity production via stock markets. Then a bank with time deposits is introduced. To find an equilibrium, we need to specify the following elements: (1) the characteristics of time deposits (active agents are willing to invest in time deposits only if they are at least as productive as company shares); (2) the agents' budget constraints need to be satisfied; (3) the demand and supply of time deposits must be in equilibrium (most of all, the supply of intermediate company stocks and time deposits must be equal to the demand, so that it is possible to achieve liquidity by selling these assets without interrupting the underlying production); and (4) the volume of time deposits must be at the optimal level, so that panics are avoided.

4.1 *Capital Markets*

BHATTACHARYA AND PADILLA [1996] and FULGHIERI AND ROVELLI [1998] characterize the optimal steady-state consumption and investment allocations that can be achieved by active agents, who are able to contact each other and trade firm stocks. The stock markets operate as follows. At each time point t there are new-born agents who are endowed with a unit of consumption good. Each new-born sets up a firm, which invests a fraction of the endowment, $I \leq 1$, in long-term production. A part of the firm's shares is retained by the new-born, who sells the remaining shares to other agents. The sales revenue and the rest of the endowment, $1 - I$, are invested by him in the stocks of firms set up by other agents. There are three age groups of firms at each time point: new firms that have started production at t , intermediate firms that started production at $t - 1$, and old firms that started production at $t - 2$. The production of old firms (R) materializes and is paid out as dividends. These firms are then closed down. If an agent encounters a consumption shock, he can obtain liquidity by selling his shares, so that liquidation of the underlying production is avoided. The following result is achieved.

PROPOSITION 2 ⁷ *In the stock markets, the steady-state competitive equilibrium satisfies*

$$P_n = 1, \quad P_i = \sqrt{R}, \quad C_1 = \sqrt{R}, \quad C_2 = R, \quad 0 < I^* = 1 - \frac{\varepsilon(R - \sqrt{R})}{R - 1} < 1.$$

Here $P_n = 1$ ($P_i = \sqrt{R}$) denotes the price of a new (intermediate) firm's share of stock, $C_1 = \sqrt{R}$ ($C_2 = R$) represents the consumption of an early (late) consumer, and I^* is the investment in production. Now $I^* < 1$, since not all of the endowment is invested in new production; instead, $1 - I^*$ is invested in intermediate stocks. Given $P_n = 1$, $P_i = \sqrt{R}$, and the value of an old firm, R , a firm's share yields a fixed return, \sqrt{R} , in every period.

Note that if the operating costs of the bank are zero, banks and capital markets yield equal returns. With positive operating costs, capital markets provide better returns.

The idea of Proposition 2 is easy to see. To begin, arbitrage is avoided only if $P_n = 1$, $P_i = \sqrt{R}$. Inserting these into (4) and (5) below, with no time deposits (every β in (4) and (5) is zero), yields $C_1 = \sqrt{R}$, $C_2 = R$. Substituting these into the economy's resource constraint, $\varepsilon C_1 + (1 - \varepsilon)C_2 = 1 - I + IR$, produces I^* .

4.2 Bank with Subordinated Time Deposits

4.2.1 Characteristics of Time Deposits

A time deposit lasts for two periods. It cannot be interrupted after the first period, but can then be resold. The bank attracts $A/2$ in new time deposits at each time point, and so the volume of outstanding time deposits is equal to A . How much does the bank pay on time deposits? The bank is established by passive agents, and it maximizes their expected utility, thereby minimizing payments on time deposits. To attract active agents, time deposits need to be at least as productive as company stocks. Suppose that the bank pays interest R on a time deposit at maturity, that is, after two periods. Let P_i^{TD} denote the resale price of an intermediate time deposit. Arbitrage is eliminated if

$$(3) \quad \frac{P_i^{TD}}{1} = \frac{R}{P_i^{TD}} = \sqrt{R}.$$

In (3) the first (second) term denotes return on a time deposit in the first (second) period, and the third term gives the one-period return on a firm's share. To avoid arbitrage, it must be that $P_i^{TD} = \sqrt{R}$. Thus, at maturity, the bank pays interest R on a time deposit, which cannot be interrupted before maturity, but which can be resold at the market price \sqrt{R} at the intermediate stage. Both time deposits and firm shares yield the same one-period return, \sqrt{R} .

The bank could pay a different return, say $r \neq R$, on a time deposit at maturity. Then, owing to perfect competition, the intermediate price of a time deposit is r/\sqrt{R} , and the price (size) of the new time deposit is r/R . Since the bank hopes that the

⁷ FULGHIERI AND ROVELLI [1998].

total volume of time deposits is A at every time point, it needs to make $AR/2r$ new time deposit contracts in every period. The bank pays r on each maturing time deposit. In every period the payments total $r \times AR/2r$, or $AR/2$. The costs of time deposits are the same as under the original contract in which the bank pays R on a time deposit at maturity. Because the promised payment at maturity has no effect on the total costs of time deposits, we can simply assume that the bank pays R units on a time deposit at maturity.⁸

4.2.2 Budget Constraints and Market Equilibrium

We begin with the following lemma.

LEMMA 1 *In equilibrium, time deposits and bank stocks yield the same return, \sqrt{R} , in each period. Therefore, each combination of these yields the very same return to an active agent: an early consumer obtains \sqrt{R} units, and a late consumer receives R units.*

The first part of Lemma 1 was proved above, and the second part is proved below.

PROOF Consider a new-born active agent. Let β_n^0 (β_i^0) denote his savings in new (intermediate) time deposits, and θ_n^0 (θ_i^0) his savings in the stocks of new (intermediate) firms. Then

$$(4) \quad \begin{aligned} (i) \quad & \beta_n^0 + \beta_i^0 P_i^{TD} + \theta_n^0 + \theta_i^0 P_i = 1, \\ (ii) \quad & \beta_n^0 P_i^{TD} + \beta_i^0 R + \theta_n^0 P_i + \theta_i^0 R = \sqrt{R}. \end{aligned}$$

The left-hand side gives the allocation and the right-hand side the total value of the portfolio. Here (i) is the value of the initial portfolio, and (ii) is the value after a period. If an agent becomes a late consumer, he can reinvest his wealth. Let β_n^1 (β_i^1) denote his savings in new (intermediate) time deposits. Then θ_n^1 (θ_i^1) is his savings in the stocks of new (intermediate) firms, and we have

$$(5) \quad \begin{aligned} (iii) \quad & \beta_n^1 + \beta_i^1 P_i^{TD} + \theta_n^1 + \theta_i^1 P_i = \sqrt{R}, \\ (iv) \quad & \beta_n^1 P_i^{TD} + \beta_i^1 R + \theta_n^1 P_i + \theta_i^1 R = R. \end{aligned}$$

Here (iii) is the initial value of the reallocated portfolio, and (iv) the value of the reallocated portfolio after a period. Thus a late consumer obtains R units. *Q.E.D.*

Both time deposits and firm shares are risk-free and yield a return of \sqrt{R} in each period. Therefore, each combination of deposits and shares yields the very same return, \sqrt{R} , in every period, with certainty. An early consumer obtains \sqrt{R} consumption units, and a late consumer R units.

Both the stock and time-deposit markets must clear. Given Lemma 1, the equilibrium holdings of these assets are not unique. We focus on active agents and define the following symbols: $\bar{\beta}_n^0$ ($\bar{\beta}_i^0$) = average amount of new (intermediate) time deposits purchased by new-borns, $\bar{\beta}_n^1$ ($\bar{\beta}_i^1$) = average amount of new (intermediate)

⁸ We thank the anonymous referee who encouraged us to investigate this subject.

time deposits bought by intermediate agents, $\bar{\theta}_n^0$ ($\bar{\theta}_i^0$) = average amount of new (intermediate) firm stocks purchased by new-borns, and $\bar{\theta}_n^1$ ($\bar{\theta}_i^1$) = average amount of new (intermediate) firm stocks purchased by intermediate agents. In equilibrium, the demand and supply are equal for time deposits and firm stocks:

$$(6) \quad \begin{aligned} & \text{(i)} \quad \alpha \bar{\beta}_n^0 + \alpha(1 - \varepsilon) \bar{\beta}_n^1 = \frac{1}{2} A, & \text{(ii)} \quad \alpha \bar{\beta}_i^0 + \alpha(1 - \varepsilon) \bar{\beta}_i^1 = \frac{1}{2} A, \\ & \text{(iii)} \quad \alpha \bar{\theta}_n^0 + \alpha(1 - \varepsilon) \bar{\theta}_n^1 = (\alpha - \frac{1}{2} A) I^*, & \text{(iv)} \quad \alpha \bar{\theta}_i^0 + \alpha(1 - \varepsilon) \bar{\theta}_i^1 = (\alpha - \frac{1}{2} A) I^*. \end{aligned}$$

Here (i) ((ii)) states that total savings in new (intermediate) time deposits are equal to the supply of new (intermediate) time deposits. Likewise, (iii) ((iv)) indicates that total savings in new (intermediate) firm stocks are equal to the supply of new (intermediate) stocks. The supply of stocks is flexible; the larger the supply of time deposits, the less can be invested directly in stocks of the firms, $(\alpha - A/2)I^*$. The term is positive because it is assumed that there are sufficiently many active agents in the economy to fulfil the supply of time deposits.

Now (i) and (iii) in (6) ensure demand–supply equilibrium for new time deposits and firm stocks. Let us investigate the markets for intermediate time deposits at time t . Time deposits are supplied by agents who encounter a consumption shock: early customers of generation $t - 1$, $\varepsilon \alpha \bar{\beta}_n^0$, and late consumers of generation $t - 2$, $(1 - \varepsilon) \alpha \bar{\beta}_n^1$. Given (i), the supply is $A/2 - (1 - \varepsilon) \alpha \bar{\beta}_n^0$. The demand consists of new-borns of generation t , $\alpha \bar{\beta}_i^0$, and the additional demand from late consumers of generation $t - 1$, $(1 - \varepsilon) \alpha (\bar{\beta}_i^1 - \bar{\beta}_n^0)$. Given (ii), the demand totals $A/2 - (1 - \varepsilon) \alpha \bar{\beta}_n^0$. Hence, the demand is equal to the supply. In the same way, one can show that the intermediate market for firm stocks clears. Therefore, the demand for liquidity can be satisfied by selling intermediate time deposits and firm shocks without needing to liquidate prematurely the underlying physical investment.

To sum up, the agents' equilibrium holdings of time deposits and firm stocks are not unique. Yet, the price mechanism ensures that there is a market-clearing solution, so that the markets for time deposits and firm stocks clear, and agents' budget constraints are satisfied. Then, both time deposits and firm stocks yield \sqrt{R} in each period. As a result, independently of his portfolio allocation decisions, an active early consumer can consume \sqrt{R} units, and an active late consumer R units. The amount of time deposits must still be determined.

4.2.3 Optimal Amount of Time Deposits

This subsection shows that the optimal amount of time deposits is the minimum amount needed to satisfy the maturity matching constraint. To begin, we have

LEMMA 2 *The bank minimizes the amount of time deposits.*

PROOF With time deposits, the bank's resource constraint is

$$(7) \quad R \left[(1 - \alpha) + \frac{1}{2} A \right] - B \left[(1 - \alpha)(2 - \varepsilon) + A \right] = \frac{1}{2} RA + (1 - \alpha) \times \left[\varepsilon \sqrt{D_2} + (1 - \varepsilon) D_2 \right].$$

The first term on the left-hand side represents production output. At each time point the bank attracts $1 - \alpha$ in new demand deposits and $A/2$ in new time deposits, and invests the funds in production. The second term gives the costs of banking, which are dependent of the amount of deposits. Passive new-borns, $1 - \alpha$, and passive late consumers of the previous generation, $(1 - \alpha)(1 - \varepsilon)$, save in demand deposits. Given time deposits A , the total amount of deposits adds up to $(1 - \alpha)(2 - \varepsilon) + A$. The right-hand side in (7) represents the payments on maturing time deposits and the payments on demand deposits. Some manipulation yields

$$R - B \left[(2 - \varepsilon) + \frac{A}{1 - \alpha} \right] = \varepsilon \sqrt{D_2} + (1 - \varepsilon) D_2.$$

This reveals that $dD_2/dA < 0$; time deposits reduce payments on demand deposits. It is optimal to minimize the amount of time deposits. *Q.E.D.*

Intuitively, each size unit, including both demand and time deposits, entails a cost of B to the bank. Active agents require the same return on time deposits as they could obtain by investing directly in firm stocks. Thus the costs of time deposits, B , are borne entirely by the depositors who save in demand deposits. Obviously, they will minimize the amount of time deposits.

The minimum amount of time deposits is determined by maturity matching, since without it the bank is vulnerable to panics. Thus the liquidation value of assets must be at least equal to the liquidation value of deposits. Now the assets consist of maturing long-term production, $R[(1 - \alpha) + A/2]$, and liquidated intermediate production, $l[(1 - \alpha) + A/2]$. The value of deposits is more complex. The late consumers of generation $t - 2$ as well as the early consumers of generation $t - 1$ withdraw their deposits with certainty. During a panic, new-borns do not save their endowments in the bank. Moreover, the late consumers of generation $t - 1$ panic and mimic early consumers by trying to withdraw their deposits. Only intermediate time deposits cannot be withdrawn. The maturity matching constraint is satisfied if

$$R \left[(1 - \alpha) + \frac{1}{2} A \right] - B[(1 - \alpha)(2 - \varepsilon) + A] - \frac{1}{2} RA \\ - (1 - \alpha) \left[\varepsilon \sqrt{D_2} + (1 - \varepsilon) D_2 \right] - (1 - \alpha)(1 - \varepsilon) \sqrt{D_2} + l \left[(1 - \alpha) + \frac{1}{2} A \right] \geq 0.$$

The first four terms together constitute a resource constraint (7) that sums to zero. The maturity matching constraint simplifies to $(1 - \alpha)(1 - \varepsilon) \sqrt{D_2} = l[(1 - \alpha) + A^*/2]$, from which the minimum amount of time deposits, A^* (which is also the optimal amount of time deposits), can be solved for:

$$A^* = \frac{2(1 - \alpha) \left[(1 - \varepsilon) \sqrt{D_2} - l \right]}{l}.$$

The stabilizing effect of time deposits is simple. Since time deposits boost the liquidation value of the bank but cannot be interrupted, they help create maturity matching. Without time deposits the value of demand deposits exceeds the liquidation value of intermediate production, $(1 - \alpha)(1 - \varepsilon) \sqrt{D_2} > (1 - \alpha)l$, which makes the bank vulnerable to panic. However, there are time deposits amounting to $A^*/2$,

so that the value of demand deposits is equal to the increased liquidation value of intermediate production, $(1 - \alpha)(1 - \varepsilon)\sqrt{D_2} = l(1 - \alpha) + lA^*/2$. Maturity matching is satisfied, and the bank can pay back deposits if a panic occurs.

4.2.4 Special Rules

This subsection confirms that a panic-free banking system can be achieved only if the maturity-matching constraint is supported with special rules.

First, it is shown that time deposits need to be subordinated to demand deposits.⁹ At the next time point after a panic, the bank cannot settle the promised payments on deposits in full, because some production has been liquidated during the panic, and the materializing value of the production is low. Suppose that in this case the bank adopts a *fair-sharing rule*. Each depositor receives an equal share, f , of the promised payment if the bank cannot settle the payments in full. That is, the bank pays fD_2 units on long-term demand deposits and fR units on long-term time deposits, $f < 1$. If there is no panic, the bank can pay the promised payments in all, $f = 1$.

Suppose that the maturity-matching requirement is satisfied, but a panic occurs. Only intermediate time deposits, which cannot be interrupted, are retained in the bank. Given maturity matching, the value of bank assets, i.e., intermediate production, erodes to zero. Thus, at the next time point, the bank has no materializing production and cannot pay anything on maturing time deposits, so it fails.

Is panic withdrawal of demand deposits rational? Suppose that one late consumer opts to sit out the panic and does not withdraw D_1 . The bank then has D_1 units of intermediate production. At the next time point, the production materializes, yielding D_1R units. Under the fair-sharing rule, D_1R units are shared equally among the agents, that is, between the late consumer with demand deposits and the agents who have maturing time deposits. As a result, the agents with time deposits can share D_1R units, whereas the late consumer with demand deposits does not obtain anything (the measure of his demand deposits is zero, and the total measure of the time deposits is positive; hence, the whole return is paid on time deposits). By joining the panic, the late consumer can obtain D_1 . Thus panic withdrawal of demand deposits is rational.

Consequently, maturity matching is a necessary but not a sufficient condition for preventing panics. It guarantees that a late consumer, who saves in demand deposits, can obtain the early consumer's allotment by withdrawing immediately. Unfortunately, it does not guarantee that the late consumer obtains a greater return by waiting for his true consumption time than by withdrawing the allotment of the early consumer at once.

Consider an identical case, except that time deposits are subordinate to demand deposits. One late consumer does not join the panic by withdrawing D_1 . The bank has D_1 units of intermediate production. At the next time point, the production

⁹ More precisely, we show that panics are avoided at least when time deposits are subordinate to demand deposits.

materializes, yielding D_1R units. Since $D_1R > D_2$, and since demand deposits are senior to time deposits, the late consumer receives the promised return, D_2 . Thus he rationally opts to sit out the panic. Since each late consumer acts in the same manner, each waits for the next time point, and thus demand deposits are retained in the bank. No intermediate production is interrupted, and it materializes at the next time point, yielding $R[(1 - \alpha) + A/2]$, which covers the operating costs as well as the payments on long-term demand deposits, D_2 , and time deposits, R (recall the resource constraint (7)). Something is even left over for the early consumers of the next generation. Thus late consumers have no reason to panic, and no intermediate production is liquidated. A conclusion follows.

LEMMA 3 *When time deposits are subordinate to demand deposits, demand deposits are risk-free and late consumers have no reason to panic by mimicking early consumers and withdrawing their deposits. Since the panic of existing depositors is avoided, no intermediate production is liquidated. Thus the bank can pay the promised returns to short-term demand deposits, D_1 , long-term demand deposits, D_2 , and maturing time deposits, $AR/2$.*

The payments on deposits are independent of whether or not the following generations will save their endowments in the bank.

(a) Consider first the payments to early consumers of generation $t - 1$ at time point t . The payments, $(1 - \alpha)\varepsilon D_1$, are based on production investment at $t - 2$. The value of the maturing production does not depend on whether the next generation (t) saves its endowments in the bank at time point t .

(b) Consider now the payments to late consumers of generation $t - 1$ at $t + 1$. The payments are based on production investment at $t - 1$. The decision of the newest generation ($t + 1$) as to whether to save in the bank has no effect on payments on maturing time deposits and long-term demand deposits. Neither has the decision of the previous generation (t) any effect on payments. To see this, two cases need to be examined. First, if generation t saves its endowments in the bank at t , the bank continues to operate normally and can pay R on time deposits and D_2 on long-term demand deposits at $t + 1$. Second, if generation t panics and does not save its endowments in the bank at t , the returns of late consumers do not change. The production output is sufficient to pay R on time deposits and D_2 on long-term demand deposits. Since a part of the output, $(1 - \alpha)\varepsilon D_1$, is reserved for early consumers of generation t , which did not even save in the bank, the bank has extra returns. Thus late consumers can be sure that their deposits are safe.

Therefore, panic withdrawals of existing deposits are avoided. It must still be shown that the new generation (t) will not panic and stay away from the bank. It will be shown that the new-borns' deposits will be safe at time points t , $t + 1$, and $t + 2$.

First, it is shown that a new-born cannot lose anything at time point t . To ensure this, we make the following assumption.

ASSUMPTION 3 *To avoid a maturity mismatch, the bank promises to pay back the deposits of new-borns at once if it cannot attract sufficient shares of both deposit types.*

The promise protects new-borns of generation t . And it has been shown that the agents of the older generation do not panic (Lemma 3). Thus the new-borns cannot lose anything at time point t . At time points $t + 1$ and $t + 2$, the bank can pay the promised returns (Lemma 3). In sum, under Lemma 3, the savings of generation t will be safe at time points t , $t + 1$, and $t + 2$. Consequently, the new-borns of each generation will save their endowments in the bank, and no panic occurs.

Since time deposits are now subordinate to demand deposits, the panic-preventing amount of time deposits changes:

$$(8) \quad A^* = \frac{2(1 - \alpha) [(1 - \varepsilon)\sqrt{D_2} - l]}{R + l}.$$

The analysis of section 4 can be summarized as follows.

PROPOSITION 3 *Panics can be prevented via maturity matching. The bank attracts not only demand deposits but also uninterruptible time deposits, which can be resold in secondary markets. Passive agents obtain liquidity by saving in demand deposits, and active agents by saving in marketable time deposits. Since time deposits entail excessive costs for the bank, it minimizes the volume of time deposits so that the maturity-matching constraint is just barely satisfied. Due to the operating costs of banking, the bank must offer a lower return to agents than they could obtain by investing directly in firms.¹⁰*

Importantly, when $B > 0$, it is possible that the panic-free solution is unachievable. When B is relatively large, the burden of operating costs may be so large that the bank is unprofitable. We give a numerical example below.

4.2.5 Numerical Example I

Assume the following economy: $R = 1.15$, $B = 0.025$, $\varepsilon = 0.3$, $\alpha = 0.35$, $l = 0.64$. The agents who can participate in capital markets obtain $R = 1.15$. Without time deposits, a bank can offer $D_2 = 1.127$, but the bank is vulnerable to panics, since $1 - \varepsilon > l$. With time deposits, the bank can offer $D_2 = 1.124$. The amount of time deposits is 0.074, and the amount of demand deposits is $2 - \varepsilon = 1.7$. The ratio of time deposits to total deposits is 4.2%.

It is clear that capital markets can supply greater returns than the bank, since banking entails costs. Moreover, the bank can pay more on demand deposits without stabilizing time deposits. Thus the bank regulator can raise the expected utility of passive agents if it can offer deposit insurance at no cost. Then the bank can avoid panics without the burden of time deposits and can pay more on demand deposits.

¹⁰ If the operating costs of banking are zero ($B = 0$), a panic-free bank can offer the same optimal allocation (\sqrt{R}, R) to passive agents as active agents can obtain by investing directly in firms. In this case the stabilization with time deposits is costless.

In this example the differences in payments on demand deposits are small, since the necessary amount of time deposits is modest and the operating costs are low.

Assume the same economy, except that $B = 0.08$, $l = 0.35$. It is possible to show that now a bank can pay positive interest on deposits, if it does not attract any time deposits ($D_2 \approx 1.016$). But the bank is now vulnerable to panic. To avoid panic, the bank needs to issue time deposits. This, however, increases the costs of banking so much that the value of bank resources is less than 1 and thus $D_2 < 1$; banking is unprofitable.

5 Panic-Free Banking with Capital Stock

This section extends the analysis by exploring whether capital stock can be adopted to eliminate panics. How does capital stock differ from time deposits in this context, if there is any difference? We proceed as follows. First, bank stocks are introduced, and then the optimal amount of capital stock is determined. The optimal amount turns out to be equal to the panic-preventing amount of time deposits. Thereafter, the agents' budget constraints and the equilibrium constraints for the markets of firm stocks and bank stocks are presented. Finally, we show that both markets clear.

5.1 Firm E

The operation of capital markets was reviewed in section 4. Each active agent set up a firm of his own and sold a part of its stock to other active agents. The sales revenue was invested in the stocks of the other firms. Suppose now that the agents coordinate their actions and merge several small firms of unit size. The size of the big firm, Firm E, is E . It has $E/2$ units of production started at $t - 2$ and $E/2$ units of production started at $t - 1$. The process continues so that at each time point old production materializes, yielding $ER/2$. The firm reinvests a part of it, $E/2$, in new production and pays out the remainder, $E(R - 1)/2$, as dividends.

5.2 The Bank

Passive agents establish a bank, which makes the following suggestion to active agents. Instead of setting up Firm E, active agents should invest the same amount, E , in capital stock in the bank, which commits to pay a fixed dividend $E(R - 1)/2$ at each time point. Since capital stock is naturally subordinate to demand deposits, the order of moves is the following at each time point: (1) the bank pays back the demand deposits, (2) the bank pays out dividends, (3) stockholders can trade stocks.

At each time point, after dividends are paid, the market price of a unit of stock is

$$P_S = \frac{\delta \frac{1}{2}(R - 1)}{1 - \delta} = \frac{1}{2}(1 + \sqrt{R}), \quad \delta = 1/\sqrt{R}.$$

It is easy to show that active agents are willing to invest in bank stocks.

LEMMA 4 *Bank stocks and firm stocks yield an equal return, \sqrt{R} , R .*

PROOF Consider an active new-born who invests in bank stocks. Given the endowment, he can purchase $1/P_S$ units of bank stocks. After the first period, the agent's wealth is

$$1 + \frac{\frac{1}{2}(R-1)}{P_S} = \sqrt{R},$$

where the first term gives the value of $1/P_S$ units of bank stock, and the second term represents dividend payment on a unit. Thus, an early consumer can consume \sqrt{R} . If the agent becomes a late consumer, he invests the dividend income in additional bank stocks. After the additional acquisition, he holds \sqrt{R}/P_S units of bank stock. At the end of the period his wealth adds up to

$$P_S \frac{\sqrt{R}}{P_S} + \frac{\sqrt{R} \frac{1}{2}(R-1)}{P_S},$$

which is equal to R and can be consumed by the late consumer. *Q.E.D.*

Until now we have investigated the characteristics of the bank stocks. We must still determine the optimal amount of capital stock. The analysis is quite similar to the analysis of time deposits in the previous section. We begin with the following lemma.

LEMMA 5 *The bank minimizes the amount of capital stock.*

PROOF The resource constraint of the bank is

$$(9) \quad R[(1-\alpha) + \frac{1}{2}E] - B[(1-\alpha)(2-\varepsilon) + E] = \frac{1}{2}ER + (1-\alpha) \times [\varepsilon\sqrt{D_2} + (1-\varepsilon)D_2].$$

The first term represents output. At each time point, 50% of capital stock is tied to intermediate production, and the remainder is invested in new production. The second term gives the operating costs. The first term on the right-hand side gives the costs of capital stock. The costs total $RE/2$, consisting of dividends, $E(R-1)/2$, and investments in new production, $E/2$. The last term gives the payments on demand deposits. Some manipulation yields

$$R - B \left[(2-\varepsilon) + \frac{E}{1-\alpha} \right] = \varepsilon\sqrt{D_2} + (1-\varepsilon)D_2.$$

It is clear that $dD_2/dE < 0$; the greater the amount of capital stock, the smaller the payments on demand deposits. The amount of capital stock is minimized. *Q.E.D.*

The intuition is the same as for time deposits. Each size unit entails operating costs. Yet, active agents require the very same return on bank stocks as they could obtain by investing directly in firms. Therefore, the operating costs are borne entirely by those depositors who save with demand deposits.

Thus it is necessary to solve for the minimum amount of time deposits. This can be done via the maturity matching constraint. The liquidation value of bank assets

covers the liquidation value of bank deposits if

$$R\left[(1 - \alpha) + \frac{1}{2}E\right] - B[(1 - \alpha)(2 - \varepsilon) + E] - (1 - \alpha)\left[\varepsilon\sqrt{D_2} + (1 - \varepsilon)D_2\right] - (1 - \alpha)(1 - \varepsilon)\sqrt{D_2} + l\left[(1 - \alpha) + \frac{1}{2}E\right] \geq 0.$$

Dividends are excluded, since they are paid out only after the withdrawal of deposits. Using this and (9), the panic-preventing amount of capital stock can be calculated:

$$E^* = \frac{2(1 - \alpha)\left[(1 - \varepsilon)\sqrt{D_2} - l\right]}{R + l}.$$

This is equal to the panic-preventing amount of time deposits, (8). The intuition for the following lemma is based on the operating costs of banking.

LEMMA 6 *Demand deposits are less productive than bank stocks.*

Now we know the amount of capital stock as well as the characteristics of the stocks. Next we focus on agents' budget constraints and portfolio allocation decisions. Given Lemma 5, active agents dislike demand deposits. Since bank stocks and firm stocks yield the same return in each period, \sqrt{R} , the agents are indifferent between them. Hence there is no unique optimal allocation decision. Yet each agent's budget constraint must be satisfied:

$$\begin{aligned} \text{(i)} \quad & \rho^0 P_S + \theta_n^0 P_n + \theta_i^0 P_i = 1, & \text{(ii)} \quad & \rho^0 \sqrt{R} + \theta_n^0 P_i + \theta_i^0 R = \sqrt{R}, \\ \text{(iii)} \quad & \rho^1 P_S + \theta_n^1 P_n + \theta_i^1 P_i = \sqrt{R}, & \text{(iv)} \quad & \rho^1 \sqrt{R} + \theta_n^1 P_i + \theta_i^1 R = R. \end{aligned}$$

Here ρ^0 (ρ^1) denotes a new-born's (late consumer's) investment in bank stocks. Recall that θ_n^0 (θ_i^0) represents a new-born's investment in stocks of new (intermediate) firms, whereas θ_n^1 (θ_i^1) is a late consumer's investment in the stocks of new (intermediate) firms. In (i)–(iv), the left-hand side gives the allocation of funds, and the right-hand side the value of the total portfolio. Independently of portfolio allocation decisions, each new-born's portfolio, (i), has the initial value 1, and after a period the value of the portfolio, (ii), is \sqrt{R} . If an agent becomes an early consumer, he can consume this. If he becomes a late consumer, he can reallocate his funds, (iii), wait for a period, and then enjoy the final portfolio, (iv), by consuming R units. Both time deposits and firm stocks are risk-free and yield \sqrt{R} in each period. Therefore, each combination of them yields the very same income with certainty. An early consumer obtains \sqrt{R} consumption units, and a late consumer gets R units.

Although each can freely allocate his funds, the demand and supply of firm stocks and bank stocks must clear. We define the following symbols: $\bar{\rho}^0$ ($\bar{\rho}^1$) = average number of units of bank stocks purchased by new-borns (late consumers), $\bar{\theta}_n^0$ ($\bar{\theta}_i^0$) = average number of new (intermediate) units of firm stocks purchased by new-borns, and $\bar{\theta}_n^1$ ($\bar{\theta}_i^1$) = average number of new (intermediate) units of firm stocks purchased by late consumers. The markets are in equilibrium if the demand and supply are equal for bank and firm stocks. Three conditions need to be satisfied:

$$\begin{aligned} \text{(10)} \quad \text{(i)} \quad & \alpha \bar{\rho}^0 + \alpha(1 - \varepsilon) \bar{\rho}^1 = E, & \text{(ii)} \quad & \alpha \bar{\theta}_n^0 + \alpha(1 - \varepsilon) \bar{\theta}_n^1 = (\alpha - E)I^*, \\ \text{(iii)} \quad & \alpha \bar{\theta}_i^0 + \alpha(1 - \varepsilon) \bar{\theta}_i^1 = (\alpha - E)I^*. \end{aligned}$$

In (10), (i) indicates that savings in bank stocks are equal to the supply. Moreover, (ii) ((iii)) states that total savings in new (intermediate) firm stocks equal their supply. Again, the number of established firms and the amount of their stocks vary according to the amount of bank stocks.

We show in Appendix A.2 that the markets for bank stocks and firm stocks clear. Therefore, the demand for liquidity can be satisfied by selling bank stocks and firm stocks without the need to liquidate prematurely the underlying physical investment.

In sum, we have seen that a bank can eliminate panics by issuing marketable shares, which yield the same return, \sqrt{R} , R , as direct investment in firm stocks. The capital stock stabilizes the bank system by increasing the liquidation value of the bank, so that the maturity matching constraint is satisfied. As a result, depositors know that the bank is always able to pay off the deposits, so that they would not have to bear the losses in a panic. Therefore, depositors have no reason to panic, and panics are avoided. A conclusion follows.

PROPOSITION 4 Capital stock helps to prevent bank panics efficiently. The needed amount of capital stock is equal to the panic-preventing amount of time deposits.

Just as in the context of time deposits, capital stock helps eliminate panics if the bank's operating costs and the required capital stock are sufficiently low. For instance, if the operating costs are zero, panics are avoided, and the stabilizing effect of the capital stock does not entail any costs. However, if the operating costs and the minimum capital stock are sufficiently high, the stabilizing effect of capital stock may be so expensive that banking is unprofitable.

6 Extensions

Up to now, it has been assumed that there are sufficiently many active agents prepared to save with time deposits or bank stocks. The opposite scenario is explored in section 6.1. Additionally, it is commonly argued that banks abuse their position as financial intermediaries by gambling with depositors' funds (e.g., MERTON [1977], NIINIMÄKI [2009]). Liquidity of deposits is needed to protect depositors from this type of moral hazard. When depositors observe moral hazard, they can immediately salvage their deposits by withdrawing them from the bank (CALOMIRIS AND KAHN [1991]). This poses an interesting question. Is the bank system subject to moral hazard if it is stabilized with time deposits that cannot be immediately withdrawn? This point is investigated in section 6.2.

6.1 Too Few Active Agents

6.1.1 Model

Initially there are no active agents, because the agents live in isolation, as in WALLACE [1988], and cannot contact each or participate in capital markets. The

concept of isolation, however, is not taken as seriously as in WALLACE [1988]. It is fairly realistic to assume that isolation is imperfect. Everyone can contact other agents and participate in capital markets, but this entails costs. To keep the analysis simple, it is assumed that the cost of market participation or the cost of breaking through the isolation erodes $1 - m$ percent of the asset return in each period. The cost is, however, at least $1 - m$ units in each period. As before, suppose that the initial value of a firm stock is one unit and that its value is \sqrt{R} after one period. Yet, after deducting the costs of market participation, the value of agent's wealth is $\sqrt{R}m, 0 < m < 1$. If he invests this in production for the second period, the value is $\sqrt{R}m \times \sqrt{R}m$, or Rm^2 . The erosion is assumed to be so severe that $\sqrt{R}m < 1$. Therefore, the erosion renders direct investment in firm stocks unprofitable, and agents live de facto in isolation. No agent will participate in the markets, and thus there are no capital markets at all. Alternatively, a bank could provide a liquid investment opportunity. Yet, as seen above, the bank is subject to panic without time deposits or stocks. Unfortunately, no-one is willing to invest in time deposits or bank stocks in the absence of capital markets. It seems that no banks can be established, and that there is no way to invest in production.

Fortunately, the scenario is more optimistic. Suppose that a bank attracts time deposits but does not issue stocks. The initial size (value) of a time deposit is a unit, and it pays D_{TD} units at maturity (after two periods). The bank chooses D_{TD} so that

$$(11) \quad \varepsilon u\left(m\sqrt{D_{TD}}\right) + (1 - \varepsilon)u\left(m^2 D_{TD}\right) = \varepsilon u\left(\sqrt{D_2}\right) - (1 - \varepsilon)u\left(D_2\right), \quad D_2 > 1.$$

An agent obtains the same expected utility by becoming an active agent and saving in marketable time deposits as by remaining passive and saving in demand deposits. Since the bank maximizes the total utility of its depositors, the interest on time deposits observes the condition $D_{TD} = D_2/m^2 > R$. Thus the interest on time deposits exceeds the long-term production. However, some time deposits are needed to prevent panics. Above, the market value of an intermediate time deposit, P_1^{TD} , was solved for from the nonarbitrage constraint

$$\frac{m P_1^{TD}}{1} = \frac{m^2 D_{TD}}{m P_1^{TD}} = \sqrt{D_2}.$$

The first term expresses the return on a time deposit during the first period, and the second term the return for the second period, whereas the third term represents the return from a demand deposit in one period. It is clear that $P_1^{TD} = \sqrt{D_{TD}}$.

Since market participation entails at least $1 - m$ units of cost, it is optimal that some agents specialize in saving in time deposits, whereas the others save only in demand deposits. Next we must list the active agents' budget constraints and indicate that the market for time deposits clears.

In equilibrium each active agent's budget constraint must be satisfied:

$$(12) \quad \begin{array}{ll} \text{(i)} & \beta_n^0 + \beta_i^0 P_1^{TD} = 1, \\ \text{(iii)} & \beta_n^1 + \beta_i^1 P_1^{TD} = \sqrt{D_2}, \end{array} \quad \begin{array}{ll} \text{(ii)} & m\beta_n^0 P_1^{TD} + m\beta_i^0 D_{TD} = \sqrt{D_2}, \\ \text{(iv)} & m\beta_n^1 P_1^{TD} + m\beta_i^1 D_{TD} = D_2. \end{array}$$

Recall that β_n^0 (β_n^1) denotes the new-born's (late consumer's) savings in new time deposits, and β_i^0 (β_i^1) symbolizes his savings in intermediate time deposits. In

(i)–(iv) the left-hand side gives the allocation of funds and the right-hand side the value of the portfolio. Since both new and intermediate time deposits yield the same risk-free return, $\sqrt{D_2}$, in every period, every allocation yields the same return to an agent. Thus each new-born's portfolio, (i), has an initial value of 1, and after a period the value of the portfolio, (ii), is $\sqrt{D_2}$. If an agent becomes an early consumer, he can enjoy this. If he becomes a late consumer, he can reallocate his funds, (iii), wait for a period, and then enjoy the final portfolio, (iv), that is, consume D_2 units. Although active agents can allocate their funds among new and old time deposits in different ways, the allocations must be such that the market for time deposits clears:

$$(13) \quad (i) \quad \alpha \bar{\beta}_n^0 + \alpha(1 - \varepsilon) \bar{\beta}_n^1 = \frac{1}{2}A, \quad (ii) \quad \alpha \bar{\beta}_i^0 + \alpha(1 - \varepsilon) \bar{\beta}_i^1 = \frac{1}{2}A.$$

Again, $\bar{\beta}_n^0$ ($\bar{\beta}_i^0$) gives the average amount of new (intermediate) time deposits purchased by new-borns, and $\bar{\beta}_n^1$ ($\bar{\beta}_i^1$) the average amount of new (intermediate) time deposits bought by late consumers. Here (i) ((ii)) states that the total savings in new (intermediate) time deposits are equal to the supply. Given (i), the markets for new time deposits are in equilibrium. It is easy to show that the market for intermediate time deposits clears, so that the demand for liquidity can be satisfied by selling time deposits. Since the proof is similar to that in section 4, it is omitted. Consequently, under the budget constraints, there is a market-clearing allocation of time deposits at the given levels of active agents and time deposits. We must still find the optimal levels of active agents and time deposits.

To determine the optimal amount of active agents, recall (i) and (iii) from (12). They describe the budget constraints of a new-born agent and a long-term consumer. Since the budget constraints of each agent are satisfied, the budget constraints are satisfied on average:

$$\bar{\beta}_n^0 + \bar{\beta}_i^0 P_{TD}^1 = 1, \quad \bar{\beta}_n^1 + \bar{\beta}_i^1 P_{TD}^1 = \sqrt{D_2}.$$

Inserting these into (13) gives the optimal share of active agents:

$$(14) \quad \alpha^* = \frac{\frac{1}{2}A^*}{I^{**}}, \quad \text{where } I^{**} = \frac{1 + (1 - \varepsilon)\sqrt{D_2}}{1 + \frac{\sqrt{D_2}}{m}}.$$

To find the optimal amount of time deposits, recall the bank's resource constraint,

$$(15) \quad R[(1 - \alpha^*) + \frac{1}{2}A] - B[(1 - \alpha^*)(2 - \varepsilon) + A] = \frac{1}{2}AD_{TD} + (1 - \alpha^*) \times [\varepsilon\sqrt{D_2} + (1 - \varepsilon)D_2].$$

The optimal amount of time deposits generates maturity matching when time deposits are subordinate to demand deposits:

$$R[(1 - \alpha^*) + \frac{1}{2}A] - B[(1 - \alpha^*)(2 - \varepsilon) + A] - (1 - \alpha^*) [\varepsilon\sqrt{D_2} + (1 - \varepsilon)D_2] + l[(1 - \alpha^*) + \frac{1}{2}A] - (1 - \alpha^*)(1 - \varepsilon)\sqrt{D_2} \geq 0.$$

Given the resource constraint, the sum of the first three terms is $AD_{TD}/2$. Given this, (14), and $D_{TD} = D_2/m^2$, it is easy to solve for the optimal amount of time

deposits:

$$(16) \quad A^* = \frac{2[(1-\varepsilon)\sqrt{D_2} - l]}{\frac{D_2}{m^2} + \frac{(1-\varepsilon)\sqrt{D_2}}{I^{**}} + l\left[1 - \frac{1}{I^{**}}\right]}.$$

The solution (D_2, α, A) is determined by (14), (15), and (16). The solution is socially optimal and it is optimal to establish a bank only if the bank can pay positive interest on deposits, $D_2 > 1$. Since this problem is difficult to analyze, we offer numerical examples below. With certain parameter values, bank formation is socially optimal, but sometimes it is not.

In sum, the agents are initially passive. But some active agents are needed to save in time deposits so that a panic-free banking system can be constructed. Thus the bank pays very high interest on time deposits in order to motivate a few agents to break through the isolation and become active. The bank determines the interest on demand deposits and the interest on time deposits so that each agent obtains the same expected consumption package $(\sqrt{D_2}, D_2)$. Thus each agent obtains a productive, liquid saving asset. Active agents save in marketable time deposits, whose nominal returns are high, but the high costs of market participation reduce the returns, so that active agents receive the same expected utility as passive agents who save in demand deposits.

The bank provides a fundamental service to economy by transforming the fixed liquidity of assets to deposits of different liquidities. The underlying assets have a high long-term return (R) and reasonable liquidation value (1). Since the long-term return is not sufficiently high and since market participation entails costs, it is not profitable to trade the underlying asset in secondary markets. The bank solves the problem by creating time deposits whose long-term return exceeds the return on the underlying asset, $D_{TD} > R$. The very high long-term return on time deposits renders them marketable. The bank finances the very high payments on time deposits by reducing long-term payments on demand deposits, $D_2 < R$. This is accepted by passive agents, because the bank provides them a return present: the demand deposits can be withdrawn early at the relatively high value of $\sqrt{D_2}$. The bank can do this without generating the risk of panic by denying early withdrawal of time deposits. This does not cause a liquidity risk to active agents, who can resell time deposits. Thus each agent receives a liquid saving asset. Demand deposits are more liquid than bank assets, $1 < \sqrt{D_2} < D_2 < R$. Time deposits are nominally less liquid than bank assets, $0 < 1 < R < D_{TD}$, but effectively liquid, since they are marketable.

Consequently, for some parameter values there is no investment, and production without a bank or a panic-free bank can be established only if the bank issues marketable time deposits. Thus the establishment of the bank generates capital markets as a by-product. This case is most likely to arise in the context of a small, unknown bank that operates on the periphery, probably in an emerging economy, where there are no natural secondary markets for time deposits. Time deposits can be made lucrative and the secondary markets can be created by paying very high interest on them. But even so, secondary markets of time deposits are likely to

be modest. It may require much time and effort to find a willing trading partner. The required high payments on time deposits reduce interest on demand deposits. Therefore, although it may be possible to establish a panic-free bank system via maturity matching, the costs may be substantial. The bank is able to pay more interest on demand deposits if the bank regulator offers deposit insurance and the bank could abandon stabilization through time deposits.

Finally, we have assumed a special cost of market participation: (1) the cost of market participation erodes $1 - m$ percent of the asset return in each period, and (2) the cost is at least $1 - m$ units to each active agent in each period. If the second part is dropped, the fixed cost of market participation disappears, and it is unnecessary that some agents specialize in saving entirely in time deposits. In the optimal solution, each agent can split his funds among demand deposits and time deposits. On the contrary, if market participation entails only the fixed cost, it is optimal that a few agents specialize in time deposits. Furthermore, to minimize the number of active agents, it is optimal that the wealthiest agents specialize in time deposits. In our model this means that late consumers save with time deposits. If the volume of late consumers is not sufficiently large, a few new-borns are also needed to save via time deposits. Consequently, the cost structure of market participation has a strong effect on agents' decisions to become active or remain passive.

6.1.2 Numerical Example II

Consider an economy $R = 1.073073$, $B = 0.025$, $\varepsilon = 0.3$, $l = 0.6$. First, suppose the erosion effect is $m = 0.9$. We can show that $D_2 \approx 1.02$, $D_{TD} \approx 1.26$, $A^* \approx 0.1074$, and $\alpha^* \approx 0.06675$. Therefore, it is possible to motivate some agents to break through the insolvency by paying high interest on time deposits. Interest on demand deposits is positive. However, it is possible to deduce from $R - (2 - \varepsilon)B = \varepsilon\sqrt{D_2} + (1 - \varepsilon)D_2$ that without time deposits the bank could pay higher interest on demand deposits, $D_2 \approx 1.036$. Thus the regulator can boost the agents' expected utility by offering deposit insurance.

Consider an identical case but where $m = 0.5$. The bank is unprofitable, and it cannot pay positive interest on demand deposits. It is impossible to prevent panic by using time deposits.

6.2 Absence of Moral Hazard

It is often argued that deposits must be liquid in order to eliminate moral hazard (e.g., CALOMIRIS AND KAHN [1991]). This section briefly indicates that moral hazard can be eliminated in the panic-free bank system, in which the bank also attracts time deposits.

Let us enrich the model by assuming a risky long-term project. It requires a unit of investment and yields \hat{R} ($\hat{R} > R$) after two periods if it succeeds. The risky project succeeds with probability s in every period. If it fails, the failure is irreversible and the value of the project is zero. Thus the final output, \hat{R} , materializes with

probability s^2 . The NPV of the risky project is assumed to be negative, $s^2 \hat{R} < 1$. If a risky project is successful, but liquidated after the first period, its liquidation value is $\hat{l} < l$.

Suppose that a banker establishes a bank in a perfectly competitive economy. If he operates as above and invests in safe long-term projects, which yield R after two periods, he earns zero returns due to competition. This tempts him to take risks. If he promises the same interest on deposits as above, but invests the funds in risky projects and the risk taking is successful, he earns profits $\hat{R} - R$. If the risk taking fails, the costs are suffered by depositors.

Suppose that the bank changes its investment strategy at time point t and begins risk taking. The new strategy is observed by depositors during the same period. How do they react at $t + 1$? The depositors who have intermediate time deposits cannot interrupt their deposits. The depositors with demand deposits know that the maturity-matching constraint is no longer satisfied, since the liquidation value of long-term risky production, \hat{l} , is less than the liquidation value of long-term safe production, l . Given this, along with Definitions 1 and 2, the agents with demand deposits panic immediately. Knowing this, the banker will never switch to risk taking. He knows that he cannot win. If he switches to risk taking, a disciplinary panic immediately occurs and the bank fails. Therefore, moral hazard is eliminated even if the bank also attracts time deposits.

7 Conclusions

The new Basel II Accord stipulates capital requirements for banks. Two key types of capital exist: tier 1 (e.g., common stock) and tier 2 (e.g., subordinate time deposits). The capital requirements are determined in order to correctly price bank risk. Their existence, however, poses the question of whether the same kind of capital requirements could also be used to prevent bank panics. Is capital stock in this use more effective than time deposits? These questions are examined in this paper.

The paper confirms that both capital stock and time deposits offer an effective option for preventing panics. Passive agents, who cannot participate in capital markets, will save in demand deposits, whereas active agents, who can participate in capital markets, prefer more productive time deposits, bank stocks, and firm stocks. Since time deposits and bank stocks are marketable, they represent effectively liquid assets for agents who can participate in capital markets. Therefore, each agent obtains a liquid saving asset. A bank can create desired liquidity even when it operates under maturity matching. If the operating costs of the bank are positive or if there are relatively few active agents in the economy, the stabilization effect of time deposits (or capital stock) entails costs. Thus, if the regulator can offer deposit insurance at low costs, it may be socially optimal to prevent panics by insuring deposits rather than by utilizing time deposits or capital stock.

Obviously, the model cannot capture all of the key features of actual banking systems, and it observes some restrictive assumptions. Finally, we will discuss this

subject. First, in the model, long-term production is assumed to be risk-free. As a result, the banking system has only one risk: panic. This makes it quite easy to eliminate panics via maturity matching. If the upcoming value of bank assets is risky, the scenario is more problematic. Consider symmetric information and a bank that operates initially under maturity matching. If the value of the bank assets slumps, it is possible that the maturity-matching constraint is no longer satisfied, which triggers a panic. Under asymmetric information, the scenario is even more problematic. A depositor cannot observe the slumped value of bank assets, and so he does not know whether the maturity-matching constraint is satisfied. Yet, if the initial amount of capital stock is sufficiently high, a depositor (almost certainly) knows that the bank is risk-free even after the slump and that the maturity-matching constraint is still satisfied. Second, it might be easier to maintain maturity matching using bank stocks than time deposits. If a bank aims to maintain maturity matching via time deposits, it needs to attract new time deposits in every period. It may be a demanding task to sell subordinate long-term time deposits during financial turmoil. Third, under asymmetric information, a bank faces the same difficulties in attracting capital stock as do standard firms (recall, e.g., MYERS AND MAJLUF [1984]). Fourth, in our study the share of early consumers is fixed. If this assumption is relaxed, the analysis is rather complex in theoretical models (e.g., GREEN AND LIN [2000], [2003]).

In sum, the model is mechanical and contains certain restrictions. As to the regulatory recommendations, our conclusions are, however, quite positive. It may be possible to prevent panics using maturity matching. The positive result is more likely to occur with capital stock than with time deposits, and the required amount of capital stock may be rather high. Obviously, more research is needed regarding risky returns of bank assets and stochastic demand of liquidity.

Appendix

A.1 Proof of Proposition 2

Given $L = 0$ and $D_2 = (D_1)^2$, the maximization problem simplifies to

$$(A1) \quad \varepsilon u(\sqrt{D_2}) + (1 - \varepsilon)u(D_2),$$

$$(A2) \quad \varepsilon\sqrt{D_2} + (1 - \varepsilon)D_2 = 1 + I(R - 1) - (2 - \varepsilon)B,$$

$$(A3) \quad 0 \leq I \leq 1,$$

$$(A4) \quad \varepsilon(R - \sqrt{R}) < B < (R - 1)/(2 - \varepsilon).$$

It is easy to see from (A1) that expected utility is maximized when D_2 is maximized. The maximum value of D_2 is determined by the resource constraint, (A2). Since resources are largest when $I = 1$, (A2) can be rewritten as $\varepsilon\sqrt{D_2} + (1 - \varepsilon)D_2 = R - (2 - \varepsilon)B$. Given this and (A4), it follows that $D_2 > 1$; interest on deposits

is positive. Furthermore, the rewritten resource constraint satisfies $D_2 - \varepsilon(D_2 - \sqrt{D_2}) = R - (2 - \varepsilon)B$. If $D_2 = R$, the consumption (right-hand side) exceeds the resources (left-hand side), due to (A4). Thus $D_2 < R$. Hence the optimal allocation satisfies $1 < D_1^* < D_2^* < R$, $D_1^* = \sqrt{D_2^*}$, and $I = 1$. *Q.E.D.*

A.2 Proof of Market Clearing

We show that the markets for intermediate bank stocks clear. Let us investigate the markets for bank stocks at time t . They are supplied by agents who encounter a consumption shock: early costumers of generation $t - 1$, $\varepsilon\alpha\bar{p}^0$, and late consumers of generation $t - 2$, $(1 - \varepsilon)\alpha\bar{p}^1$. The demand consists of new-borns of generation t , $\alpha\bar{p}^0$, and the additional demand by late consumers of generation $t - 1$, $(1 - \varepsilon)\alpha(\bar{p}^1 - \bar{p}^0)$. The supply is equal to the demand if $\varepsilon\alpha\bar{p}^0 + (1 - \varepsilon)\alpha\bar{p}^1 = \alpha\bar{p}^0 + (1 - \varepsilon)\alpha(\bar{p}^1 - \bar{p}^0)$, which simplifies to $0 = 0$. Thus the demand is equal to the supply. In the same way, it is possible to show that the intermediate market for firm stocks clears.

It is, for example, possible that a few active agents specialize in investing only in bank stocks whereas the others favour firm stocks. In this case $\bar{p}^0 = e/P_S$ and $\bar{p}^1 = e\sqrt{R}/P_S$, where e denotes the share of active agents who specialize in bank stocks. It is easy to get $e = E/2I^*$. *Q.E.D.*

References

- ALLEN, F., AND D. GALE [1998], "Optimal Financial Crises," *The Journal of Finance*, LIII, 1245–1284.
- AND — [2004], "Financial Fragility, Liquidity, and Asset Prices," *Journal of the European Economic Association*, 2, 1015–1048.
- BHATTACHARYA, S., P. FULGHIERI, AND R. ROVELLI [1998], "Financial Intermediation versus Stock Markets in a Dynamic Intertemporal Model," *Journal of Institutional and Theoretical Economics*, 154, 291–319.
- AND J. PADILLA [1996], "Dynamic Banking: A Reconsideration," *The Review of Financial Studies*, 9, 1003–1032.
- CALOMIRIS, C., AND C. KAHN [1991], "The Role of Demandable Debt in Structuring Optimal Banking Arrangements," *The American Economic Review*, 81, 497–513.
- CHEN, Y., AND I. HASAN [2006], "The Transparency of the Banking System and the Efficiency of Information-Based Bank Runs," *Journal of Financial Intermediation*, 15, 307–331.
- AND — [2008], "Why do Bank Runs Look like Panic?" *Journal of Money, Credit and Banking*, 40, 536–546.
- DIAMOND, D. [1997], "Liquidity, Banks and Markets," *Journal of Political Economy*, 105, 928–956.
- AND P. DYBVIIG [1983], "Bank Runs, Deposit Insurance and Liquidity," *Journal of Political Economy*, 91, 401–419.
- FREIXAS, X., B. PARIGI, AND J.-C. ROCHET [2000], "Systemic Risk, Interbank Relations, and Liquidity Provision by the Central Bank," *Journal of Money, Credit and Banking*, 32, 611–638.
- FULGHIERI, P., AND R. ROVELLI [1998], "Capital Markets, Financial Intermediaries, and Liquidity Supply," *Journal of Banking & Finance*, 22, 1157–1179.

- GREEN, E., AND P. LIN [2000], "Diamond and Dybvig's Classical Theory of Financial Intermediation: What is Missing?" *Federal Reserve Bank of Minneapolis Quarterly Review*, 24, 3–13.
- AND — [2003], "Implementing Efficient Allocations in a Model of Financial Intermediation," *Journal of Economic Theory*, 109, 1–23.
- MERTON, R. [1977], "An Analytic Derivation of the Cost of Deposit Insurance and Loan Guarantees," *Journal of Banking and Finance*, 1, 3–11.
- MISHKIN, F. S. [2007], *The Economics of Money, Banking, and Financial Markets*, 8th ed., Pearson: New York.
- MYERS, S., AND N. MAJLUF [1984], "Corporate Financing and Investment Decisions when Firms Have Information that Investors do Not Have," *Journal of Financial Economics*, 13, 187–221.
- NIINIMÄKI, J.-P. [2003], "Maturity Transformation without Maturity Mismatch and Bank Panics," *Journal of Institutional and Theoretical Economics*, 159, 511–522.
- [2009], "Does Collateral Fuel Moral Hazard in Banking?" *Journal of Banking and Finance*, 33, 514–521.
- QI, J. [1994], "Bank Liquidity and Stability in an Overlapping Generations Model," *The Review of Financial Studies*, 7, 389–417.
- [2003], "Liquidity Provision, Interest-Rate Risk, and the Choice between Banks and Mutual Funds," *Journal of Institutional and Theoretical Economics*, 159, 491–510.
- ROCHET, J.-C., AND X. VIVES [2004], "Coordination Failures and the Lender of Last Resort: Was Bagehot Right after All," *Journal of European Economic Association*, 2, 1116–1147.
- VON THADDEN, E.-L. [1997], "Intermediated versus Direct Investment: Optimal Liquidity Provision and Dynamic Incentive Compatibility," *Journal of Financial Intermediation*, 7, 177–197.
- [1999], "Liquidity Creation through Banks and Markets: Multiple Insurance and Limited Market Access," *European Economic Review*, 43, 991–1006.
- WALLACE, N. [1988], "Another Attempt to Explain an Illiquid Banking System: The Diamond and Dybvig Model with Sequential Service Taken Seriously," *Federal Reserve Bank of Minneapolis Quarterly Review*, Fall, 3–16.
- [1996], "Narrow Bank Meets the Diamond–Dybvig Model," *Federal Reserve Bank of Minneapolis Quarterly Review*, Winter, 3–13.

J.-P. Niinimäki
Research Department
Bank of Finland
P.O. Box 160
00101 Helsinki
Finland
E-mail:
J-P.Niinimaki@windowslive.com

The Political Economy of Preindustrial Korean Trade

by

HUN-CHANG LEE AND PETER TEMIN*

Preindustrial Korea had little foreign trade in spite of the advantage of being a small peninsular country. We present a theory of political economy to show that the preindustrial Korean policy of suppressing private trade, like that of China, only can be explained by noneconomic factors such as the consideration of externalities and rulers' incentives, bounded rationality of policymakers, and the path dependence of history. It was a rational or bounded-rational decision to increase total gains, that is, economic and noneconomic gains, from trade under the east Asian geopolitics. (JEL: N 75, F 14)

1 Introduction

One of the central questions of economic history is why the industrial revolution began in Europe rather than Asia. If we trace the roots of industrialization back to the Neolithic revolution, following DIAMOND [1997], then agriculture began both in Europe and in Asia. Agricultural knowledge was communicated across the broad expanse of Eurasia; but Western Europe and China were the only two urbanized societies that had the potential to industrialize by the eighteenth century (GOODY [1996]). Historians of technology from NEEDHAM [1981] to MOKYR [1990] have reminded us that China, not Europe, was the leader in technical progress throughout many of the centuries before industrialization. Why then did Europe industrialize first?

ACEMOGLU, JOHNSON, AND ROBINSON [2005] argued that the growth of Atlantic trade played a central role in the rise of the West.¹ It follows that the restrictive Chinese trade policy played a central role in the fall of her superiority. The ratio of merchandise exports to GDP in China has been estimated to have been 0.7% around 1870, while those of Asia, western Europe, and the world were 1.7%, 8.8%, and 4.6% (MADDISON [2001]). If China had pursued economic gains from trade actively, it

* Korea University, Seoul (corresponding author) and MIT Department of Economics, Cambridge, MA. We thank Byoung-Heon Jun, Peter I. Yun, In-Song Gill, and members of the Harvard Economic History Workshop for useful comments. We also thank anonymous referees for constructive suggestions. All errors are, however, our own.

¹ Trade in this paper means international trade.

might have arrived at Europe before Vasco da Gama's voyage, strengthened its naval power, and enjoyed trade gains between East and West. Early modern history would have been different.

Why did China adopt a restrictive trade policy? DIAMOND [1997] argued that geography made the difference. Europe's highly articulated coastline, replete with inlets and peninsulas and even a few large islands, gave rise to political fragmentation, competition, and trade that stimulated industrialization. LANDES [1972] and [1998] emphasized many of the same factors, emphasizing the role of trade and the "expansion of Europe." China, by contrast, has a smooth coastline. Lacking geographic protection, rival governments could not survive, and the Chinese Empire had a monopoly of political and economic power. The dead hand of monopoly stifled any explorations that might have led to industrialization.

A look at the map, however, shows that part of the Asian coastline is as indented as the European coastline. The South China Sea is called Asia's Mediterranean (MANGUIN [1993], DENG [1997]), and the area around Korea and Japan, where – as in Europe – there were smaller countries and political competition, has been called the East Asian Mediterranean (YUN [2002]). It would be an obvious question why these similar geographic conditions did not lead to economic results similar to those in Europe if we did not know that trade was much less active in Asia than in Europe during the early modern period. The ratio of trade to GDP in Chosŏn Korea was at almost the same low level as that in Ming and Qing China (LEE [2004]).² But this observation just takes the question back one step. Why was there so little international trade among the smaller states of Asia in the complex geography in and around the Korean peninsula? This is the question we answer here.

To sharpen the question, we note that Korea stood to benefit greatly from trade because it is a small country; it has experienced great economic success in recent years with its export-push strategy. Korea's dependence on trade has increased rapidly, and the ratio of merchandise exports to GDP has been about 30% since the 1980s. By contrast, the preindustrial Korean states never tried to foster private trade. Private maritime trade was active only in the ninth century (during the Unified Silla period). During the long Chosŏn dynasty the maritime trade of Korea was prohibited. This trade inactivity was a serious obstacle to economic growth.

Why did Korea not participate in maritime trade vigorously to exploit its geographical advantage? Why did it become active in trade in the ninth century and inactive afterward? Why did Korean governments have a propensity to suppress private trade? In order to answer these questions, we need to have a theory of preindustrial trade and to understand the Chinese world order or tributary trade system. In the next section we propose a theory of preindustrial trade, which we hope extends

² FRANK [1998] argued that China had been the center of the world economy before the eighteenth century. But the case of Korea, which had had a very intimate relationship with China, seemed to reveal that the Chinese world order did not develop an international division of labor comparable to Europe's. China had little motive and intention to become the center of the world economy. It just wanted to maintain its political suzerainty.

beyond Korean history and sheds some light on the comparative analysis of trade policy. We use this model to explain the pattern of Korean preindustrial trade.

2 *A Theory of Preindustrial Trade*

Economic theory says that trade exists when there are net economic gains and rational actors act to maximize them. If we want to explain preindustrial trade, especially by China and Korea, we need to modify the theory. The assumption of rational actors need not be modified; people in general prefer an advantageous outcome to a less advantageous one. We assume all actors were rational, but the ruler limited the extent of everyone else's activities. We assume that preindustrial trade policy was an outcome of rational choices by rulers.

Trade produces both economic gains and noneconomic effects. In the industrial period the economic gains from trade are enormous owing to the development of markets and technology, while its noneconomic effects often are negligible compared to the economic benefits. The situation in the preindustrial period was different, even opposite, because the preindustrial era was different from the industrial era in that the economy often was not governed by the market, and the polity typically was not democratic. Economic gains from trade might be small, owing to high transaction costs and small trade volumes. The noneconomic effects from trade could be substantial, however, because trade might affect diplomacy, internal politics, and even culture.

The rulers of preindustrial states attached great importance to border security. Wars and conquests broke out frequently in the preindustrial era, and the frontier was far more fluid than in the industrial era. The most important object of diplomacy therefore often was border security, and trade might be subordinated to diplomacy.

In addition, there were internal political concerns. Every policy, including trade policy, ultimately aimed to consolidate the current ruler's political power. The distribution of economic gains from trade might affect the distribution of power; nobles and local elites could increase their power by amassing trade gains, or they could connect to a foreign trading power and defy state authority. Ideologies that would threaten state authority or cultures that would destabilize society also might enter the country.

These factors explain why preindustrial states had a propensity to administer trade – sometimes directly carrying out trade, sometimes imposing severe limitations on private trade. First, trade had externalities, so markets or free trade did not ensure the maximization of total net gains. State intervention had a meaning. Second, rulers in nondemocratized societies in general were inclined to maximize their gains rather than the country's gains, so they wanted to control trade.

Our preindustrial trade theory may appear to be a special case because it presupposes these externalities, but we argue that modern trade theory is the special case because it ignores externalities. Viewed broadly, economic acts have externalities; the economic domain and other domains are interrelated. This is why an interdisci-

plinary approach is needed. All the preindustrial states in northeast Asia of which we have knowledge considered other aspects as well as economic aspects when making decisions about trade. In the modern world, the reluctance of the European Union and the United States to reduce agricultural protection under pressure from the World Trade Organization can only be understood as the result of externalities from trade (GROSSMAN AND HELPMAN [2002]).

It is not easy to measure diplomatic, military, internal political, and cultural aspects. There must have been some method of comparing different aspects and calculating total net benefit or some criteria and principles to guide decision, since decisions to prohibit, to limit, or to encourage trade were made. The problem may be easier in a nondemocratic polity because it is necessary only to consider the rulers' incentive, for rulers in nondemocratized society were inclined to maximize their gains rather than the country's gains. Maximizing the country's economic gain might not guarantee maximizing the rulers' gain, owing to distribution and dynastic problems.

As a result, rulers internalized externalities and did not encourage trade to the level needed to maximize economic gains. Salient examples are the trade policies of Ming and Qing China, Chosŏn Korea, and Tokugawa Japan, which prohibited private maritime trade, and those of Chosŏn Korea and Tokugawa Japan, which prohibited trade with Western countries. Because they emphasized political logic, rulers might sacrifice economic logic, having a negative effect on economic growth. Even in modern democratic societies like the United States and the European Union, politicians restrict international trade in agricultural goods to preserve rural society (an externality) and their own political power (as in less democratic nonindustrial polities).

It is difficult to measure total net benefit covering diplomatic, defense, cultural goods and services, and internal political aspects. SIMON [1982] argued that human beings exhibit bounded rationality, owing to the limitation of their knowledge and computational ability, even in the purely economic domain. The limitation is far more severe in a choice problem that covers various domains or a long time span, as in the case of premodern human beings. We observe in history two major procedures to cope with the limitation of human rationality.

The first was to follow the principle of priority. People determined the rank of importance among competing goals, pursued first an objective of high priority; then they pursued a goal of a low priority within the limits of not hurting that of higher priority. This is a method often used by statesmen and diplomats. The highest priority of the foreign policy of preindustrial China and Korea was state security, that is, border security, which made sense when war and conquest broke out frequently and the frontier was fluid. If trade affected the security problem, or if the security problem could be managed through trade, it was rational to subordinate an economic action such as trade to diplomacy in order to deal with the security problem, because a security risk incurred a huge cost, far exceeding trade gains in the short term. For example, one politician in fifteenth-century Chosŏn argued that giving generous gifts to the Ryukyu (now Okinawa) ambassador did not make sense because they

did not have the potential to invade Korea, whereas giving generous gifts to the Japanese made sense because they had this potential.³ Only slightly less important was the need to promote domestic tranquility, that is, to prevent independent groups from threatening the state's power.

The second procedure was to devise institutions or policies to internalize externalities. Institutions produce stable patterns of behavior, and they can avoid inconsistent acts arising from difficult choice problems in various domains. This mode of operation also saves the costs of calculation arising from difficult choice problems. The prominent example of such an institution is the Chinese tribute system. In actual diplomatic practice, the most important aspects of the tribute system were mutual political recognition and the conduct of regular diplomatic exchanges that were Chinese (meaning Confucian) in concept, ritual, and rhetoric. All foreign rulers were regarded as outer vassals whose embassies were expected to appear at the Chinese imperial court with local products as tributes. In return, the Chinese emperor gave them imperial gifts and invested them with titles and official seals. According to Sino-centric terminology, all items coming to China were recorded as "tribute" of foreign rulers, and all items sent to foreign rulers were listed as "bestowals" of the emperors of China.

China, as the largest and the most culturally and economically advanced state in Asia, claimed universal rule, and used the tribute system to maintain its suzerainty. The rulers of China usually declared themselves ready to sacrifice economic substance and refused to acknowledge any dependence on trade; trade consequently remained formally subordinate to diplomacy (FAIRBANK [1968]). Foreign rulers strategically sought recognition as tributaries in a hierarchical Chinese world order. By this act they could maintain peaceful relations with China and might consolidate their political position. Economic benefits from trade and the import of high culture compensated for the tributary states' lower status in the tribute system. "For Central Asia, relations with China meant trade; for China, the basis of trade was tribute" (FLETCHER [1968, p. 209]).

Why did China adopt a tribute system under which trade was subordinated to diplomacy and discourage private trade? China's economic gains from trade would have been relatively small because it was so large (WILLS JR. [1968]). Trade manipulation also contributed to the defense of China's borders and the establishment of its suzerainty, which were the main foreign policy objectives of the Chinese state. China faced nomadic tribes of formidable military power, such as the Xiongnu and the Mongols, in the steppe of Central Asia. These tribes, however, urgently wanted Chinese goods, and trade manipulation contributed to the peace of China's borders. The tribute trade rituals contributed also to establishing the suzerainty of the Chinese empire.

Therefore, even if China incurred economic losses in tribute trade, its diplomatic and military gains from trade were big enough to compensate for any economic

³ *Chosŏn wangjo silok* (The veritable records of the Chosŏn dynasty), King Sejo 1386.

loss. The idea that China was the center of civilization, and the Confucian teaching that governance by virtue with an economically generous attitude was the royal road to becoming king, played some role in devising and justifying the tribute system. However, ideologies played an ancillary role; building up the tribute system was rational without considering them. Historically, the need for a countermeasure against the Xiongnu threat was the decisive factor in building up the tribute system (YÜ [1967]).

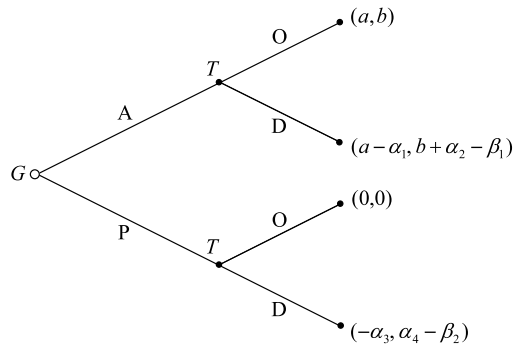
If the rulers restricted trade to preserve their own political power in despotic politics, their welfare could collide with social welfare. If one could count the total net benefit from trade and devise institutions or policies to internalize externalities, however, it seems obvious that total welfare would increase compared to the situation where only economic net gains are counted. This proposition is true in the short run, but it may not always be true in the long run. Institutions or policies that reflected the socioeconomic necessities at the time of devising gave rise to corresponding ideologies and formed new interest groups. As time went by, the socioeconomic background changed, but the ideology and the interest structure became rigid. In order to reform institutions, a fundamental change in socioeconomic background might be needed to exert strong pressure upon a dominant ideology and interest group. Institutions have inertia, and history is path-dependent (DAVID [1985]).

The tribute system lasted for almost two millennia. It may have had a positive effect on the social welfare of China in its early stages and played a role in the long duration of the huge Chinese Empire, but many scholars think that it had a negative effect on economic development at the late stages of its existence. After the adoption of European cannon in the seventeenth century, northern horse-riding tribes were not a serious threat to China, and the opportunity cost of tribute trade became large as trade between East Asia and Europe grew remarkably. During 1750–1875, when the pan-Asian trading ring was integrated into the world market, China lost her status as an industrial exporter and her favored position in the balance of trade (DENG [1997, pp. 116–125]). By the eighteenth century it was almost obvious that maintaining the tribute system was irrational in economic and political terms. Before the forced door opening to Western countries in the middle of the nineteenth century, however, China did not change the tribute trade system.

This phenomenon cannot be explained without path dependence. The group gaining the most from the tribute system was composed of rulers, and rulers objected to changing the tribute system, because they feared losing both control over trade and suzerainty. Confucianism overemphasized the cost of cultural debasement from contact with Western countries, especially permeation of the Christian religion, and underestimated the economic benefit from trade. The utility function of the Confucians also attached a greater value to culture and a smaller value to the economy than do those of modern economic men. Without considering Confucianism it is difficult to explain the irrational adherence to the tribute system at its end.

We now present a simple game-theoretical model to explain the pattern of preindustrial trade and clarify the reasons why preindustrial rulers often limited and

Figure 1
Government's Decision with Private Trade



sometimes even prohibited private trade.⁴ The order of the game proceeds as follows. First, the state rulers (or government, G) decide whether to allow (A) or to prohibit (P) private trade. Then the traders (T) decide whether to obey (O) or to disobey (D) the order, and the game ends (see Figure 1). The government chooses the more advantageous strategy, considering the response of traders or trading counterparts.

In Figure 1, a is government's gains from trade, and the b is traders' gains from trade. If official trade exists, a and b are the increments in gains from allowing private trade. The ruler or government gets tariff revenues, and traders earn profits that are part of the economy's gains from trade. Here a represents total gains including also the externalities mentioned above.

The α_i represent losses of the ruler or gains of traders arising from disobedience. They include economic and noneconomic losses or gains. Disobedience means, economically, illegal trade and, politically, defying state authority. To disobey when trade is allowed means for the traders to defy the government, refuse to follow its rules, and even possibly to overthrow it. If we confine our attention to economic gains or losses, α_1 is the loss of tariff revenue due to smuggling or other illegal activity. If α_1 includes a risk of losing political power, it may well be larger than a . The costs include both the damage to a country's diplomatic relations, as in the tribute system, and the damage to its internal stability.

A similar duality applies to α_3 . If we confine our attention to economic gains or losses, α_3 measures the cost of trying to eliminate smuggling. If we take a broader

⁴ The political economy of international trade has come to be a growing field in both economics and political science. In economics, GROSSMAN AND HELPMAN [2002] have collected many of the papers they have written on the subject, dealing with the complex problems of trade policies in democratic countries. In political science, MILNER [1997] used game theory to explore international aspects of trade policy.

view, α_3 may include costs to the ruler of chronic disobedience by his subjects, such as piracy. If pirates can become dissidents, then the costs may be high indeed.

The β_i represent the expected costs of punishments, that is, the probability of being detected times the amount of penalty for disobedience. They are not decision variables in this model, but the effects of government power. The stronger the government, the more likely that traders who disobey the rules will be punished.

The extensive-form game in Figure 1 can be solved by backward induction. Traders obey if $\alpha_2 < \beta_1$ and $\alpha_4 < \beta_2$. They disobey if the inequalities go the other way. Stated differently, if the gains from disobedience are smaller than the expected losses from it, they obey, and vice versa. Their decisions do not depend on a and b at all, that is, on the conventional gains from trade.

We assume the traders' estimates of the α_i and β_i are drawn from a probability distribution of traders, and the government maximizes its expected payoff. We denote the probability that traders disobey as p when private trade is allowed and q when it is prohibited. The expected payoff to the government when it allows trade is $(1 - p)a + p(a - \alpha_1) = a - p\alpha_1$. The expected payoff when the government prohibits trade is $-q\alpha_3$; even if private trade is prohibited, the government incurs the expected cost $-q\alpha_3$. If $a - p\alpha_1 > -q\alpha_3$, government chooses the strategy of allowing trade. If $a - p\alpha_1 + q\alpha_3 < 0$, it chooses the strategy of prohibiting trade. If official trade exists and $a - p\alpha_1 + q\alpha_3 < 0$, only the policy of allowing official trade is allowed. Note that b , the traders' gains from legitimate private trade – what we conventionally think of as the gains from trade – do not enter into the ruler's decisions.

Then how can we explain the appearance of policies to prohibit or limit trade? Our simple model can generate an important conclusion: they cannot be explained by economic factors alone. Assume initially that α_1 includes only economic gain and loss. Then $-\alpha_1$ means the decrease in fiscal revenue arising from a decrease in the legal trade volume owing to smuggling. Then a conclusion can be easily gotten. Whatever the amount of smuggling, the legitimate trade cannot be below zero. That is, $a - \alpha_1 > 0$. Therefore $a - p\alpha_1 + q\alpha_3$ is above zero. The ruler always allows trade. This makes sense; the ruler's gains from trade, a , dominate his decision. It is strengthened if there are costs to the ruler to illegal trade when it is prohibited and if the state is weak so that many traders choose to violate this ban. We therefore have the proposition that if trade was prohibited, there must have been noneconomic factors that affected trade policy.

Now expand the interpretation of the model to include noneconomic gains and losses. We cannot rule out the possibility that $a - p\alpha_1 + q\alpha_3 < 0$ when noneconomic gains and losses are important. If the government is very weak and inefficient, it may anticipate that official trade could decrease dramatically if private trade were allowed. Traders might accumulate resources that could be used to challenge the ruler's authority. Ideas brought in by traders might diminish support for the ruler. The ruler might have much more than a at stake in choosing a trade strategy. If these potential losses, represented by $-\alpha_1$, are very large, $a - p\alpha_1 + q\alpha_3$ may be below zero. If the government is weak so that traders do not fear punishment and p is high,

then this conclusion may be even stronger. The ruler's decision depends entirely on the costs of permitting trade; the costs of enforcing a trade ban – no matter how high – do not decrease the attractiveness of a trade ban.

3 *The Pattern of Preindustrial Korean Trade*

The gradual flow of cultural and technological influences from China during the first millennium B.C.E. helped Korean states to form. These states developed to the point where their existence was known in China around the fourth century B.C.E. The most advanced among them was Old Chosŏn, which had established itself in present-day southern Manchuria and the northern part of the Korean Peninsula. This first ancient state of Korea resisted incorporation into the Chinese world, and Han China destroyed it. After the fall of Old Chosŏn in 108 B.C.E., China set up command posts in the occupied territory. This event, as well as the subsequent political structure of Korea, is shown in Table 1, with Chinese dynasties appended for reference.

Table 1
Dynasties of Korea and China

Korean dynasties		Chinese dynasties	
Old Chosŏn	–108 B.C.E.	Han	206 B.C.E.–220
		Six Dynasties	222–589
Three Kingdoms		Sui	581–618
Koguryŏ	37 B.C.E.–668	Tang	618–907
Paekche	18 B.C.E.–663	Song	960–1279
Silla	57 B.C.E.–935	Khitans Liao	947–1125
		Jurchen Chin	1122–1234
Koryŏ	918–1392	Mongol Yuan	1271–1368
		Ming	1368–1644
Chosŏn	1392–1910	Qing	1636–1911

The Chinese colonies lasted until the beginning of the fourth century C.E., when they were conquered by local Korean tribal states, which ironically had derived much of their high culture from contact with these outposts of Chinese civilization. After the fall of the Chinese colonies, the Korean peninsula was divided among the three states of Koguryŏ, Paekche, and Silla. “In the Sui period, only one of China's neighbors, Koguryŏ in northern Korea and southeastern Manchuria, had any claim to be a ‘state’ with a mainly sedentary population and stable institutions” (FRANKE AND TWITCHETT [1994, p. 4]). Three great Sui expeditions against Koguryŏ in 612–614 ended so disastrously that they contributed to the collapse of the Chinese dynasty. The Tang dynasty was no more successful in a series of big expeditions

between 644 and 659, but Tang and its ally Silla managed to destroy Paekche in 663 and Koguryō in 668. The Chinese attacks on Old Chosŏn and Koguryō revealed its intolerance of Korean states that represented an independent Manchurian power.

Tang China tried to bring the entire Korean peninsula under Chinese imperial control, but Silla successfully repulsed a Tang invasion in 676. Tang was forced to accept Silla as an autonomous tributary state ruling over the southern two-thirds of the Korean peninsula. Korea succeeded in maintaining this political autonomy from China because of both its political and cultural development and its geographical distance and isolation. It was forced to accept a tributary relation with China, unlike Japan, because it shared a common border with China. Korea then maintained peaceful diplomatic relations with China for twelve centuries. Such a long period of peaceful diplomatic relations is unique in world history. Cultural and economic interchanges also lasted without interruption, and no country absorbed Chinese institutions and culture more successfully than Korea.

Korea has remained a basically unified country since its unification by Silla, except for the existence of Parhae (699–926), founded by the people of Koguryō. “In the whole world only China among existing nations can claim a clearly longer history as a unified political entity” (FAIRBANK, REISCHAUER, AND CRAIG [1978, p. 287]). Three successive dynasties, Silla (–935), Koryō (918–1392), and Chosŏn (1392–1910), had very long reigns, partly due to their peaceful relations with China, which were due in turn to the tribute system. Their territories were confined to the Korean peninsula, unlike Old Chosŏn, Koguryō, and Parhae.

The Koryō dynasty pursued a northern expansion policy, and its founder proclaimed the desire to recover the ancient territories of Koguryō. By the fifteenth century Korea had almost the same territory as that of today. In 1712 the frontier between Korea and China was fixed. Before Silla repelled the Tang invasion in 676, the frontier was unstable and fluid; between 676 and 1712, the frontier was stable but not fixed. Border security was a serious concern to Korean dynasties, which was one important reason for accepting the Chinese tribute system.

The Korean states chose to accommodate the foreign policy of the Chinese tribute system, mainly because China was militarily strong and advanced in culture, economy, and technology. But tribute system was not fully institutionalized and consolidated before the Ming Dynasty. Before the fifteenth century Korean states followed pragmatic foreign policies, and were neither “loyal vassals” nor “model tributary,” as often described by the historians of China (YUN [1998, p. 228]).

Confucianism became the ruling ideology in Korea at the beginning of the Chosŏn dynasty. By the middle of the sixteenth century, the Korean leading elites began to develop an ideological belief in the inherent moral correctness of the tribute system. One of them criticized the foreign policies of former dynasties because they were merely fearful of Chinese military power and did not truly submit in their hearts (YI [1986], “subyu” (supplement)). When the Japanese invaded Chosŏn in 1592, Ming China sent troops to expel the invaders, validating the tribute system for the Chosŏn elites. Chosŏn Korea came to serve Ming China with its heart, not only

because China was militarily strong and culturally and economically advanced, but also because China was the cradle and center of Confucianism.

When Western countries demanded trade with Chosŏn in the nineteenth century, peacefully at first and forcibly later, Chosŏn refused. Most Korean elites did not anticipate any economic benefit from trade with Western countries, and they feared that the influx of Western religion accompanying trade would undermine Confucian ethics. They thought Westerners were eager to pursue economic gains because they were morally inferior, whereas tribute relations were based on moral virtue, not on economic calculation. Confucianism played an important role in prolonging the tributary relation during the Chosŏn dynasty.

The Chinese tribute system was imitated in the relationship between countries surrounding China. For example, Koryŏ and Chosŏn Korea received Japanese and northern Jurchen emissaries according to the rules of the tribute rituals when they were divided and weak. Chinese merchants and the Jurchens were instructed to present tribute to the king in a traditional Palgwanhwae ceremony, a tribute ritual intended to strengthen royal authority (OKUMURA [1979]). The Chosŏn court also regularized its foreign relations by implementing a hierarchy consisting of China at the top, then Korea, and then the Japanese and the Jurchens. By the early fifteenth century, this policy helped Chosŏn to realize its objective of border security (ROBINSON [1992]).

Korea signed an unequal treaty to enter into free trade relations with Japan in 1876, and signed unequal treaties to open its ports to Western countries after 1882. Korea realized it could not resist the foreign military forces demanding diplomacy and trade any longer. It rapidly accumulated knowledge of the modern world, and recognized the importance of trade to economic development. The Korean government tried to transform the tributary relation with China into a modern international relationship, but the Qing government opposed it. China's decisive defeat in the war with Japan in 1894 put an end to the tribute system (FAIRBANK, REISCHAUER, AND CRAIG [1978, p. 616]).

There are reliable serial trade statistics for Korea only after it opened its doors to the modern world in 1876. The ratio of trade volume to GDP was estimated to have been 20% in 1911 (KIM (ed.) [2006]). Before the opening of the doors, there are only fragmentary Korean trade statistics. The Japanese reported that Korean trade in the early 1870s was about 3 million yen (KANG [1962]). It is the first total trade volume data. The 3 million yen could buy about 600 thousand Japanese sŏk (1 Japanese sŏk = 180 liters) of milled rice, which amounts to about 1.5% of GDP.⁵ Korea, notwithstanding the advantage of being a small peninsular country, had a very low level of trade just before opening the doors to the modern world in 1876.

Before the door-opening, the Korean trade volume was the largest during the late seventeenth century and the early eighteenth century, owing to transit trade between Japan and China based on the massive inflow of Japanese silver. TASHIRO [1981]

⁵ The total production of milled rice was 9–10 million Japanese sŏk, and it was estimated to be 25–30% of GDP (LEE [2004]).

estimated the trade volume between Korean and Japan during 1684–1710, and *Tongmungwanji*, the records of Korean government about the diplomatic relations with China, recorded that about 500–600 thousand silver taels flowed to China annually during the period. From these statistics, we can roughly estimate that the total annual trade volume at its peak was about 2 million silver taels. It could buy about 700–900 thousand Japanese sōk, which amounts to about 2.5% of GDP (LEE [2004]).

Before the sixteenth century, it is almost impossible to estimate the total trade volume, but we can ascertain the trend of trade with some confidence. The ratio of trade to GDP might have exceeded 1% after the late sixteenth century, owing to the increase in Japanese silver inflow. Before then, it might have exceeded 1% only in the ninth century (LEE [2004]).

Korean in its preindustrial history sometimes showed the potential to develop trade and become a maritime power. But it had serious impediments to trade, and as a result a very slowly growing trend of trade. The first ancient state of Korea, Old Chosŏn, did not agree to be incorporated into Han China's world order and sought to serve as an intermediary in the trade between Han China and the outlying territories beyond its northeastern borders, which provoked a Han invasion. This state showed the ability to become a maritime power (SHIBA [1992, p. 8]), but Chinese foreign policy would not tolerate it.

During the third and fourth centuries A.D. the southeast coast region developed external trade. Iron produced in this region was distributed in the Korean Peninsula and Japan. This open system of external trade collapsed in the sixth century owing to the Silla advance. The consolidation of the Three Kingdoms shrunk external trade among local powers in the Korean peninsula, so the formation of territorial states had transformed the open-trade system into an administered trade system (YI [1998]). We find no evidence that private maritime trade independent of the emissary traffic was allowed by the Three Kingdoms. It seems that a maritime ban was implemented.

The unified Silla government was more active in pursuing trade gains than the contemporary Japanese government. During the ninth century, when the state power of unified Silla weakened, private trade by Koreans prospered. Silla merchants organized private coastal trade in China and engaged in trade with Japan, dominating the East Asian sea trade. The most distinguished figure, Chang Po-go, controlled the flourishing trade with China and Japan. This vigorous seaborne trade of private traders was unique in the history of premodern Korea. Chang Po-go involved himself in the political strife of the capital and finally was assassinated. After his death, some smaller merchants continued their activities, but private maritime trade gradually weakened.⁶

⁶ “His death and the subsequent disappearance of his maritime commercial empire very probably marked the passing of the high-water mark of Korean mastery over the seas lying between China, Korea and Japan. Control over the ocean commerce of this part of the world began to shift to the hands of the Chinese and then some centuries later to traders and pirates from West Japan” (REISCHAUER [1955, p. 294]).

The Koryŏ government also came to prohibit the private maritime trade independent of the emissary traffic, and implemented a maritime ban. Korean merchants did not go to Japan after the Koryŏ dynasty, as the interest of the Korean dynasties in maritime trade diminished. Beginning in the latter part of the eleventh century, Japanese vessels came to Korea, presented tribute, and carried out trade. The weakening of state authority and the rise of local powers in eleventh-century Japan brought about maritime advances (TANAKA [1975]). Private trade with China grew gradually in the late Koryŏ dynasty.

The Chosŏn government prohibited all the private trade with China at first. Such a strengthened policy against private trade can be explained by the attitude of Ming China and the rise of Confucianism. Ming consolidated the tribute system, implemented a maritime ban, and criticized the private trade of Korean embassies. The power elites who founded the Chosŏn dynasty criticized private trade by powerful men that weakened state discipline, based on the doctrine of Confucianism that became the ruling ideology of the new dynasty (SUGAWA [1997]). Before the Koryŏ dynasty, Confucianism had little to do with the adoption and maintenance of the tribute system, but it played an important role in consolidating it during the Chosŏn dynasty.

However, the Chosŏn court colluded in private trade with China attendant on tribute traffic and authorized border-market trade. Private trade grew remarkably owing to the silver imports in the sixteenth and seventeenth centuries.

Korean vessels could not go abroad for trade during the Chosŏn dynasty. The ambassadors of Chosŏn to Ming turned to the land route to avoid the possibility of shipwreck on the sea route, which shows that reducing transportation costs was secondary to the goal of avoiding risks to ambassadors. Sea traffic with China also was not allowed, even though the main trade route to China before the Song Dynasty had been by sea. The movement of the Chinese capital to Beijing also decreased the incentives for maritime traffic. Chosŏn maintained the maritime ban even after Ming relaxed the policy. Private trade with Japan was accomplished by Japanese vessels, and vessels of Korean merchants were not permitted to go to Japan. By the time when the Korean government allowed Koreans to go overseas for trade in 1882, Koreans had lost completely any competitive edge in maritime trade.

These serious setbacks of Korean trade were caused mainly by political and diplomatic factors, not by economic factors. The preindustrial rulers had incentives to administer trade and monopolize trade gains. All the preindustrial Korean states preferred official trade to private trade, which was often suppressed. This tendency was strengthened by incorporation into the Chinese tribute system. For most of the preindustrial period Korean states had to accommodate Chinese foreign policy due to Korea's geopolitical location. China was a big, strong, and culturally advanced state that claimed universal rule, and Korea, as one of the smaller neighboring states, naturally accepted the Chinese tribute system and its own inferior position.

Accepting the Chinese tribute system had profound influences on the pattern and trends of Korean trade. Under the tribute system, China's rulers attached a pivotal importance to gift exchange, which symbolized its suzerainty and was also an

economic exchange of real value. This tribute trade was, in principle, reciprocal exchange expressing ties of amity. Reciprocal exchanges often were accomplished by the evaluation of exchanged items, however, showing that this gift exchange was not completely independent of market forces.

China allowed tribute embassies to carry out trade at the capital and at the border to procure embassies' expenses or goods desired by the state. If the purpose of trade was to procure embassies' expenses or state demanding goods, it may be classified as official trade. Otherwise, it is considered private trade. The borderline between the two types was not always clear.

Under the tribute system, private trade independent of tribute traffic was more severely restricted than that attendant on it. Ming China prohibited private maritime trade independent of tribute traffic after 1381, although this policy was relaxed in 1567. Qing China revived the maritime ban during 1656–1684. Even when Ming and Qing China lifted the maritime ban, it imposed restrictions on private maritime trade. Trade at land frontier markets independent of tribute traffic was sometimes allowed.

The ups and downs of trade were largely due to those of private trade, and official trade remained stable. Official trade even increased during the time of trade setbacks caused by political and diplomatic factors. The transformation of the open trade system into an administered trade system in the Three Kingdoms period meant organizing trade into an official entity. Incorporation into the tribute system entailed organizing trade into gift exchange, with trade attendant on tribute traffic. The weakening of state authority in China and Korea resulted in the shrinkage of tribute trade and the prosperity of private maritime trade in the ninth century. Though the Chosŏn and Ming governments at first intended to prohibit all private trade, Chosŏn was eager to dispatch tribute emissaries frequently, one of whose goals was trade. During the late seventeenth and early eighteenth centuries, when the private trade with China was at its peak prior to 1876, gift exchange occupied about ten percent of total exports, and official exports attendant on tribute traffic occupied about the same portion (LEE [2004]).

Throughout preindustrial history, the demand for Chinese goods embodying high technology and culture remained the strongest impetus for Korea to trade with China. Because luxury fabrics and handcrafted goods were eagerly sought after for consumption, Silla's king handed down a decree in 834, minutely regulating such ostentatious displays of wealth. Chinese silk of high quality had been the largest import of preindustrial Korea.

At first, much of Korea's export to China consisted of raw materials, but gradually a marked increase in handcrafted articles occurred (KIM [1934]). The Three Kingdoms had the technology to produce silk of high quality. Unified Silla produced even higher-quality silk with its own technology and exported it to China and Manchuria (HINO [1968], [1972]). But the technological gap in silk narrowed during the Unified Silla and reversed afterwards. The absorption of Chinese technology through trade also enabled Koryŏ to produce world-famous porcelain. But Chosŏn did not achieve any impressive technological progress in porcelain. The active attitudes toward trade

of the Three Kingdoms and Unified Silla were beneficial to technological transfer, while the passive attitudes toward trade of Koryō and Chosōn were not (LEE [2004]).

In the early Chosōn dynasty the central export item to Japan was cotton textiles, while the main import items from Japan were copper, sulfur, gold, silver, sapanwood, and pepper produced in Southeast Asia. Cotton exports decreased after the seventeenth century owing to Japanese import substitution. Rice exports were a substitute for part of the cotton exports, and ginseng exports prospered.

A large quantity of silver was imported from Japan after the middle of the sixteenth century. Korea exported it to China, in return importing silk, which was reexported to Japan. Korea earned big gains from the intermediate trade. The total of silver imports during the peak period of 1684–1710 amounted to 189 metric tons, and their value was 58% of total imports from Japan (TASHIRO [1981]). The cessation of silver importing around the middle of the eighteenth century reduced private trade with Japan and China. The cultivation of ginseng, which grew wild in Korea, then rose to sustain private trade with China after the late eighteenth century.

4 Explanations for Suppressing Private Trade

Korean trade history, as discussed above, presents several questions worth exploring. What factors made preindustrial states approve official trade or administered trade? First, the rate of margin from official trade exceeded that from private trade. Second, official trade might yield noneconomic gains, because it was effective in internalizing externalities. Third, official trade, which let rulers secure command of trade, does not have the negative externality represented by α_1 .

China's trade policies were exogenous to Korea, and incorporation into the Chinese tribute system gave Korea a strong incentive to suppress private trade, for the following reasons. First, gains from private trade became smaller, because China under the tribute system restricted private trade. Second, private trade had a negative externality, because it might violate the restrictive rule of the tribute system. Therefore $a - p\alpha_1 + q\alpha_3$ became smaller.

Under the tribute system the most institutionalized form of official trade was gift exchange. The tribute trade of gift exchange was inefficient for pursuing economic gains. It might not have met the demands of trading partners. The transportation cost of exchanging items by sending them to the capital and then transferring them from it was considerable, and the trade activities of government officers crowded out that of merchants. The goods of official trade were procured largely by taxes and from state-owned workshops, which crowded out market production.

Then why did rulers continue to maintain substantial amounts of gift exchange? There also were counterbalancing forces. The gift exchange allowed ancillary trade, which provided trade gains. Because the trade was carried out by tribute emissaries who had to go to China for diplomacy, there were no large additional transportation costs, and ancillary trade was advantageous for procuring emissaries' expenses. The government procured tribute articles via taxes rather than markets, because the

government wanted to firmly control the source of tribute items. It is not because they were irrational that East Asian countries preferred tribute trade.

The most important *raison d'être* for the gift exchange was efficiency in internalizing externalities. It included voluntary transfers from China and compulsory transfers to China to internalize externalities, the most important of which were diplomatic and military. There were also compulsory transfers from Song China to the Khitan Liao and the Jurchen Chin, which were the price of peace. Koryŏ was well compensated for its gift of a local present by Song China, because the latter wanted to ally with the former to deflect the northern military threat. After 1079 the Song court did not even evaluate the Koryŏ tribute and continued to fix the imperial bestowal to Koryŏ at ten thousand bolts of silk. But when Koryŏ surrendered to Mongol Yuan, the latter determined tribute items and quantities by itself, and Korea's compulsory transfer to Yuan China was substantial. When Koryŏ sent an envoy to the Yuan court to assert its autonomy in 1356, one of its demands was that Koryŏ should decide the amount and frequency of the tribute (CHŎNG et al. (eds.) [1973, pp. 774f.]). After Koryŏ asserted its autonomy, the amount of tribute decreased drastically. Ming demanded a heavy burden of tribute, reminiscent of the Mongol exploitation, to test Korea's sincerity. After confirming Korea's sincerity, Ming alleviated the burden. After the decisive defeat in 1636, Chosŏn's forced transfer to Qing was substantial. As they resumed normal relations, the tribute trade also recovered reciprocity. The average value of various Korean tribute items per year, measured in silver money, was 50,000 liang (1 liang (tael) = 1.3 ounces) around 1630, 130,000 liang during 1637–1644, 70,000 liang during 1645–1735, and 60,000 liang after 1736, while that of Qing's gift measured in silver money was 50,000 liang (CHUN [1970], [1971]).

Because the gift exchange alone could not satisfy the trade needs of the state, other forms of official trade took place. When Korea or China had difficulties in acquiring certain articles or sufficient amounts from tribute trade, it also relied on market exchange. What China desperately endeavored to acquire were horses to defend her against invasion of nomadic tribes. Because China could not acquire sufficient horses from tribute trade, it relied on an administered market. The Chinese government often predetermined the purchase prices of horses for defense. The predetermined prices under administered trade varied in the long run, reflecting market forces, and often had an element of transfer to internalize externalities. Though horses were the most important merchandise imports of China, Korea's tribute was nominal and horse exports were few, except when Ming purchased about ninety thousand horses from Korea during 1374–1429. The quantity and price were set by the Ming, and the purchase of horses was in fact compulsory. The price paid to Korea was not much different from that to inner Asia. But Korea had a comparative disadvantage in horse production, while inner Asia had a comparative advantage. Ming purchased Mongol horses at horse fairs for 5–7 liangs, while tribute horses from the Mongols were evaluated at 10 liangs in 1597. The Mongols participated in the trade voluntarily, and subsidies were given to the tribute of Mongols through the high price of horses. The internal price of a big horse in Korea was 90 liangs in 1401 – much higher than that of Mongol horses. Korea barely satisfied the

Chinese demand for horses by commandeering them from civilians, and regarded them as tribute as well (SERRUYS [1975, pp. 255–268], NAM [1960]). Whereas the Ming's purchase prices of Mongolian horses might have been the monopsony price reflecting market forces, those of Korean horses also had an element of compulsory transfer. The prices predetermined in the trade of authorized border markets with Qing China were also unfavorable to Korea.

Official trade was inefficient in expanding. Therefore the good way to increase rulers' gain from trade, a in terms of our model, was allowing both official trade and private trade. China and Korea allowed private trade for the most part of their history, but administered it. A free-trade system appeared in China and Korea only after the unequal treaties forced by Britain and Japan in the nineteenth century. Trade attendant on tribute traffic was never prohibited, but it had predetermined trading dates, places, and participants and avoided certain items so as not to damage rulers' positions and arouse negative externalities. Private trade that was independent of tribute traffic was more severely limited than trade attendant on tribute, because it was more difficult to monitor and posed a threat to the official trade, including tribute trade. In terms of our model, α_1 was large.

Why did preindustrial states sometimes choose to prohibit private trade, while promoting official trade? It appears that $a - p\alpha_1 + q\alpha_3 > 0$ when only allowing official trade, but $a - p\alpha_1 + q\alpha_3 < 0$ when allowing private trade. Unlimited private trade also would encroach on official trade. If official trade provides noneconomic gains, the potential losses, represented by $-\alpha_1$, become large. If the government is weak so that traders do not fear punishment and p is high, one may anticipate that official trade could decrease dramatically if private trade were allowed. And the preindustrial rulers in Korea and China often thought that the marginal increase in gains from allowing private trade was not big.

Independent maritime trade was more severely limited than independent land trade; in fact, it was the main target of prohibition. It was economically efficient and vital to trade development, but it was difficult to superintend, and apt to arouse negative externalities such as information leakage, coastal insecurity owing to pirates, and the rise of decentralizing maritime power. Maritime traders could amass great fortunes and threaten state authority, and it was impossible to prevent maritime traders from interacting with people of foreign, hostile countries. In terms of our model, a from independent maritime trade was big, and α_1 from it also was very big. Rulers in China and Korea implemented maritime trade bans because they evaluated α_1 as very high and calculated $a - p\alpha_1 + q\alpha_3 < 0$. The main reason why Ming China implemented the maritime ban was to blockade the rise of naval resistance power and defend itself from Japanese and Chinese pirates, and Manchu Qing implemented it to destroy Chinese naval resistance, especially that under Zheng Chenggong (LI [1990, pp. 80f.], WANG [2000, pp. 21, 31f.]).

Korea had an incentive to escape from the Chinese tribute system because a small peninsular country could earn big gains from trade, as mentioned in the introduction. Then why did the Three Kingdoms accommodate the tribute system and suppress private trade, even though the Chinese states were weak and the tribute system was

not consolidated? They accepted it to enlist diplomatic and military support from the Chinese dynasties to win the fierce military conflicts among them. Rulers might have thought that additional economic gains from private trade were not considerable because tribute trade and the official trade attendant on tribute traffic existed. They also might have considered possible negative externalities from private trade. Moreover, there had been a diplomatic principle from ancient times in northeast Asia that subjects did not have the right of diplomacy, which provided an ideology for the tribute system. A policy behind the principle was to capture or monopolize the gains from trade (KIM [1935], ARANO [1988]). Private trade independent of tribute traffic violated this principle. In terms of our model, $a - p\alpha_1 + q\alpha_3 < 0$ for independent private trade.

Private maritime trade became active in northeast Asia after the middle of the eighth century as state power in China and Korea weakened. This suggests that if states or rulers had not intervened, private maritime trade would have prospered in northeast Asia. When state power decreased, the expected costs of punishment for independent maritime trade, β_2 , decreased, so q increased. This might have changed the sign of $a - p\alpha_1 + q\alpha_3$ and led the state to give up prohibiting private maritime trade.

The active and open attitude of Unified Silla toward foreign relations played some role in the active pursuit of trade gains and provided a beneficial environment for the rise of maritime adventurers. The Three Kingdoms were active in maritime traffic owing to internal competition, and this active tradition was inherited by the Unified Silla. Hostile relations with Japan and Parhae led to an active foreign policy, and Korean emigration was vigorous during the Unified Silla Period. Korea dominated the maritime trade among Korea, China, and Japan, and Korean pirates harassed the Japanese coast.

After the transition to the Koryŏ dynasty, the situation reversed itself. Koryŏ became inactive in maritime advance, and the leadership of maritime trade in East Asia was transferred from the Koreans to the Chinese and Japanese. How can this reversal be explained? The founder of the Koryŏ dynasty and his successors, who inherited ongoing maritime power, must have known of the big gains from maritime trade, because ninth-century Korea witnessed the prosperity of private maritime trade. They must have estimated a to be big, but $a - p\alpha_1 + q\alpha_3 < 0$ for independent maritime trade nevertheless. What made α_1 so big? The most important factor seems to have been the lesson that the rise of maritime power threatened the state power of the Unified Silla dynasty. We infer from the cessation of independent maritime trade that the rulers of the Koryŏ dynasty attached greater importance to the rulers' incentive to monopolize diplomacy and trade than to the economic gains from unrestricted trade. Naval bases decreased and influences of maritime origin were defeated in the power struggle at an early stage of the Koryŏ dynasty.

The Chosŏn dynasty's strong policy against private trade can be explained in part by the rise of Confucianism. Confucians generally underestimated the economic role of commerce, a , and overemphasized the adverse effect of private economic motives on society and politics, α_1 .

The policy against private trade became less strict in the sixteenth and seventeenth centuries. The external reason was that the attitude of the Chinese government toward private trade became less strict. The internal reason was that markets gradually grew, owing mainly to population increase. The population in Korea was roughly estimated as more than five million around 1400 and about eighteen million around 1800 (KWON AND SHIN [1977]). The growth of the market and the consequential trade motive increased the cost of trying to eliminate smuggling, α_3 . And the Korean government also became less strict. β_2 decreased to become smaller than α_4 , so q increased. Therefore, the sign of $a - p\alpha_1 + q\alpha_3$ changed. The late Chosŏn court allowed private trade and levied taxes on it; a grew with the expansion of it. The growth of markets influenced the trade policy of preindustrial Korea toward more openness, while the consideration of externalities and rulers' incentives often played a role in suppressing private trade. The latter force was predominant in Korea for most of the period before the door-opening to Western countries.

Even after Ming and Qing China removed their maritime ban, however, Korea clung to this policy. The rulers of the Chosŏn dynasty did not anticipate big gains from independent maritime trade. In terms of the model, they evaluated a as small. But in theory Korea's economic gains from trade would not be small, because it had the advantage of being a small peninsular country. Why did this divergence between theory and actuality occur? First, the factor endowment and industrial composition of Korea were not so different from those of China. Korea, like China, produced almost all the necessary goods like food, salt, cloth, and iron. Therefore, the economic gains from trade might not have been big in static terms. Active trade would have transformed the composition of production and developed markets, however, which would have realized bigger gains from trade over time. Luxury and intermediate trade would have prospered, as in the ninth century, through private traders seeking profits. Trade gains probably would have been considerable in dynamic terms. Chosŏn rulers seem to have counted trade gains only in static terms, however, and estimated them to be small as a result. To them, the prosperity of private maritime trade during the ninth century was a fact of the dim past. It was difficult for preindustrial rulers to calculate trade gains in dynamic terms because bounded rationality in the case of calculating dynamic efficiency was serious in the preindustrial age. Even modern human beings are often irrational in coping with future problems (ARIELY [2008]).

Second, the tribute system and Confucianism lowered a from independent private trade. The Chosŏn court wanted to import books, medicine, and goods indispensable to military defense. Since trade connected to tribute traffic satisfied the state's desire for trade fairly well, there was no urgent need to foster private trade. The Chosŏn court could not conceive of trade as a means of increasing a state's wealth. Trade was just a means to acquiring goods indispensable to the state. This way of thinking by the Chosŏn court reflected Confucianism. It had deep concern to stabilize the economy, but little concern to grow it. It was afraid that an influx of luxuries such as silk might damage social discipline, and tried to regulate it. Confucianism aimed to establish a harmonious and stable agricultural society, and

rulers governing it were afraid that a flourishing maritime trade might be harmful to that aim.

Third, rulers in the Chosŏn dynasty evaluated not only a as small, but also α_1 as large. The Chosŏn court also worried about the possibility that maritime merchants would leak state secret information and induce invasions by foreign countries. After two devastating wars in the mid-Chosŏn period, it worried about leaks more seriously. Korean rulers prohibited independent maritime trade because they counted its total net benefit as negative.

The main reason why Korea prohibited trade with Westerners (including Americans) in the nineteenth century was the threat of debasement of Confucian culture by permeation of the Christian religion. Korean elites also thought that the exchange of Korean silver and gold for Western clothes was unfavorable to Korea. Only a few men of foresight thought trade with Westerners was beneficial to the Korean economy and culture, and they had no influence on policy (LEE [2003]).⁷ Because the Chosŏn court evaluated a as very small and α_1 as very big, it turned out that $a - p\alpha_1 + q\alpha_3 < 0$. When the Tokugawa bakufu in Japan prohibited trade with Europeans except the Dutch in the seventeenth century, it attached a bigger importance to trade than did the Korean government. The main reason for the prohibition was to blockade the alliance of Catholic countries and internal Catholic powers that might menace the bakufu's authority, and to put the right of trade under its firm control.

5 Conclusions

Modern trade can be fully, but not completely, explained by economic logic. But the preindustrial Korean policy of suppressing private trade can be explained only by noneconomic factors such as the consideration of externalities and rulers' incentives, bounded rationality of policymakers, and the path dependence of history. We have shown that these factors can also explain Chinese trade policy under the tribute system (LEE AND TEMIN [2005]).

It cannot be said that the Korean and Chinese governments were ignorant of the benefits from trade or that they did not pursue economic gains from trade. They tried to maximize total net benefits from trade, including economic ones. Because bureaucratic states developed earlier than markets in China and Korea (unlike Europe), the motives to restrict trade that were due to the noneconomic factors exerted a stronger influence on their trade policy than the motives to promote trade that were due to the growth of markets. In terms of our model, a in the economic sense was small, and a and α_1 in the noneconomic sense were big. But the growth of markets and trade motives widened the range of private trade, resulting in "the eclipse of the tribute system by trade" in the Southern Song period (SHIBA [1983, p. 110]), the late Qing period (FAIRBANK [1953, p. 33]), and the late Chosŏn period,

⁷ Che-Ga Pak argued in his famous book, *Bukhaku ūi*, in the late eighteenth century, that Korea was poor because its ships did not go abroad for trade, and that the Korean government should foster maritime trade to remedy its poverty (PAK [1971]).

though the policy that subordinated trade motives to diplomacy and internal politics with higher priority was not abandoned until the compulsion of free trade by modern military powers.

In Europe, where there was no hegemonic country like China after the fall of Roman Empire, feudal societies appeared, and self-governing city-states emerged in the context of decentralized states. Feudal lords had strong incentives to participate in trade actively in order to survive and expand in the competition between them, as did European nation states when they grew, because of the strong competition in their environment. In such a situation, rational rulers would not adopt or continue for long institutions or policies under which economic interests were sacrificed for the sake of noneconomic goals, because this strategy led to the weakening of the state power. Rulers were more concerned with trade gains that would attract funds for winning competition when rivalry became violent. By comparison, rulers of the centralized state in Korea did not have strong incentives to promote trade, because of the weakness of internal and external competition. If China had been divided into several countries, the political situation of East Asia would have been changed into one like Europe's, and we cannot rule out the possibility that rulers of East Asia would have behaved like European mercantilists.

How can the European trade policy be explained in our model? First, the economic gains associated with a became large, because European countries were small. Second, the externalities of a also became large, because the rulers needed to finance funds for military competition from trade gains. Third, the negative externalities of private trade associated with α_1 became small, because there was no suzerainty worth defending. Fourth, there appeared a new form of α_3 . If a country prohibited private trade, its military rival could benefit by traders changing their trading place.

The tribute system played an important role in Korea's long peaceful relation with China. After Silla repulsed the Tang invasion to bring the entire Korean peninsula under its imperial control in 676, no state situated in the center of China invaded Korea, and Ming China even dispatched a large armed force to aid Korea against a Japanese invasion. The tribute system was devised to take total benefit into consideration; it satisfied both diplomatic and economic needs of east Asian countries.

The tribute system prevented Korea and China from reaping dynamic gains from the development of trade, because they limited private trade. This is an important factor to explain why east Asia lagged behind western Europe in the early modern period. How could this undesirable result have come about? First, because of bounded rationality and path dependence, the Chinese and Koreans could not adjust institutions rapidly to maximize dynamic efficiency in altered circumstances. Second, rulers internalized externalities more than needed to maximize social welfare, which made things worse.

In retrospect, the Chinese beginning with the Han Dynasty and the Koreans beginning with the Unified Silla period chose peace and stability, incurring the opportunity cost of dynamic economic development, whereas Europeans after the fall of Rome chose competition and dynamic economic development, incurring the

opportunity cost of less peace and stability (FAIRBANK, REISCHAUER, AND CRAIG [1978, p. 151]). The last phase was worse for Asians, but the overall performance is more complex. Asians and Europeans both acted rationally, but with bounded rationality. The difference in histories between them reflected mainly their different geopolitical situations. The most talented Chinese who devised the tribute system could not have imagined that it would last for such a long time and become a fetter on dynamic economic development, causing China to lag behind the Western barbarians.

References

- ACEMOGLU, D., S. JOHNSON, AND J. ROBINSON [2005], "The Rise of Europe: Atlantic Trade, Institutional Change and Economic Growth," *The American Economic Review*, 95, 546–579.
- ARANO, Y. [1988], *Kinsei Nihon to Higashi Ajia (Japan and East Asia in Early Modern)*, Tokyo Daigaku Suppankai: Tokyo.
- ARIELY, D. [2008], *Predictably Irrational: The Hidden Forces that Shape our Decisions*, HarperCollins Publishers: New York.
- CHŎNG, IN-JI et al. (eds.) [1973], *Koryŏ sa (History of Koryŏ Dynasty)*, Vol. 1, Asea nunhwasa: Seoul.
- CHUN, HAE-JONG [1970], *Hanjung kwangaesa yŏn'gu (Studies in the History of Sino-Korean Relations)*, Ilchogak: Seoul.
- [1971], "Ch'ŏngdae Hanjung Kwangae ŭi Ilgochal (The Change of Qing Attitudes toward Korea)," *Dongyanghak*, 1, 229–245.
- DAVID, P. A. [1985], "Clio and the Economics of QWERTY," *The American Economic Review*, 75, 332–337.
- DENG, K. G. [1997], *Chinese Maritime Activities and Socioeconomic Development, c. 2100 B.C.–1900 A.D.*, Greenwood Press: Westport, CT.
- DIAMOND, J. [1997], *Guns, Germs and Steel: The Fate of Human Societies*, Norton: New York.
- FAIRBANK, J. K. [1953], *Trade and Diplomacy on the China Coast*, Harvard University Press: Cambridge.
- [1968], "A Preliminary Framework," pp. 1–19 in: FAIRBANK (ed.) [1968].
- (ed.) [1968], *The Chinese World Order: Traditional China's Foreign Relations*, Harvard University Press: Cambridge, MA.
- , E. O. REISCHAUER, AND A. M. CRAIG [1978], *East Asia, Tradition & Transformation*, Houghton Mifflin: Boston.
- FLETCHER, J. F. [1968], "China and Central Asia, 1368–1884," pp. 206–224 in: FAIRBANK (ed.) [1968].
- FRANK, A. G. [1998], *ReORIENT: Global Economy in the Asian Age*, University of California Press: Berkeley.
- FRANKE, H., AND D. TWITCHETT [1994], "Introduction," pp. 1–42 in: H. Franke and D. Twitchett (eds.), *The Cambridge History of China, Vol. 6: Alien Regimes and Border States*, Cambridge University Press: Cambridge.
- GOODY, J. [1996], *The East in the West*, Cambridge University Press: Cambridge.
- GROSSMAN, G. M., AND E. HELPMAN [2002], *Interest Groups and Trade Policy*, Princeton University Press: Princeton, NJ.
- HINO, K. [1968], "Kokusai Koryu Shi Jukara Mita Man-sen no Orimono (Silk Fabrics in Manchuria and Korea as seen from the History of International Exchange)," *Chōsen Gakuhō*, 48, 239–257.

- [1972], “Kokusai Koryu Shi Jukara Mita Man-sen no Orimono (Silk Fabrics in Manchuria and Korea as seen from the History of International Exchange),” *Chōsen Gakuhō*, 63, 97–128.
- KANG, DOK-SANG [1962], “Yissi Chosen Kaikoku Chyokugo ni okeru Cho-Ni Boeki no Tenkai (Trade between Korea and Japan after the Port-Opening of Yi Dynasty),” *Rek-ishigaku Kenkyu*, 265, 1–18.
- KIM, NAK-NYEON (ed.) [2006], *Hankuk ūi Gyōngje Sōngjang (Economic Growth in Korea) 1910–1945*, Seoul National University Press: Seoul.
- KIM, SANGGI [1934], “Kodae ūi Muyōk Hyōngtae-wa Namal ūi Haesang Paljōn-e Chuihaya (Status of Ancient Trade and Maritime Expansion of the Late Period of Silla Dynasty),” *Chin-tan Hakpo*, 1, 86–112.
- [1935], “Kodae ūi Muyōk Hyōngtae-wa Namal ūi Haesang Paljōn-e Chuihaya (Status of Ancient Trade and Maritime Expansion of the Late Period of Silla Dynasty),” *Chin-tan Hakpo*, 2, 115–133.
- KWON, TAI-HWAN, AND YONG-HA SHIN [1977], “Chosōn Wangjo Sidae Ingu Ch’ujōng-e kwanhan yōn’gu (On Population Estimates of the Yi Dynasty, 1392–1910),” *Dong-A Mun-Hwa*, 14, 289–330.
- LANDES, D. S. [1972], *The Unbound Prometheus*, Cambridge University Press: Cambridge.
- [1998], *The Wealth and Poverty of Nations*, W. W. Norton & Company: New York.
- LEE, HUN-CHANG [2003], “Chosōn Chunghuki Silhakja ūi Haero Muyōk Yuksongron (A Study on Maritime Trade Promotion Policies Proposed by ‘Practical Learning’ in Chosōn Dynasty),” pp. 227–265 in: Sōnggo Yi Sōngmu Kyosu Jōngnyōn Ginyōn Non-chong Ganhengwuwonhwoi (eds.), *Chosōn Sidae ūi Sasang-kwa Munhwa (Ideology and Culture in Chosōn Dynasty)*, Chipmundang: Seoul.
- [2004], “Hanguk Chōnkūndae Muyōk ūi Yuhyōng-gwa gū Pyōndong-e kwanhan Yōn’gu (A Study on the Pattern and Change of the Pre-Industrial Korean Trade),” *Kyngje Sahak (Review of Economic History)*, 36, 83–122.
- AND P. TEMIN [2005], “Trade Policies in China under the Tribute System as Bounded Rationality,” pp. 279–311 in: T. Y. Wang, K. Xu, and M. Wan (eds.), *Zheng He Yuanhang yu Shijie Wenming (Zheng He Voyages and World Civilization)*, Peking University Press: Beijing.
- LI, JIN-MING [1990], *Mingdai Haiwai Maoyishi (History of Foreign Trade in Ming Dynasty)*, Zhongguo shehuikexue chubanshe: Beijing.
- MADDISON, A. [2001], *The World Economy: A Millennial Perspective*, OECD Publishing: Paris.
- MANGUIN, P. [1993], “Trading Ships of the South China Sea,” *Journal of the Economic and Social History of the Orient*, 36, 253–280.
- MILNER, H. V. [1997], *Interests, Institutions, and Information: Domestic Politics and International Relations*, Princeton University Press: Princeton, NJ.
- MOKYR, J. [1990], *The Lever of Riches*, Oxford University Press: New York.
- NAM, DO-YOUNG [1960], Ryōmal sōncho Majyōngsang-ūro pon Taemyōng kwange (Relations between Koryō and Ming, as Viewed from their Policies on Horses),” *Dong Kook Sa Hak*, 6, 26–74.
- NEEDHAM, J. [1981], *Science in Traditional China and the West*, Harvard University Press: Cambridge, MA.
- OKUMURA, S. [1979], “Koryō ni okeru Palgwanhwaeteki chīzyo to kokusai kankyo (The Order of Palgwanhwa and International Environment in Koryō Dynasty),” *Chosen-shikenkyukai*, 16, 71–99.
- PAK, CHE-GA [1971], *Bukhak ūi (Discourse on Northern Learning)*, Ŭlyu nunhwas: Seoul.
- REISCHAUER, E. O. [1955], *Ennin’s Travels in Tang China*, Ronald Press Company: New York.
- ROBINSON, K. R. [1992], “From Raiders to Traders: Border Security and Border Control in Early Chosōn, 1392–1450,” *Korean Studies*, 16, 94–115.

- SERRUYS, H. [1975], *Sino-Mongol Relations during the Ming III: The Trade Relations: The Horse Fairs (1400–1600)*, Institut Belge des Hautes Études Chinoises: Brussels.
- SHIBA, Y. [1983], “Song Foreign Trade: Its Scope and Organization,” pp. 89–115 in: M. Rossabi (ed.), *China among Equals: The Middle Kingdom and its Neighbors, 10th–14th Centuries*, University of California Press: Berkeley.
- [1992], “Koshiron (On Port Market),” pp. 1–34 in: Y. Arano, M. Ishii, and S. Murai (eds.), *Ajia no naka no Nihonshi (Japanese History among Asia) III*, Tokyo Daigaku Suppankai: Tokyo.
- SIMON, H. A. [1982], *Models of Bounded Rationality: Behavioral Economics and Business Organization*, Harvard University Press: Cambridge, MA.
- SUGAWA, H. [1997], “Koryō koki ni okeru shogyo seisaku no Tenkai (The Development of Commercial Policy during the Late Koryo Period),” *Chosen bunkakenkyushitsu kiyō*, 4, 25–45.
- TANAKA, T. [1975], *Chusei taikai kankeishi (History of Foreign Relations in Middle Ages)*, Tokyo Daigaku Suppankai: Tokyo.
- TASHIRO, K. [1981], *Kinsei Nichō tsūko boeki shi no kenkyū (Studies on the History of the Diplomatic Relations and Trade between Japan and Korea in 17th and 18th Centuries)*, Sobun Sha: Tokyo.
- WANG, GUNGWU [2000], *The Chinese Overseas: From Earthbound China to the Quest for Autonomy*, Harvard University Press: Cambridge, MA.
- WILLS JR., J. E. [1968], “Qing Relations with the Dutch,” pp. 225–256 in: FAIRBANK (ed.) [1968].
- YI, HYUN-HAE [1998], *Hankuk Kodae ūi Saengsan-kwa Kyōyōk (Agricultural Production and Trade in Ancient Korea)*, Ilchokak: Seoul.
- YI, I [1986], *Yulgok chōnsō (Collection of I Yi’s Writings)*, 38 vols., Sōnggyun’gwan Taehakkyo Taedong Munhwa yōn’guwōn: Seoul.
- YÜ, YING-SHIH [1967], *Trade and Expansion in Han China: A Study in the Structure of Sino-Barbarian Economic Relations*, University of California Press: Berkeley.
- YUN, MYEONG-CHEOLL [2002], *Hanminjok ūi Haeyanghwaldong-kwa Tonga Chijunghae (Maritime Activities of Korean Nation and Eastasian Mediterranean)*, Hakyeoun: Seoul.
- YUN, P. I. [1998], “Rethinking the Tribute System: Korean States and Northeast Asian Interstate Relations, 600–1600,” Ph.D. Dissertation, University of California, Los Angeles.

Hun-Chang Lee
 Department of Economics
 Korea University
 Seongbuk-gu, Anam-dong
 Seoul, 136-701
 Korea
 E-mail:
 lee hc@korea.ac.kr

Peter Temin
 Gray Professor Emeritus of Economics
 MIT Department of Economics
 50 Memorial Drive, Room E52-280A
 Cambridge, MA 02142-1347
 U.S.A.
 E-mail:
 ptemin@mit.edu