particular) is believed to account currently for nearly 30 percent of foreign currency supply in the parallel market (Novaes, 1990).

34 The extent to which traders engage in fake invoicing is typically measured by partner country trade-data comparisons. To investigate the scale of under-invoicing or over-invoicing of exports, for instance, one would need to look at the ratio of exports to major partner countries, as shown by domestic data, to the corresponding imports as recorded in partner country data. When this ratio is less than unity, the evidence points to under-invoicing of exports. To be able to make these partner-country comparisons, however, it is important to adjust the trade data for transport costs, timing of transactions, and classification of transactions. See MacDonald (1985), Gulati (1988), and Arslan and van Wijnbergen (1989) for recent attempts to use these procedures to estimate the degree of under- and over-invoicing in foreign trade transactions.

35 See Arslan and van Wijnbergen (1989) for econometric evidence supporting this proportion in the case of Turkey, and Kamin (1991*b*).

36 The welfare effects of foreign exchange restrictions have been analyzed by Greenwood and Kimbrough (1986). Using a choice-theoretic cash-in-advance general equilibrium model, they examine how the imposition of foreign exchange controls affects decision-making by private agents, notably the decision to evade the restrictions by purchasing foreign currency illegally in the parallel market. They show that while foreign exchange controls may improve the trade balance and the balance of payments of an economy with parallel markets, they unambiguously lower economic welfare. This is because foreign exchange controls essentially place a quota on imports, thus raising their domestic relative price in the same manner as a tariff would.

37 Policies of active repression of parallel markets have been attempted by some countries (Guyana in 1980, Tanzania in 1983, or Algeria in May 1990). It has proved difficult to maintain a punitive stance against well-entrenched informal activities.

2

Models of Informal Financial Markets

1 Models of Informal Credit Markets

1.1 McKinnon-Shaw Models of Financial Repression

1.1.1 An Overview of McKinnon-Shaw Models

The first systematic analyses of financial markets in developing countries to take seriously the special characteristics of financial institutions in such countries were by McKinnon (1973) and Shaw (1973).[1] The so-called McKinnon-Shaw school, which coined the term "financial repression" to describe such characteristics, represents the currently dominant strain of thinking about developing-country financial markets, and the forceful criticism of financial repression by economists in the McKinnon-Shaw tradition has provided the intellectual underpinnings for a recent movement toward financial liberalization in many parts of the Third World.

According to McKinnon (1973), the defining feature of underdevelopment is fragmentation – i.e, a situation in which agents face different prices and do not have access to the same technology. This fragmentation has largely been the product of government policy designed to favor certain activities or certain classes of agents at the expense of others. In turn, intervention has often been justified by the pursuit of social goals that are inhibited by the improper functioning of capital markets. In the absence of private finance, public policy has relied on transfers of income to targeted activities. Since fiscal constraints have often precluded direct income transfers through the budget, governments have resorted to less

direct and more inefficient means of redistributing income – that is, by altering the structure of relative prices. A standard example is the protection of infant industries. With a well-functioning capital market, a new activity that is likely to incur short-run losses without public support, but that holds the prospect of yielding large future returns, would be able to cover its short-run losses by borrowing against its anticipated future profits. If the financial means to do so are not available, however, and if government lending or the payment of production subsidies is not feasible, a recourse to trade restrictions may be perceived as a third-best option.

The key problem is the inadequacy of financial intermediation. Because there is no one to intermediate between savers and those with profitable investment opportunities, self-financing is perceived by McKinnon to be the rule in many developing countries. In this setting, initial endowments determine the opportunity to undertake new investments. This presents serious efficiency problems, because there is no mechanism to equalize the return to capital across activities. Such problems are aggravated in the presence of indivisibilities in investment technologies, since individuals with small initial capital endowments who happen to face highly profitable, but large, investment opportunities, will not be able to avail themselves of such opportunities – even over time – by relying on their own resources.

For reasons that essentially amount to the presence of uncertainty, wealth holders in developing countries are assumed by McKinnon to resist holding financial assets other than money. Such financial intermediation as does exist, therefore, is conducted by the banking system. However, the banking system performs inefficiently as a financial intermediary. The reason is "financial repression" – i.e., administratively imposed usury ceilings on loan interest rates and strict collateral requirements, which together force banks to concentrate their portfolios on safe, low-yielding assets. An additional component of financial repression is the imposition of large reserve requirements, generally enacted for fiscal motives (see McKinnon and Mathieson, 1981). For both reasons, banks pay low interest rates to savers.

Financial repression is perceived to have a number of adverse macroeconomic consequences. First, in the developing-country

financial context (in which primary securities are largely absent) money is held as a store of value, rather than simply for transactions purposes. Consequently, the real return on money represents the marginal return to saving, and a low real interest rate on money, consisting of a low nominal interest rate on deposits, often combined with a high rate of inflation, tends to depress private saving. Because the accumulation of money balances and physical investment are complementary, as in McKinnon (1973), the reduction of saving reduces investment in physical capital. Moreover, low administratively-determined interest rates on bank loans ensure that credit will be rationed by non-price means, implying that many low-return projects will be financed, artificially depressing the aggregate rate of return on investment. Financial repression thus reduces both the quantity and quality of new investment, and thereby also the economy's rate of growth. Economists in the McKinnon-Shaw tradition emphasize an additional feedback from low growth to lower saving through a standard life-cycle mechanism, which in turn magnifies the adverse effects described above.

The emergence of informal financial markets to fill the credit gap and serve as an alternative vehicle for saving is acknowledged by the McKinnon-Shaw school, but such markets are perceived as inefficient, limited solutions to the deficiencies of financial intermediation aggravated by financial repression, and are not incorporated into the formal analytical framework.

The policy prescriptions that follow from the McKinnon-Shaw analysis are straightforward: to wit, interest rates should be freed from administrative restrictions, and restrictions on bank lending that take the form of high reserve requirements or excessively stringent collateral requirements should be eased. In short, what is needed is plentiful credit at high real interest rates. Such policies would increase total saving, improve the efficiency of investment, and either because of the complementarity of money and capital or because of the increased availability of credit (discussed below), increase total investment. The disincentive effect of high real interest rates on investment that is featured in industrial-country macroeconomics is not likely to materialize – at least over a fairly broad initial range of real interest rates – because under the credit-rationed regime there are plenty of high-yield investment projects that remain unexploited. The increase in the quantity and quality

of investment that would ensue would of course generate favorable growth effects, producing positive feedbacks by inducing additional saving.

1.1.2 *The Link Between Interest Rates and Investment*

The most controversial point in the McKinnon-Shaw analysis is the link between high real interest rates and private investment. This link was formulated differently in McKinnon (1973) and Shaw (1973). As already indicated, for the former, saving in the form of money (hoarding) and investing in physical capital are perceived as complementary. The reason is that hoarding serves as a "conduit" to accumulate savings, permitting the undertaking of lumpy investment expenditures in an environment of self-finance. A high real interest rate performs two functions. First, it discourages agents from devoting their savings to investment in low-yield projects. These are viewed by McKinnon as alternatives to higher-yield lumpy investments. Second, it increases the income of savers who are accumulating to finance these higher-yield projects, and thus it facilitates the accumulation process. For Shaw, the mechanism linking high interest rates to investment operates through "financial deepening". Unlike McKinnon, Shaw allowed for the possibility of external finance of investment. Higher interest rates offered by the banking system would increase the flow of real resources into banks, referred to as financial deepening, and thus increase the availability of funds for investment.

As Molho (1986) has shown, these mechanisms are mutually consistent, and can be integrated into a single model. However, neither is without difficulty. The McKinnon "complementarity" hypothesis claims to establish a positive relationship between the rate of return on money and the quantity (not just the quality) of investment. However, a high rate of return on money favors future, high-yield investment at the expense of current, low-yield investment. In other words, it induces a substitution of future for current investment. In the aggregate, the effect on total investment would be a washout, except for the possible disincentive to consumption provided by the higher yields available on saving in the

form of financial assets. In short, the McKinnon analysis underlying the complementarity hypothesis would appear to establish a link between real interest rates on money and the *ex ante*, or desired, quantity of saving, as well as between the real loan interest rate and the quality of aggregate investment, rather than between the real return on money and the quantity of investment.

As is well known, positive effects of increases in real interest rates on saving depend on the dominance of substitution over income effects. Moreover, the link between higher real deposit rates and *ex ante* private saving is explicitly attributed by McKinnon to the role of money as a store of value in the absence of other financial assets. This link becomes problematic if the existence of alternative stores of value, such as may be provided by informal markets, relegates money to a transactions role.

Even if an increase in the real deposit interest rate increased *ex ante* private saving, however, whether aggregate investment would prove to be higher *ex post* depends on general equilibrium interactions. One argument for a direct one-for-one conversion of desired saving into realized investment is provided by Galbis (1977) and Fry (1982), and is described in figure 2.1 (adapted from Fry, 1982). Positing a saving schedule SS that is a positive function of the real formal interest rate, and an investment schedule II with a negative slope, they argue that, while in a neoclassical model the equilibrium interest rate r^* would be determined by the point of intersection of the schedules SS and II, under financial repression an interest rate ceiling below the market-clearing level, such as at r_1, ensures that there will be an excess demand for loanable funds, given by $(I_1 - S_1)$. Under these circumstances the *ex post* quantity of investment will be determined by the short side of the market – i.e., it will be equal to the *ex ante* supply of saving S_1. In this context, any increase in the desired volume of saving, such as would be brought about by a higher real formal interest rate at r_2, would be translated one-for-one into increased realized investment. This would be the case as long as the interest rate is subject to a ceiling below its market clearing level, since for interest rates below r^* *ex post* investment would simply be tracing out *ex ante* saving along the locus SS.

While the foregoing is the most widely accepted mechanism linking higher real interest rates to increased investment in financially

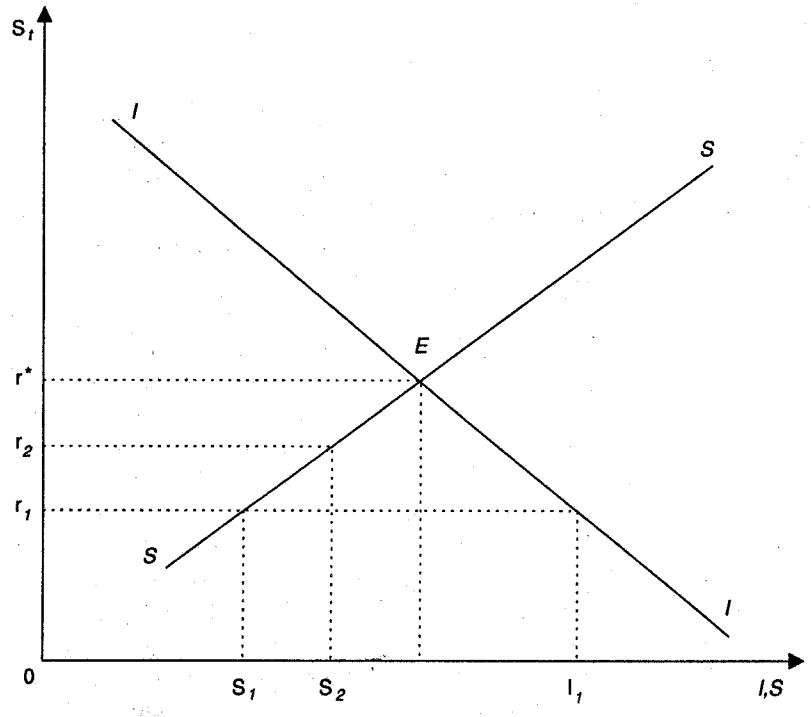

Figure 2.1 Determination of realized investment under financial repression

repressed economies, it suffers from serious logical flaws. In any realistic model of saving and investment in such countries, the *SS* and *II* schedules would both depend on a host of other variables in addition to the real formal interest rate, some of which – including the interest rate in the informal sector – will be endogenous to the general-equilibrium system and will therefore in general be affected by changes in administered interest rates. Adjustments in these non-interest determinants of saving and investment will cause shifts in the saving and investment schedules, so the *ex post* level of investment cannot be simply read off of the *ex ante* saving schedule. The familiar saving-investment diagram, therefore, is an incomplete partial-equilibrium analysis which may not necessarily provide a useful approximation to the

effects of changes in controlled interest rates on realized aggregate investment.

With regard to Shaw's financial deepening mechanism, the effects of changes in administered deposit rates on the quantity of investment depend, of course, on the extent to which resources drawn into the financial system are themselves either directly drawn out of investment (i.e., through substitution of hoarding for low-yield investment, as in the discussion of McKinnon above) or reduce investment through indirect means. The latter (as we shall see in section 1.2), may indeed occur in the presence of informal credit markets. In addition, the effectiveness of "financial deepening" in increasing total investment depends on the disposition of banks' newly-acquired resources. In a credit-rationed environment, increased saving by some households may be channeled by the financial system into increased consumption by others. The outcome depends on a comparison between the rate of time preference among households who would like to shift consumption from the future to the present, but are prevented from doing so by credit rationing under financial repression, and the marginal product of capital.[2] More fundamentally, for the nation as a whole, total investment will rise when the rate of return on money increases only if total saving rises and/or foreign investment by domestic residents decreases as a result. This depends on the extent to which an increase in the rate of return on money increases the marginal return to saving or affects the margin between sending funds abroad versus keeping them at home. These issues can only be addressed in the context of a full general-equilibrium model. In this case, it must be a model that does justice to the role of informal financial markets in the context of financial repression.

1.1.3 Short-run Stabilization Under Financial Repression

The discussion up to this point does not properly address the issue of short-run macroeconomic stabilization in a financially repressed economy. What is needed for this purpose is a short-run macroeconomic model that incorporates financial repression. McKinnon addressed the issue of inflation stabilization, arguing

(see McKinnon, 1973, chapter 7) that the traditional orthodox "tight credit" approach to combating inflation involved unnecessary output costs because it aggravated financial repression, creating adverse supply-side effects on output through the role of credit in financing working capital. His preferred strategy to combat inflation is to increase the deposit rate of interest while restricting the rate of monetary expansion. This increases the real return on money and thereby the real demand for money, exerting downward pressure on the price level through this means as well as by expanding output. The latter follows from an increased supply of credit for working capital. This argument, however, was not developed in the context of a formal model.

Formal dynamic general equilibrium models designed to study stabilization in the context of financial repression have, however, been developed by a number of later writers in the McKinnon-Shaw tradition. Comprehensive overviews of such models have recently been provided by Fry (1982, 1988), who singles out the contributions of Kapur (1976), Mathieson (1980), and Galbis (1977), as well as a more detailed model of his own.

Kapur (1976) links the level of output to the stock of working capital in fixed-coefficients Harrod-Domar fashion, and assumes that banks finance all net additions to working capital as well as a fraction of any nominal increases in the stock of working capital required to offset the effects of inflation. All bank lending is devoted to the financing of working capital. Since both the public's currency/deposit ratio and the banks' reserve ratio are fixed, the real flow of credit for working capital, $\dot{L}/P$, is given by:

$$\dot{L}/P = \hat{M}q(M/P), \tag{2.1}$$

where $\hat{M}$ is the rate of growth of high-powered money, taken to be a policy variable, q is the loan to money ratio, and M/P is the real money stock. An increase in the policy-determined nominal rate of return on deposits increases the demand for money. Since Kapur expresses the rate of inflation as a function of the stock excess supply of money, the rate of inflation falls on impact. This reduces the share of real credit expansion required to offset inflationary erosion of the stock of working capital and thus increases the rate of expansion of working capital. Consequently, the rate of growth of real output rises.

Mathieson's (1980) model follows that of Kapur fairly closely, except that the "disequilibrium money" inflation function used by Kapur is replaced by an investment function which expresses the share of investment in output as a function of the difference between the rate of return on physical capital and the real loan rate. In addition, because Mathieson's main concern is with policies that can be undertaken to sustain the solvency of the financial system upon liberalization, he does not impose a zero-profit condition tying the loan rate to the deposit rate. Instead, the two interest rates are treated as separate policy instruments that can be manipulated by the authorities to move controlled interest rates toward market-clearing levels while subsidizing financial institutions with a large share of their portfolios in previously-contracted low-interest loans.

In the Kapur-Mathieson analytical framework, an increase in the administered deposit rate increases the flow of real resources to the formal banking system. These are then channeled into additional investment, which has a constant marginal product. There is no Keynesian problem of deficient aggregate demand, because reduced consumption is offset by increased investment on a one-for-one basis. The rate of growth of output increases not because more productive investment is undertaken, as in McKinnon, but because the mix of aggregate demand changes, increasing the share of investment – i.e., the quantity, not the quality of investment rises.

By contrast, Galbis (1977) develops a model which holds the saving rate constant, and examines the implications for growth of improving the quality of aggregate investment by raising the deposit rate. Galbis' model contains two sectors: a traditional sector in which capital has a low (and constant) marginal product, and which relies on self-finance. The rate of investment in this sector depends positively on the marginal product of capital and negatively on the real deposit rate, thereby embodying the contemporaneous substitutability between money and capital alluded to above. The second sector has a high marginal product of capital and relies on external finance. For this sector investment depends positively on the marginal product of capital and negatively on the real loan rate. Both sectors have constant saving ratios out of income. Thus the real flow of deposits from the first sector is the difference between the sector's saving and investment, and is

therefore a positive function of the real deposit rate. This real flow of deposits determines the supply of loans available to the high-productivity sector. The demand for loans by the latter is in turn equal to the difference between its desired investment and saving. The real loan interest rate adjusts to clear the flow loan market. In this setting, an increase in the deposit interest rate causes the traditional sector to reduce investment and increase deposit accumulation. Thus, the supply of loans to the high-productivity sector increases, the real loan interest rate falls, and investment in the high-productivity sector rises. This improvement in the quality of investment raises the aggregate growth rate, without changing the aggregate saving rate.

Although each of the papers described above focuses on the macroeconomic implications of raising the artificially-controlled deposit rate in a repressed economy, advocates of the McKinnon-Shaw framework have also considered the effects of other policy tools under financial repression. McKinnon and Mathieson (1981), for example, examine the steady-state effects of raising deposit rates, but also of altering required ratios, lifting capital controls, and pegging the nominal, rather than the real, exchange rate. They argue that financial repression reflects a need to finance fiscal deficits through seignorage. In this context, maintaining ceilings on deposit interest rates simply raises the steady-state inflation rate, because by discouraging the holding of deposits it shrinks the base for the inflation tax. Liberalizing interest rates thus permits the government to finance a given deficit with a lower steady-state inflation rate. Given such liberalization, raising required reserve ratios may lower the steady-state inflation rate over some range, but will eventually increase it by lowering returns to depositors and thus again shrinking the base of the inflation tax. Freeing capital flows is perceived to be a mistake, for similar reasons. The option to move assets abroad is likely to be exercised by private agents under financial repression, thereby reducing the demand for base money. In fact, this reasoning leads McKinnon and Mathieson to advocate the prohibition of foreign currency holdings by domestic residents, such as would arise in parallel currency markets. Finally, with regard to management of the nominal exchange rate, the need to finance a continuing fiscal deficit through seignorage precludes the adoption of a fixed nominal exchange rate, while the reduced

size of the financial system under financial repression ensures that a free market for foreign exchange would be thin and unstable. Thus, McKinnon and Mathieson endorse real exchange rate targeting for financially repressed economies.

Analytical work in the McKinnon-Shaw framework has thus addressed the macroeconomic effects of a number of financial policy instruments in an economy which contains no large, well-organized formal securities markets, relying primarily on the banking system for formal financial intermediation, but simultaneously maintaining ceilings on deposit and/or loan interest rates, high reserve requirements, and controls on international capital movements. As such, analysts in this tradition have tried to be faithful to developing-country financial realities, at least insofar as the formal financial sector is concerned. The key policy prescription of this school is to increase controlled interest rates in the formal sector, and the claim is that such a policy would have favorable macroeconomic effects by increasing the quantity of saving and investment, as well as by improving the quality of aggregate investment in repressed developing economies. The analysis that underlies this prescription, however, is often partial equilibrium in nature. More importantly for our purposes, though informal markets for credit and foreign exchange are acknowledged by the McKinnon-Shaw school as likely to arise under financial repression, the formal dynamic general equilibrium models that have been developed along McKinnon-Shaw lines have not integrated the roles that these markets may play in mediating the effects of financial policies in repressed economies.

1.2 Neo-Structuralist Models

1.2.1 An Overview

An alternative perspective on developing-country financial markets is provided by the Neo-Structuralist school. Like proponents of the McKinnon-Shaw framework, Neo-Structuralists take the view that developing-country macroeconomics differs in fundamental ways from the standard textbook industrial country model.[3] Moreover, several structural features of the macroeconomic models used by

proponents of the Neo-Structuralist framework are held in common with the McKinnon-Shaw framework. These include the specification of the formal financial sector as consisting essentially of the banking system, and the role of working capital in determining the supply-side response of the economy to changes in monetary policy instruments, as in the Kapur-Mathieson model.[4] Of particular interest for the present purpose, however, is that a key point of divergence between the Neo-Structuralist and McKinnon-Shaw frameworks, is precisely the treatment of informal credit markets.[5] As indicated above, the existence of such markets is acknowledged in the McKinnon-Shaw analysis, but they are perceived as inefficient, limited in scope, and not central to the transmission of shocks from the financial to the real sector. By contrast, informal credit markets occupy a prominent place in Neo-Structuralist models. Such markets play two important roles in these models. In commodity markets, they are taken to determine the marginal cost of funds relevant for private spending and saving decisions. In financial markets, they represent an alternative mode of financial intermediation available to private savers.

The argument for according an important commodity-market role to the interest rate prevailing in informal credit markets is straightforward. It consists of the observation that, when interest rates in the formal market are administered at below market-clearing levels, agents cannot borrow and lend without limit at such interest rates.[6] The marginal cost of funds, therefore, is given by the curb market rate. Van Wijnbergen (1982) provides empirical evidence for this role of the curb market rate in the case of Korea, where the real curb rate turns out to have significant explanatory power in estimated consumption and investment functions.

The omission of informal credit markets also provides the basis for a Neo-Structuralist critique of the McKinnon-Shaw analysis of the financial sector in repressed economies. It is argued by van Wijnbergen (1983*b*), for example, that the results obtained in the McKinnon-Shaw framework are based on the hidden assumption that the inflow of resources into the banking system induced by an increase in deposit interest rates comes out of "unproductive" assets like cash or gold, rather than out of productive physical capital or out of other assets that provide intermediation between savers and investors. One such asset that is widely available to savers in the

developing world is, of course, lending in the informal market.

The role of such loans is modeled in standard Tobin portfolio fashion. To take the example of van Wijnbergen (1983*b*), household financial wealth (A) consists of three assets: currency (CC), bank deposits (D^p), and curb market loans (L^p):

$$A = CC + D^p + L^p. \tag{2.2}$$

The asset demand functions in this model are given by:

$$CC = f^c(-\pi, r_L, r_d, y)A, \tag{2.3}$$

$$D^p = f^d(-\pi, r_L, r_d, y)A, \tag{2.4}$$

$$L^p = f^L(-\pi, r_L, r_d, y)A, \tag{2.5}$$

where π is the expected rate of inflation, r_L is the real rate of return on curb market loans, r_d is the real rate of return on deposits, and y is the level of real GDP. The three assets in the model are assumed to be gross substitutes – i.e., partial derivatives with respect to their own rate of return are positive, while those with respect to the real rate of return on competing assets are negative. An increase in real GDP increases the demand for currency and deposits, but reduces the demand for curb market loans. The usual portfolio adding-up constraints apply to these partial derivatives. Banks are subject to a required reserve ratio of μ, and loan out a fraction b of their remaining funds. This fraction depends directly on the exogenous bank loan interest rate r_b:

$$L^b = b(r_b)(1 - \mu)D^p. \tag{2.6}$$

Loans from banks and from the curb market are perfect substitutes from the point of view of firms, who borrow to finance real working capital, which in turn is an increasing function of the real product wage w and the level of output:

$$L^d = L^d(w, y). \tag{2.7}$$

The financial market sector of van Wijnbergen's model is completed by the equilibrium condition requiring equality between the total (formal and informal) supply of loans and the demand for loans:

$$L^p + b(1 - \mu)D^p = L^d. \tag{2.8}$$

Equation (2.8) captures the essence of the Neo-Structuralist critique of the McKinnon-Shaw financial framework. An increase in the administered deposit interest rate r_d will increase the demand for deposits D^p, because the partial derivative of D^p with respect to r_d is positive, according to equation (2.4). From (2.8), this indeed increases the supply of loans, as in the Kapur-Mathieson model. But whether the total supply of loans to finance working capital, given by the left-hand side of (2.8), increases or not depends on several factors. First, it depends on the magnitude of the induced reduction in L^p. Using equation (2.5), this is given by the partial derivative of f^L with respect to r_d, which is determined by the degree of substitutability between loans and deposits. If loans are closer substitutes than currency for deposits, the partial of f^L with respect to r_d will be large and L^p will fall. Notice that the reduction in L^p will reduce the supply of loans to finance working capital on a one-for-one basis, since there are no "leakages" out of the informal loan market into other forms of asset holding by intermediaries in this market.

The formal banking system, on the other hand, is taken to provide inferior intermediation, so that each unit increase in D^p does not result in an additional unit of resources available for the financing of working capital. This is so for three reasons. First, the formal sector is subject to legal reserve requirements, whereas the informal sector is unregulated. This accounts for the coefficient $(1 - \mu)$ on D^p. Second, the formal banking system is also assumed to hold free reserves, whereas informal credit market institutions do not, accounting for the presence of the factor b multiplying $(1 - \mu)D^p$ in (2.8). Together, these factors ensure than an additional unit of deposits leads to only $b(1 - \mu)$ units of additional lending. Finally, the formal banking system is often subject to limits on aggregate credit expansion for stabilization purposes. In terms of equation (2.8), if such a constraint were binding, b would fall to offset any increases in D^p, leaving $b(1 - \mu)D^p$ unchanged, in the face of an increase in r_d. In this last case, the effect of an increase in r_d on the supply of loans would be unambiguously negative, amounting to a reduction in L^p without an offsetting increase in $b(1 - \mu)D^p$.

1.2.2 *Critical Assumptions in the Financial Sector*

In this setting, then, an increase in deposit interest rates offered by the formal banking system could actually decrease the supply of loans available to investors. The reason is that the transfer of one unit of saving from the informal credit market, where it would have been available to finance one unit of investment, to the formal financial system, where it would finance at most only $b(1 - \mu)$ units of investment, would actually diminish the efficiency of financial intermediation in the economy. Whether this is indeed the outcome depends on the extent to which funds drawn into the banking system come out of curb market lending, rather than out of "unproductive" assets, as well as on the extent to which banks can expand total credit when they receive an increase in their deposits.

A key empirical question, then, is whether bank deposits are closer substitutes in household portfolios for curb market lending or for "unproductive" assets. Typically such assets are represented by cash in Neo-Structuralist models, as is the case in the van Wijnbergen model described above. *Ex ante*, one might expect that, since bank deposits are likely to be more liquid than curb market loans, and since the risk attributes of cash and deposits are likely to be more similar than those of deposits and curb loans, cash would be a much closer substitute for deposits than would curb loans. In this case, an increase in the deposit interest rate would draw resources primarily out of cash hoards, rather than out of curb lending. This effect could be sufficiently strong as to offset the role of bank reserves in syphoning resources out of the lending stream, thereby increasing the availability of resources for investment, in spite of the relative inefficiency of the formal banking system as a financial intermediary under financial repression.[7] If so, the Neo-Structuralist critique of the McKinnon-Shaw analysis of the effects of raising deposit rates, while theoretically correct, need not be empirically important. However, somewhat surprisingly, Van Wijnbergen (1982) finds, in the case of Korea, that curb market loans are closer substitutes for time deposits than are cash. Unfortunately, little evidence is available on this issue from other countries, and Korea may not constitute a representative case.

A second key assumption in the analysis just described is that, while institutions in the formal financial system hold substantial reserves, thereby providing imperfect intermediation between savers and investors, institutions in the informal market do not, and are thus able to convert their liabilities into loans on a one-for-one basis. Notice that, if this asymmetry were not present – i.e., if the factor $b(1-\mu)$ applied equally to both types of intermediaries – the existence of an "unproductive" third asset would imply, via the adding-up constraint and the gross substitutes assumption, that an increase in the deposit interest rate would indeed increase the supply of loans.

Owen and Solis-Fallas (1989) have recently questioned whether this assumed asymmetry is plausible, citing in particular the possibility that the value of the loans extended for "productive" purposes by financial intermediaries in the informal credit market will fall short of the value of such loans in households' portfolios, due to the use of informal credit for consumption purposes and/or the existence of significant transactions costs in such markets. While neither of these phenomena can invalidate the Neo-Structuralist analysis,[8] it is indeed legitimate to ask why informal credit market institutions do not hold reserves equal in magnitude to those held by institutions in the formal sector. One straightforward answer, of course, is provided by the McKinnon-Shaw school – i.e., financial repression. The latter consists not only of the ceilings placed on formal-market interest rates, but also of the imposition of large reserve requirements on the financial institutions within the regulatory reach of the authorities – those in the formal sector. In part, then, the large reserves held by banks are due to such reserve requirements which, as indicated previously, fulfill the function of helping to finance fiscal deficits by increasing the revenue from seignorage. By definition, institutions in the informal sector are not subject to such requirements. Moreover, since the liabilities of intermediaries operating in the informal credit market are less liquid than those of banks, the demand for free reserves by such institutions can also be expected to be less than those of banks. On both counts, then, the Neo-Structuralist assumption that the informal market provides more direct intermediation than the the formal banking system in a repressed economy may well be justified.

The final assumption to be considered relates to the composition

of household portfolios and its implications for the nature of the transmission mechanism linking the financial and real sectors. In particular, while Neo-Structuralist models acknowledge the existence of "inflation hedges" such as real estate or foreign currency as potential "unproductive" assets, they do not generally examine the implications of such assets either for the effects of changes in controlled interest rates on loan availability in particular or, more broadly, for the transmission mechanism linking financial policy variables with the real sector in repressed economies.[9]

To examine the potential role that such assets could play, consider the effects of increasing administered deposit rates in a model which incorporates a third asset in the form of an "inflation hedge", rather than currency. Such an asset differs from currency in that its nominal price is endogenous. The nominal return from holding the asset consists of the expected rate of change in its nominal price. In a developing-country financial setting such assets could consist of real estate, precious metals, or foreign exchange. An increase in the interest rate on deposits would lead households to substitute away from both curb market loans and inflation hedges. In the context of a Tobin-type portfolio model, the latter are likely to be less close substitutes for deposits than is currency, both because inflation hedges are less liquid than currency and deposits, as well as because the real returns on inflation hedges are not likely to be highly correlated with those on currency and deposits in the face of variable inflation, especially when nominal interest rates on deposits are controlled. Since an increase in deposit rates is more likely to attract funds out of the curb loan market when the third asset is not a close substitute for deposits, replacing currency with an inflation hedge in the role of third asset increases the likelihood that an increase in the deposit rate decreases the supply of curb-market loans.

In addition, however, the financial market repercussions of increasing deposit rates are much more complex when an inflation hedge replaces currency as the third asset in household portfolios. Since the inflation hedge has a variable nominal price and nominal rate of return, the portfolio reallocations triggered by the change in deposit rates will in general alter both the contemporaneous and future price of the inflation hedge. The effects on the loan interest rate, which is typically the variable of interest (as representing the

marginal cost of funds) in Neo-Structuralist models will be determined simultaneously with changes in the current and future prices of the inflation hedge, and thus cannot be inferred simply from the *ceteris paribus* effect of the change in the deposit rate on the excess supply of loans.

Moreover, in such a setting it is no longer possible to single out the interest rate in the informal credit market as the marginal cost of funds relevant for saving and investing decisions, and thus as the sole mechanism for transmission of monetary policy changes from the financial to the real sector. Since inflation hedges represent an alternative store of value, the rate of return on such assets will also be relevant for those decisions, and changes in the real value of the stock of the asset held as an inflation hedge will in general affect the behavior of private saving.

These considerations turn out to be of particular relevance in the context of financial repression. As pointed out earlier, capital controls tend to be an important component of financial repression in developing countries (recall that such controls are in fact advocated as part of the proper management of a repressed economy by McKinnon and Mathieson). In the presence of such controls, parallel currency markets tend to arise in which foreign exchange is traded at market-determined prices in well-organized markets that behave in many ways very much like traditional asset markets in industrial countries.[10] Foreign exchange thus becomes an ideal inflation hedge, and parallel currency markets have to be incorporated into an adequate treatment of informal credit markets in developing countries.

2 Models of Parallel Currency Markets

Parallel markets for foreign exchange have been analyzed and modeled from a number of different perspectives. "Real trade" models have emphasized the effect of trade taxes on smuggling activities and illegal currency transactions. The monetary approach has highlighted the role of money market disequilibria in determining the behavior of parallel exchange rates. The portfolio/currency substitution framework stresses the role of asset composition in the determination of supply and demand for foreign exchange. Finally, recent developments in the analysis of models of formal two-tier exchange rate systems has focused on inter-market leakages and yield predictions which are qualitatively similar to those derived from currency substitution models. This section reviews these alternative approaches and highlights their major implications.[11]

2.1 Smuggling and Real Trade Models

Following the early partial equilibrium analyses of Boulding (1937), Bronfenbrenner (1947) and Michaely (1954) of a consumption commodity market subject to price control and rationing, "real trade" models of the determination of the parallel market premium on foreign exchange focus on the parallel market for foreign currency itself and neglect its interactions with the rest of the economy. Specifically, the market is modeled as reflecting the demand for foreign currency to purchase illegal imports and the supply of foreign currency derived from illegal sources. Martin and Panagariya (1984), McDermott (1989), Sheikh (1976), and Pitt (1984), for instance, emphasize the role of smuggling and underinvoicing of exports as the main sources of foreign exchange supply, whereas Culbertson (1975) stresses the resale of officially allocated foreign exchange.

To examine the implications of this class of models, consider a small, open economy, in which domestic importers and exporters face given world prices.[12] In each period, both agents must choose the amount of goods to be transacted through "official" channels, and the amount to smuggle in (for importers) or out (for exporters). Both categories of agents face a government customs agency whose purpose is to enforce tax laws at the borders. "Checks" are conducted randomly by the agency, because of prohibitive administrative costs. When an offender is caught, he must pay a fine, or go to jail. In what follows, we will assume that the penalty imposed is confiscation of a portion of the stock of smuggled goods.

Consider first the representative importer. Formally, let us denote by q_m the quantity of the good to be imported through legal means, and let $\tilde{q}_m$ be the quantity to be smuggled in through illegal channels. By transacting through the official market, the

importer has access to foreign exchange at the (fixed) official exchange rate of $\bar{e}$, but must pay an *ad valorem* tariff given by $(\tau_m - 1)p_m^*$, where p_m^* denotes the world price. If the importer chooses to transact through the parallel market, he must acquire foreign exchange at a more depreciated rate denoted s, and face a probability π of being caught by the customs agency, in which case a portion θ of the amount of goods smuggled is confiscated. Denoting the domestic price of imports by p_m, expected profits Σ_m can be written as a weighted average of profits that would be realized in the two alternative states of nature:

$$\sum_m = \pi[p_m\{q_m + (1-\Theta)\tilde{q}_m\} - \bar{e}\tau_m p_m^* q_m - sp_m^*\tilde{q}_m] + (1-\pi)[p_m(q_m + \tilde{q}_m) - \bar{e}\tau_m p_m^* q_m - sp_m^*\tilde{q}_m], \qquad 0 < \Theta < 1$$

or, rearranging terms,

$$\sum_m = p_m[q_m + (1 - \Theta\pi)\tilde{q}_m] - p_m^*(\bar{e}\tau_m q_m + s\tilde{q}_m). \quad (2.9)$$

The detection technology available to the tax enforcement agency is such that the probability of getting caught is an increasing function of the smuggling ratio, $\delta_m \equiv \tilde{q}_m/q_m$:

$$\pi = \pi(\delta_m), \quad \pi' > 0, \quad \pi'' > 0, \quad \pi(0) = 0 \quad \pi(\delta_m^c) = 1, \quad (2.10)$$

which implies that the probability of catching offenders is zero when no smuggling activities occur, rises at an accelerating rate, and reaches the value 1 when the smuggling ratio is high enough $(\delta_m \geqslant \delta_m^c)$.

Let $\rho = s/\bar{e}$ denote (one plus) the parallel market premium. Dividing all terms in equation (1) by $\bar{e}p_m^*$ yields:

$$\sum_m^* = \mu_m[q_m + (1 - \Theta\pi)\tilde{q}_m] - \tau_m q_m - \rho\tilde{q}_m, \quad (2.11)$$

where $\Sigma_m^* \equiv \Sigma_m/\bar{e}p_m^*$ denotes profits in terms of the import good valued at the official exchange rate, and $\mu_m \equiv p_m/\bar{e}p_m^*$ the price markup.

The importer determines first the optimal amounts of the good to be imported legally and illegally so as to maximize profits, for given values of the tariff rate, the parallel market premium and the price markup μ_m. From (2.10) and (2.11), these conditions are given by:[13]

$$\mu_m + \mu_m(-\Theta\tilde{q}_m)(-\pi'\tilde{q}_m/q_m^2) - \tau_m = 0,$$

$$\mu_m(1 - \Theta\pi) + \mu_m\tilde{q}_m(-\Theta\pi'/q_m) - \rho = 0,$$

or rearranging terms,

$$\mu_m(1 + \theta\pi'\delta_m^2) = \tau_m, \quad (2.12a)$$

$$\mu_m(1 - \Theta\pi)(1 - \Omega_m) = \rho, \quad (2.12b)$$

where $\Omega_m \equiv \Theta\pi'\delta_m/(1 - \Theta\pi) > 0$.

Equation (2.12*a*) indicates that $1 \leqslant \mu_m \leqslant \tau_m$, so that the optimal domestic price is in general greater than the foreign price valued at the official exchange rate but *less* than the after-tax price. Equation (2.12*b*) requires that $\Omega_m < 1$ - a condition which can alternatively be expressed as a requirement on the detection technology *per se*, or on the proportion of output that would be confiscated by the tax enforcement agency, in case the importer is caught.

The zero-profit condition can be used to solve for the equilibrium domestic price markup. This yields:

$$\mu_m = (\tau_m + \rho\delta_m)/[1 + (1 - \Theta\pi)\delta_m]. \quad (2.13)$$

Equation (2.13) indicates that the domestic price markup is a weighted average of the parallel market premium and the tariff rate, with the weight on the latter falling monotonically as the equilibrium smuggling ratio increases.

Equations (2.12*a*) and (2.12*b*) form a system which determines the optimal combination of the price markup and the smuggling ratio consistent with profit maximization by importers. Since these equations are non-linear, a solution cannot be derived explicitly. The equilibrium can, however, be represented graphically as shown in the right-hand panel of figure 2.2.[14] The locus L_mL_m corresponds to equation (2.12*a*), while S_mS_m corresponds to equation (2.12*b*). The former curve has a negative slope, while the latter has a positive slope. The intersection of the two curves (point E_m) determines the optimal combination (μ_m, δ_m). It can formally be shown that the equilibrium solution always exists if $\tau_m > \rho$. More generally, existence of the solution means that an importer will tend to smuggle only if the tariff is so high that it pays to purchase

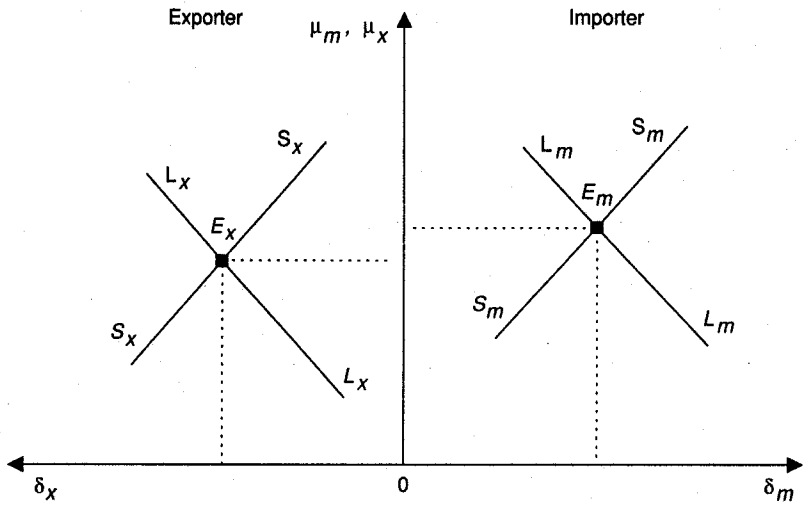

Figure 2.2 Parallel market premia and domestic prices

foreign exchange in the parallel market at a premium, given the possibility of getting caught by the customs enforcement agency ($\pi\tau_m > \rho$).

Consider now the behavior of the exporter. Conceptually, the problem faced by exporters is similar to that faced by importers: they must decide on the proportion of exports to sell abroad through legal and illegal channels, taking into account the likelihood of getting caught – and therefore the possibility of sanctions – the cost of producing export goods, and the tariff rate. Assuming the same detection technology as above, and denoting expected profits of the exporter by Σ_x, the problem can formally be stated as:

$$\sum_x = \pi[p_x^*\{\bar{e}(1-\tau_x)q_x + (1-\Theta)s\tilde{q}_x\} - p_x C(q_x, \tilde{q}_x)] + (1-\pi)[p_x^*\{\bar{e}(1-\tau_x)q_x + s\tilde{q}_x\} - p_x C(q_x, \tilde{q}_x)],$$

where τ_x denotes the export tax rate, $C(q_x, \tilde{q}_x)$ the cost function associated with the production of the export good, and $\pi = \pi(\delta_x)$, with $\delta_x \equiv \tilde{q}_x/q_x$. Rearranging terms yields:

$$\sum_x = p_x^*[\bar{e}(1-\tau_x)q_x + (1-\Theta\pi)s\tilde{q}_x] - p_x C(q_x, \tilde{q}_x). \quad (2.14)$$

Dividing expression (2.14) by $\bar{e}p_x^*$ yields:

$$\sum_x^* = (1-\tau_x)q_x + (1-\Theta\pi)\rho\tilde{q}_x - \mu_x C(q_x, \tilde{q}_x), \quad (2.14')$$

where $\Sigma_x^* \equiv \Sigma_x/\bar{e}p_x^*$, and $\mu_x \equiv p_x/\bar{e}p_x^*$. As before, exporters determine the levels of legal exports and smuggled exports so as to maximize profits. Assuming, for simplicity, that $C(q_x, \tilde{q}_x) = q_x + \tilde{q}_x$, the first-order conditions are given by:[15]

$$(1-\tau_x) - \Theta\rho\tilde{q}_x(-\pi'\tilde{q}_x/q_x^2) - \mu_x = 0,$$
$$\rho(1-\Theta\pi) - \Theta\rho\tilde{q}_x(\pi'/q_x) - \mu_x = 0,$$

or, rearranging terms:

$$\mu_x = (1-\tau_x) + \Theta\rho\pi'\delta_x^2, \quad (2.15a)$$
$$\mu_x = \rho(1-\Theta\pi)(1-\Omega_x), \quad (2.15b)$$

where $\Omega_x = \Theta\rho\pi'\delta_x/(1-\Theta\pi) > 0$.

Using the zero profit condition at the optimum, the domestic price markup can be derived as:

$$\mu_x = [(1-\tau_x) + (1-\Theta\pi)\rho\delta_x]/(1+\delta_x), \quad (2.16)$$

which indicates that the domestic price is a weighted average of the export tariff rate, and the product of the probabilty of getting caught and the premium. We now require that $\Omega_x < 1$ – a condition which obtains if the smuggling ratio is not "too high", or if the penalty rate Θ is not "too low."

Equations (2.15*a*) and (2.15*b*) constitute a nonlinear simultaneous system in δ_x and μ_x whose solution can be, as before, determined graphically. Equation (2.15*a*) is represented as the curve L_xL_x in the left-hand panel of figure 2.2, while equation (2.15*b*) is shown as curve S_xS_x.[16] The former curve has a positive slope while the latter has a negative slope. Intersection of the two curves (at point E_x) determines the optimal combination (δ_x, μ_x).

Table 2.1 summarizes the effects of changes in tax rates, the premium, the shape of the detection technology and the penalty rate on the equilibrium values of smuggling shares and domestic prices for importers and exporters. The table indicates, for instance, that an increase in the parallel market premium raises the domestic

Table 2.1 Summary of comparative static effects of trade taxes, the premium and detection technology

Variable	Importers		Exporters	
	δ_m	μ_m	δ_x	μ_x
τ_m	+	+	...	...
τ_x	...	...	−	+
ρ	−	+	+	+
π'	−	+	−	+
Θ	−	+	−	+

Note: '...' indicates that the comparative-static effect is not defined.

price of the imported good (direct effect), but lowers the smuggling ratio, because demand falls.

We can now proceed with the flow determination of the parallel market premium. The flow demand for foreign exchange in the parallel market is the sum of demand by (identical) importers for the purpose of smuggling imports in; flow supply is given by the aggregate sum of supplies from individual exporters resulting from smuggling out a proportion of their sales abroad. The equilibrium value of the premium is therefore the value which equates total demand and total supply. For simplicity, we normalize the fixed number of exporters and importers to unity.

From the first-order optimization conditions for importers (equations 2.12*a* and 2.12*b*), we have:

$$\rho = \tau_m m(\delta_m), \tag{2.17}$$

where $m(\cdot) = (1 - \Theta\pi)(1 - \Omega_m)/(1 + \Theta\pi'\delta_m^2)$.

Similarly, from equations (2.15*a*) and (2.15*b*):

$$\rho = (1 - \tau_x)x(\delta_x), \tag{2.18}$$

where $x(\cdot) = 1/[(1 - \Theta\pi)(1 - \Omega_x) - \Theta\pi'\delta_x^2]$.

Since there are no capital flows between the economy and the rest of the world, current account equilibrium in the long-run imposes two additional conditions. First, the value of legal exports must be equal to the value of legal imports, in foreign currency terms. Second, the value of illegal exports must be equal to illegal

imports, also in foreign currency terms. Formally, these conditions imply:

$$\bar{e}p_x^*q_x = \bar{e}p_m^*q_m, \qquad sp_x^*\tilde{q}_x = sp_m^*\tilde{q}_m. \tag{2.19}$$

Combining these equations yields:

$$\delta_x = [1 - \Theta\pi(\delta_x)]\delta_m. \tag{2.20}$$

Equations (2.17), (2.18) and (2.20) form a system which allows the simultaneous determination of the export smuggling ratio, the import smuggling ratio, and the parallel market premium. This is illustrated in figure 2.3. The curves *MM* and *XX* in the upper panel of the figure depict equations (2.17) and (2.18), respectively. The curve *SS* in the lower panel of the figure depicts equation (2.20). The curve *MM* is downward sloping, while curve *XX* is upward sloping. The equilibrium obtains at point *E*. Using these values, the solution for the detection probability and the domestic prices can be inferred.

The diagram can be used to analyze the effects of commercial trade policies and customs enforcement rules on the parallel market premium. An increase in the import tariff rate τ_m raises the demand for illegal foreign exchange and pushes *MM* outwards, raising the premium. A fall in the export tariff rate τ_x pushes *XX* outwards, and reduces the premium. The implications of changes in the other parameters of the model – the probability of getting caught, and the penalty rate – can similarly be inferred graphically (see Macedo, 1987).

The analysis developed above can be summarized as follows. An importer will tend to smuggle if the tariff is so high that it pays to purchase foreign exchange in the parallel market at a premium, given the possibility of getting caught by the customs enforcement agency. Similarly, under the same detection technology, the incentive to smuggle exports out will exist when the subsidy (or tax rate) on exports is smaller than the parallel market premium weighted by the probability of success in smuggling. In a "real trade" framework, planned smuggled imports are interpreted as the flow demand for foreign currency in the parallel market, while successfully smuggled exports are interpreted as the flow supply of foreign currency. The long-run parallel market premium is then determined by the equilibrium conditions for legal and illegal trade. In

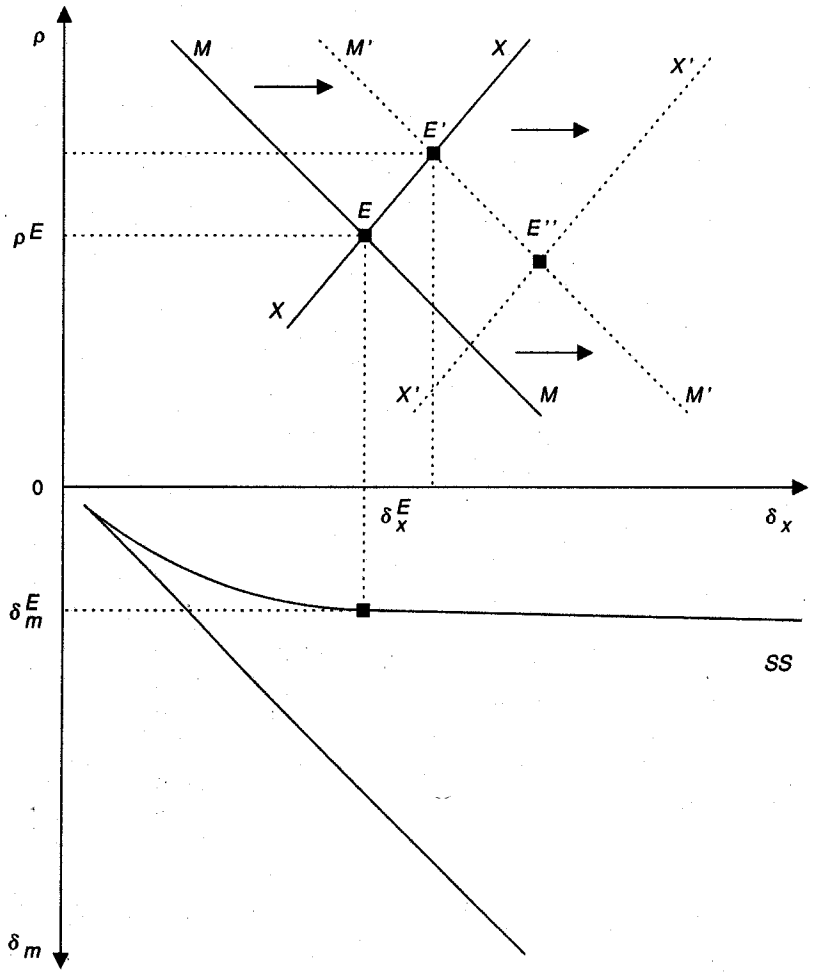

Figure 2.3 Flow determination of the parallel market premium in the real trade model

the long-run equilibrium, where legal exports equal legal imports and successfully smuggled exports pay for planned smuggled imports, the premium can be expressed as a weighted average of import and export tariff rates, and is therefore determined – as is the smuggling ratio – by the structure of tariff barriers.

Real trade models provide an adequate framework for an analysis of the impact of trade restrictions – as opposed to exchange controls – on the parallel market exchange rate. The basic limitation of the approach is that, since the only reason to deal in foreign currency is to buy imported goods, the sole purpose of black market activity is to enable smuggling to take place. This, however, assumes away the portfolio motive which has been identified as a critical component of the demand for foreign currency. Moreover, although the approach provides a useful analysis of the long-run determinants of the exchange rate differential, there is no mechanism providing a satisfactory explanation of the short-run behavior of the premium (which is taken as given by exporters and importers in most models). The approach can, however, be extended in this respect. Macedo (1987) for instance extends the model developed above and shows that whereas the premium is in the long-run determined by the structure of trade taxes, in the short-run it is given by the requirement of portfolio balance.

2.2 The Monetary Approach

The monetary approach, developed initially by Blejer (1978), emphasises the role of monetary factors in the behavior of parallel market exchange rates. Blejer outlines a model of the premium by grafting a flow parallel market for foreign currency into a monetary model of the balance of payments in which the rate of devaluation of the official exchange rate depends on the inflation differential with the rest of the world. In Blejer's model, an increase in the domestic money stock – initiated, say, by a rise in net domestic credit – results in an *ex ante* disequilibrium between supply and demand in the money market.[17] As excess cash balances are worked off by agents, domestic prices rise. This reduces the demand for domestic goods, raises the demand for foreign goods and foreign currency, and entails a depreciation of the parallel exchange rate. This, in turn, increases the differential between the official and the parallel rate, thereby raising the incentive to underinvoice exports, to smuggle exports, or to divert remittances through unofficial markets. Although the increase in the illegal supply of foreign exchange tends to reduce the initial upward pressure on the

parallel rate, a higher stock of money will in general be associated with a depreciation of the free exchange rate. Therefore, a restraint on the rate of growth of domestic credit is a key policy instrument for preventing official reserve losses when there exists a parallel market for foreign currency.

The monetary model provides important insights into the relationships between monetary policy and the behavior of parallel market exchange rates. In addition to the empirical results reported by Blejer for Brazil, Chile, and Colombia, the model has been tested for India (Biswas and Nandi, 1986) and Turkey (Olgun, 1984), with generally good results. Blejer's formulation suffers, however, from an important limitation. The model assumes that the demand for foreign exchange on the unofficial market arises only because the public desires to alter the composition of its portfolio of financial assets and not for the purpose of carrying out international purchases of goods. By doing away with the existence of restrictions on trade (like tariffs and quotas), it is further assumed that no foreign exchange is demanded in the parallel market to pay for goods that are imported into the home country without declaration at the border. Thus, all the current account needs are assumed to be satisfied by the official foreign exchange market. This assumption may be particularly inadequate in view of the exchange controls on both the current and capital account operations and the quantitative restrictions on foreign trade imposed by most developing countries, which divert a substantial part of the transactions demand for foreign exchange – as emphasized in real trade models – from the official market to the parallel market.

Agénor (1991) has developed a model which extends Blejer's (1978) monetary model so as to deal with some of the restrictive features described above. The model provides a synthesis between the monetary approach and the currency substitution framework described below, and considers explicitly stock-flow interactions. Econometric estimates of the model using quarterly data for a group of 12 developing countries confirm the role of monetary disequilibria, as well as changes in the official exchange rate and expected rates of return, as major determinants of the behavior of the parallel market exchange rate.

A full macroeconomic model based on the monetary approach has also been discussed by Agénor (1990*a*). The model, which is estimated using cross-section annual data for eight countries, incorporates illegal trade transactions, foreign exchange rationing, currency substitution features, and forward-looking rational expectations. The simulation results indicate that anticipated expansionary credit and fiscal policies have a positive impact on real output and prices, a negative effect on foreign assets, and are associated with a depreciation of the parallel exchange rate. The analysis also shows that the adjustment process following a temporary shock is inversely related to the degree of rationing in the official market for foreign exchange. The higher the degree of rationing is, the lower will be the offsetting effect on the money supply coming through the balance of payments, and the higher the rate of depreciation of the parallel exchange rate generated by an expansionary policy.

2.3 Portfolio and Currency Substitution Models

The portfolio-balance approach, developed by Dornbusch *et al.* (1983), Frenkel (1990), and Macedo (1982, 1985) stresses the role of asset composition in the determination of the parallel market exchange rate, in contrast to the "flow" approach typical of smuggling models. The general observation underlying this class of models is that foreign exchange is a financial asset – even in countries with a low level of capital market development. Loss of confidence in the domestic currency, fears about inflation and increasing taxation, and low real interest rates give rise to a demand for foreign currency, both as a hedge and a refuge for funds, and as a means of acquiring and hoarding imports. Expectations are taken to play a key role in determining short-term supply and demand shifts and in accounting for the volatility of parallel exchange rates. Although the partial equilibrium formulation of Dornbusch *et al.* (1983) assumes the existence of domestic and foreign interest-bearing assets, the essential features of the approach can be captured by models where domestic agents hold in their portfolios only non-interest bearing domestic and foreign money.[18] This class of models, based on the "currency substitution" hypothesis, provides considerable insight into the short- and long-run behavior of parallel market exchange rates.[19]

In all these models, output is exogenous and the desired proportion between domestic and foreign currencies is given by a "liquidity preference function" (Calvo and Rodriguez, 1977) which depends on the expected – and, under perfect foresight, actual – rate of depreciation of the parallel market exchange rate. Private capital transactions are usually ignored, so that the reported current account is equal to the change in central bank reserves, which determines in turn – together with an exogenously determined rate of growth of domestic credit – changes in the domestic money stock. The unreported current account determines the change in the stock of foreign currency held in private agents' portfolios.[20] The flow supply of foreign exchange in the parallel market usually derives from under-invoicing exports. The propensity to under-invoice, when endogenous, is assumed to depend positively on the level of the premium. The probability of detection is also assumed to rise as fraudulent transactions increase, and this translates into a rising – but at a decreasing rate – marginal under-invoicing share.

Portfolio balance implies that at each instant the domestic currency value of the stock of foreign assets is equal to a desired proportion of private wealth. In the short run, the parallel market rate moves so as to set portfolio demand equal to the existing stock of foreign currency, implying that flow demand and supply may diverge at any given moment. The determination of the parallel exchange rate at any instant of time is made, therefore, through the portfolio balance equation, with the stock of foreign currency assumed fixed. In the long run, the parallel rate and the private sector holdings of foreign currency are determined by the requirements of both portfolio and current account equilibrium.

Although there remain important differences between the individual formulations,[21] some general conclusions can be derived from this class of models. In a fixed-exchange rate regime, an expansionary fiscal and credit policy generates a depreciation of the parallel exchange rate, a rise in prices, an appreciation of the official real exchange rate, and a decline of the relative price of exports surrendered via the official market relative to those that use the parallel market. As a consequence, the proportion of export proceeds repatriated at the official exchange rate falls, and foreign reserves decline.[22] Eventually, the central bank will "run out" of reserves, and a balance of payments crisis will ensue. At this point, the inconsistency between expansive macroeconomic policies and a pegged official exchange rate will become unsustainable, and corrective measures will need to be implemented – for instance, in the form of a parity change. This process leading to a "devaluation crisis" has been well documented by Edwards (1989), and Edwards and Montiel (1989).

An important issue in the context of currency substitution models is the effect of devaluation on parallel market premia. Consider first a fixed-exchange rate regime in which cross transactions exist between foreign exchange markets and the under-invoicing share depends endogenously on the premium (Kamin, 1991*b*). In this regime, the *long-run* impact of a once-and-for-all official devaluation on the parallel rate is ambiguous. The effect will in general depend on the degree to which fraudulent transactions react to changes in the premium, the rationing scheme imposed by the central bank, and the elasticity of export volumes to changes in relative prices. The greater the response of the under-invoicing share to the change in the premium, the smaller the central bank's marginal propensity to resell officially remitted foreign exchange, and the smaller the response of export volumes, the more likely it will be that the parallel market rate will depreciate – less than proportionally, so that the premium falls – in response to a parity change.[23]

The *short-run* behavior of the parallel exchange rate and the premium in response to a devaluation reflects the typical behavior of asset prices. Consider first the case where the devaluation is unexpected. The parity change causes a decline in the flow supply of foreign currency to the private sector (since the premium falls) at the initial parallel rate.[24] For current account balance to be maintained, a depreciation of the parallel rate is required. At the moment of the devaluation, the parallel rate depreciates sharply. Subsequently, current account losses of foreign currency drive the unofficial exchange rate up still further until it reaches a new long-run equilibrium at the same moment foreign currency holdings reach their new steady-state level.

Suppose now that the devaluation is anticipated, that is, it is announced before being implemented. The announcement of the future devaluation raises immediately the anticipated – and actual, since expectations are rational – rate of depreciation of the parallel

exchange rate, so that the free rate depreciates and foreign currency holdings rise. After the initial jump, the parallel rate continues to depreciate while private agents accumulate foreign currency in their portfolios until the economy hits a new saddle path at the instant the devaluation actually occurs. From this point on, the parallel rate continues to depreciate while foreign currency holdings now decline, since the unofficial current account deteriorates following the devaluation. At the date of the announcement of the future devaluation, the under-invoicing share jumps upwards, and grows as the parallel market rate depreciates. When the devaluation is effectively enacted, the premium and the under-invoicing share fall sharply, but then recover partially, since the parallel market rate continues to depreciate, until reaching its new steady-state level.

This description of the transmission mechanism of a parity change provides an interesting explanation for the seemingly puzzling empirical results on 60 devaluation episodes in developing countries described by Kamin (1988). The study shows that prior to the typical devaluation, the growth rates of exports and imports fall sharply, while the current account balance and reserve levels deteriorate markedly. Immediately following the devaluation, exports recover strongly and the current account improves (contrary to what a "*J*-curve" model would predict), while imports continue to fall – albeit at a slower pace – and rebound sharply in the second year after the devaluation. A rationale for this sequence is as follows (Kamin, 1991*b*). Continuous inflation and hence appreciation of the real (official) exchange rate lead to increases in the parallel market premium, increases in export under-invoicing, and reductions in officially measured exports. In turn, this drop in export proceeds leads to reserve losses and declines in imports as the authorities tighten foreign exchange allocations. The expectation that the deteriorating external balance will prompt an official devaluation induces a speculative rise in the parallel market rate which further reinforces the need for official exchange rate adjustment. Following the devaluation, the parallel market premium falls, reducing under-invoicing and increasing officially recorded exports. Improved reserve flows allow the authorities to expand progressively sales of foreign exchange, so imports increase as well.

Consider now the case of a crawling-peg regime. The official and parallel exchange rates depreciate at the same rate in the steady state, thereby leaving the spread unaffected. An increase in the rate of devaluation of the official exchange rate leads to an equivalent rise in the rate of depreciation of the parallel rate, and this generates a portfolio shift away from domestic money holdings. If the official and parallel foreign exchange markets are effectively segmented, the supply of foreign currency is fixed in the steady state, so that only an increase in the premium can restore portfolio equilibrium. The increase in the steady-state level of the premium caused by a higher rate of official exchange rate depreciation has been emphasized by Dornbusch (1986) and Pinto (1989). It is important to note that the steady-state premium does not depend on the *level* of the official exchange rate, but only on its rate of change. This implies that discrete, one-shot devaluations will reduce the premium only temporarily, in the absence of fundamental changes in fiscal and monetary policies. This result has important implications regarding attempts at unifying the official and parallel foreign exchange markets (see Agénor, 1990*c*).

A source of ambiguity in the long-run effects on the premium of an increase in the rate of crawl relates to the role of the exchange rate differential as an implicit tax on exports (Pinto, 1989, 1991, and Kharas and Pinto, 1989).[25] On the one hand, a higher rate of devaluation raises the rate of depreciation of the parallel market rate, making foreign currency holdings more attractive. This, by itself, would raise the premium. At the same time, however, for a given real fiscal deficit, a smaller domestic currency base is required to generate a given amount of revenues from the inflation tax, creating therefore the ambiguity. Whether the premium rises or falls depends upon the inflation elasticity of the share of domestic currency holdings in total financial wealth. If this is less than unity, raising the rate of devaluation of the official exchange rate raises the unit yield of the inflation tax, and lowers the premium. Otherwise, the premium will actually rise. Thus, an acceleration of the official rate of devaluation does not necessarily lower the premium; the outcome depends crucially on the inflation elasticity of the demand for foreign currency. In turn, this elasticity rises with the rate of inflation – that is, the propensity to shift into foreign currency to avoid the inflation tax becomes stronger as the rate of growth of domestic prices rises. This results in a "seignorage

Laffer curve", with the (unit) yield of the inflation tax rising for inflation rates below the seignorage-maximizing level of inflation and falling above it. A similar reasoning yields a U-shaped curve linking the steady-state premium and the inflation rate, representing the trade-off between the tax on exports and the inflation tax.

Overall, therefore, the impact of a devaluation on the parallel market premium is theoretically ambiguous, whether one considers an officially fixed exchange rate system or a crawling peg regime. These predictions of currency substitution models have been recently subjected to formal empirical tests.[26] The evidence supports in general the presumption that parallel market rates depreciate, but less than proportionally, in response to a devaluation of the official exchange rate, and that the premium falls initially. However, the evidence also suggests that this reduction will only be temporary if fiscal and credit policies are maintained on an expansionary course, implying that a devaluation, by itself, cannot permanently lower the premium. Studies by Edwards (1989), Edwards and Montiel (1989), and Kamin (1991*b*) on a large sample of devaluation episodes in developing countries have well documented these facts.[27] Similarly, in the empirical model presented in Agénor (1990*a*), a once-and-for-all devaluation of the official exchange rate is associated in the short run with an output contraction, a rise in the inflation rate, an increase in net foreign assets, and a less-than-proportional depreciation of the parallel rate. In the long run, the official devaluation results in a permanently higher price level and a more depreciated parallel exchange rate, but has no effect on the premium. The econometric results presented in Agénor (1991) also support the view that the parallel rate depreciates less than proportionally following an official parity change.

2.4 Models of Dual Exchange Rate Markets with Leakages

There have been some major developments, over the past few years, in the literature dealing with the properties of formal two-tier exchange rate regimes which may prove useful for understanding the behavior of parallel currency markets.[28] The recent

analytical literature has focused on the possibility of illegal cross-operations between the commercial and financial exchange markets; see for instance Bhandari (1988), Bhandari and Decaluwé (1986, 1987), Bhandari and Végh (1990), Gros (1987, 1988), and Guidotti (1988). Gros (1988) has shown that a divergence between the two exchange rates induces in the short run a flow of arbitrage activity, the magnitude of which depends on both the costs of evading exchange controls and the size of the exchange rate differential. Bhandari and Végh (1990) have developed an optimizing model in which the coefficient of leakage is endogenously determined by utility-maximizing agents. Several aspects of dual exchange rate models with leakages are relevant for the analysis of illegal or quasi-legal markets for foreign exchange in developing countries.[29] In models of economies with dual legal markets with a floating "financial" exchange rate, the floating rate plays a role similar to the parallel market exchange rate. Moreover, formal economic modeling of legal and illegal parallel markets for foreign exchange may look very similar, since risk and transaction cost functions may be indistinguishable.

Consider, for example, the model of a legal, dual exchange rate system with leakages developed by Bhandari (1988). The model, which explicitly recognizes both private (fraudulent) and officially-sanctioned cross transactions between the two exchange markets, is based on a stochastic, rational expectations approach. Both the real and financial sectors of the economy are explicitly considered. Asset accumulation equations in the model are specified in beginning-of-period, *ex-ante* equilibrium, rather than in end-of-period, *ex-post* terms for analytical tractability. Although there is no currency substitution, agents may hold foreign-currency denominated bonds in their portfolios. The interest parity condition, properly modified to reflect leakages between markets from repatriated interest receipts, is assumed to hold continuously.

The conclusions of the model can be readily reinterpreted for an economy in which the dual exchange rate system is composed of a legal and an illegal market. A positive supply shock raises output and causes excess demand in the money market, necessitating an increase in the yield on domestic assets to restore equilibrium. As a consequence, the parallel exchange rate must appreciate in order to maintain uncovered asset yields in line. A permanent increase

in the foreign price level is associated with an appreciation of the parallel exchange rate and – as a result of the partial offset provided by the movement in the free rate – leads to a less than equi-proportionate rise in domestic prices. In the long run, a devaluation of the official exchange rate – anticipated or unanticipated – leads to an equi-proportionate depreciation of the parallel exchange rate, thereby leaving the spread unaffected. In the short run, a once-and-for-all unanticipated devaluation leads initially to an increase in prices and real output. At the same time, the reported current account improves, since the premium falls, leading to reserve accumulation and an expansion of the money supply, while the unreported current account deteriorates, leading to a fall in the stock of foreign assets held by the private sector. The increase in domestic prices reduces the real money stock, thus necessitating an increase in the domestic asset yield to re-equilibrate the market. This, in turn, requires a rise in the expected rate of depreciation of the parallel exchange rate, which brings an immediate depreciation of the free rate. The higher the degree of compulsory cross-transactions, or the higher the penalty costs associated with fraudulent cross-transactions, the greater will be the rate of depreciation of the parallel exchange rate. Overall, the devaluation leads to a less-than-proportionate depreciation of the parallel exchange rate, implying therefore a fall in the premium. The conclusions regarding the short- and long-term effects of an official devaluation are qualitatively very similar to those derived from currency substitution models.

3 An Assessment

This review of existing approaches to modeling informal markets for credit and foreign exchange suggests two essential implications. First, it is critically important to distinguish between flows and stocks in the analysis of such markets. In particular, loans through the informal credit market and holdings of foreign exchange must be viewed as components of a portfolio of assets, and as such, as responding to relative rates of return and a wealth constraint. Second, parallel exchange rates and informal interest rates have pervasive effects on other aspects of the economy, influencing

income, prices, as well as both the supply of, and the demand for, domestic goods. The next chapter dwells on the first implication in the context of a generalized portfolio model which simultaneously accounts for both types of informal markets. The analysis will then be extended to account for output and price effects in a more disaggregated macroeconomic framework.

Notes

1 See also McKinnon (1981) and Fry (1988).
2 See Montiel (1986).
3 In fact, one of the seminal papers by one of the most prominent Neo-Structuralists is entitled, "IS-LM in the Tropics: Diagrammatics of the New Structuralist Macro Critique", (Taylor, 1981).
4 In the Neo-Structuralist literature, the emphasis on the supply-side role of working capital is based on Cavallo (1977).
5 See Buffie (1984) for an explicit distinction between the two approaches based on their treatment of the informal loan market.
6 See van Wijnbergen (1983*b*) and Buffie (1984).
7 Nevertheless, as long as some funds are drawn out of informal credit markets, and as long as reserve requirements in the formal banking system remain high, the quantitative effects of raising deposit rates on the availability of funds for investment would be diminished.
8 Lending for consumption purposes is irrelevant, since the stock L^p presumably refers to the net stock of curb market loans held by the household sector, as is true for the other asset demands. Lending within the sector is already netted out. The existence of transaction costs would, under competitive conditions, be reflected in a wedge between borrowing and lending rates in the informal credit market, not between the assets and liabilities of institutions operating in that market.
9 An exception is Taylor (1983), who considers "gold" (representing a coterie of "unproductive" assets) in this role. However, the price of "gold" is taken to be a state variable in Taylor's model. Since this prevents the price of gold from adjusting discontinuously in response to new information, it rules out conventional asset-price dynamics, which may represent an important aspect of the transmission mechanism.
10 For formal tests of efficiency of parallel currency markets in developing countries, see Appendix A and Ghei and Kiguel (1992).

11 In addition to the discussion provided here, it is worth noting that there have been some recent attempts to integrate informal markets in goods and foreign currencies in Computable General Equilibrium models; see Azam and Besley (1989), and Nguyen and Whalley (1989). See also Charemza and Ghatak (1990) for a disequilibrium approach.
12 The analysis here follows Macedo (1987), and abstracts from the effect of uncertainty on agents' behavior. The treatment of risk in real trade models of smuggling is examined by Sheikh (1989).
13 It can be checked that the second-order conditions for an optimum are satisfied if $\pi(\cdot)$ is a convex function, that is, if $\pi'' > 0$ as assumed above.
14 For simplicity, all curves are represented as linear functions in the figure.
15 The second order conditions are also verified here.
16 As before, the curves are represented as straight lines for convenience.
17 In Blejer's model, flow monetary disequilibrium is measured as the difference between the expansion of the domestic-credit component of the base (and variations in the money multiplier) and the changes in the demand for real cash balances.
18 The analytical formulation used by Dornbusch *et al.* is not particularly adequate for developing countries with underdeveloped financial systems. Moreover, the process of currency substitution – whereby foreign-currency denominated money balances increasingly substitute for domestic money as a store of value, unit of account, and medium of exchange – has gained importance in many developing countries over the past few years.
19 See Dornbusch (1986), Edwards (1989), Edwards and Montiel (1989), Kamin (1991*b*), Kharas and Pinto (1989), Kiguel and Lizondo (1990), Lizondo (1987, 1991), Pinto (1989, 1991), and Samiei (1987).
20 In terms of the "real trade" model developed above, the rate of change of the stock of foreign assets would be given by, from equation (2.11),

$$\dot{F} = s[(1 - \Theta\pi)p_x^*\tilde{q}_x - p_m^*\tilde{q}_m] \equiv F(\overset{+}{\rho}, \overset{+}{\tau}_x, \overset{-}{\tau}_m).$$

21 Edwards and Montiel (1989) for instance consider a three-good economy and develop a fairly general analytical framework, but they assume that foreign currency holdings remain constant – excluding therefore an important source of dynamics.
22 In addition to its impact on the propensity to under-invoice exports, an increase in the premium – without an equivalent increase in domestic prices – may generate a positive wealth effect on aggregate demand, which may cause further deterioration of the current account of the balance of payments.

23 Nowak's (1984) result, according to which an official devaluation will be associated with an appreciation of the parallel exchange rate, depends critically on the assumption that the central bank does not accumulate foreign exchange (see Kamin, 1991*b*).
24 An increase in the parallel market rate, given the official exchange rate, increases the share of exports channeled through the unofficial market for foreign currency via under-invoicing or smuggling, and thus increases the flow supply of foreign exchange. Conversely, import demand will fall, as well as the share of imports channeled through the parallel market (as a result of over-invoicing or smuggling), which will turn decrease the flow demand for foreign currency.
25 This is because there are two prices at which foreign exchange can be bought and sold, exports whose proceeds are repatriated at the official exchange rate are taxed relatively to other exports.
26 Dornbusch *et al.* (1983) present empirical tests of their model for Brazil, while Phylaktis (1991) considers the case of Chile. The results show a significant impact of the interest rate differential – as well as, for Chile, the degree of capital restrictions – on the parallel market premium. Fishelson (1988), using the actual rate of depreciation of the parallel market rate as a proxy for the expected rate of official devaluation, tests the Dornbusch *et al.* model for a group of 19 countries over the period 1970–79. More recently, Kaufman and O'Connell (1990) have provided estimates of the model for Tanzania, over the period 1967–88. Portfolio factors are shown to affect the behavior of the parallel market premium mainly in the short run, while flow factors play a predominant role in the long run.
27 The recent experience of Argentina also provides a good illustration of these propositions. Following the devaluation of December 1989 the premium dropped immediately, but, because of the lack of financial discipline, the free market rate rose quickly to 1,230 australes, bringing the premium back to 23 percent. See Kamin (1991*a*) for a further analysis.
28 The early strand of literature was based on the assumption that the dual exchange markets were effectively segmented, implying that the freely floating exchange rate would ensure a zero capital account. As a result, a major channel of transmission of external disturbances – via movements in foreign assets – was completely eliminated.
29 Note that in a dual rate system (with a fixed commercial rate and a freely floating financial rate) an illegal parallel market may persist, owing to the retention of (some) capital controls.