The New Theory of Foreign Direct Investment: Merging micro-level and macro-finance

Book Review

Katheryn N. Russ

University of California, Davis

> "Foreign direct investment is a form of international capital flows (p.8)." Assaf Razin and Efraim Sadka, *Foreign Direct Investment: An analysis of aggregate flows*. Princeton: Princeton University Press, 2007.
>
> "…capital flows are important and interesting, but I also believe that they can be largely separated from the real decisions about the location of production and the direction of trade (p.xv)." James R. Markusen, Multinational Firms and the Theory of International Trade. Cambridge, MA: The MIT Press, 2002.

I. Introduction

What is the origin and purpose of a multinational firm? How is it different from its native counterparts? In what countries and industries should one expect to see more of them? What is their impact on the pattern of trade and the distribution of wealth across countries? With the enormous growth of foreign direct investment (FDI) in the last 25 years, these questions have found voices in two divergent literatures, both rooted in the general equilibrium modeling of the 1950s. Mundell (1957) first mathematically modeled cross-border capital flows in a Heckscher-Ohlin framework, simultaneously illustrating the two ideas underpinning these separate views of FDI in studies confirming or contradicting them for 5 decades afterward: (1) that flows of capital and goods are substitutes and (2) that all else equal, capital flows should be

related to relative endowments when there are barriers to trade, flowing from capital abundant countries to capital poor ones. In the 1980s, it became the work of theorists like Wilfred Ethier, Gene Grossman, Elhanan Helpman, James Markusen, and Assaf Razin to pinpoint exactly why foreign direct investment differs from the way Mundell and neoclassical growth models envisioned it, as it became clear that FDI was increasing, but between rich countries and in tandem with intra-firm trade.

Some of these theorists, like Markusen, chose to use a "micro-level" approach, focusing on the structure of fixed costs, relative country size and endowments, and preference conditions that compel firms to expand operations overseas. Others, like Razin, soon after with Efraim Sadka, chose a "macro-finance" path, embedding sunk costs, taxes, and other distortions as fulcrums to analyze the behavior cross-border capital flows and their implications for country welfare.[1] In Markusen's recent text, *Multinational Firms and the Theory of International Trade* and Razin and Sadka's treatise, *Foreign Direct Investment: Analysis of Aggregate Flows*, these giants in the field bring us up to date and carry us forward into the next generation of modeling, simulation, and empirical analysis of the multinational enterprise (MNE). In particular, Markusen gives us a roadmap to match theories of the firm with observed aggregate flows, while Razin and Sadka derive new microeconomic fundamentals to formalize the more descriptive characterization of FDI common in international finance. To do so, the authors begin at very different places, but use tools with more overlap than one might expect. They leave the reader with whispers of a mechanism to reattach locational choice and intra-firm trade flows with a rigorous theory of capital flows, as well as a plentitude of promising direct suggestions and methodologies for

[1] See Russ (2004) for a detailed discussion of the literature on FDI and cross-border capital flows in the 1980s.

other fertile avenues of future research. Below, I discuss the way that their philosophies are manifest in these helpful manuals for modeling and empirics, as well as how their narratives relate to recent trends in the literature.

II. Illlustrating the micro-level versus macro-finance approach

The principal stylized fact motivating both of their studies is the same: the majority of FDI is characterized by two-way flows between industrialized nations, with very little flowing into the poorest countries of the world despite the lure of low relative wages. Nonetheless, they replicate this fact using theoretical frameworks which illustrate the completely different philosophical stances of the literatures they have helped originate and nurture—one focusing on scale economies arising from nonrival headquarters services and the other on the efficient allocation of physical capital. Markusen, for instance, offers multiple ways to generate two-way flows between similar countries, all based solidly in classical trade theory and industrial organization, all derived using the trifecta of horizontal FDI modeling: firm-level scale economies, trade costs, and imperfect competition. With these tools, he engineers the proximity motive that compels a firm to establish a plant abroad, allowing it to serve the foreign market while circumventing trade costs.

Markusen's two most basic general equilibrium models and more sophisticated extensions presented later in the text generate larger flows of FDI between large countries similarly endowed with the factor most intensively used by multinational firms. He thus captures two-way flows between industrialized countries and to some degree the very low levels of FDI directed to poorer countries, which presumably have smaller endowments of skilled labor or

other resources important to producing special headquarters services that can be applied to production in plants overseas. The first version (in Chapter 5) is a general equilibrium model of oligopolistic competition and the second (in Chapter 6) a Dixit-Stiglitz variation with monopolistic competition.[2] Consumers benefit from extra varieties (in oligopoly due to decreasing markups and in monopolistc competition due to an intrinsic love of variety itself), so given the increasing returns to firm scale and the cost of shipping exports, some firms from each country will be able to profit from investing abroad. Just as Krugman 1979 (with an endogenous markup) has a scale effect from increasing market size and Krugman 1980 (with a constant CES-based markup) does not, a larger market—arising either by assumption or through eliminating restrictions on international trade and investment—lowers the markup and increases firm scale in the oligopolistic model but has no markup or scale effect in the CES environment. He thereby emphasizes the role of industrial structure as an issue of central importance in measuring the determinants and welfare effects of FDI flows.

Razin and Sadka also offer several different models that produce two-way flows of FDI between countries with similar endowments of capital and labor. However, instead of proximity, they propose a Ricardian engine of comparative advantage in investing. The approach is a bit shocking in its innovation (comparative advantage in *investing,* not production), but well supported by a rich empirical literature on "cream-skimming" and the positive correlation of foreign takeovers with target firms' post-takeover productivity, beyond even the studies cited in the text. In one model, they assume that upon entry, each firm gets a noisy signal about how efficient its manufacturing process is, so that the owner is not certain what the optimal level of

[2] Chapter 5 is based on Markusen and Venables (1998) and Chapter 6 on Markusen and Venables (2000).

capital investment might be. Every firm's true productivity lies within a fixed interval from the signal it observes. The size of this interval is known and identical across firms. At this point, the original owner can hold onto the firm or sell it to a domestic or foreign investor. Any owner intending to produce in the following period must decide whether to pay a screening cost to find out the true productivity of the firm and adjust its capital stock to the optimal level, or to simply retain the initial level of firm capital exogenously endowed upon entry. Naturally, only owners observing a signal above an endogenous threshold level will choose to bother screening and adjust the capital stock through additional investment.

The Ricardian twist stipulates that foreign investors have a lower screening cost in one of two or more industries. Therefore, each country is somewhat specialized in the sense that it has expertise that helps its residents to discern the true productivity of firms worldwide in a particular industry relatively cheaply, a type of comparative advantage. A lower screening cost increases the value of a potential target for foreign investors, allowing them to bid more than domestic investors to acquire a firm in the industry of foreign expertise. It seems quite plausible to imagine Germany better able to pick winners in the US pharmaceutical (or beer or auto) industry while the US is able to cherry pick among targets in Germany's retail sector. Thus, one might see two-way flows as Germany invests in US pharmaceutical firms and the US in German retail chains. Like Ricardian gains from trade in the traditional sense, there are real gains arising from comparative advantage in this model. The presence of foreigners with a relatively low screening cost for any industry has the real consequence of increasing final output and improving economic efficiency by having more firms acquired and screened for potential optimization of their capital stock than would be the case under capital controls.

The previous model is introduced in partial equilibrium, with the two-way flows suggested as a reasonable extension by the authors. It stops short of explaining why capital does not flow from capital-rich to capital-poor countries as the Mundell and older neoclassical models would suggest. To do this, the authors explicitly construct a general equilibrium model,[3] retaining the assumption that firms are heterogeneous but for simplicity removing any uncertainty about their idiosyncratic efficiency levels. The screening cost is now just a fixed "setup" cost, but due to superior expertise in management or R&D, it is again lower for foreign firms than domestic firms in one industry, again producing two-way flows. Capital in the model is perfectly mobile between countries but labor is not. The result is that the returns to capital can equalize across countries with different relative labor endowments, squelching further capital flows despite a wedge persisting between the wage rates. This is the explanation the authors offer for why low-wage countries remain capital-poor, amending traditional theories of capital flows.

Clearly, the philosophical underpinnings of these two sets of models differ. In particular, Markusen is quite clear in stating at several points that FDI flows are not about transfers of physical capital, refraining from including any representation of capital in his production technologies, whereas Razin and Sadka use capital flows as a conceptual starting point. Markusen uses sophisticated models of competition to motivate two-way flows of FDI, Razin and Sadka appeal to a Ricardian logic. Markusen explains small flows to developing countries by pointing to their scarce supplies of skilled labor, Razin and Sadka suppose it is due to an excess of (undifferentiated) labor. Markusen rigorously models free entry and ex ante identical

[3] It is a general equilibrium model in the sense that there are labor market clearing conditions and endogenous wage rates. It is not clear how consumers behave or whether there is any output in the first period that can be used to expand (endogenously) the capital stock in the second. I assume that there is no output in period one.

firms, while Razin and Sadka assume an exogenously fixed number of heterogeneous firms. Markusen thinks about firms deciding whether to invest abroad. Razin and Sadka think about investors of an unspecified nature—are they individuals, firms, institutions? It does not matter.

Nonetheless, both theoretical approaches rely on the conjecture that it is some special technical or managerial know-how that generates the two-way flows of FDI between industrialized countries that are the dominant feature of any objective description of the data. Markusen and other micro-level theorists envision this as a nonrival resource yielding firm-level economies of scale that encourage expansion through the establishment of overseas plants. Razin and Sadka and macro-finance theorists call it "intangible capital." By any name, it can be used to improve the productivity of multiple plants simultaneously at relatively low cost (compared to the costs of starting a new firm or making investment decisions without this special resource) and can be generated through specialization, learning, R&D, or innovation. In addition, the source of financing for the investment—be it mutual funds, bank loans, equity issues in overseas stock markets—is indeterminate in both sets of models, set aside as an issue separate from the economically meaningful decisions of how to manage the firm. There is a fundamental meeting of the minds on these key issues.

Notwithstanding, the authors of the two works are interested in a somewhat different set of stylized facts. First, Markusen considers the potential for intra-industry FDI flows to be important in matching any model to the data, not just as a clever mechanism to generate two-way flows. Second, over and above the predominance of flows between large industrialized countries, which can be explained using models of horizontal FDI, Markusen wants to

understand why countries with very low wages do not attract investment even for fragments of the manufacturing process that seem to require only unskilled labor. He explains this using the unified "knowledge capital" model in Chapters 7 through 9, which allows for vertical fragmentation of production within a single firm but spread across two countries. By supposing that even when tasks using unskilled labor require some fixed input of technical or managerial know-how from skilled labor, he is able to show that countries where skilled labor is extremely scarce will attract almost no FDI of any kind. Third, although he declares the horizontal motive to appear more empirically relevant than vertical fragmentation of firm production across countries, he views the vertical aspect of the micro-level literature's knowledge capital model as important to generate the complementarities between trade and investment that have driven the enormous growth of intra-firm trade. Finally, he considers the facts that multinationals are prevalent in industries intensive in R&D and skilled labor to be of central importance in his models. He eschews the idea of FDI flows being connected to endowments of physical capital as being unsupported in the data, focusing instead on skilled and unskilled labor.

With superior screening, the Razin and Sadka models described above replicate the oft-cited stylized fact that firms taken over by foreign investors are more productive than their domestic counterparts. In later chapters, they capitalize on their expertise in the public finance aspects of FDI to revisit the role of taxation in multinational firm behavior. They offer a new stylized fact: source country corporate tax rates impact the discrete decision of whether to establish a foreign affiliate, but not the magnitude, while host country corporate tax rates affect both this discrete decision and the magnitude of flows. They provide a straightforward theoretical

intuition rationalizing this result. If the source country's tax rate is sufficiently high compared to a potential host country, then a firm will decide to establish a foreign affiliate. The amount of production activity transferred (or attributed) to the affiliate then clearly depends on how high the host country's effective tax rate is, as the host country's stance on deductions for depreciation, etc., affects the amount of capital invested by the parent in its affiliate. The magnitude of these effects and the potential gain from their proposal of policy coordination is not clear, but in light of Markusen's assertion that taxes "appear to be of secondary importance (p.6)" among the determinants of FDI, perhaps could be profitably discussed.

The texts offer substantive but distinct roadmaps to help the reader explore any of these issues on her own through simulation and econometric analysis. Markusen's text presents strategies to address nonlinear programming (discrete choice) problems using GAMS. Razin and Sadka rigorously examine the implications of a number of econometric specifications when taking their models to data on aggregate FDI flows. In particular, they illustrate the advantages of using a Heckman selection model to avoid problems generated by country-pair observations with zero FDI flows. Silva and Tenreyro (2006) discuss why the Poisson quasi-maximum-likelihood (PQML) specification might be superior in these situations, but recent montecarlo experiments by William Martin and Cong S. Pham (2008) suggest that the Heckman selection method may perform much better when there are many zero-observations, as is generally the case with large panel datasets. Because the issue appears to remain somewhat arguable, Razin and Sadka's Chapter 8, comparing different methodologies (though not the PQML), is particularly relevant at this point in time.

III. Unexpected parallels

Razin and Sadka are also preoccupied by the form capital flows. In particular, they wish to account for why portfolio investment is more volatile and less prevalent relative to FDI in developing countries. To this end, they suppose that there is some split between a manager's knowledge of a firm's efficiency and the owner's assessment. The only way for an owner to have full knowledge of a firm's true efficiency level is to buy a controlling interest in the firm. With superior information, a controlling owner can optimize the firm's capital stock through additional capital investment. The alternative is to purchase the firm as a portfolio investor. Without being able to observe the firm's true efficiency level, portfolio investors instruct the firm manager to adjust the firm's capital stock according to an expected level of efficiency, which inevitably is less profitable than optimizing according to the true efficiency level. The market knows that controlling owners have superior knowledge of a firm's profitability. Selling one's interest as a controlling owner naturally engenders a degree of suspicion in the marketplace which selling as a portfolio investor does not—a surprising and compelling application of the classic lemon problem.

Of course, the lemon problem would be irrelevant if there were no shocks in the economy that might motivate an investor to sell her interest in the firm. But the authors have planted liquidity shocks in the model. The shocks fall idiosyncratically across investors, with some investors consistently more prone to liquidity shocks than others. These "illiquidity-prone" agents are less entranced by the idea of purchasing a controlling interest through direct investment. Controlling ownership permits better management and higher profits, but

liquidation of the direct investment incurs a lemon cost when buyers insist on paying less than they would to portfolio investors, not knowing if the direct investor is cashing out to cover (for instance) a hypothetical margin call or because she has inside information that the firm got a bad efficiency draw. Thus, investors with a lower propensity for liquidity shocks engage in FDI while illiquidity-prone investors such as open-ended mutual funds engage in foreign portfolio investment, making portfolio investment much more volatile than FDI. If the liquidity shocks are correlated with an aggregate shock, then the economy where aggregate shocks are more prevalent attracts very few of the illiquidity-prone investors who prefer portfolio investment. Thus, Razin and Sadka offer a rigorous theoretical justification of why FDI is large as a proportion of total capital inflows in countries where there is less transparency about firm productivity and more macroeconomic volatility.

While attacking a different question, Razin and Sadka's depiction of the impact of asymmetric information on the choice of FDI versus portfolio investment is striking in its similarity to Markusen's exposition of the impact of asymmetric information on the choice of FDI versus licensing. There is an incompleteness in the FDI/portfolio investment model: why can't an investor about to purchase a controlling interest in a firm buy a put option to sell at a price above the market's skeptical lemon-loss price before the purchase takes place? There appears to be an unexploited arbitrage opportunity. Markusen's problem is a bit different in form, but with a similar incompleteness. A firm considers setting up an office to sell its product abroad, but is uncertain about market conditions in the host country. It does not if the size of the market will make the sunk cost involved in establishing its own affiliate for marketing and distribution worthwhile. A local agent could be contracted to market and distribute the

product, but would extract some of the producer's profits in return, as there is an incentive to minimize its marketing efforts for a given payment. Because it can not credibly stipulate the local agent's effort level within the contract, the firm must pay the agent an additional "information rent" to induce him to exploit the market's full sales potential, which the agent knows but the firm can only guess. The agent acts as an arbitrageur, in a sense, insuring the firm (for an extra fee) against the possibility that the foreign market is too small to make investment in an affiliate profitable. An arbitrageur in the Razin and Sadka model could serve a similar role, even without superior information, insuring the potential direct investor against the possibility of a lemon-loss in the event of a liquidity shock, but receiving an extra rent because it knows the direct investor's outside option involves the lemon-loss. If one imagines a dynamic extension, liquidity shocks are not altogether different from demand shocks that could act in a manner very similar to Markusen's demand uncertainty.

The case also illustrates an unexplored bridge between the micro-level and macro-finance approaches: where does the role of the investor stop and the firm begin? In the Markusen models, the firm makes investment decisions. In Razin and Sadka's world, the investor runs the firm, at times through a manager.[4] This begs the question, should the firm be insuring itself by contracting an agent, or should the direct investor be insuring herself by purchasing a put option? Either approach should result in the same behavior on the part of the affiliate, but they currently are split into two separate literatures.

IV. Recent advances and questions for the future

[4] As an illustration, in the literature of open-economy macroeconomics, the consumer is the investor and the firm's manager acts in the interest of the consumer (since the manager generally is given the consumer's discount factor when profit maximizing, be it constant or stochastic).

Beginning with studies by Joshua Aizenman (1992 and 1993) and continuing with a host of new theoretical papers and some empirical work, it is becoming clear that multinational firms play an important but little understood role in the transmission of macroeconomic shocks across countries. To the degree that there are nontraded goods and that asset markets are not completely effective at insuring against country-specific shocks, the multinational firm may be an important mechanism for risk sharing. Like factor-price equalization in trade models, risk-sharing is often used as a benchmark in international macroeconomics to judge whether capital or other resources are distributed efficiently across countries. Silvio Contessi (2007) shows that multinational activity increases the comovement of consumption relative to output across countries. Natalia Ramondo and Veronica Rappaport (2008), prove that even complete asset markets may not be sufficient to allow perfect risk sharing when there are nontraded goods and countries are asymmetric in size or the volatility of shocks[5] and that the activities of multinational firms can help fill this gap. The treatises by Markusen and Razin and Sadka suggest new questions in this context. With different factor intensities, are some industries better than others as conduits for the risk-sharing properties of multinational production? It is known that a lack of transparency retards the deepening of credit markets, but is it also a barrier to risk sharing insofar as it inhibits FDI? Pol Antras, Mihir A. Desai, and C. Fritz Foley (2007) suggest that it is. Is firm behavior associated with vertical fragmentation (inside or outside the firm) more effective than horizontal FDI as a hedge against country-specific shocks, or does it exacerbate their effect? A few papers, notably Bergin, Feenstra, and Hanson (2008) and a series of papers by John McLaren, are beginning to construct frameworks to answer these

[5] See also Russ (2007) for a discussion of risk sharing and the asymmetric volatility of shocks.

questions looking at vertical fragmentation outside the firm, but the specific characteristics of multinationals proposed in the texts discussed above—knowledge capital and specialized screening, asymmetric information and internalization or controlling ownership, industrial structure, intra-industry versus inter-industry flows of FDI—have not yet entered the discussion.

Similarly, to the degree that there exists a relationship between the return on net foreign assets and exchange rate behavior as suggested by the recent valuation effect literature originated by Pierre-Olivier Gourinchas and Helene Rey (2007), Razin and Sadka's liquidity model and Markusen's model of endogenous markups are likely to have implications for the behavior of exchange rates in a dynamic model. Lilia Cavallari (2007), for example, links multinational activity with deviations from purchasing power parity. Valderrama and Smith (2008) have begun to uncover the business cycle properties of FDI versus portfolio or equity investment in a small open economy in a model with perfect competition and a representative firm, but the business cycle implications of choices made by firms that behave as modeled by Markusen and direct investors challenged in the ways described by Razin and Sadka have yet to be explored. Antras and Helpman (2004) have pushed forward the micro-level theory of internalization—is there a flip side relating to the investor's choice of FDI versus portfolio investment? That is, are there beneficial external economies for the agent investing in an MNE instead of purchasing securities from separate firms listed in separate markets, over and above any boost in profitability arising from internalization? We know that MNEs act as financial intermediaries in a direct way by channeling investment funds between branches in different countries. Does the internalization of diverse production and sales activities by the multinational firm also act as a

form of financial intermediation when it forces long-term cross-border investment due to Razin and Sadka's lemon-liquidation cost?

There is also a question of whether FDI in different industries—nontraded versus tradable, or financial versus nonfinancial, intermediate versus final goods—has different implications for a country's business cycle? On the issue of financial versus nonfinancial industries, at least, Cetorelli and Goldberg (2008) and de Blas and Russ (2008) suggest this could well be true. If so, which market structures and which industries should be the focus of a unified theory of FDI in an open economy? Finally, if firms are earning revenues in one currency and paying dividends in another, then Russ (2007) suggests that the source of financing, omitted for the sake of clarity in past models, may motivate a complex relationship between FDI and exchange rate fluctuations. Outside of a few papers like Froot and Stein (1991), the relevance of the source of financing is a topic with little empirical or theoretical guidance to date.

In short, from fragmented literatures, a unified theory of foreign direct investment is slowly emerging. On the foundations of general equilibrium, sunk costs, asymmetric information, and economies of scale, micro-level modeling engines are slowly merging with macro-finance questions to form a new theory of FDI. All of the questions mentioned above regarding risk-sharing, business cycle transmission and behavior, exchange rate behavior, industry-specific considerations, and the structures and sources of financing could have nontrivial implications for policy analysis as well as our basic understanding of the role of the multinational in the global economy. It is fortunate that the scholars who originated the literatures are providing

the field with manuals to help guide the way forward. Despite different philosophical beginnings, they are leading us down converging paths.

Antras, Pol, Mihir A. Desai, and C. Fritz Foley (2007), 'Multinational Firms, FDI Flows and Imperfect Capital Markets." *NBER Working Paper No. 12855*.

Antras, Pol and Elhanan Helpman (2006), 'Contractual Frictions and Global Sourcing.' *NBER Working Paper No. 12747*.

Bergin, Paul R., Robert C. Feenstra and Gordon H. Hanson (2008), 'Outsourcing and Volatility.' Manuscript, University of California, Davis.

Cavallari, Lilia (2007), 'A Macroeconomic Model of Entry with Exporters and Multinationals.' Berkeley Electronic Journal of Macroeconomics 7(1): 32.

Cetorelli, Nicola and Linda Goldberg (2008), 'Banking Globalization, Monetary Transmission, and the Lending Channel.' Manuscript, Federal Reserve Bank of New York.

Contessi, Silvio (2007), 'International Macroeconomic Dynamics, Endogenous Tradability, and Foreign Direct Investment with Heterogeneous Firms.' Manuscript, Federal Reserve Bank of St. Louis.

de Blas, Beatriz Perez and Katheryn N. Russ (2008), 'FDI in the Banking Sector: Why lending costs fall when spread proxies increase.' Manuscript, University of California, Davis.

Froot, Kenneth A. and Jeremy C. Stein (1991), 'Exchange Rates and Foreign Direct Investment: An imperfect capital markets approach.' *Quarterly Journal of Economics* 106: 1191-1127.

Gourinchas, Pierre-Olivier and Helene Rey (2007), 'International Financial Adjustment.' Journal of Political Economy 115(4): 665-703.

Martin, William and Cong S. Pham (2008), 'Estimating the Gravity Model When Zero Trade Flows Are Frequent.' Manuscript presented at the 2008 Econometric Society North American Summer Meetings, Carnegie Mellon University, Pittsburg, PA, 21 June.

Ramondo, Natalia and Veronica Rappaport (2008), 'The Role of Multinational Production in Cross-Country Risk Sharing.' Manuscript, University of Texas at Austin.

Russ, Katheryn N. (2007), 'The Endogeneity of the Exchange Rate as a Determinant of FDI.' Journal of International Economics 71:344-372.