THE WORLD BANK

Research Observer

Volume 25 • Number 2 • August 2010

INSIDE

ISSN 0257-3032 (Print)
ISSN 1564-6971 (Online)
www.wbro.oxfordjournals.org

OXFORD JOURNALS

THE WORLD BANK

Research Observer

The World Bank Research Observer is intended for anyone who has a professional interest in development. *Observer* articles are written to be accessible to nonspecialist readers: contributors examine key issues in development economics, survey the literature and the latest World Bank research, and debate issues of development policy. Articles are reviewed by an editorial board drawn from across the Bank and the international community of economists. Inconsistency with Bank policy is not grounds for rejection.

The journal welcomes editorial comments and responses, which will be considered for publication to the extent that space permits. On occasion the *Observer* considers unsolicited contributions. Any reader interested in preparing such an article is invited to submit a proposal of not more than two pages to the Editor. Please direct all editorial correspondence to the Editor, *The World Bank Research Observer*, 1818 H Street, NW, Washington, DC 20433, USA.

For more information, please visit the Web sites of the *Research Observer* at www.wbro.oxfordjournals.org, the World Bank at www.worldbank.org, and Oxford University Press at www.oxfordjournals.org.

THE WORLD BANK

Research Observer

Volume 25 • Number 2 • August 2010

Ideas and Innovation in East Asia

Milan Brahmbhatt • Albert Hu

The generation, diffusion, absorption, and application of new technology, knowledge, or ideas are crucial drivers of development. The authors examine the exceptionally fast growth in domestic innovation efforts in Korea, Taiwan (China), Singapore, and China, drawing on information about R&D as well as patent and patent citations data. They also use the World Bank Investment Climate Surveys to investigate sources of technological innovation in the other middle- and low-income East Asian economies. They then evaluate the role of three main channels for knowledge flows to East Asia—international trade, acquisition of disembodied knowledge, and foreign direct investment. Results from estimating an international knowledge diffusion model using patent citations data show that, while East Asian innovations continue to draw heavily on knowledge flows from the US and Japan, citations to the same or to other East Asian economies are quickly rising, indicating the emergence of national and regional knowledge stocks as a foundation for innovation. JEL codes: F2, L2, O3

Knowledge creation and diffusion is at the heart of economic growth and development. Romer (1993), for example, stresses the importance of overcoming 'idea gaps' relative to 'object gaps' in the process of development, that is of overcoming barriers to the creation and absorption of productive ideas versus gaps in the availability of objects such as factories or raw materials.[1] No other part of the developing world has seen more dynamism and success in knowledge diffusion and creation over the last three decades than in the more advanced emerging market economies of East Asia. We aim to document the patterns and trends of knowledge diffusion and creation in these dynamic East Asian economies, to relate them to the economic forces that have led to greater economic integration in the region, and to place them in a comparative global context. To sharpen the comparative perspective, we contrast, whenever possible, statistical findings on East Asia with those on other parts of the world. We do not attempt to link

The World Bank Research Observer

doi;10.1093/wbro/lkp017 Advance Access publication November 11, 2009 25:177–207

innovation outcomes to specific public policies in East Asian economies, although we are hopeful that the analysis may be suggestive of fruitful hypotheses for future research in this regard.

A working definition of innovation is needed before we embark on our analysis. The economic and technological heterogeneity of East Asia necessitates a definition of innovation that encompasses the introduction of new or improved goods, services, production processes, and marketing methods, as well as better modes of business organization in general. In particular, a key component of innovation in East Asia is what can be broadly termed 'knowledge diffusion'. This involves absorbing and adapting existing knowledge and applying it in a productive and commercially viable way. It includes knowledge flow that crosses national borders as well as diffusion of technology within those borders. We are agnostic about whether such flows represent an externality, a knowledge spillover, or that all the private rents have been appropriated.[2]

East Asia as a region has outspent every other region on R&D (as a proportion of GDP) over the last decade. Within the region, disparities in spending have widened, though. Newly Industrialized Economies (NIEs) such as Korea, Singapore, and Taiwan (China) now devote 2 percent or more of GDP to R&D, comparable to the level in developed countries. China has seen a rapid intensification of R&D spending that has brought it close to its official target of 1.5 percent of GDP. On the other hand middle-income economies such as Indonesia, the Philippines, and Thailand spend a miniscule 0.1–0.2 percent of GDP on R&D. Patenting per capita in the NIEs has grown at a pace four times that in the developed world and is now approaching average developed-country levels; on the other hand, patenting in most middle- and low-income economies of the region remains negligible.

Notwithstanding the difficulty of establishing causal relations, a distinctive and common feature of those East Asian economies that have made significant progress in technological innovation is their exceptionally high engagement in international trade. The robust econometric evidence on capital goods imported as a conduit of technology diffusion has also been corroborated from surveys of Southeast Asian firms who report new machinery as the primary source of innovation. While the econometric evidence is mixed, a rich body of case study literature argues that East Asian firms may have derived significant technological benefits from exports under long-term contract manufacturing agreements with foreign multinational corporations, as part of the latter's global production networks.

Recent decades have also seen deepening East Asian economic integration, reflected in increasing flows of goods, services, and investment between the region's economies, largely driven by growing vertical specialization within the region. This, and the increasing pre-eminence of the R&D capability of East Asia's NIEs, raises the possibility of increasing technology diffusion within the region.

Using patent citations made by the U.S. Patent and Trademark Office (USPTO) as a measure of knowledge flow, we find that, while East Asia continues to draw heavily on knowledge flows from the US and Japan, citations to other "compatriot" patents from the same East Asian economy or to other East Asian economies are quickly rising, indicating the emergence of East Asian national and regional knowledge stocks which are now providing an indigenous or regional foundation for innovation and cross-border knowledge flows.

A key insight of the technological innovation literature that effective knowledge diffusion is dependent on the absorptive capacity of the firms at the receiving end finds support in our patent citation analysis. While Japan and the United States remain key sources of knowledge diffusion for Korea, Taiwan (China), and Singapore, economies such as China and Malaysia, who have weaker indigenous R&D capability, cite Korean and Taiwan (China) patents more intensively, controlling for existing stocks of patented knowledge.

Our analysis is subject to two caveats. First, while we characterize the geographical focus of the study as East Asian, we are mainly dealing with the more advanced and middle-income economies of East Asia. Second, it is beyond the scope of this paper to establish causal linkage from public policy to the patterns of knowledge creation and diffusion that we document, although we do observe correlations between the two in our analysis of the East Asian experience.

The rest of the paper is organized as follows. We provide, in a comparative context, an overview of technological innovation in East Asia using R&D and patent statistics and evidence drawn from the World Bank's Investment Climate Surveys (World Bank, 2006a). We then survey the theories and empirics on channels of international knowledge diffusion and summarize East Asia's experience with these channels. We use patent citations to investigate the intensity of knowledge diffusion both from Japan and the United States to East Asia, and within East Asia. We conclude by briefly drawing implications for policies and institutions that may help foster domestic innovation as well as absorption of knowledge from abroad.

Technological Innovation in East Asia: An Overview

East Asia has made rapid progress in technical change over the past decade. But the progress has been uneven. While Korea, Taiwan (China) and, to a lesser extent, Singapore are closing in on the world technology frontier in areas they excel at, and while China is going through a rapid process of R&D intensification, the rest of East Asia, particularly Southeast Asia, remains far behind, both in absorbing existing technologies and in creating new ones. We will use

conventional measures, including patents and R&D, to provide an overview of the current status of technological innovation in East Asia.

R&D in East Asia

Total world spending on R&D amounted to $830 billion in purchasing power parity (PPP) terms in 2002 (table 1),[3] of which some 78 percent was performed by developed countries, much higher than their 59 percent share in world GDP (in PPP terms). However, the developed country proportion has fallen over the last decade, as developing or emerging economies raised their share from 13 percent in 1992 to 22 percent in 2002. East Asia has been at the heart of the rise in developing country R&D, contributing almost three-quarters of the increase and quintupling over the decade (in nominal terms) to reach $112 billion in 2002 or 13.5 percent of the world total. R&D intensity in East Asia—the ratio of R&D spending to GDP—rose from 0.7 percent in 1992 to 1.2 percent in 2002.

Table 1. Research and Development Expenditures

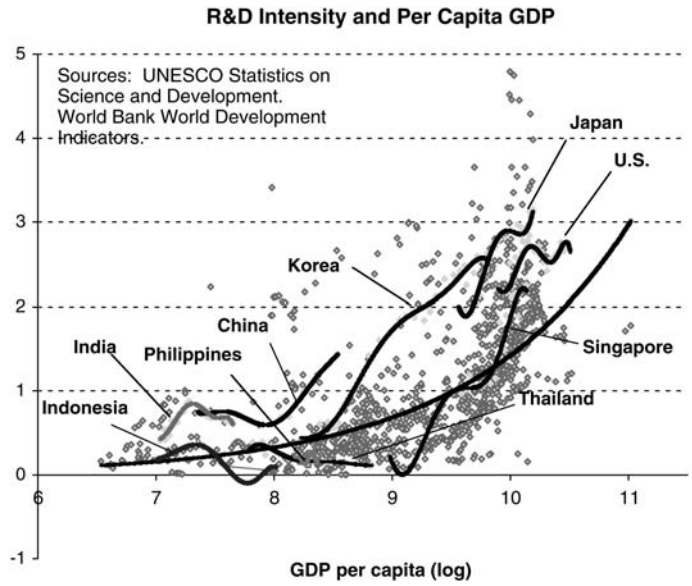
	R&D spending 2002		*R&D (% GDP*)*	
	US$ bill (PPP)	*% of world*	*1992*	*2002*
East Asia	111.7	13.5	0.7	1.2
NIEs	36.4	4.4	1.6	2.2
Hong Kong	1.1	0.1	0.3[b]	0.6
Korea	20.8	2.5	1.9	2.5
Singapore	2.2	0.3	1.2	2.2
Taiwan, China	12.2	1.5	1.8	2.3
Southeast Asia	3.3	0.4	0.1	0.2
Indonesia	0.3	0.0	0.1[c]	0.1[a]
Malaysia	1.5	0.2	0.4	0.7
Philippines	0.4	0.0	0.2	0.1
Thailand	1.1	0.1	0.2	0.2
China	72.0	8.7	0.8	1.2
World	829.9	100.0	1.7	1.7
Developed	645.8	77.8	2.3	2.3
Japan	106.4	12.8	2.9	3.1
USA	275.1	33.1	2.6	2.6
Developing	184.1	22.1	0.6	0.9
Latin America	21.7	2.6	0.5	0.6
Emerg. Europe	30.3	3.7	1.0	1.2

* Regional data are sum of R&D divided by sum of PPP GDP. [a]2001; [b]1995; [c]1994.
Source: UNESCO (2004, 2006).

East Asian economies differ widely in R&D performance. Korea, Singapore, and Taiwan (China) now devote 2.2–2.5 percent of GDP to R&D spending, comparable to levels in the United States and the upper end of the scale among developed economies. On the other hand R&D spending in economies such as Indonesia, the Philippines, and Thailand is only 0.1–0.2 percent of GDP, among the lowest of all economies for which we have data. In-between these two extremes is China, where R&D spending has risen at 20 percent a year over the last decade to reach 1.4 percent of GDP by 2004, or $109 billion in PPP terms.[4] R&D spending in Malaysia also accelerated after the mid-1990s, reaching 0.7 percent of GDP by 2002.

Figure 1 shows a scatter plot of panel data for R&D intensity and the logarithm of per capita GDP for a sample of developed and developing economies between the mid-1970s and the early to mid-2000s. A fitted regression line—obtained by regressing R&D to GDP ratio on the logarithm of per capita GDP and the square of it—is plotted, as well as the time paths of a small sample of countries. R&D intensity not only increases with per capita GDP but does so at an accelerating pace. Among the East Asian economies for which we have data, China, Korea, and Singapore have demonstrated an R&D intensification—the acceleration of R&D expenditure in relation to GDP—trajectory that is much steeper than that elsewhere in the world. On the other hand R&D intensity in Indonesia, the Philippines, and Thailand has been decelerating, both before and after the

Figure 1. R&D Intensity and Per Capita GDP, mid 1970s to mid 2000s



financial crisis of the late 1990s. The three countries have also been investing less in R&D than their level of economic development would predict.

Patenting in East Asia

The limitations of patents as an indicator are well known: not all inventions are patentable and not all patentable inventions are patented. Of the inventions that are patented, the vast majority do not find their way into commercial utilization, so the distribution of the economic and technological significance of patents is highly skewed. Nevertheless an empirical regularity that arises from the literature on patents and R&D is the high correlation between the two.[5] The relatively standard process that all patent applications have to go through makes the counts of patents granted by national patent offices to inventors of different nationalities a useful tool for the enormously difficult task of cross-country comparison of technological innovation. For this study we use patents granted by the USPTO.[6] While this might induce bias if the propensity to seek U.S. patents varies across countries, the United States has been the most important export market for East

Table 2. USPTO Patents Granted, annual averages

	Number of patents		*Patents per 100,000 people*		
	1990–04	*2000–04*	*1990–04*	*2000–04*	*% change*
East Asia (9)	2,239	12,108	0.14	0.72	17.6
NIEs	2,159	11,601	2.93	14.74	17.5
Hong Kong	184	616	3.15	9.32	11.4
Korea	633	4,009	1.44	8.67	19.7
Singapore	36	382	1.09	9.87	24.6
Taiwan (China)	1,307	6,593	6.30	30.17	17.0
Southeast Asia	31	140	0.01	0.04	15.3
Indonesia	6	15	0.00	0.01	8.8
Malaysia	13	64	0.07	0.28	15.3
Philippines	6	18	0.01	0.02	10.4
Thailand	6	43	0.01	0.07	20.9
China	48	368	0.00	0.03	22.9
World	107,361	182,523	1.98	2.95	4.1
Developed (21)	104,170	168,017	12.88	19.58	4.3
Japan	22,647	35,687	18.23	28.54	4.6
USA	59,024	97,104	23.00	33.56	3.9
Developing					
Latin Amer. (11)	173	368	0.04	0.08	6.3
Emerg. Europe (9)	205	348	0.07	0.12	5.6

Source: USPTO (2004).

Asia. It also has the most open and dynamic innovation system in the world. The USPTO patent and patent citations data have been well developed and have been subject to rigorous analysis. Therefore we believe USPTO patents are the best available patent statistics for us to assess the current status of technological innovation across East Asia in relation to the rest of the world.

USPTO Patents Granted to East Asia

The number of USPTO patents granted to East Asian economies averaged some 12,108 per year in 2000–04, more than five times the number a decade earlier, in 1990–94.[7] Over the same period the number of patents registered by selected Latin American countries increased from 173 to 368. Table 2 also shows patent counts scaled by population. In the early 1990s the number of patents per 100,000 people in East Asia was 0.14, which is 2–3 times the levels in Latin America and in Emerging Europe. By 2000–04, East Asian patents per capita had risen to 0.72, some 6–9 times the levels in the other two regions. The vast majority of these patents are generated by the NIEs, Taiwan (China) and Korea in particular, which by 2004 had become the 4th and 5th biggest recipients of USPTO patents in the world, after the United States, Japan, and Germany.

US patenting by East Asian economies varies enormously across the region. Taiwan (China) now generates some 30 patents per 100,000 people, about as many as Japan and the United States, the best performers among the developed economies. Korea, Hong Kong, and Singapore generate around 8–10 patents per capita per year, similar to the performance of the developed OECD countries in the mid-1980s, although only about half the average OECD level today. Further down the scale, Malaysia generates 0.2–0.3 patents per capita, similar to Korea in the mid-1980s. Finally, countries such as China, Indonesia, the Philippines, and Thailand bring up the rear with patents per capita in the 0.01–0.07 range, although patenting in China is rising very rapidly from a low base.[8]

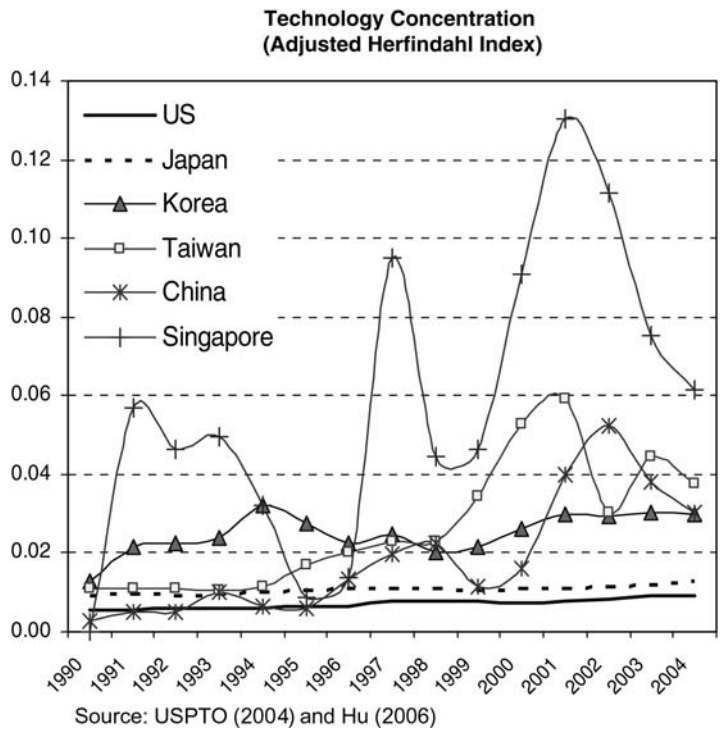
In which Technologies Is East Asia Innovating?

Is patenting in East Asia broadly diversified or concentrated in particular sectors? The USPTO classifies patents into one of around 480 technology classes. Figure 2 shows adjusted Herfindahl indexes of concentration across these technology classes for several major Pacific Rim economies.[9] An index level of 1 represents concentration in just one technology, while (in this case) an index of around 0.002 would mean equal distribution across all classes. Japan and the United States have the most diversified patent portfolios, whose composition has also been quite stable over time. Of the four East Asian economies, Singapore has the most technologically concentrated patent portfolio. Taiwan (China) had been

patenting more evenly across technology areas than Korea did before 1999, but the trend has been reversed since. China's USPTO patents have been relatively diversified. In recent years the change in the degree of concentration of Chinese patents seems to track those of Taiwan (China) and Singapore.

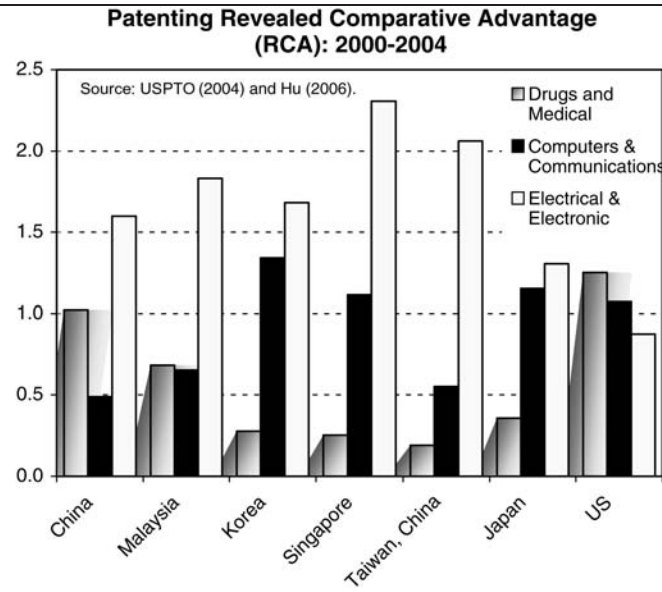
Jaffe and Trajtenberg (1999) distill the lengthy list of USPTO patent classes into six broad groups: chemicals; computers and communications; drugs and medical; electrical and electronic; mechanical; all other. We use these highly aggregated groups to obtain a more concrete understanding of the technological composition of the East Asian economies' USPTO patents. A major area of concentration in East Asia is electrical and electronics. The median share of patenting in this technology area for the seven East Asian economies in 2002–04 is 38 percent, ranging from a low of 25 percent in Hong Kong to 45–50 percent in Taiwan (China) and Singapore. The second most important area of concentration is computers and communications, with a median East Asian share of 15 percent, ranging from a low of 12 percent for China and Malaysia to 25–30 percent for Korea and Singapore. The share of East Asian patenting in these two areas has generally risen since the early 1990s.

Figure 2. East Asian Technological Concentration, Selected Economies, 1990–2004



The high shares of East Asian patenting in these sectors may just reflect the high technological opportunity and propensity to patent in these sectors worldwide. To put these concentration measures in a global perspective, we construct a Revealed Comparative Advantage (RCA) measure. This is computed as the share of a technology group in a given economy's total patents as a ratio of the same share for the world as a whole. A ratio above 1 suggests that an economy patents disproportionately in the technology group relative to the world average. Drawing a parallel with the corresponding concept in international trade empirics, we consider this economy as having a comparative advantage in that particular area of technology. The RCA measures for five East Asian economies, Japan and the United States are plotted in figure 3. The RCA ratio for electrical and electronics is well above unity for all five East Asian economies. For Singapore it is over two, suggesting that the country has twice as large a share of patents in this technology area as an average country. The electrical and electronics RCA for the United States, on the other hand, is below unity. That East Asia has been patenting disproportionately in electrical and electronics accords with the general perception of the region's strength in this area. Notable cases of excellence include Korea in DRAM technology and LCD manufacture and Taiwan (China) in the wafer foundry industry, and in testing and packaging services. By comparison most East Asian economies show a distinct revealed comparative disadvantage in the drugs

Figure 3. Patenting Revealed Comparative Advantage, Selected Economies and Technologies, 2000–2004



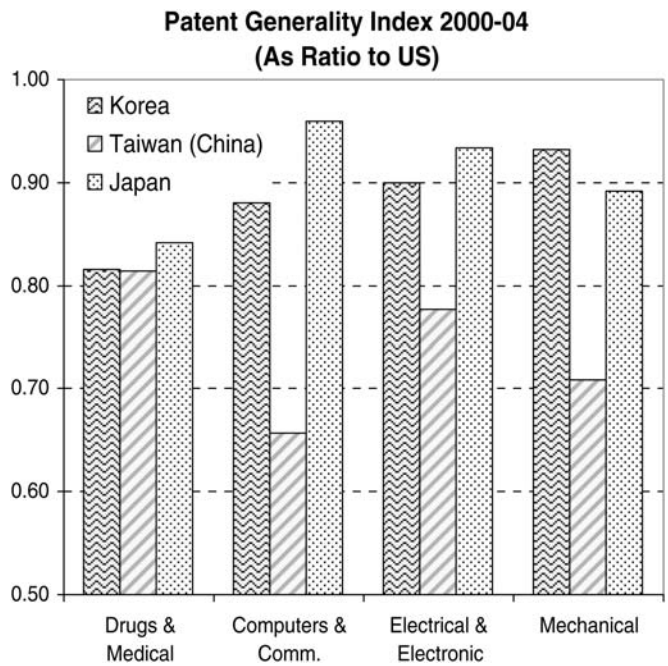
and medical sector, with only China achieving parity with the world average. In computers and communications, Korea and Singapore patent more than the world average, alongside Japan and the United States.

How Good is East Asian Patenting?

Although the volume of patenting in economies such as Korea and Taiwan (China) equals or exceeds that in most developed economies, is the same also true of the *quality* of their innovations? The technological or economic value of patents can vary enormously. In fact the distribution of patent values is highly skewed. A survey of the realized economic value of patents in Germany and the United States, for example, found that the top 10 percent of patents accounted for over 80 percent of economic value (Scherer and Harhoff 2000). Thus a simple count of patents may not provide an adequate summary of the quality of the underlying innovations.

A particularly useful feature of patents is that they contain citations to the previous patents and scientific literature, which define the "prior art" to which the patent is making an original contribution. Trajtenberg, Henderson, and Jaffe (1997) suggest measuring the quality of patents with indexes of patent

Figure 4. Patent Generality, Selected Economies and Technologies, 2000–2004



"generality" and "originality" based on patent citations.[10] A patent is said to have greater generality and impact if it is cited by patents from a wider range of technology classes. It is also said to be more basic or original if it cites patents from a wider range of patent technology classes.

Since the absolute magnitude of the indexes is hard to interpret, we plot, in figures 4 and 5 respectively, the originality and generality measures of Korean, Japanese, and Taiwan (China) patents from four technology groups relative to the corresponding measures of U.S. patents in the same technology group. U.S. patents have higher generality and originality indexes across all technology fields than those of Korea, Japan, and Taiwan (China).[11] Japanese patents generally achieve quality ratings that are 80 to 90 percent or more of U.S. quality ratings. Korea, in particular, is close to Japan in most technology areas, even matching or exceeding it in some. Taiwan (China) tends to achieve somewhat lower generality and originality scores, but still scores 70 to 80 percent of U.S. levels.

Investment Climate Surveys

Our analysis using R&D and patents statistics is somewhat skewed toward the higher-income and newly industrialized economies in East Asia. What is arguably

Figure 5. Patent Originality, Selected Economies and Technologies, 2000–2004

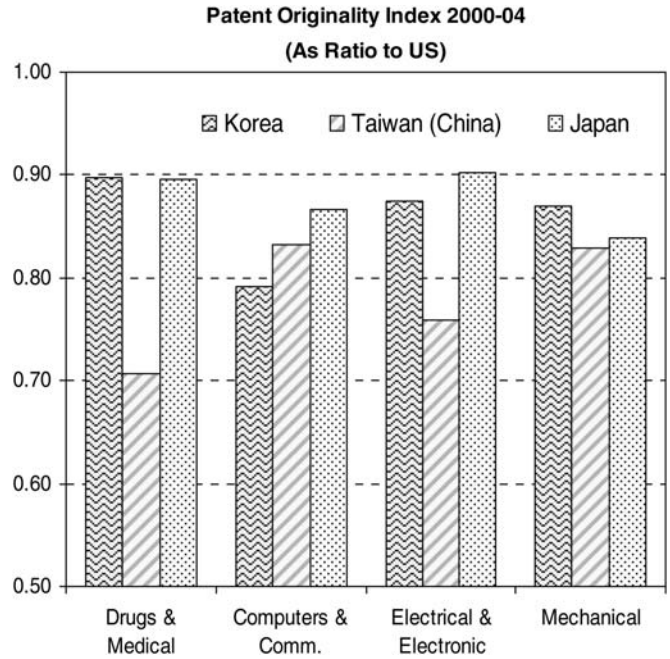


Table 3. Most Important Source of Technological Innovation, 2003 (% of Firms)

Source of innovation	*Cambodia*	*Indonesia*	*Malaysia*	*Philippines*	*Thailand*	*Average*
Embodied in new machinery	42.1	48.7	49.9	43.0	33.1	43.4
Cooperation with clients	11.9	15.1	8.6	9.7	17.2	12.5
By hiring key personnel	14.5	17.9	11.4	14.2	3.0	12.2
Developed within the firm locally	16.1	4.7	7.2	8.3	19.4	11.1
Transferred from parent	6.0	2.7	11.0	4.3	11.8	7.2
Developed with supplier	1.6	7.0	5.2	5.0	7.2	5.2
Other	7.8	3.9	6.7	15.5	8.2	8.4

Note: Malaysia is 2002, Thailand is 2004.
Source: World Bank Investment Climate Surveys.

more important in low- and middle-income economies, which are far from the world technology frontier, is absorptive capacity and channels to effect technology transfer from developed countries. Therefore R&D and patent statistics will not capture the main technological innovation activities in these economies. The challenge is the scarcity of meaningful measures of these diffusion related activities. The World Bank Investment Climate Surveys collected information on technological innovation activities in firms around the world. The survey data provide a window to examine the nature and sources of innovation in this less developed part of East Asia.

Table 3 shows that the single most important source of knowledge for firms in these East Asian economies was technology embodied in new machinery or equipment. In Indonesia and Malaysia, close to half of the technological innovation took the form of new machinery. The next two most cited sources of innovation—those developed in cooperation with client firms and the hiring of key personnel—were cited by 12–13 percent of firms, while innovations developed or adapted within the firm were cited by 11–12 percent. One should be cautioned against interpreting these different sources as exclusive to each other. Development within the firm locally, for example, may well be related to adopting new capital equipment. The Investment Climate Survey data obviously reaffirm the importance of embodied technology transfer in the form of new machinery for low- to middle- income countries.

Channels of Technology Diffusion: The Literature and the East Asian Experience

With developed countries performing 80 percent of world R&D, for a country that is far from the world technology frontier, absorbing and adapting existing,

advanced technology for its own use is often a more effective way of closing the technology gap than relying on the country's indigenous R&D. However, the extent and quality of knowledge absorption from abroad is dependent on the absorptive or learning capacity of the domestic economy, and conducting R&D can be an effective way of acquiring that capacity (Cohen and Levinthal 1989). Empirical evidence, albeit largely based on developed country data, indicates a large magnitude of the benefits of international knowledge flows.[12] This is especially relevant for most East Asian economies—including low- and middle-income countries—because of their generally exceptionally high levels of engagement in world trade and direct investment flows.

Technology Transfer through Imports

Capital goods become a conduit of international technology diffusion when they are imported by countries where the technology embodied in the capital equipment is new and more advanced. Coe and Helpman (1995) identify international R&D spillover through imports by relating a country's total factor productivity to its own R&D stock and a weighted sum of its trade partners' R&D stock, with the weights being the shares of the country's imports from its trade partners. They find that about one-quarter of the total benefits of R&D investment in a G7 country accrues to its developed-country trade partners from 1971–90. Coe, Helpman, and Hoffmeister (1997) refine and extend this approach to studying R&D spillover from developed to developing countries over the same period of time. They find that a 1 percent increase in the R&D stock of industrialized countries raises output in developing countries by 0.06 percent, half of which is contributed by the United States. A cross-country comparison shows that most East Asian economies have exceptionally high levels of imports (including capital goods imports) for economies at their level of per capita income.

Learning through Exports?

Developing country firms can acquire new technology and learning capabilities by exporting to the developed country market and engaging with customers in that market. A large literature of case studies has developed which affirms the beneficial effect of exporting on technical progress in East Asia, particularly for firms in the electronics industry.[13] As developed country firms outsource manufacturing activities to developing countries, to take advantage of the latter's low labor cost, a result of trade liberalization measures, these firms often provide their developing country subcontractors with technology, capital equipment, and training so that the latter can produce to the specifications demanded by customers in the developed country markets. Some 70–80 percent of Korea's electronics

exports were under this type of contract manufacturing by 1990, while over 40 percent of Taiwan (China)'s computer hardware exports were of this form at around the same time. Over the past 15 years China has seen such contract manufacturing become central to the enormous expansion of its manufactured exports.

In contrast to the compelling tone of the case study literature, econometric evidence on the learning effect of exports is rather inconclusive when the nonrandom choices that firms make regarding exports is taken into consideration. Clerides, Lach, and Tybout (1998) find little evidence for learning benefits from exports using plant-level data from Colombia, Mexico, and Morocco. However, Kraay (2006) and Aw, Chen, and Roberts (1997) find evidence that past export experience helps explain current productivity for Chinese and Taiwan (China) firms.[14]

Reconciling the gap between the case studies literature and the econometric literature, Pack (2006) observes that export data do not typically distinguish exports under long term contract manufacturing relationships, which are more likely to involve technology transfer from other types of exports; so it is perhaps not surprising that econometric studies based on generic export data arrive at only mixed results regarding export learning effects.[15]

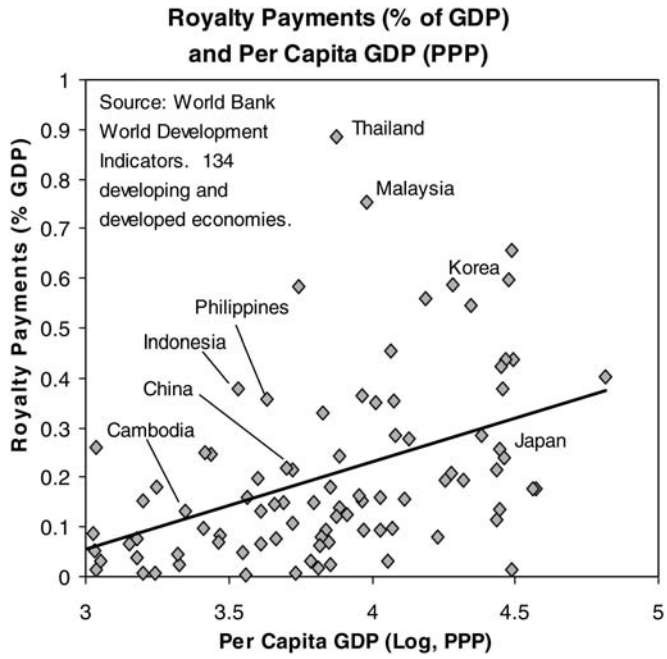
Technology Licensing and Disembodied Knowledge Flows

Firms can also purchase disembodied external knowledge through acquisition of patents, non-patented inventions, licenses, disclosures of know-how, trademarks, designs, patterns, and other consultancy and technological services. Royalty payments abroad provide a rough measure of this form of technology transfer. Figure 6 indicates that royalty payments abroad by East Asian economies are also generally much higher relative to other economies at similar income levels. Firms may also derive disembodied knowledge flows through technological spillovers, benefiting from open source information such as scientific, technical, and industry journals; informal contacts and communications through networks of researchers and specialists; trade and industry associations; and trade fairs.

Foreign Direct Investment

The literature on foreign direct investment (FDI) as a channel of technology diffusion makes a distinction between horizontal and vertical diffusion. The former occurs between competing firms in the same industry, whereas the latter represents knowledge spillovers between firms from different industries that are vertically linked.

Figure 6. Royalty Payments and Per Capita GDP, 2003



Gorg and Greenaway (2004) review 40 studies on horizontal productivity spillovers in manufacturing industries worldwide and conclude that only eight find unambiguous evidence of positive horizontal spillovers, mostly for developed economies.[16] Some authors (for example Aitken and Harrison 1999; Konings 2001) have attributed the lack of evidence for technology spillover from FDI to the competitive pressure FDI has imposed on the domestic market, which may dominate any technology spillover that may have occurred. Other authors (Kokko, Tansini, and Zejan 1996; Glass and Saggi 2002; Borensztein, De Gregorio, and Lee 1998; Lipsey 2000) argue that technology spillover is dependent on the absorptive capacity of the local firms: the smaller the gap between foreign and domestic firms, the greater the spillover.

Several studies have reported affirmative evidence on vertical technology spillover from FDI. Javorcik (2004), for example, finds that a one standard deviation increase in foreign presence in the purchasing sector of the economy in Lithuania is associated with a 15 percent rise in output of local firms in supplying sectors.[17]

Keller (2002) examines the impact of international trade, FDI, and disembodied knowledge flows (for example through direct communication) as channels of knowledge flow.[18] He finds that all three channels are significant, but that imports are the most important, explaining about two-thirds of the estimated

impacts, while FDI and disembodied flows explain about one-sixth each on average.[19]

Historically East Asian economies have embraced FDI with different levels of enthusiasm. Korea and, to a lesser extent, Taiwan (China) have tended to restrict FDI while emphasizing licensing of foreign technology and upgrading of domestic technological capabilities, including through domestic R&D and strengthening of technical education and labor force skills. Singapore's economic development has largely been driven by FDI. In-between these two extremes are China, which has drawn large FDI inflows, and middle-income Southeast Asian economies such as Malaysia, Thailand, Indonesia, and the Philippines (since the 1980s).

Deepening East Asian Economic Integration

East Asia has seen a deepening of intraregional economic integration over the past decades. According to a recent study by the Asian Development Bank (ADB 2008), the share of intraregional trade in total trade for integrating Asia—basically East Asia plus India—has steadily increased from a fifth in the early 1950s, to about a third in the 1990s, to over a half today. The report goes on to observe that "Asia now trades more with itself than either the EU or North America did at the outset of their integration efforts." But unlike the EU and North America, where intraregional trade is driven by final goods trade, in East Asia it has been driven by intermediate goods due to vertical specialization (IMF 2007). The increasing flows of goods, services, and investment between East Asian countries has generated potential opportunities for more intensive knowledge flow within East Asia.

Knowledge Flows to, from, and within East Asia: Patent Citations

The large empirical literature on international knowledge diffusion largely adduces evidence of knowledge flows by investigating correlations between activities that are expected to facilitate knowledge flows—for example, foreign trade, FDI, or technology licenses—and economic productivity. This section looks at evidence from patent citations, which provide a unique and more direct window onto flows of knowledge between inventors, firms, and economies.[20]

Patent Citations and Knowledge Diffusion

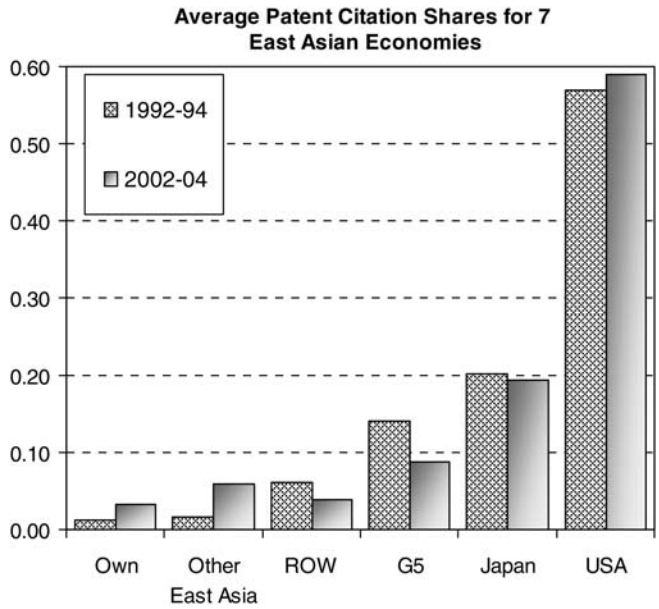
We use citations made by USPTO patents to other USPTO patents to trace knowledge flow within and beyond East Asia. An applicant for a USPTO patent

has a legal duty to disclose any knowledge of prior art related to his or her invention in the form of citations to a previously granted USPTO patent among other forms of publications of existing inventions, including patents granted in other jurisdictions and scientific journal articles. Patent citations therefore play an important legal function in delimiting property rights. On the other hand, the decision regarding which patents to cite ultimately rests with the patent examiner, who is supposed to be an expert in the area and hence to be able to identify relevant prior art that the applicant misses or conceals. All this makes patent citations perhaps less arbitrary than, say, academic journal citations, but it does also open the door to the possibility that citations may have been inserted by the patent examiner, which therefore do not necessarily track knowledge flow. In addition the citations data we use only include those made to USPTO granted patents, thus citations made to patents granted by other patent offices are not considered. To the extent that the propensity to cite non-U.S. granted patents varies across the East Asian countries that we investigate, our knowledge flow measures may be biased. However, we are not aware of any study that validates this concern.

A number of studies have directly or indirectly confirmed the usefulness of patent citations as an indicator of knowledge diffusion, despite their noise content. Trajtenberg (1990), for example, finds a robust correlation between citation-weighted patent counts and consumer surplus from the invention and diffusion of computer tomography. In a direct attempt to address the noisiness of patent citations as an indicator of knowledge flow, Jaffe, Trajtenberg, and Fogarty (2000) report results from a survey of inventors who have cited other inventors in their patent applications. They find that citations are a noisy indicator of knowledge flow, in the sense that knowledge flow is much more likely to have occurred where a citation is made; however, many citations also occur in the absence of any knowledge flow.[21]

Figure 7 shows the share of various foreign economies in patent citations made by seven East Asian economies (as a group).[22] The United States remains by far the largest source of citations for East Asian innovators, providing close to 60 percent, this proportion having risen slightly between 1992–94 and 2002–04. Japan is the second largest source, contributing close to 20 percent. (Korea is an interesting exception to this general pattern: its reliance on US citations is substantially lower than other East Asian economies—around 45 percent—while its reliance on Japan is higher, around 33 percent). The share of other G5 economies, defined here as comprising Canada, France, Germany, Italy, and the United Kingdom, is lower, less than 10 percent, having fallen over the last decade. Perhaps most interesting, the share of citations made by East Asian economies to other East Asian economies, while still low, is rising fast, picking up from 1.7 percent of citations in

Figure 7. Patent Citations by Seven East Asian Economes, 1992–1994 and 2002–2004

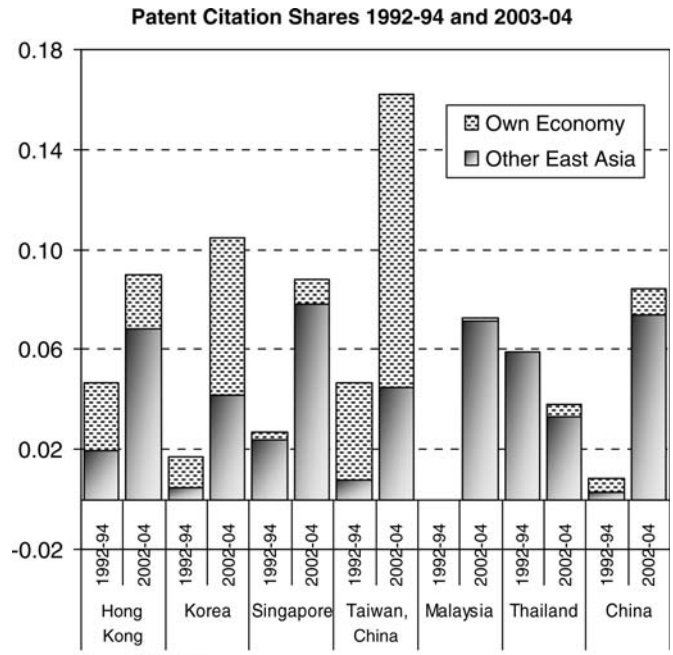


1992–94 to 5.9 percent in 2002–04. Most of these intra-East Asian patent citations are to those held by Korea and Taiwan (China), the two largest innovators in the region. Figure 7 also indicates that the share of citations by inventors in an East Asian economy to patents granted to other inventors in the same economy (referred to as "compatriot citations") is also rising, reaching 3.3 percent on average in 2002–04.

Figure 8 provides a closer look at the rise of intraregional and compatriot knowledge flows for individual East Asian economies. The share of citations to other East Asian economies, typically to Korean and Taiwan (China) patents, is highest—around 7–8 percent—in China, Hong Kong, Malaysia, and Singapore. The share of compatriot patent citations is highest in Korea—around 6 percent—and Taiwan (China), where it is already over 10 percent.

Such raw citation shares provide useful information on the gross flows of knowledge between economies but say little about the *intensity* of knowledge relationships. It is not surprising that East Asian economies should have large shares of citations to U.S. patents, simply because the United States is by far the largest generator of patents worldwide, and thus of potential citations. Even Japan, which produces almost as many patents per capita as the United States, still has over 40 percent of its citations to the latter. Researchers have therefore developed a *citation frequency* measure of how intensively patents in one country

Figure 8. Patent Citations Within East Asia, 1992–1994 and 2003–2004



Source: Hu (2006)

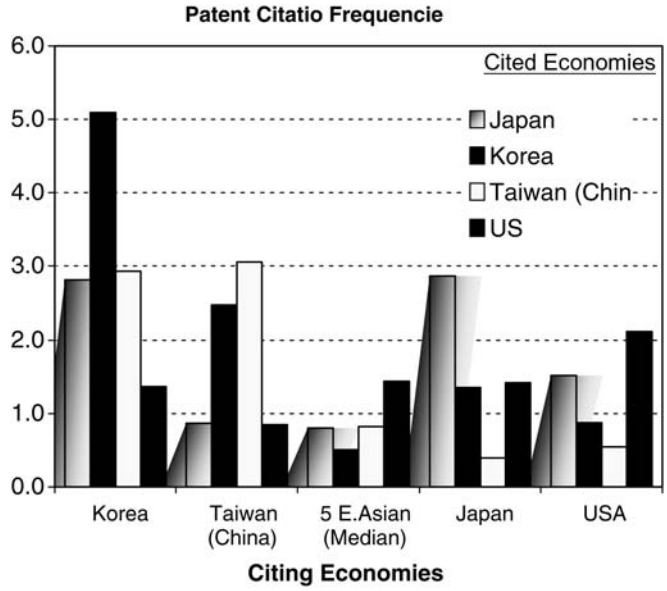
cite patents from another, after controlling for the size of the potential pool of citations between the two:[23]

$$CF_{A,B} = \frac{C_{A,B}}{N_A N_B}$$

where $CF_{A,B}$ is the number of citations made by country A's patents to those of country B, $C_{A,B}$, divided by the product of the potential number of citing patents in country A (N_A) and potential number of citable patents in country B (N_B).

Figure 9 shows patent citation frequencies between the United States, Japan, Korea, and Taiwan (China) in the electrical and electronics technology field. There are several striking features. Each of these four economies cites compatriot patents from the same economy much more intensively than patents in the rest of the world. Thus, after controlling for the fact that the potential pool of citable electrical and electronics patents in Korea is much smaller than the potential pool in the United States, Korean patents cite other Korean patents almost five times as intensively as they do U.S. patents. This finding is consistent with earlier findings that there is a significant degree of *geographical localization* in knowledge spillovers. Jaffe, Trajtenberg, and Henderson (1993) find that the frequency of citations from a patent in one American state to other patents from the same

Figure 9. Patent Citation Frequencies, Electrical and Electronics, 2004



state is higher than from other states, while Jaffe and Trajtenberg (1999) confirm that citation frequencies within OECD economies are much higher than those from one OECD economy to another.

A major reason for geographical localization of knowledge spillovers is the *tacitness of knowledge*. Complex scientific and technical knowledge often cannot be easily codified and can only be fully communicated if accompanied by face-to-face interaction. Tacitness and geographical localization help explain the economic usefulness of cities and industrial clusters, which facilitate face-to-face interactions and knowledge spillovers. At the national level these findings provide more evidence for the value of domestic R&D and innovation efforts, since they suggest that it is easier for local residents to absorb the knowledge spillovers from local innovations than it is from foreign ones. Jaffe, Trajtenberg, and Henderson (1993) also find that localization within U.S. states fades away over time, while Audretsch and Feldman (1996) find that geographical clustering is greatest in industries with high R&D intensity and high employment of skilled labor, as well as in industries at an early stage of their life cycle, when more of the knowledge about that industry is still in the heads of skilled workers and less of it has been codified in manuals and protocols.

Figure 9 also provides evidence for the high intensity of intra-East Asian cross-border knowledge flows. The citation frequency from Korean to both Taiwan (China) and Japanese patents is more than twice as high as the citation frequency

for U.S. patents. Reciprocating, Taiwan (China)'s citation frequency to Korea is also near three times its frequency with Japan and the US, while Japan's frequency to Korea is almost as high as its citation frequency to U.S. patents. These trends again confirm the growing regional dimension in East Asian knowledge flows.

Estimating the Knowledge Diffusion Model

The preceding analysis of patent citations data, while revealing interesting patterns, is "partial" in the sense that it has not fully accounted for all the factors that may determine the intensity of knowledge flow as measured by citation frequency. In this section we investigate how the intensity of knowledge flow between two countries varies, controlling for the age structure of a country's patent portfolio and how close the two countries are in technology space. A particular patent is more likely to be cited the longer it has been around, other things being equal, since the likelihood that the knowledge embodied in the patent has been observed and diffused would be higher. On the other hand, knowledge becomes obsolete over time and thus becomes less citable. To incorporate these two counteracting forces of knowledge diffusion and knowledge decay, we estimate a double exponential knowledge diffusion model,[24] the details of which we describe in the appendix.

It is also more likely for two countries to cite each other if there is a greater overlap in the technology fields that the two countries' inventors work in, everything else being equal. To measure how close two countries are in technology space, we construct a technology proximity measure for each pair of countries. When a patent is granted, the USPTO assigns it to one of over 400 technology classes on the basis of the technological nature of the invention. The technology proximity measure is essentially a correlation between two countries in the technology-class distributions of their patents.

Our primary interest in estimating the knowledge diffusion model lies with the country-pair fixed effects, that is, after the effects of knowledge decay and diffusion, and technology proximity, are controlled for, how much more intensive is the knowledge flow between one pair of countries than another? We report the country-pair estimates in table 4.[25] The estimates are scaled by the effect of the United States citing itself for ease of interpretation. For example, if the country-pair estimate for country X citing country Y is 0.5, this would imply that X cites Y half as frequently as the United States cites itself.

Confirming our finding of intensive compatriot patent citations obtained using "raw" citation frequencies earlier, the diagonal elements of the upper half of table 4 are all bigger than the off diagonal elements. In the case of Korea, localization of knowledge flow is 16 percent higher than that in the United States. Both

Table 4. Citation Frequencies: Estimated Country-Pair Fixed Effects*

Citing economies	*Cited economies*			
	USA	*Japan*	*Korea*	*Taiwan (China)*
USA	1.00	0.57	0.38	0.29
Japan	0.46	0.80	0.44	0.23
Korea	0.46	0.70	1.16	0.69
Taiwan (China)	0.26	0.25	0.71	0.83
China	0.36	0.31	0.44	0.41
Hong Kong	0.45	0.41	0.42	0.40
Malaysia	0.44	0.32	0.53	0.57
Singapore	0.95	0.60	0.93	1.63
Thailand	0.66	0.33	0.27	0.10

* Normalized by "US citing US" = 1.
Source: Hu (2006).

Korea and Taiwan have become sources of knowledge diffusion. Japan cites Korea at 44 percent of the frequency at which the United States cites itself. This is quite close to the frequency at which Japan cites the United States. Taiwan (China) cites Korea more intensively than it cites the United States and Japan. Korea reciprocates by citing Taiwan (China) almost as intensively as it cites Japan, and more intensively than it cites the United States. Korean and Taiwan (China) patents appear to be more important as a source of knowledge for both China and Malaysia than those of the United States and Japan. An explanation is that, given the absorptive capacity of China and Malaysia, it is easier for them to build upon the technologies of Korea and Taiwan (China) than those of the United States and Japan. Between the two city economies, Singapore shows an exceptionally high citation frequency to Taiwan (China), and also to Korea, which significantly exceeds or equals (also high) the citation frequencies to Japan and the United States. Thailand appears to be the anomaly by citing the United States and Japan more than it cites Korea and Taiwan (China).

Concluding Remarks

We have analyzed the patterns of the generation and diffusion of ideas in East Asia. There seem to be four groups of economies sorted by their ability to generate and absorb new ideas. The first group consists of the NIEs that are at an advanced stage of transition from imitation to innovation. Most notably, Korea and Taiwan (China) have emerged as centers of technological innovation and sources of knowledge diffusion for the rest of the economies in the region, based

on conventional measures of innovation and patent citations analysis. The second group consists of China, which is fast moving up the technology ladder, with heavy investment in both indigenous innovation and absorptive capacity. The third group includes middle-income economies, including Thailand and Malaysia. They have had varying degrees of success with absorbing technology from developed countries, but have not shown clear signs of graduating from imitation to innovation. The fourth group consists of low-income economies in East Asia, which are at the beginning of the diffusion stage.

While East Asian economies have adopted different approaches to technological innovation, there are common lessons to be drawn. First of all, everyone starts with absorbing existing technologies from abroad. The economies' successes with technical change is largely determined by their ability to avail themselves of channels of knowledge diffusion. These include capital goods import, exports to developed-country markets, hosting FDI, and movement of personnel who embody new and advanced technology. The relative significance of a particular channel may vary between countries, but the overall importance of these channels as catalyst of knowledge diffusion is universal.

Direct evidence in the form of robust causal relation has been hard to come by. However, there is an overwhelming amount of evidence based on correlations that is consistent with the hypotheses that international trade, FDI, and movement of people have accelerated intra- and inter-regional knowledge diffusion. Thus, policies that foster greater participation in trade and foreign investment, and which facilitate the easier movement of people involved in innovation—for example scientists, engineers, technical personnel, and business people more generally—are likely to accelerate knowledge diffusion.

Successful adoption of new technology may also depend on the technology gap. When the gap is too wide, the adopting economy may not have the absorptive capacity to learn and adopt the technology for its own use. The result that patents by China and Malaysia cite Korea and Taiwan (China) more intensively than they cite Japan and the United States corroborates this hypothesis.

The varying success East Asian economies have had with moving beyond pure imitation also indicates that exposing oneself to knowledge diffusion alone is not sufficient. The findings reported in this paper, explicitly and implicitly, lend support to the hypothesis that indigenous R&D is indispensable in the process of technological innovation in developing countries and that R&D and knowledge diffusion interact and reinforce each other. When a country is far from the world technology frontier, returns to absorptive activities are clearly higher than those to indigenous R&D. But as the country narrows the technology gap, R&D becomes productive both in strengthening the country's ability to absorb more advanced existing technology and in innovating.

Policymakers in developing countries are ultimately interested in what institutions and public policies lead to greater knowledge diffusion and generation. Despite the challenge of establishing causality in empirical development research, we would certainly want to explore in a more direct way the roles of trade, FDI, and the movement of people in the diffusion of new technology and thus make the connection between the patterns of knowledge diffusion we have documented and the determinants of such patterns. From such work more practical policy recommendations are likely to follow.

Appendix: Estimating the Knowledge Diffusion Model

The double exponential knowledge diffusion model is specified as:

$$CF_{iT,jtg} = (1 + TD_{iT,jtg})\alpha(ij, T, t, g)\exp(1 - \beta_1(T - t))(1 - \exp(-\beta_2(T - t)) \quad \text{(A1)}$$

where i and j denote citing and cited countries respectively; citing patents are granted in year T and cited patents in year t; g is one of the six main technological fields. The left-hand side is the citation frequency of patents of country i granted in year T citing country j's patents that are granted in year t in technological area g, that is the empirical frequency of a patent from the group defined by iT citing a patent from the group with the characteristics of jtg.

The probability that a patent is cited by another patent depends on, among other things, the likelihood that the cited patent comes to the knowledge of the inventor of the citing patent and the relevance of the knowledge embodied in the cited patent to the citing patent. The former increases with the lag between the grant dates of the citing and cited patents (T–t): the longer the cited patent has been around, the more likely it becomes known to the inventor of the citing patent, whereas the latter diminishes with the lag. As new knowledge emerges, as the wide adoption of the old knowledge reduces the economic rent accruable to the proprietary knowledge embodied in the cited patent, or both, the likelihood is reduced that the cited patent remains relevant and a prior art to a potential citing patent. The double exponential model in equation A1 captures these two processes with β_1 measuring the speed of obsolescence and β_2 the speed of diffusion.

The first term of the right-hand side of equation A1 contains the technology distance between the citing patents and the cited patents. The technology distance variable is defined as:

$$TD_{iT,jtg} = V_{iT}\,'V_{jtg}$$

where V_{it} is a 428-element vector of patent class shares of country i's USPTO patents granted in year t. TP is bounded between 0 and 1 and monotonically

increasing in the similarity between two economies' patent portfolio, which we use to measure the technological proximity between the two economies. The closer the potentially citing patent is to the potentially cited patent in the technology space, the easier it is for the citing patent to capture knowledge spillover from the cited patent and therefore the likelihood of citation increases. Finally, $\alpha(ij, T, t, g)$ represents a number of fixed effects we are interested in estimating:

$$\alpha(ij, T, t, g) = \exp(\sum_i \sum_j \alpha_{ij} D_{ij} + \sum_T \alpha_T D_T + \sum_t \alpha_t D_t \sum_g \alpha_g D_g).$$

For each set of fixed effects (Ds), one reference case is left out in the estimation. The citing–cited country-pair specific effect is estimated with the α_{ij}s. For example, with the United States citing itself as the reference group, α_{TWUS} would measure how much more intensively Taiwan (China) cites the United States relative to the United States citing itself. The citation frequency may also vary with the grant year of the citing patents; and this is captured by α_Ts. With both the effects of citing year and citation lag included, we are not able to estimate a full set of cited year effects of α_ts. Instead we group the cited year ts into groups and estimate the group effects. Lastly, we also allow the average citation frequencies of the six main technology fields to differ.

Instead of estimating all the country-pair effects in our citation database, which would lead to an explosion of the number of parameters to be estimated and overtax the data's identifying capability, we choose to be selective in the number of citing and cited countries to model. For cited countries, we include the United States, Japan, the G5, Korea, and Taiwan (China) in view of their dominance in patent numbers and as a source of citations. All seven East Asian economies and the United States and Japan are included as citing countries, the latter for comparison and benchmarking purposes.

The estimation results show that the model explains 99 percent of the variation in the citation frequency. Technology distance has a large impact on citation frequency: two patents being in the same technology class are 47 times more likely to cite each other than otherwise.

Funding

Albert Hu acknowledges financial support from the National University of Singapore Academic Research Fund (R-122-000-091-112).

Notes

Milan Brahmbhatt (corresponding author) is Economic Adviser in the Poverty Reduction and Economic Management Network of the World Bank; email address: Mbrahmbhatt@worldbank.org.

Albert Hu is Associate Professor in the Department of Economics at the National University of Singapore; email address: ecshua@nus.edu.sg. The authors thank Li Jia, Lim Zong Liang, and Radu Tatucu for invaluable research assistance. All opinions expressed in this paper are those of the authors and not necessarily those of the institutions with which they are affiliated.

1. See also Romer (1990a, 1990b) and Aghion and Howitt (1992, 2006).

2. The OECD's 2005 *Oslo Manual* defines innovation broadly as "the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organisation or external relations." Innovation activities are defined as "all scientific, technological, organisational, financial and commercial steps which actually, or are intended to, lead to the implementation of innovations" (OECD 2005).

3. UNESCO (2005, 2006). There is R&D data available for a number of economies through 2004 and 2005; but 2002 seems to be the most recent year for which there is comprehensive data for the world as a whole.

4. It is worth noting that the absolute value of China's R&D in PPP terms is particularly affected by the unusually large disparity between its PPP exchange rate (as calculated by the World Bank and other researchers) and its market exchange rate. Thus China's R&D expenditures in 2004 at market exchange rates were, according to data from UNESCO (2006), $23.8 billion, or only 21 percent of the PPP figure. By comparison Korea's R&D spending in 2003 was $22.8 billion in PPP terms and $16 billion at market exchange rates, or 70 percent of the PPP figure. In Malaysia R&D at market rates was 42 percent of R&D in PPP terms. Note, however, that while this issue is relevant for measuring absolute levels of R&D, it does not affect R&D intensity (the ratio of R&D to GDP), since both the numerator and denominator of that ratio use the same conversion rate. The recent 2007 revision of PPP exchange rate factors by the International Comparison Project should however lead to a downward revision in the PPP R&D figure for China in particular.

5. See Griliches (1992) for an early survey. Among recent studies, Bottazi and Peri (2005) examine the short and long run dynamics of knowledge production in OECD countries and find long run elasticities of patenting on R&D and the stock of foreign knowledge of around 0.8 and 0.6 respectively. Bosch, Lederman, and Maloney (2005) investigate the relationship between patenting and R&D worldwide, including developing economies. They find a significant relationship between patenting and R&D at the global level, although the estimated elasticity of patenting with respect to R&D for OECD economies (around 1) is substantially higher than among developing economies.

6. Issues and pitfalls in the use of patents as innovation indicators are discussed in Griliches (1992), Hall, Jaffe, and Trajtenberg (2001), Jaffe and Trajtenberg (2002) and Jaffe and Lerner (2004). Levin and others (1987) and Cohen, Nelson, and Walsh (2000) document the importance for U.S. firms of alternative methods of protecting proprietary knowledge such as secrecy, lead times, moving down the learning curve, and provision of sales and manufacturing services that complement the innovation.

7. This discussion draws on the NBER Patent Citation Database (http://www.nber.org/patents/, described in Hall, Jaffe, and Trajtenberg 2001), updated through 2002 by Bronwyn Hall (http://elsa.berkeley.edu/~bhhall/patents.html) and through 2004 by Albert Hu (Hu 2006). The use of U.S. patents may be justified by the fact that creators of commercially valuable inventions have a strong incentive to take out a patent in the United States, given its position as the largest market in the world. Close to 50 percent of patents granted by the USPTO in 2000–04 were to foreigners. Nevertheless, there is a large home bias in patenting (inventors are more likely to patent in their home jurisdiction than elsewhere), and inventors in different economies may also face different incentives to patent in the United States (for example depending on the level of exports to that country). These factors could introduce biases that need to be adjusted for. We concentrate on patents and patent citations for eight East Asian economies: China, Hong Kong, Korea, Malaysia, the Philippines, Singapore, Taiwan (China), and Thailand, together with two of their largest economic partners, the United States and Japan.

8. Hu and Jefferson (2005) suggest several reasons for the acceleration in Chinese patenting: (i) the acceleration in China's R&D spending noted above; (ii) the strengthening of China's Patent Law in 1992 and 2000; (iii) the vast influx of foreign direct investment to China, which has greatly increased the market value of intellectual property for both foreign and domestic firms; (iv) the rapid relative growth of "complex industries" like electronics and machinery, which involve many separately patentable subproducts and processes; (v) the acceleration of enterprise reform after the mid-1990s, which has greatly strengthened private property rights vis-à-vis state owned enterprises.

9. Bronwyn Hall suggested correcting for small sample bias when calculating Herfindahl indexes for patents-based concentration measures using the following formula: $\hat{\eta} = (N{\cdot}HHI - 1)/(N - 1)$, where HHI is the conventional Herfindahl index, $HHI = \sum_{j=1}^{J} (N_j/N)^2$, where j indexes the number of patent classes and N the total number of patents.

10. The indexes are calculated as: $1 - \sum_{i=1}^{N} S_i^2$ where S_i is the share of patent citations made to patent class i for originality and the share of patent citations made by patents from class i to the patent concerned for generality. The second term of the formula is the Herfindahl index of concentration. The indexes are thus negatively correlated with the concentration of patent citations.

11. Some East Asian economies have few if any patents in some technology fields, resulting in few citations with which to compute generality or originality indexes. The discussion therefore focuses on Korea and Taiwan (China), economies with sufficient patenting activity for meaningful measurement.

12. Eaton and Kortum (1996), for example, estimate that foreign technology accounts for at least 80 percent of domestic productivity growth in most OECD countries, the only exceptions being the United States and Japan, the two world leaders. Bottazi and Peri (2005) estimate that a 1 percent increase in U.S. R&D leads to a 0.35 percent rise in the number of USPTO patent applications from other OECD countries, with a lag of between 5 and 10 years.

13. See, for example, Hobday (1995, 2000), Kim (1997), Matthews and Cho (2000), Kim and Lee (2002), and Nabeshima (2004).

14. For surveys of the literature see Hoekman and Smarzynska-Javorcik (2006) and Tybout (2006).

15. Another possibility is that learning which is necessary for serving the export market takes place before firms actually start exporting (Hallward-Driemer, Iarossi, and Sokoloff 2002; Tybout 2006). Kim (1997) describes Samsung's efforts to master production of microwave ovens in the 1970s in response to a prospective order from J.C. Penney, with improvements in productivity preceding actual export flows. Nabeshima (2004) observes that to be selected as an Original Equipment Manufacturer (OEM) supplier firms already need to possess production and technological capabilities that allows them to meet quality, cost, and delivery requirements.

16. While 22 studies find positive and significant horizontal spillover effects, Gorg and Greenaway (2004) challenge the results of the 14 which do not use panel data, because such studies are unable to deal with problems of reverse causality. Haskel, Pereira, and Slaughter (2002) and Griffith, Redding, and Simpson (2003) find small but significant effects for the UK, while Keller and Yeaple (2003) find large, significant ones for the United States. Javorcik and Spatareanu (2003) find evidence for horizontal spillovers in Romania.

17. See also Blalock and Gertler (2005) and Saggi (2002).

18. Keller (2002) proxies disembodied knowledge flows by bilateral language skills—the proportion of the population in the recipient country who speak the language of the spillover sender country. The study looks at knowledge flows at the industry level among countries at the world's technology frontier—the G7 industrialized economies.

19. See also Xu and Wang (1999) and Schiff, Wang, and Olarreaga (2002)

20. The discussion draws on data from the NBER Patent Citations Database (Hall, Jaffe, and Trajtenberg 2001). See endnote 24 for further details.

21. Hall, Jaffe, and Trajtenberg (2001) delineate the conceptual, operational, and modeling issues that users of the data may come across.

22. All the citations that we analyze are "non-self' citations, which exclude citations made to a patentee's own patents.

23. For further details see Jaffe and Trajtenberg (2002) and Hu (2006).

24. The model was first proposed and estimated in Caballero and Jaffe (1993) and again estimated in Jaffe and Trajtenberg (1999) and Hu and Jaffe (2003). The model makes two identification assumptions regarding the underlying structure of the knowledge diffusion process. One is proportionality, that is more highly cited patents are more highly cited at all lags. The second assumption is stationarity, which means that the citation lag distribution does not change over time. While these may be restrictive, they are needed to identify the model.

25. Full results are reported in Hu (2006) and are available upon request.

References

ADB (Asian Development Bank). 2008. *Emerging Asian Regionalism: A Partnership for Shared Prosperity.* Manila: Asian Development Bank.

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth through Creative Destruction." *Econometrica* 60(2): 323–52.

_____. 2006. "Appropriate Growth Policy: A Unifying Framework." *Journal of the European Economic Association* 4:269–314.

Aitken, Brian J., and Ann E. Harrison. 1999. "Do Domestic Firms Benefit from Direct Foreign Investment? Evidence from Venezuela". *American Economic Review* 89:605–18. Audretsch, David, and Maryann Feldman. 1996. "R&D Spillovers and the Geography of Innovation and Production." *American Economic Review* 86(3): 630–40.

Aw, Ben Yan, Xiaomin Chen, and Mark J. Roberts. 1997. "Firm-Level Evidence on Productivity Differentials, Turnover, and Exports in Taiwanese Manufacturing." NBER Working Papers 6235, National Bureau of Economic Research, Cambridge, MA.

Blalock, Garrick, and Paul J. Gertler. 2005. "Foreign Direct Investment and Externalities: The Case for Public Intervention." In Theodore H. Moran, Edward M. Graham, and Magnus Blomstrom eds., *Does Foreign Direct Investment Promote Development?* Institute for International Economics, Washington, DC: 73–106.

Borensztein, Eduardo, José De Gregorio, and Jong-Wha Lee. 1998. "How Does Foreign Direct Investment Affect Economic Growth?" *Journal of International Economics* 45(1): 115–35.

Bosch, Mariano, Daniel Lederman, and William F. Maloney. (2005). "Patenting and Research and Development: A Global View." World Bank Working Paper 3739.

Bottazi, Laura, and Giovanni Peri. 2005. "The International Dynamics of R&D and Innovation in the Short Run and the Long Run." NBER Working Paper 11524.

Caballero, Ricardo J., and Adam B. Jaffe. 1993. "How High Are the Giants' Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth." In Olivier Jean Blanchard, and Stanley Fischer eds., *NBER Macroeconomics Annual 1993*. Cambridge, MA: MIT Press: 15–74.

Clerides, Sofronis K., Saul Lach, and James R. Tybout. 1998. "Is Learning by Exporting Important? Micro-Dynamic Evidence from Colombia, Mexico, and Morocco." *Quarterly Journal of Economics* 113(3): 903–47.

Coe, David T., and Elhanan Helpman. 1995. "International R&D Spillovers." *European Economic Review* 39(5): 859–87.

Coe, David T., Elhanan Helpman, and Alexander Hoffmeister. 1997. "North–South R&D Spillovers." *Economic Journal* 107:134–49.

Cohen, Wesley M., and Daniel A. Levinthal. 1989. "Innovation and Learning: The Two Faces of R&D." *The Economic Journal* 99:569–96.

Cohen, Wesley, Richard Nelson, and John Walsh. 2000. "Protecting their Intellectual Assets: Approbriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)." NBER Working Paper 7552.

Eaton, Jonathan, and Samuel Kortum. 1996. "Trade in Ideas: Patenting and Productivity in the OECD." *Journal of International Economics* 40(3/4): 251–78.

Glass, Amy Jocelyn, and Kamal Saggi. 2002. "Intellectual Property Rights and Foreign Direct Investment." *Journal of International Economics* 56(2): 387–410.

Gorg, Holger, and David Greenaway. 2004. "Much Ado about Nothing? Do Domestic Firms Benefit from Foreign Direct Investment?" *The World Bank Research Observer* 19(2): 171–97.

Griffith, R., S. Redding, and H. Simpson. 2003. "Productivity Convergence and Foreign Ownership at the Establishment Level." CEPR Discussion Paper 3765.

Griliches, Zvi. 1992. "The Search for R&D Spillovers." *Scandinavian Journal of Economics* 9:29–47.

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools." CEPR Discussion Paper 3094.

Hallward-Driemer, Mary, Giuseppe Iarossi, and Kenneth L. Sokoloff. 2002. "Exports and Manufacturing Productivity in East Asia: A Comparative Analysis with Firm-Level Data." NBER Working Paper 8894.

Haskel, J., S. Pereira, and M. Slaughter. 2002. "Does Inward Foreign Direct Investment Boost the Productivity of Firms?" NBER Working Paper 8724.

Hobday, Michael. 1995. *Innovation in East Asia: The Challenge to Japan*. London: Edward Elgar.

_____. 2000. "East versus Southeast Asian Innovations: Comparing OEM and TNC-led Growth in Electronics." In Linsu Kim, and Richard R. Nelson eds., *Technology, Learning and Innovation: Experiences of Newly Industrializing Economies*. Cambridge: Cambridge University Press.

Bernard Hoekman, and Beata Smarzynska-Javorcik eds. 2006. *Global Integration and Technology Transfer*. Washington, DC: World Bank and Palgrave Macmillan.

Hu, Albert. 2006. "Knowledge Flow in East Asia and Beyond." Background Paper, World Bank.

Hu, Albert, and Adam B. Jaffe. 2003. "Patent Citations and International Knowledge Flow: The Cases of Korea and Taiwan." *International Journal of Industrial Organization* 21(6): 849–80.

Hu, Albert, and Gary Jefferson. 2005. "A Great Wall of Patents: What is Behind China's Patent Explosion?" *Journal of Development Economics*, 90(1): 57–68.

IMF (International Monetary Fund). 2007. *Regional Economic Outlook: Asia and Pacific*. Washington, DC: International Monetary Fund.

Jaffe, Adam, and Josh Lerner. 2004. *Innovation and its Discontents*. Princeton, NJ: Princeton University Press.

Jaffe, Adam, and Manuel Trajtenberg. 1999. "International Knowledge Flows: Evidence from Patent Citations." *Economics of Innovation and New Technology* 8(1–2): 105–36.

_____. 2002. *Patents, Citations and Innovations: A Window on the Knowledge Economy*. Cambridge, MA: MIT Press.

Jaffe, Adam, Manuel Trajtenberg, and Michael Fogarty. 2000. "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors." *American Economic Review* 90(2): 215–18.

Jaffe, Adam, Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographical Localization of Knowledge Spillovers as Evidenced by Patent Citations." *The Quarterly Journal of Economics* 108(3): 577–98.

Javorcik, Beata Smarzynska. 2004. "Does Foreign Direct Investment Increase the Productivity of Domestic Firms? In Search of Spillovers through Backward Linkages." *American Economic Review* 94(3): 605–27.

Javorcik, Beata Smarzynska, and Mariana Spatareanu. 2003. "To Share or Not to Share: Does Local Participation Matter for Spillovers from Foreign Direct Investment?" Policy Research Working Paper 3118, World Bank, Washington, DC.

Keller, Wolfgang. 2002. "Knowledge Spillovers at the World's Technology Frontier". Centre for Economic Policy Research Discussion Paper 2815.

Keller, Wolfgang, and Stephen Yeaple. 2003. "Multinational Enterprises, International Trade and Productivity Growth: Firm Level Evidence from the United States." CEPR Discussion Paper 3805.

Kim, Linsu. 1997. *From Imitation to Innovation: Dynamics of Korea's Technological Learning*. Harvard: Harvard Business School Press. Kim, Youngbae, and Byungheon Lee. 2002. "Patterns of Technological Learning Among the Strategic Groups in the Korean Electrical Parts Industry." *Research Policy* 31(4).

Kokko, Ari, Ruben Tansini, and Mario C. Zejan. 1996. "Local Technological Capability and Productivity Spillovers from FDI in the Uruguayan Manufacturing Sector." *Journal of Development Studies* 32(4): 602–11.

Konings, Josef. 2001. "The Effects of Foreign Direct Investment on Domestic Firms." *Economics of Transition* 9(3): 619–33.

Kraay, Aart. 2006. "Exports and Economic Performance: Evidence from a Panel of Chinese Enterprises." In H. Hoekman, and B. Smarzynska-Javorcik eds., *Global Integration and Technology Transfer*. Washington, DC: World Bank and Palgrave Macmillan: 139–61.

Levin, Richard C., Alvin K. Klevorick, Richard R. Nelson, and Sidney G. Winter. 1987. "Appropriating the Returns from Industrial Research and Development." Brookings Papers on Economic Activity 3.

Lipsey, Robert E. 2000. "Interpreting Developed Countries' Foreign Direct Investment." NBER Working Paper 7810, National Bureau of Economic Research, Cambridge, MA.

Matthews, John A., and Dong-Sung Cho. 2000. *Tiger Technology: The Creation of a Semiconductor Industry in East Asia*. Cambridge: Cambridge University Press.

Nabeshima, Kaoro. 2004. "Technology Transfer in East Asia: A Survey." In M. Shahid Yusuf, Anjum Altaf, and Kaoru Nabeshima eds, *Global Production Networking and Technological Change in East Asia*. Washington, DC: World Bank; New York: Oxford University Press.

OECD. 2005. *Oslo Manual. Guidelines for Collecting and Interpreting Innovation Data*, 3rd edn. Paris: OECD.

Pack, Howard. 2006. "Econometric versus Case Study Approaches to Technology Transfer." In H. Hoekman and B. Smarzynska-Javorcik eds., *Global Integration and Technology Transfer*. Washington, DC: World Bank and Palgrave Macmillan.

Romer, Paul. 1990a. "Are Nonconvexities Important for Understanding Growth?" *The American Economic Review* 80(2): 97–103.

_____. 1990b. "Endogenous Technological Change." *Journal of Political Economy* 98:S71–S102.

_____. 1993. "Idea Gaps and Object Gaps in Economic Development." *Journal of Monetary Economics* 32(3): 543–73.

Saggi, Kamal. 2002. "Trade, Foreign Direct Investment, and International Technology Transfer: A Survey." *World Bank Research Observer* 17(2): 191–235.

Scherer, F.M., and D. Harhoff. 2000. "Technology Policy for a World of Skew-distributed Outcomes." *Research Policy* 29:559–66.

Schiff, Maurice, Yanling Wang, and Maurice Olarreaga. 2002. "Trade-Related Technology Diffusion and the Dynamics of North–South and South–South Integration." World Bank Policy Research Working Paper 2861.

Trajtenberg, Manuel. 1990. "A Penny for Your Quotes: Patent Citations and the Value of Innovations." *Rand Journal of Economics* 21(1): 172–87.

Trajtenberg, Manuel, Rebecca Henderson, and Adam B. Jaffe. 1997. "University versus Corporate Patents: A Window on the Basicness of Invention." *Economics of Innovation and New Technology* 5(1): 19–50.

Tybout, James R. 2006. "Plant and Firm-Level Evidence on 'New' Trade Theories." In H. Hoekman, and B. Smarzynska-Javorcik eds., *Global Integration and Technology Transfer.* Washington, DC: World Bank and Palgrave Macmillan.

UNESCO (United Nations Educational, Scientific, and Cultural Organization). 2004. "A Decade of Investment in Research and Development (R&D): 1990–2000." *USI Bulletin on Science and Technology Statistics* 1 (April).

_____. 2005. *UNESCO Science Report 2005*. Paris: UNESCO Publishing.

_____. 2006. "Statistics on Research and Development." Data table, updated on May 3.

USPTO (United States Patent and Trademark Office). 2004. *Patents by Country, State, and Year—All Patent Types (2004)*.

World Bank. 2006a. *World Bank Group Enterprise Surveys*. (https://www.enterprisesurveys.org/).

_____. 2006b. *World Development Indicators 2006.*

Xu, Bin, and Jianmao Wang. 1999. "Capital Goods Trade and R&D Spillovers in the OECD." *Canadian Journal of Economics* 32(5): 1258–74.

Impact Assessments in Finance and Private Sector Development: What Have We Learned and What Should We Learn?

David McKenzie

Until recently rigorous impact evaluations have been rare in the area of finance and private sector development. One reason for this is the perception that many policies and projects in this area lend themselves less to formal evaluations. However, a vanguard of new impact evaluations on areas as diverse as fostering microenterprise growth, microfinance, rainfall insurance, and regulatory reform demonstrates that in many circumstances serious evaluation is possible. The purpose of this paper is to synthesize and distill the policy and implementation lessons emerging from these studies, use them to demonstrate the feasibility of impact evaluations in a broader array of topics, and thereby help prompt new impact evaluations for projects going forward. JEL codes: O16, O17, C93

The recent external review of World Bank research noted that "perhaps the most important role of Bank research is to learn what works, and to widely disseminate the results" (Banerjee and others 2006, p. 148). Rigorous impact evaluations, which compare the outcomes of a program or policy against an explicit counterfactual of what would have happened without the program or policy are one of the most important tools that can be used, along with appropriate economic theory, for understanding "what works." Despite this, until recently impact evaluations have been rare, especially outside the areas of health and education.[1] This is now particularly apparent in the area of finance and private sector development, where the recent financial crisis has prompted renewed attention to knowing what works in terms of getting finance to consumers and firms, and in getting the private sector growing again.[2]

The World Bank Research Observer

doi;10.1093/wbro/lkp011 Advance Access publication December 6, 2009 25:209–233

One reason for the lack of impact evaluations in this area is the perception that many finance and private sector development (hereafter FPD) policies and projects lend themselves less to formal evaluations.[3] Changes in laws or regulations may occur at an economy-wide level, or a large loan may only be given to one or two banks or firms. However, in many cases it is still possible to formally evaluate FPD policies or projects. Regulations may be implemented in some regions and not others, or apply only to firms of a certain industry or size. Generally available programs or policies may have a low take-up that can be raised through targeted interventions. And in a non-trivial number of cases it will indeed be feasible to implement a randomized experiment. The purpose of this paper is to demonstrate the feasibility of such impact evaluations, distill the lessons of these new evaluations for policymakers and practitioners, and help prompt new impact evaluations for projects going forward.

I begin by highlighting policy and implementation lessons from four areas where impact evaluations are beginning to emerge: microenterprises, microfinance, rainfall insurance, and regulatory reform. I use impact evaluations in these areas to illustrate various methods which are possible when evaluating FPD reforms, as well as to note some of the key challenges to their effective use. I then discuss several reasons why these policy areas are at the forefront of FPD impact evaluations, which leads to a discussion of where the biggest opportunities appear to be going forward for new knowledge generation of what works.

Many of the examples discussed here will come from randomized experiments, which have increasingly become the preferred method of evaluation for many development economists (Duflo and Kremer 2005). Randomized experiments offer many advantages for evaluation, chief among them being that they ensure that the only reason why some firms, consumers, or other units are subject to a policy or program and others are not is pure chance. This also makes the results easy to communicate to policymakers.

However, recently there has been a debate about whether the profession is overemphasizing randomization (Rodrik 2008; Deaton 2009; Ravallion 2009; Imbens 2009). Many of the issues discussed, such as for whom the treatment effect is identified, and whether the results are generalizable to other settings, are also important considerations in using non-experimental methods. There are three lessons from this debate that I consider important for the discussion in this paper. The first is that we must not let methodological purity determine which questions to try and address: just because a policy can't be randomized does not mean we should give up on trying to understand whether it is working or not. Indeed I consider a range of approaches that can be used for ensuring more rigorous impact evaluation. Second, studies need to go beyond a simplistic black-box approach of "Does this work or not?" to try and understand why and how it works, and for whom. Third, I agree with Imbens (2009) who argues that, given

the question which one is interested in answering is possible to answer with randomization, there is little to gain and much to lose by not randomizing. Randomization is not always feasible, but I do not know of a single study that has credibly argued that they could have randomized but chose not to do so because of a belief that they would get a more rigorous assessment of impact by not so randomizing.

What Have We Learned?

I begin by reviewing a selection of recent research in four areas of private sector and financial research. These serve as examples of both the types of substantive questions that can be answered through formal impact evaluations, as well as good illustrations of some of the research methodologies that can be used to help answer such questions.

Raising the Incomes of the Self-employed

Self-employment accounts for a large share of the labor force in most developing countries. For example, Gollin (2002) reports that in Ghana, Bangladesh, and Nigeria 75–80 percent of manufacturing workers were self-employed. Self-employment is particularly important among the poor. Banerjee and Duflo (2007) find that between 47 and 69 percent of urban households who live on less than US$2 per day in Peru, Indonesia, Pakistan, and Nicaragua own a non-agricultural business. A central question for policymakers is then how to raise the incomes of these poor businesses, and whether in fact the typical microenterprises owned by the poor have any ability to grow.

In the absence of market failures, a standard model of firm size determination (for example Lucas 1978) would suggest that the answer is no—the reason for firms being small in such models is that their owners have low entrepreneurial ability. Of course market failures are pervasive in many developing countries, with restrictions on access to credit being a notable example. However, an influential branch of theory suggests that in the presence of credit constraints, the prospects of microenterprise growth from small investments is low, due to production non-convexities (Banerjee and Newman 1993). The argument is that the profitable investments facing a business are lumpy (for example a new machine), and that, without sufficient access to external credit, individuals who start a business too small will be trapped in poverty, earning low returns. Conversely, if these non-convexities are not important, then if small firms are operating well below the optimal production point (given their entrepreneurial ability), we might expect the returns to additional capital investment to be particularly high.

However, assessing the extent to which a lack of capital hampers the income growth of microenterprises is complicated by the fact that firm owners with more capital stock or greater access to credit are likely to differ in a host of other ways from owners with less capital, such as in terms of entrepreneurial ability in the Lucas model. Two recent randomized experiments in Sri Lanka and Mexico (de Mel, McKenzie, and Woodruff 2008a; McKenzie and Woodruff 2008) illustrate one approach to impact evaluation which can resolve this problem and credibly identify the impact of additional capital on firms. Grants of between US$100 and US$200 were given to a randomly selected subset of poor microenterprises in each country. The authors then compared the profits of firms which randomly received these grants to those which didn't to determine the extent to which grants raise business incomes.[4] Their results challenge the somewhat conventional wisdom that subsistence firms have no scope for growth (see table 1 for a summary of the key results from studies discussed in this paper). The authors find the grants do substantially raise incomes for the average firm receiving a grant, and estimate real returns to capital of 5.7 percent per month in Sri Lanka and 20 percent per month in Mexico, much higher than market interest rates in both countries. They explore heterogeneity in the treatment effects in an attempt to understand why the returns are so high. They find returns to be highest for high ability, credit-constrained firm owners, which is consistent with the view that credit market failures prevent talented owners from getting their firm to its optimal size.

These randomized experiments show grants work in raising incomes for the average microenterprise owner. In the particular research studies, the grants were not part of a government or NGO program, but rather given out by the researchers and funded through research grants. However, there are several cases where governments have employed grants as a way of raising the incomes of the self-employed. An example is the Microemprendimientos Productivos program in Argentina, which provided financial support in the form of in-kind grants to finance inputs and equipment to beneficiaries, with the aim of helping them obtain a sustainable source of income and reduce their dependence on welfare payments (Almeida and Galasso 2007). The Mexican Jovenes con Oportunidades program provides grants to young people for completing the last few years of schooling, with these grants kept in bank accounts that can be accessed for paying for further study or for starting a business. Grants to microenterprises are also more common in disaster recovery situations, such as that following the Indian Ocean tsunami of December, 2004 (de Mel, McKenzie, and Woodruff 2008c).

A question which faces policymakers who wish to give grants to raise the incomes of microenterprises is whether these grants should be in the form of unrestricted cash or be made in kind, as was the case with the Argentine

Table 1. Summary of Main Findings

Study	*Policy of program studied*	*Main results*
Panel A: Results that largely confirm or support conventional wisdom		
Bruhn (2008); Kaplan, Piedra and Seira (2007)	Business registration reform in Mexico	Reform increases the number of registered firms and employment. Less in line with conventional wisdom, Bruhn shows this is from new entry, not formalization, of existing firms.
Giné and Love (2006)	Bankruptcy reform in Colombia	Reducing reorganization costs improves efficiency of bankruptcy process, with more viable firms more likely to be reorganized and less viable firms to be liquidated.
Oh and others (2008)	Credit guarantee policy in Korea to support SMEs during crisis	Guarantees improves survival rates, sales growth and employment growth.
de Janvry, McIntosh, and Sadoulet (2008)	Introducing a credit bureau in Guatemala	Awareness of the bureau leads to a modest and temporary increase in repayment rates and to microfinance groups ejecting worst-performing members.
Panel B: Results that challenge or overturn conventional wisdom		
de Mel, McKenzie, and Wordruff (2008a,b)	Conditional and unconditional grants to micro enterprises in Sri Lanka	Returns to capital are high for male-owned firms, but zero for female-owned films. No difference between conditional and unconditional transfers.
Giné and Karlan (2008)	Removing group liability in microfinance groups in the philippines	No change in default rates when joint liability removed, and faster client growth in converted branches.
Karlan and Zinman (2008)	High interest rate consumer loans in South Africa	High interest loans let marginal recipients be more likely to keep their jobs, have higher incomes, and experience less hunger.
Giné and Yang (2009); Cole and others (2008)	Offering rainfall insurance to farmers in Malawi and India	Take-up is extremely low so that insurance leads to little risk mitigation or changes in farmer behavior.
Cole, Sampson, and Zia (2009)	Financial literacy training in Indonesia	Program has zero impact on the general population. but increased bank account use for financially illiterate. However, small cash payments have much more effect than financial education.

program.[5] In the randomized experiments in Sri Lanka and Mexico, half of the grants were given as cash and the other half as raw materials and equipment for the businesses (chosen by the owner). The authors find in both studies that there is no difference between the two forms of grant: they result in approximately the

same change in capital stock and same increase in business profits. If business owners have profitable opportunities to expand they will invest additional cash in these opportunities. If they don't, then any inputs or equipment they provide will crowd out the investments they would have made on their own and they can sell excess capital stock if it is not yielding a return. This suggests that policymakers can achieve the same results with the cheaper and easier to administer cash grants.[6]

Impact evaluations are not only useful for showing what works, but also what doesn't. This can guide new policy experiments. A first example of this from the microenterprise experiments is that while the grants succeeded in raising the incomes of male business owners, the average return to capital for women receiving the grant in Sri Lanka was zero (the Mexican study contained only men). Grants alone thus did not work in raising the incomes of self-employed women. In follow-up work, de Mel, McKenzie, and Woodruff (2008b) combine the experimental results with several theoretical models to try and understand why the grants did not work for raising business income for women. They find that women did not invest smaller grants in the business, while the larger grants invested in the business had low returns. They speculate that a possible explanation for this is inefficient household use of resources, with other household members capturing a share of the income and working capital held by women, leading women to use fixed business assets as a store of value rather than simply for production. They also find returns to be particularly low in business sectors dominated by women. This has led to ongoing field experiments designed to determine the impact on business profits of getting women to shift into sectors which both men and women work in, as well as a replication of the study in Ghana to understand whether the same gender differences emerge in a country with much higher female participation rates in self-employment.

A second example of what doesn't work from this same body of work is that, although the one-time grants succeed in raising the incomes of poor male business owners, they do not lead to significant employment creation. A comparison of the characteristics of microenterprise owners with those of wage workers and owners of firms with five or more employees suggests that only one-quarter to one-third of them have attributes such as ability, motivation, and ambition similar to that of larger firm owners (de Mel, McKenzie, and Woodruff 2008d). The key question for policymakers is then how to unleash the employment-creating potential of these select microenterprise owners. Providing access to credit, business training, and business development services have been the typical ways in which governments have tried to do this. However, to date there has been little rigorous evaluation of such business training programs,[7] something which ongoing evaluations hope to correct.

Rethinking the Central Precepts of the Microfinance Movement

The previous section demonstrated that one-off grants can raise the incomes of the average microenterprise owner. Grants to certain vulnerable groups, and perhaps even large sections of the poor, may be sustainable as part of a government social protection program (the Oportunidades program in Mexico covers 5 million households, almost one-quarter of Mexico's population).[8] However, in terms of finance and private sector development policies, most of the focus on households and microenterprises has been through microfinance. The most famous example of microfinance is that of the Grameen Bank. The model of microfinance most strongly associated with it is group lending to women at low interest rates. Recent impact evaluations (along with the success of microfinance institutions such as Banco Compartamos in Mexico, which offers individual loans at quite high interest rates to both men and women) give strong reasons to question this archetypal model of microfinance as necessarily the best way to expand access to finance to the poor and to help the small business sector to go forward.[9]

Many microfinance organizations focus almost exclusively or largely on female borrowers. For example, 97 percent of Grameen Bank's seven million borrowers are women,[10] as are 70 percent of FINCAs borrowers,[11] and 65 percent of ACCIÓN's five million clients.[12] While part of this reflects a social mission, many of the justifications are economic in nature. Women are argued to be poorer than men on average (for example Burjorjee, Deshpande, and Weidemann 2002; FINCA 2007), to have less collateral, and hence to be more credit-constrained (for example Khandker 1998; Armendáriz and Morduch 2005). But if this is the case, when women do receive access to credit, it should generate higher returns than when men receive access. The experimental evidence from Sri Lanka (and supporting non-experimental evidence from Mexico and Brazil) in de Mel, McKenzie, and Woodruff (2008b) provides a reason to question this extensive focus on women and a suggestion that more products need to be developed to fit the needs of urban male clients.

Group liability is often hailed as one of the central innovations of the microfinance movement, mitigating both the adverse selection and moral hazard problems which can give rise to credit market failures. The idea is that borrowers who know they will be liable for the debts of others in their group will have an incentive to screen others so that only reliable people will join their group, and then to monitor their group members to ensure they invest their funds wisely and exert enough effort. However, as Giné and Karlan (2008) note, group liability has several pitfalls which may cause it to be disliked by many borrowers. It may be particularly troublesome for small business owners, who might be discouraged from undertaking somewhat risky but high return projects by other group

members, may need different size loans or different loan periods from other group members, and find frequent group meetings costly in terms of time. Finally, there is also a concern that group liability loans are less useful for establishing credit records in credit bureaus than individualized loans, making graduation to larger loans more difficult (de Janvry, McIntosh, and Sadoulet 2008).

Giné and Karlan (2008) carried out a randomized experiment with a microfinance bank in the Philippines to investigate the extent to which group lending really reduces the moral hazard problems. Half of the group-lending centers of the bank were randomly chosen to be converted to individual liability. The authors find no change in default rates after one and three years in the converted centers, and there is faster client growth in the converted branches. These results suggest that group liability is not that important for reducing moral hazard, but, since the converted loans were all initially screened by groups, the authors cannot say anything about the importance of groups for screening out bad risks. In ongoing work, the authors are examining this issue, comparing newly formed groups to new individual loan clients.

The third precept of microfinance that has been strongly challenged by recent impact evaluations is the belief that serving the poor requires low interest rates. Muhammad Yunus (2007) states "a true microcredit organization must keep its interest rate as close to the cost of funds as possible," criticizing the high interest rates being charged by Banco Compartamos. This lies at the heart of the debate on commercialization of microfinance (Cull, Demirgüç-Kunt, and Morduch 2009; Harford 2008). However, the high returns to capital for many microenterprises in Sri Lanka and Mexico suggest the ability to repay loans at rates significantly higher than market interest rates. The problem, especially for urban business owners seeking individual loans, is often one of access rather than interest rate. In follow-up work in Sri Lanka, de Mel, McKenzie, and Woodruff (2009b) find that few of the high return microenterprises qualify for a loan from microfinance banks, which lend on a basis of physical collateral and not on whether the owner's business shows high prospects for growth.

The most striking evidence that high interest rate loans can improve welfare comes from a study of consumer loans in South Africa. Karlan and Zinman (2008) conducted a randomized experiment with a microlender, in which applicants who were marginally rejected for consumer loans were randomly selected into two groups, one of which received a second look and hence a higher probability of getting a loan. The loans were for four months at a monthly interest rate of 11.75 percent (equivalent to an APR of 200 percent). Despite these high interest rates, the authors find that six to twelve months later the marginal loan recipients were more likely to have kept their job, had higher incomes, and experienced less hunger. This is not to argue that the customers wouldn't have been even better off had loans been available at lower interest rates. But at the existing

rates, not only did the customers benefit, but these marginal loans appear to have been profitable for the bank.

This study illustrates well some of the pros and cons of trying to build policy on the basis of a randomized experiment. The impacts estimated are credible and easily understood by policymakers; and they are the impacts for marginally rejected consumers, a group of interest certainly to the bank. However, the fact that this group can benefit a lot from additional access to high interest rate credit is not informative about whether poorer individuals, who are far from the creditworthy cutoff, would stand to benefit from high interest rate loans—other studies are needed to look at this question.

To be sure, these existing impact evaluations consist of only a couple of rigorous studies from a couple of countries, and it will be important to see if the results are repeated in replication studies. Nevertheless, the results do suggest reasons for questioning the structure of the prototypical microfinance product. Moreover, despite the rampant expansion of microfinance worldwide and the tremendous amount of attention this has received in the media, to date there has been little rigorous impact evaluation of the welfare effects of the basic microfinance product.[13] Several large-scale randomized trials of microfinance are currently nearing completion. The first preliminary results from a randomized trial involving 2,400 households in India were recently presented by Esther Duflo.[14] While the full results are not yet available, two points to note are: first, take-up was only 17.5 percent—that is, most households offered a loan did not want one; second, the preliminary results show very modest impacts, with no significant effects on health or education, and relatively little use for business purposes. As more results become available from this and other impact evaluations of microfinance, there will be a new impetus toward policy efforts in the microfinance domain.

Insuring Poor Farmers

Missing credit markets are one important reason why firms in developing countries are less productive than they could be. However, reluctance to take up credit may be linked to the existence of another important market failure: the lack of an insurance market. This may be particularly important in occupations such as farming, which is subject to substantial income risk from rainfall variation during the growing season. One solution which has been proposed and introduced in a number of countries is Rainfall Index Insurance, which links payouts to rainfall at local rain gauges.

An important question of interest is then whether offering this rainfall insurance works in increasing the use of credit by risk-averse farmers. A randomized experiment conducted by Giné and Yang (2009) among farmers in Malawi finds

evidence that it does not. The authors worked with the Malawian farmers' association, financial institutions in Malawi, and the Commodity Risk Management Group of the World Bank to offer smallholders credit to purchase high-yielding seed varieties. Farmers in some localities were randomly selected to be just offered credit, while those in other localities were offered a bundle of credit and insurance. Take-up of the credit was 33 percent for farmers offered the loan without insurance, and only 17.6 percent for those who were offered a loan bundled together with rainfall insurance.

Take-up rates of rainfall insurance have also been low elsewhere—Giné, Townsend, and Vickrey (2008) report a take-up rate of only 4.6 percent for one product in India. In a cross-sectional non-experimental setting, they find that risk-averse households are actually less likely, not more likely, to purchase the insurance, especially when they are unfamiliar with other types of insurance and the insurance provider. They attribute this to uncertainty about the insurance product, which as a new technology requires some risk and trust to participate in. In follow-up randomized experiments in India, Cole and others (2008) investigate the sensitivity of the take-up decision in relation to the price, the presence of an endorsement from a third trusted party, the means of presentation, and the liquidity constraints. Their results are consistent with the view that, in addition to price and liquidity, trust and financial literacy influence take-up to a significant degree.

These studies have several implications for efforts to develop better insurance products for the poor. In addition to finding that price matters, the findings on trust and financial literacy suggest scope for modifying implementation and marketing in a way which will boost demand. To the extent that poor farmers are unable to understand complicated insurance products, as an introductory product to get people used to the idea of insurance, a simpler product design with fewer thresholds and payment schedules may be preferred to a more complicated product that offers more complete insurance.[15] For example, a product that pays out if rainfall is below 150 mm during the specified period and does not if rainfall is above is simpler to understand than the more standard product which in the example of Cole and others (2008, p. 9) "pays zero when cumulative rainfall during a particular 45 day period exceeds 100 mm. Payouts are then linear in the rainfall deficit relative to this 100 mm threshold, jumping to Rs. 2000 when cumulative rainfall is below 40 mm." It would be interesting in future impact evaluations to compare the take-up and efficacy of simpler designs to more complex designs.

Second, the authors find take-up to be much higher in villages where a positive past insurance payout has occurred. They conclude from this that it would be useful to modify the contracts to ensure they pay out a positive return with sufficient frequency as to engender trust in the population, whereas the standard

contracts pay out very rarely. The trade-off here is that for the same insurance premium, more frequent payouts mean smaller amounts can be paid out each time, resulting in less complete coverage of catastrophic losses to compensate for greater coverage of more common losses. Third, since liquidity constraints mattered a lot for take-up, they suggest that it might be beneficial to bundle the insurance product together with a loan. The results in Malawi shows this results in less credit uptake than if pure loans were offered, though it might offer greater insurance uptake than if insurance alone was offered, and would not preclude offering a separate loan-only product.

Learning from Regulatory Reform

The impact evaluations profiled above used randomized experiments to randomly offer the program to selected individuals, firms, banking branches, farming localities, and the residents of slum areas. However, this approach to evaluation may not be possible with some forms of FPD projects, such as reforms in the regulatory environment. Nevertheless, in many cases rigorous impact evaluation is still possible. We illustrate this through consideration of two recent studies which have conducted impact evaluations of regulatory reforms.

The view that burdensome regulations are an important barrier to private sector development was famously expressed by de Soto (1989), who calculated that it would take 289 days, 11 permits, and over $1,000 to register legally a small business in Peru. This emphasis on regulatory reform has been further spurred by the World Bank's Doing Business project, which ranks countries each year on both the overall ease of doing business and on the extent of reforms undertaken in the previous year. The 2009 report notes that almost 1,000 reforms recorded in the areas measured by Doing Business have occurred in the past six years, with the most common reform being one which makes it easier to start a business by reducing the costs and number of procedures needed. Yet despite the huge number of reforms, there is almost no rigorous impact evaluation of these reforms.

An exception is found in Bruhn (2008) and Kaplan, Piedra, and Seira (2007), who study the impact of business registration reform in Mexico. The reform was organized by a federal agency, but implemented at the municipal level since many business registration procedures were set locally. Due to staffing constraints, the federal agency could not implement the reform in all priority municipalities at once, but instead staggered the reforms, introducing them first in some municipalities and then later in others. Among the municipalities identified as priorities for implementation, there was no specification of which should go first. This allows the authors to use the municipalities in which the reform was introduced later as a control group for the municipalities in which it was introduced earlier, using a

difference-in-differences estimation methodology. This estimation essentially looks at the period where the first few municipalities had reformed and others had yet to. It then compares the change in the number of registered businesses (or in other outcomes of interest) for those municipalities where the reform was introduced early to the change in these same outcomes for municipalities where the reform was introduced later. This is an estimation strategy that is likely to be applicable in understanding a number of other regulatory reforms, which might be phased in over time.[16]

The headline result from both Bruhn (2008) and Kaplan, Piedra, and Seira (2007) is that the reform succeeded in increasing registrations. This is where the most simplistic measures of impact would stop. For example, World Bank (2008) reports that following a reduction in the minimum capital requirement there was an increase in new company registrations of 55 percent in Georgia and 81 percent in Saudi Arabia.[17] But we need to go beyond simply asking whether a reform led to more businesses and actually understand how and why it worked. In the specific example of business registration reform, an important question of interest is whether these new registrations are the results of existing firms registering or of new firms starting up. Bruhn (2008) finds that the increase in registrations comes from new entry, not from the conversion of existing informal firms.

This result suggests there may be a group of the potentially self-employed, for whom the burden of registering is a barrier to business formation, and that once this pool of pent-up demand is exhausted there may be much less long-term impact. The results here do not support de Soto's (1989) view that existing informal small business owners are individuals who wish to become formal, but are stymied by high barriers to registration. They are more consistent with the view that the majority of informal businesses are informal by choice, because becoming formal offers no benefit to them. Indeed, McKenzie and Sakho (2009) estimate that for Bolivian small firms, there are huge gains to becoming formal for the subset of informal firms who don't know how to become formal, but that becoming formal would be costly to the remainder of informal firms.

We also want to know what the consequences of these reforms are for the economic outcomes we ultimately care about, such as employment generation, consumer welfare, and economic growth. Bruhn (2008) finds the Mexico reform increased employment by 2.8 percent after the reform and benefited consumers by decreasing prices by 0.6 percent, most likely as a result of additional competition. However, in doing so, it reduces the income of incumbent registered business owners. Since municipal-level GDP is collected only every five years, it is not possible to look at the overall impact on economic growth.

Although in some cases reforms might be introduced in a staggered fashion into some regions of the country first, a more common experience is for the reform to be introduced for the entire country at once. But even in this situation,

it is often the case that the reform only applies to, or should theoretically only have consequences for, a subset of the population. One special case of this is when the policy only applies to firms above (or below) some particular size threshold. A relatively common example of this occurring is in the area of labor regulation, where employment protection rules might apply only to firms above a certain number of workers.[18] For example, both Italy's employment protection legislation and Sri Lanka's Termination of Workmen Act place much more onerous requirements on firms with 15 or more employees. In some circumstances this might allow evaluation of the effects of the reform by comparing firms just above the threshold to those just below, a regression discontinuity design. This is done for Italy by Leonardi and Pica (2006). However, in practice such regulations will often cause firms to sort themselves around the size threshold, making this approach to evaluation more challenging. Abidoye, Orazem, and Vodopivec (2008) find some evidence that this is the case in Sri Lanka, with firms slower to grow from 14 to 15 workers than from 13 to 14 workers or from 15 to 16 workers.

More typically reforms introduced at the country level may affect only some firms or industries, but not others.[19] This allows for a difference-in-differences estimation strategy in which unaffected firms or industries are used as a comparison group for those affected by the reform. An example of this is seen in Giné and Love (2006), who evaluate the impact of a bankruptcy reform in Colombia, which reduced the costs of reorganizing a bankrupt firm. Their goal is to see whether the law change led to distressed, but viable, firms being more likely to reorganize when they would have previously liquidated. Since active, non-bankrupt firms are not affected by the law, the authors use a difference-in-differences strategy to compute the difference in the characteristics of bankrupt firms selecting into reorganization rather than liquidation after the law was reformed relative to the characteristics of active firms, relative to this same difference pre-reform. They find that lowering the costs of reorganization led to an improvement in the efficiency of the bankruptcy procedure, with more viable firms now more likely to be reorganized than liquidated relative to the pre-reform situation.

Lessons for Implementation of Impact Evaluations

The impact evaluations summarized above have begun to yield important policy lessons for work with microenterprises, microfinance, rainfall insurance, and regulatory reform. These are all important components of finance and private sector development policy, yet they only cover a fraction of the important policy tools and research areas in the FPD domain. The questions which then arise are

why these few areas have been at the forefront of evaluation efforts to date, and what lessons they hold for other evaluations going forward.

Why Have these Subject Areas Dominated Evaluation Efforts to Date?

A substantive reason why these topics have been at the forefront of evaluation efforts is that they have close ties with important bodies of theoretical work in development economics, and that in many cases the theory suggests reasons both why the policy may have its intended effect and why it may not work in practice. For example, in the grants to microenterprises, one body of theory suggested returns to capital may be very low due to non-convexities, while another body suggested returns could be high due to credit constraints with convex production technologies. Likewise there are theoretical reasons why group lending may have benefits, as well as reasons why it may deter certain types of borrowers. These cases where the impact of the program is theoretically uncertain motivate empirical studies to see what happens in practice.

A more practical reason is that these studies are all in areas where evaluation is most feasible for a variety of reasons. The first is one of sample size. The policies studied are ones where the units of analysis are consumers or firms, allowing the comparison of the impacts on many affected units with a control group of many other units. The second is one of data availability. The regulation studies relied on unusually good existing databases in Mexico (a quarterly labor-force panel survey and administrative data from the Mexican social security system) and a comprehensive database on the universe of bankruptcy cases in Colombia. The other studies were designed as ex ante evaluations, with data collection designed by the researchers. The randomized experiments done to date have generally been conducted by researchers working with NGOs or funding the programs through research grants. This has limited study to either programs which have been run by NGOs willing to work with researchers, or to projects which are cheap enough for research grants to fund.

Going forward this calls for a need for continued close interaction between theory and evaluation—we want to know not just whether or not something works, but why and how? It also suggests that widespread rigorous evaluation of the many other types of FPD programs and policies implemented by governments and supported or advised by international financial institutions requires a much greater commitment to evaluation and, in particular, to planning ahead so the evaluation process (including data collection) can start before the program is implemented. It also suggests unexploited benefits exist from small modifications in currently collected sources of data which do not presently have policy evaluation in mind. For example, surveys of firms should include questions on participation in particular types of policies or projects (for example Does your firm

participate in a business cluster developed by the government under its regional clustering program?) and should include enough identifying information to link with administrative records on banks, firms, or consumers participating in such programs. And unfortunately even when such data is collected, access to the microdata is often limited in many countries, so greater data accessibility is also needed.

Evaluation of Many FPD Programs Is Possible

The studies highlighted above have demonstrated a variety of methods that can be used for evaluating FPD policies—randomized experiments, difference-in-differences, and regression discontinuity designs. There are a variety of other evaluation methods available, which when used carefully can also be informative as to policy impacts. We highlight here three of these other methods which are also likely to be useful in evaluating a broad array of FPD policies.[20]

Propensity-score matching is a commonly used method for estimating a treatment impact. An example in the FPD literature is seen in Oh and others (2009), who evaluate the impact of a credit guarantee policy used by the Korean government to support small and medium enterprises in the aftermath of the Asian financial crisis. The authors use plant-level panel data on manufacturing firms and match those firms which received credit guarantees to similar firms which did not, finding that the guarantee program positively affected both survival rates and sales and employment growth of the firms receiving the guarantees. A concern with propensity-score matching is that it assumes the process of which firm receives a guarantee and which does not can be adequately captured by a set of observable variables which the firms are matched on. How plausible this is will be a judgment call in any given setting and will benefit from detailed knowledge of how the program was actually implemented. In general the literature has found the results to be closer to those obtained in an experimental setting when a rich set of data can be used for the matching, including multiple periods of pre-program data to control for existing trends. The data used by Oh and others (2009) don't meet this criteria, with only data from one year (2000) for a relatively limited set of firm characteristics being used. This suggests one should be cautious in accepting their results.

A second method is the control function approach which involves explicitly modeling how unobservables which affect the outcome are related to the observables, including the choice of participation in a program or policy regime. This approach is used along with propensity-score matching by Fajnzylber, Maloney, and Rojas (2006) to look at the impact of access to credit, training, and membership in business associations on microenterprises in Mexico. Traditionally these methods have relied heavily on functional form assumptions and distributional

assumptions, such as joint normality, which can lead to significant bias when these assumptions do not hold and as a result such methods have fallen out of favor in much of development economics. However, recently semiparametric approaches have been developed which rely less on these assumptions, but which still need an exclusion restriction to hold (see Heckman and Navarro-Lozano 2004 for a review and comparison to matching). This need for an exclusion restriction takes us back to the need for answering the underlying question needed for evaluation: thinking of an exogenous reason why some firms, consumers, or other units participate in a program and others do not.

A third method which is likely to be applicable for a wide variety of FPD evaluations is an encouragement design (Diamond and Hainmueller 2007). This can be useful when evaluating a program that is implemented at the country level, such as a change in regulation or in policy. The basic idea behind this design is that firms (or other units of interest) are randomly divided into a treatment and a control group. While the program is available to all, the treatment group receives additional encouragement to participate in the program—for example they might receive marketing visits to make them more aware of the program. If the encouragement is successful it yields a difference in program take-up rates between the two groups which can then be used in evaluating the impact of the program. More precisely, what can be estimated is the impact of the program on units which would take up the program when offered encouragement, but which wouldn't otherwise.

An example of an encouragement design being successfully used is seen in de Janvry, McIntosh, and Sadoulet (2008), who examine the impact of the introduction of a credit bureau in Guatemala. While the credit bureau is in place for everyone, knowledge of its implementation was found to be almost non-existent in surveys conducted soon after its implementation. The authors therefore randomly informed a subset of 5,000 microfinance borrowers about the existence of the bureau and how it works. They found this awareness of the bureau led to a modest and temporary increase in repayment rates and to microfinance groups ejecting their worse-performing members.

The International Finance Corporation (IFC) has recently attempted encouragement designs in two evaluations: an ongoing evaluation of business registration in Lima, Peru, where firms receive encouragement to register, and an evaluation of an alternate dispute resolution (ADR) project in Macedonia. The preliminary results from the Macedonia project also illustrate the potential downside of this approach in estimating the impact—the encouragement might not encourage very many units to take up the program.[21] The Macedonia project tried several methods of encouraging use of the ADR mechanism, but found that none of these encouragement methods succeeding in raising use. This prevents an estimation of the effect of the ADR on firms. Nevertheless, a finding that no one

wants to participate in a program, even when encouraged, is in and of itself a useful result for understanding the likely program impact. More detailed analysis of why firms do not take up the program can then be used to improve the program going forward.

The Importance of Take-up

A key difference between evaluation of most FPD programs and many impact evaluations in education and health lies in take-up. In programs such as vaccination campaigns or getting children to school programs, the goal of the program is to have all eligible individuals participate. And in the case of cash transfers, participation can be close to universal. In contrast, universal take-up is not the goal of most FPD programs, and, even when it is a goal, it is seldom the reality. Not all households or firms will want or need a loan, nor will they register formally or wish to purchase insurance. This is evident in some of the studies profiled above, such as take-up rates of 17.5 percent for microfinance, 5 to 33 percent for rainfall insurance, and no increase in the number of firms in the informal sector registering to become formal when regulations changed.

Program take-up which is less than universal offers both challenges and opportunities for impact evaluation. Learning what the level of take-up is, and which characteristics predict take-up, can be useful for refining and modifying the policy to enable it to reach its goals in the future better. For example, the low take-up of risk-averse individuals in the rainfall insurance papers, coupled with the fact that take-up was much higher when there had been a recent pay-out in the village or when there was an endorsement from a trusted third party, can help guide marketing efforts and product design in the future. It can lead to revealing other market or government failures where policy action is required.[22] Finally, take-up rates and characteristics can also be useful for gauging the potential market for taking pilot trials to scale.

However, low take-up also offers several challenges to attempts to evaluate FPD programs rigorously. The first is one of power to detect the program effect. For example, one of the ultimate goals of the work on rainfall insurance is to find out if it allows farmers to farm more efficiently and so protect their households against negative shocks. However, because few farmers purchased the insurance, and those who did only purchased enough to cover a trivial fraction of their crops, the existing studies do not allow the researchers to determine the impacts on production and household welfare.

There are two solutions to this problem of power. The first is to employ a very large sample size, so that the resulting sample will still contain enough firms or households which take-up the program to enable the researchers to detect a program impact of a given size. However, the downside of this is that it can be

very expensive. For example, consider a program such as a new loan product or business training that aims to raise the profits of microenterprises undertaking the program by 25 percent. A randomized experiment, which offered the program to half the firms and used a single follow-up survey to estimate this impact, would require a sample size of 670 firms if the take-up was 100 percent; however, it would need a sample size of 2,700 with a 50 percent take-up, and a sample size of 67,000 with 10 percent take-up.[23] An example of a randomized experiment with sample sizes of this magnitude is seen in Karlan and Zinman (2009), who randomized 58,000 direct mail offers issued by a South African lender, with 8.7 percent of those contacted applying for a loan.

The second solution to the problem of low power from low take-up is to restrict study to a group of units for whom take-up would be much higher. For example, a business training program could be advertised to all eligible firms or microfinance clients, and then the number of slots available in the program could be randomly allocated among the group of interested firms. A related example is seen in Karlan and Zinman (2008), in which consumers first apply for loans and then the pool of marginally rejected candidates (all of whom wanted a loan) is then randomly assigned to receive a second look at getting a loan. The advantage of this second approach is that it requires much smaller samples to detect a treatment impact. The downside is that of external validity—the program impact estimated will apply only to the self-selected group of individuals or firms which expressed interest in the program, and not to the general population. In some cases however this might be precisely the impact of interest—for instance, policymakers might want to know what the effect of their loan program is on firms interested in taking up credit.

The second challenge offered by low take-up is one of interpretation of program impact. Consider evaluating the impact of microfinance on microenterprise profitability in a situation where take-up of loans is only 10 percent. With a randomized experiment, comparison of the mean profits of firms offered the microfinance treatment to those which were not offered the microfinance treatment yields the average intention-to-treat effect. This is the impact on firms of being offered credit. This in itself is a parameter of interest, but in most cases we would also like to go further and know what the impact of the credit was if it was actually taken up. The standard approach is to instrument receipt of microfinance with the randomly determined offer of credit. However, if the impact of receiving credit varies by firm, what is recovered is known as a local average treatment effect (LATE) (Angrist and Imbens 1994). This is the average effect of receiving credit for firms which would take up the microfinance treatment when offered. If firms which stand to benefit more from credit are the ones who take it, this will overstate the gain in profit which the average firm would receive if it got microfinance.[24] What this means in practice is that there needs to be care taken in

interpreting program effects with low take-up and in deciding whether the parameter estimated is in fact one of policy interest. Researchers can also go further in understanding the underlying observable sources of heterogeneity in the take-up decision and in treatment effects.

What Should We Learn?

The previous sections have shown that evaluation of FPD programs and policies is possible in a wide variety of contexts and that the small number of evaluations to date are yielding useful lessons for both policy design and future evaluation efforts. The question is then where should we go from here? While we have argued that there is much greater scope for serious evaluation than is currently being realized, two general areas are particularly attractive for increased efforts.

The first is for more evaluations in the areas that have been at the forefront of existing efforts: microfinance, microenterprises, insurance, and regulatory reform. We noted that there are a number of features of these policy domains that lend themselves to rigorous evaluation. Yet there are currently only a handful of rigorous studies. More are needed on a wider range of policies in a number of different institutional settings, in order to learn what works, where, and why.

The second general area where there appears to be unexploited gains to be made from impact evaluation is in looking at the effects of other programs and policies that are widely used to benefit large numbers of consumers and firms. Three such important policy areas where evaluation seems possible, yet is currently almost non-existent, are financial literacy and consumer protection, business training, and policies to enhance the Small and Medium Enterprise (SME) sector.

The subject of financial literacy has received increased policy attention in recent years, with worldwide efforts underway to roll out financial literacy training. For example, Citi Foundation is four years into a ten year, $200 million, global program of financial education, operating in 65 countries and a number of governments have also developed programs in this area.[25] The recent global financial crisis has also turned attention to issues of consumer protection and the possible macroeconomic consequences of consumers entering into credit transactions that they do not fully understand. Financial literacy programs are ripe for evaluation efforts, since, despite the increasing amounts of money, there are always groups of consumers that receive the program and others that do not. The challenge for evaluation is making these two groups as similar as possible and in measuring the outcomes.

One preliminary study in Indonesia has found teaching financially illiterate individuals about the benefits of bank accounts did lead to an increase in bank

account use among this group, with no increase for those who were already financially literate (Cole, Sampson, and Zia 2009). Moreover, they find small incentive payments to have a much larger effect on getting individuals to open bank accounts and to be three times as cost-effective as financial education in this regard, suggesting a need for some skepticism in judging some of the lofty claims of proponents of financial education. This fledgling effort provides a good base for future evaluations to build on, with the ultimate goal of finding out under what circumstances such programs work, when they don't, and what the consequences on consumer welfare are.

A second area which is ripe for experimentation and impact evaluation lies in business training programs. Many microfinance organizations, NGOs, and governments worldwide offer short courses to budding or existing microenterprises to teach them the basics of running a small business. Public sector funding of such programs may be justified from a poverty alleviation standpoint, since, even if the programs worked and had large benefits, credit constraints and risk aversion might prevent poor people participating. Again in these programs one can design impact evaluations by comparing firm owners who are offered the training to similar individuals that are not offered the training. Several randomized experiments currently in the field are attempting to do this.

The last area I wish to stress as being particularly full of unexploited possibilities for impact evaluations lies in policies directed at the SME sector. These include SME lending policies, trade credit policies, management training, and sector-specific technical assistance. These programs are typically carried out by governments and international financial institutions (IFIs) rather than NGOs, and are too expensive usually for researchers to fund the program on offer themselves.[26] As a result, there is a real knowledge gap—and an opportunity to be grasped. If governments and operations staff at IFIs can work with researchers in evaluating the many projects being implemented, it should be possible to evaluate rigorously many of the policies being carried out for SMEs and to learn where modifications of existing strategies are needed.

Conclusions

I have surveyed the existing literature on impact evaluations in finance and private sector development with two main aims. The first was to draw emerging policy lessons and implementation lessons from the slowly growing set of rigorous impact assessments that have been carried out in areas such as microfinance, microenterprise growth, rainfall insurance, and regulatory reform. The second aim was to use the lens of these existing evaluations to demonstrate some of the different strategies for possible evaluation and to argue that much more impact

evaluation is possible than has currently been attempted.[27] Hopefully policymakers and operational staff reading this paper will agree with this message and join together with researchers to try and understand better what works and why.

Notes

David McKenzie is a Senior Economist in the Finance and Private Sector Development team of the Development Research Group at the World Bank. I thank Miriam Bruhn, Asli Demirgüç-Kunt, Xavier Gine, Bilal Zia, the editor, and three anonymous referees for useful comments and discussions. All opinions are of course my own and do not necessarily reflect the views of the World Bank.

1. For example, the Development Impact Evaluation (DIME) Initiative has until recently focused on topics in health and education. See http://web.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTDEVIMPEVAINI/0,,menuPK:3998281~pagePK:64168427~piPK:64168435~theSitePK:3998212,00.html (accessed February 4, 2009).

2. See also the recent World Bank Policy Research Report on Access to Finance which calls for more impact evaluation (Demirgüç-Kunt, Beck, and Honohan 2008).

3. A second reason may be that research on FPD has historically worried less about the challenges of identification that are a prime concern of the labor and applied microeconomics literature. Financial economists are much less likely to be exposed to impact evaluation methods in their graduate classes than health, education, or labor economists. A further purpose of this paper is thus to better expose practitioners in the FPD field to the ideas and possibilities of impact evaluations.

4. Comparing profits requires knowing how to measure the profits of microenterprises which are usually informal and keep few records. Impact evaluations have been useful for learning what works in this regard too (de Mel, McKenzie, and Woodruff 2009a).

5. This parallels the debate in the conditional cash transfers literature as to whether the conditions attached to cash grants matter (Fiszbein and Schady 2009). Our finding of no differential effect of conditioning does not immediately carry over to other forms of conditioning, such as conditioning on school attendance or health clinic visits, since firm owners can undo the conditioning of being required to spend the money on their business more easily than they can undo other types of conditions—for example, in theory they could devote less time to school work at home if children attend school more, though this seems less likely.

6. Conditional grants may still be preferred from a political economy perspective, since grants may be easier to sell to the public if they are conditional on the recipients "using them properly."

7. An exception is Karlan and Valdivia (2008), who find that business training increases the sales and repayment rates of female microfinance clients in Peru.

8. See http://www.oportunidades.gob.mx/Wn_Inf_General/Padron_Liq/Cober_Aten/index.html (accessed February 5, 2009).

9. See Cull, Demirgüç-Kunt, and Morduch (2009) for a description of the heterogeneity in the microfinance sector and the debate generated by the successful stock offering of Banco Compartamos. Karlan and Morduch (2009) provide an excellent overview of recent research on access to finance.

10. http://www.grameen-info.org/bank/index.html (numbers as of May, 2007, accessed August 15, 2007).

11. http://www.villagebanking.org/site/c.erKPI2PCIoE/b.2604299/k.FFD9/What_is_Microfinance_What_is_Village_Banking.htm (accessed August 15, 2007).

12. http://www.accion.org/about_key_stats.asp (all clients 1976–2006, accessed August 15, 2007).

13. See Armendáriz and Morduch (2005) for a summary of different non-experimental approaches that have been used to measure impact. The most well known of these is Pitt and Khandker (1998), who employ a regression discontinuity design. There is some debate as to the

extent to which the regression discontinuity applied in practice (see the discussion in Armendáriz and Morduch).

14. Presentation by Esther Duflo at the Innovations for Poverty Action 2008 Microfinance Conference at Yale University. Discussion of these results is covered at http://www.philanthropyaction.com/nc/the_real_impacts_of_micro_credit/ (accessed February 5, 2009).

15. This is not to preclude also offering the more complicated products at the same time and letting farmers choose between them. An alternative would be better financial education to teach the participants how to learn about this product. Cole and others (2008) implemented brief (5 to 10 minute) training sessions on this, which they found had no effect.

16. Note that the validity of this difference-in-differences estimation strategy relies on an assumption that the municipalities which reform later are a good comparison group for what would have happened to the earlier reform municipalities in the absence of early reform. Bruhn (2008) carries out a number of checks on pre-existing trends and municipality characteristics to argue this is the case. This strategy will be less applicable if countries decide to, for example, first introduce the reform in the capital city or business capital, and then roll the reform out to progressively smaller cities.

17. Note that these numbers for Georgia and South Africa are not even the true impact on the number of registrations, since they are a simple before–after comparison and do not control for pre-existing trends or concurrent events in the economy.

18. Priority lending also may have size thresholds. See Banerjee and Duflo (2008) who study a reform in India which increased the maximum size limit for firms to be eligible for priority-sector lending. They then use a triple-difference evaluation strategy, comparing the change in the rate of changes in outcomes before and after the reform for firms that were newly eligible for priority lending compared to firms that were already eligible.

19. Another example is seen in Kugler, Jimeno, and Hernanz (2005), who study a reform of Spain's labor law, which applied only to some demographic groups such as young workers, older workers, women under-represented in their occupations, and disabled workers, but not other groups.

20. For a good recent general reference to different estimation strategies for impact evaluation, see Imbens and Wooldridge (2008). Instrumental variables is another common technique for evaluation which we won't explicitly discuss here, but McKenzie and Sakho (2009) provide an example in FPD.

21. This discussion of the Macedonia results is based on correspondence with Alexis Diamond of the IFC.

22. For example, de Mel, McKenzie, and Woodruff (2009b) worked with a regional development bank to try and help microenterprises obtain loans. Despite 62 percent of firms showing up for information meetings, only 10 percent received loans. One reason was that in the absence of a credit bureau, applicants had to travel to other institutions and obtain letters from them attesting that they had no outstanding loans, thereby increasing the cost to applicants of applying for loans. This experience highlights the need for credit bureaus to cover microfinance.

23. These calculations were made using the sampsi command in STATA assuming: a constant treatment effect, a coefficient of variation of 1 (which is in line with what one typically sees in microenterprise data after trimming outliers), that the treatment has no effect on the variance of profits, a power of 0.90, and a test significance level alpha of 0.05.

24. See Heckman, Urzua, and Vytacil (2006) and Deaton (2009) for more discussion on interpretation of treatment effects when the take-up decision is a choice which is related to the individual unit's program effect. Ravallion (2009) also discusses some related issues in the use and interpretation of experiments.

25. See http://www.citigroup.com/citi/financialeducation/ (accessed February 10, 2009).

26. A nascent effort to evaluate a few of the IFC's programs has been underway for a few years. IFC (n.d.) describes some of these efforts. However, to date these efforts have to my knowledge not resulted in any working papers or published articles.

27. The Finance and Private Sector Development team of the Development Research Group has recently introduced a new impact note series to try and disseminate better the results of new impact evaluations which do occur. See http://econ.worldbank.org/programs/finance/impact for the latest in FPD impact evaluations.

References

The word *processed* describes informally reproduced works that may not be commonly available through libraries.

Abidoye, Babatunde, Peter Orazem, and Milan Vodopivec. 2008. "Firing Cost and Firm Size: A Study of Sri Lanka's Severance Pay System." Iowa State University Working Paper 08014.

Almeida, Rita, and Emanuela Galasso. 2007. "Jump-starting Self-employment? Evidence among Welfare Participants in Argentina." IZA Working Paper 2902.

Angrist, Joshua D., and Guido Imbens. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467–75.

Armendáriz, Beatriz, and Jonathan Morduch. 2005. *The Economics of Microfinance.* Cambridge, MA: MIT Press.

Banerjee, Abhijit, and Esther Duflo. 2007. "The Economic Lives of the Poor." *Journal of Economic Perspectives* 21(1): 141–67.

_____. 2008. "Do Firms Want to Borrow More? Testing Credit Constraints Using a Directed Lending Program." Processed. MIT.

Banerjee, Abhijit, and Andrew Newman. 1993. "Occupational Choice and the Process of Development." *Journal of Political Economy* 101:274–98.

Banerjee, Abhijit, Angus Deaton, Nora Lustig, and Ken Rogoff. 2006. "An Evaluation of World Bank research, 1998–2005." (http://siteresources.worldbank.org/DEC/Resources/84797-1109362238001/726454-1164121166494/RESEARCH-EVALUATION-2006-Main-Report.pdf) accessed February 4, 2009).

Bruhn, Miriam. 2008. "License to Sell: The Effect of Business Registration Reform on Entrepreneurial Activity in Mexico." World Bank Policy Research Working Paper 4538.

Burjorjee, Deena M., Rani Deshpande, and C. Jean Weidemann. 2002. "Supporting Women's Livelihoods Microfinance that Works for the Majority: A Guide to Best Practices." United Nations Capital Development Fund (UNCDF), Special Unit for Microfinance. Processed UNCDF.

Cole, Shawn, Thomas Sampson, and Bilal Zia. 2009. "Valuing Financial Literacy." Processed. World Bank.

Cole, Shawn, Xavier Giné, Jeremy Tobacman, Petia Topalova, Robert Townsend, and James Vickrey. 2008. "Barriers to Household Risk Management: Evidence from India." Processed. World Bank.

Cull, Robert, Asli Demirgüç-Kunt, and Jonathon Morduch. 2009. "Microfinance Meets the Market." *Journal of Economic Perspectives*, 23(1): 167–92.

Deaton, Angus. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." Processed. Princeton University.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2008a. "Returns to Capital: Results from a Randomized Experiment." *Quarterly Journal of Economics* 123(4): 1329–72.

_____. 2008b. "Are Women more Credit Constrained? Experimental Evidence on Gender and Microenterprise Returns." *American Economic Journal: Applied Economics* 1(3): 1–32.

_____. 2008c. "Rebound: How Enterprises Recover from a Natural Disaster." Processed. World Bank.

_____. 2008d. "Who Are the Microenterprise Owners? Evidence from Sri Lanka on Tokman v. de Soto." In Joshua Lerner, and Antoinette Schoar eds., *International Differences in Entrepreneurship*. Boston, MA: NBER.

_____. 2009a. "Measuring Microenterprise Profits: Must We Ask How the Sausage Is Made?" *Journal of Development Economics* 88(1): 19–31.

_____. 2009b. "Getting Credit to High Return Microenterprises: The Results of an Information Intervention." Processed. World Bank.

Demirgüç-Kunt, Asli, Thorsten Beck, and Patrick Honohan. 2008. *Finance for All? Policies and Pitfalls in Expanding Access.* World Bank Policy Research Report. Washington, DC: World Bank.

Diamond, Alexis, and Jens Hainmueller. 2007. "The Encouragement Design for Program Evaluation." (http://www.ifc.org/ifcext/rmas.nsf/AttachmentsByTitle/Encouragement/$FILE/The+Encouragement+Design+for+Program+Evaluation.pdf, accessed February 10, 2009).

Duflo, Esther, and Michael Kremer. 2005. "Use of Randomization in the Evaluation of Development Effectiveness." In Osvaldo Feinstein, Gregory K. Ingram, and George K. Pitman eds., *Evaluating Development Effectiveness*. New Brunswick, NJ: Transaction Publishers: 205–32.

Fajnzylber, Pablo, William Maloney, and Gabriel Montes Rojas. 2006. "Releasing Constraints to Growth or Pushing on a String? The Impact of Credit, Training, Business Associations and Taxes on the Performance of Mexican Micro-Firms." World Bank Policy Research Working Paper 3807. Washington, DC.

FINCA. 2007. "Frequently Asked Questions." (http://www.villagebanking.org/site/c.erKPI2PCIoE/b.2394157/k.8161/Frequently_Asked_Questions.htm, accessed August 15, 2007).

Fiszbein, Ariel, and Norbert Schady. 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty*. Policy Research Report. Washington, DC: The World Bank.Giné, Xavier, and Dean Karlan. 2008. "Peer Monitoring and Enforcement: Long Term Evidence from Microcredit Lending Groups with and without Group Liability." Processed. World Bank.

Giné, Xavier, and Inessa Love. 2006. "Do Reorganization Costs Matter for Efficiency? Evidence from a Bankruptcy Reform in Colombia." World Bank Policy Research Working Paper 3970.

Giné, Xavier, and Dean Yang. 2009. "Insurance, Credit, and Technology Adoption: Field Experimental Evidence from Malawi." *Journal of Development Economics*, 89(1): 1–11.

Giné, Xavier, Robert Townsend, and James Vickrey. 2008. "Patterns of Rainfall Insurance Participation in Rural India." *World Bank Economic Review* 22(3): 539–66.

Gollin, Douglas. 2002. "Getting Income Shares Right." *Journal of Political Economy* 110(2): 458–74.

Harford, Tim. 2008. "The Battle for the Soul of Microfinance." *Financial Times* December 6.

Heckman, James, and Salvador Navarro-Lozano. 2004. "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models." *The Review of*

Economic and Statistics 86(1): 30–57.

Heckman, James, Sergio Urzua, and Edward Vytacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics* 88(3): 389–432.

IFC (n.d.) "Innovations in Impact Evaluation in IFC." IFC Monitor: Results Measurement for Advisory Services. (http://www.ifc.org/ifcext/rmas.nsf/AttachmentsByTitle/Innovationsmonitor/$FILE/Innovations2.pdf, accessed February 10, 2009).

Imbens, Guido. 2009. "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." Processed. Harvard University.

Imbens, Guido, and Jeffrey Wooldridge. 2008. "Recent Developments in the Econometrics of Program Evaluation." Processed. Harvard University.

Janvry, Alain de, Craig McIntosh, and Elisabeth Sadoulet. 2008. "The Supply- and Demand-side Impacts of Credit Market Information." Processed. UCSD.

Karlan, Dean, and Jonathan Morduch. 2009. "Access to Finance." In M. Rosenzweig, and D. Rodrik eds., *Handbook of Development Economics*, Volume 5. forthcoming.

Karlan, Dean, and Martin Valdivia (2008) "Teaching Entrepreneurship: Impact Of Business Training On Microfinance Clients and Institutions", Mimeo. Yale University.

Karlan, Dean, and Jonathan Zinman (2008) "Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts", *Review of Financial Studies*, forthcoming.

_____. 2009. "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment." *Econometrica*, forthcoming.

Kaplan, David, Eduardo Piedra, and Enrique Seira. 2007. "Entry Regulation and Business Start-ups : Evidence from Mexico." World Bank Policy Research Working Paper 4322.

Khandker, Shahidur R. 1998. "Using Microcredit to Advance Women." World Bank Premnote (November) 8. (http://www1.worldbank.org/prem/PREMNotes/premnote8.pdf, accessed August 15, 2007).

Kugler, Adriana, Juan Jimeno, and Virginia Hernanz. 2005. "Employment Consequences of Restrictive Permanent Contracts: Evidence from Spanish Labor Market Reforms." Processed. University of Houston.

Leonardi, Marco, and Giovanni Pica. 2006. "Effects of Employment Protection Legislation on Wages: A Regression Discontinuity Approach." Processed. University of Milan.

Lucas, Robert E. 1978. "On the Size Distribution of Business Firms." *Bell Journal of Economics* 9(2): 508–23.

McKenzie, David, and Christopher Woodruff. 2008. "Experimental Evidence on Returns to Capital and Access to Finance in Mexico." *World Bank Economic Review* 22(3): 457–82.

McKenzie, David, and Yaye Seynabou Sakho. 2009. "Does It Pay Firms to Register for Taxes? The Impact of Formality on Firm Profitability." *Journal of Development Economics*, forthcoming.

Oh, Inha, Jeong-Dong Lee, Almas Heshmati, and Gyoung-Gyu Choi. 2009. "Evaluation of Credit Guarantee Policy Using Propensity Score Matching." *Small Business Economics*, 33(3): 335–51.

Pitt, Mark, and Shahidur Khandker. 1998. "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Economy* 106(5): 958–96.

Ravallion, Martin. 2009. "Should the Randomistas Rule?" *The Economists' Voice*, February 2009. (www.bepress.com/ev).

Rodrik, Dani. 2008. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" Processed. Harvard University.

Soto, Hernando de. 1989. *The Other Path: The Economic Answer to Terrorism*. New York, NY: Basic Books.

World Bank. 2008. *Doing Business 2009*. Washington DC: World Bank.

Yunus, Muhammed. 2007. "Remarks by Muhammad Yunus, Managing Director, Grameen Bank." *Microcredit Summit E-News* 5(1).

Scale Economies and Cities

Indermit S. Gill • Chor-Ching Goh

This paper summarizes the policy-relevant insights of a generation of research on scale economies. Scale economies in production are of three types: internal economies associated with large plants, localization economies that come from sharing of inputs and infrastructure and from greater competition among firms, and urbanization economies that are generated through diversity and knowledge spillovers. The benefits (and costs) of localization and urbanization are together called "external (dis) economies" because they arise due to factors outside any single household, farm or firm. The empirical literature yields some stylized facts. Internal scale economies are low in light industries and high in heavy industries. External scale economies are amplified by economic density and dissipate with distance from places where economic activity is concentrated. Scale economies are most visibly manifest in towns and cities. To simplify somewhat, towns allow firms and farms to exploit internal scale economies, medium-sized cities help firms in an industry exploit localization economies, and large cities and metropolises provide urbanization economies to those who locate within or nearby. Scale economies have implications for policy makers. The first is that because urban settlements rise and thrive because market agents demand their services, they should be seen as creatures of the market, not creations of the state. The second is that because settlements of different sizes provide differing services, towns, cities, and metropolises are more often complements for one another, not substitutes. Third, as a corollary, policymakers should aim to improve the functioning of urban settlements, and not become preoccupied with their size. JEL codes: O1, O4, O3, R3, R11

Even a casual examination of the economic geography of most countries reveals that the topography of production and population is bumpy. People, firms, jobs, services, and opportunities tend to agglomerate in relatively small parts of a country, frequently leaving the remaining areas to agriculture or to nothing. In the United States, half of the nation's GDP is produced on just 4 percent of the

The World Bank Research Observer

doi;10.1093/wbro/lkp022 Advance Access publication November 26, 2009 25:235–262

land area (Easterly and Levine 2001). As one zooms in to a smaller spatial scale, this pattern is replicated. New York is the most economically dense state in America. This in turn is because of the extreme concentration of production in the metropolitan area of New York. And the metro area contains New York County, the densest county in the nation, as well as three of the next ten densest counties (Ciccone and Hall 1996).

Besides being bumpy—even spiky—the spatial distribution of economic activity is sticky. New York has been the largest U.S. city for over a century (Kim and Margo 2004, p. 2994). No new urban agglomeration has emerged in France since 1911, and no city that was relatively big has died (Eaton and Eckstein 1997, pp. 447–8). Once established, a dense agglomeration of economic activity tends to persist. A country's economic geography exhibits inertia.

The spikiness and stickiness of a country's economic geography can be explained by a single factor: scale economies. Scale economies can be internal or external to firms and farms. Internal scale economies are the cost advantages that a firm reaps by producing in large plants. As Adam Smith recognized during his visit to a pin factory more than two centuries ago, they include the productivity advantages from the division of labor which is made possible by large scale production. Internal increasing returns create a tendency for a firm to locate its activities in one or a few locations, thereby contributing to bumpiness.

But internal economies cannot, by themselves, explain the existence of cities, because they cannot explain why many firms might choose to base themselves in a single location and often stay there for long periods. For this, it is necessary to turn to external economies—the benefits of scale that arise outside a firm or industry. These are the advantages that a bookshop on Dadabhai Naoroji Road in Mumbai gets by being close to other book stalls. They are also the advantages which financial firms derive from locating in large and diverse cities such as São Paulo or Shanghai. Agglomeration economies can take the form of either "localization" economies (the benefits from colocation of similar producers) or "urbanization" economies (the benefits from conglomeration of diverse producers). Firms benefit from locating close to other firms in either the same or different industry. And unless all of them move elsewhere together, they would lose these benefits.

However, the logic can only be stretched so far. For people and firms, being in cities comes at a price. Traffic in central London moves at only 11 miles per hour (Shaffer and Santos 2004), the same speed as horse drawn carriages a hundred years ago. Bangkok is notorious for congestion and Beijing for pollution. Housing in Delhi is amongst the most expensive in the world. Crime and slums are prevalent in many developing countries' primary cities. These are the costs of density, the *diseconomies* of agglomeration, which limit the productivity advantages to firms, and the welfare benefits to inhabitants, when they locate in dense urban areas. Combined with the differing propensities of activities to benefit from

agglomeration economies, they help to explain why economic activity within a country is not restricted to a single center, but is instead spread across centers of differing sizes. Good policies can relax the constraints generated by the congestion and overcrowding of land and resources; bad policies can result in the diseconomies quickly stifling the benefits that come from scale economies.

But caveats notwithstanding, the benefits of producing large amounts in a single plant or place have only increased as transport costs have fallen in the two centuries since Adam Smith visited the pin factory. The potential of scale economies and how access to world markets helps to exploit them can be seen in the city of Dongguan in China's Pearl River delta. In 2005 one plant there manufactured 30 percent of the magnetic recording heads used in disk drives worldwide. In several products, Dongguan's factories account for over 40 percent of global production. Of the parts and components used in personal computers, 95 percent can be sourced within Dongguan city (Gill and Kharas 2007).

Places like Dongguan have rushed headlong into the world of scale economies. While economic crises slow down and even temporarily reverse these transformations, millions of people in the developing world enter this new realm every year, and the implications—for them and for policymakers—are enormous. This paper provides a brief synopsis of work by a generation of economists who have sought to understand these scale economies. It summarizes the evidence of scale economies in production, focusing on "agglomeration economies," whose exploitation requires locating in areas that are densely populated by other producers. Finally, we discuss the lessons for policymakers in the developing world. Our main conclusions are:

- The sectoral transformations that countries must undergo as they develop have a spatial "dual." As they transition from agricultural to industrial to service-oriented modes of production firms and workers leave behind not just their villages and their agrarian occupations, but also a world where scale does not matter so much. They enter not just settlements that are larger and denser, but also a world in which producers, distributors, and consumers enjoy agglomeration economies.
- Research over the last generation indicates that different types of human settlements facilitate agglomeration economies for different forms of production. An oversimplified (but not incorrect) generalization is that market towns facilitate plant-level scale economies in marketing and distributing produce, medium-sized cities provide localization economies for manufacturing industries, while the largest cities provide urbanization economies that foster the growth of business, government, and educational services.
- Because of coordination failures associated with spatial transformations, policymakers have often seen cities as constructs of the state. In reality, like firms

and farms, cities and towns are creatures of the market. Just as firms and farms deliver final and intermediate goods and services, towns and cities deliver agglomeration economies to producers and workers. While a longer discussion of policy implications is outside the scope of this survey, two clear lessons emerge by recognizing the importance of agglomeration economies. The first is that settlements of different sizes serve functions that are complementary. Large, medium, and small cities should not be seen as substitutes—the focus should be on function, not size. The second is that it gets harder to keep settlements well-functioning as urbanization advances, but reasonable land markets, policies that address negative externalities like congestion and pollution, and universal provision of basic services remain important at every stage of the spatial transformations that are necessary for development.

Theoretical Advances

Adam Smith introduced scale economies, factor mobility, and transport costs as central to understanding the nature and causes of the wealth of nations. But until the 1980s most economists were happier to anchor their inquiries to another concept introduced in *The Wealth of Nations*: the "invisible hand" of perfect competition.[1] Constant returns led to convenient characterizations of the economy where all firms and workers were identical, so one firm or worker could be considered representative of all. Scale economies were inconvenient—they required acknowledging that specialization differentiated people and products.

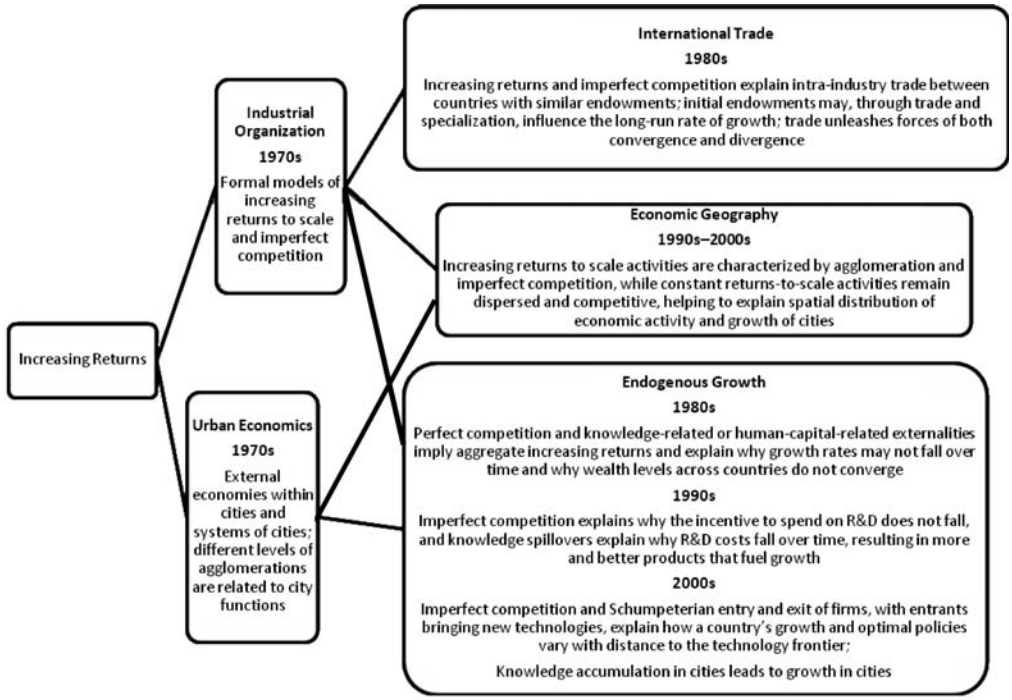
But perfect competition is an artificial construct. It assumes infinitesimally small firms with little or no influence over market prices, even in the immediate vicinity of a firm. The assumption of constant returns to scale also implies the problem of "backyard capitalism"(Krugman, 1991). That is, if small-scale production is as efficient as large-scale, every household would produce a range of goods and services in its backyard. Occasionally, the contradiction between these two phenomena—internal scale economies on the one hand, and perfect competition on the other—would surface. Because of the technical difficulties of modeling increasing returns to scale, though, it would quickly be buried again.[2]

Then, during the 1970s, a simple model of increasing returns to scale (Dixit and Stiglitz 1977) opened the door for researchers to the same realm that so many firms and workers had inhabited since the Industrial Revolution. By the late 1980s scale economies were standard features of the explanations for international trade. By the early 1990s growth theorists had accepted the need to incorporate imperfect competition among firms into aggregate formulations of an

economy. By the mid-1990s theorists were beginning to show how these ideas could be used to understand the spatial distribution of economic activity, including the rise of towns and cities (figure 1).

The literature on the microeconomic foundations of agglomeration economies has flourished in the last two decades by combining models of scale economies and insights about urban economics that highlight tensions between the benefits and costs from concentration (Mills 1967; Dixit 1973; Henderson 1974a). In general, researchers have recognized that economic growth has different impacts on firms and workers depending on their sector and location. The underlying reason is the love for variety in consumption and the economies of scale in

Figure 1. Recognizing the Importance of Scale Economies: 30 Years of Theoretical Advance



	Key Publications
Industrial Organization	Spence (1976); Dixit and Stiglitz (1977)
Urban economics	Mills (1967); Diamond and Mirrlees (1973); Henderson (1974a)
International trade	Krugman (1980, 1981); Ethier (1982); Helpman and Krugman (1987); Grossman and Helpman (1995)
Economic geography	Krugman (1991b); Fujita, Krugman, and Venables (1999); Baldwin and Martin (2004)
Endogenous growth	Romer (1986, 1990); Lucas (1988); Grossman and Helpman (1991); Aghion and Howitt (1992, 2005); Rossi-Hansberg and Wright (2007); Duranton (2007)

Source: Adapted from Gill and Kharas (2007) with suggestions from Vernon Henderson and John McCombie.

production; the proximate causes are product differentiation, monopolistic power, specialization, and location externalities. The formal recognition of scale economies, externalities, and imperfect competition makes economic theory conform more closely to the world which policymakers inhabit. The main insight is that activities with increasing returns due to reasons external to a firm tend to be concentrated in cities, while those displaying constant returns remain more dispersed.

A Guide to Scale Economies

Countries develop through sectoral shifts out of traditional subsistence-based agricultural activities into manufacturing and services. Along the way firms rather than farms become the dominant production unit. The production of differentiated manufactured goods and services increases as a share of the economy's output relative to agricultural produce. Production technology shifts away from constant returns to increasing returns to scale. And over time, scale-augmenting technical change generates scale economies. Imperfect competition becomes the rule.

The benefits of greater scale in production can be either internal or external to a producer. Internal scale economies are the *within-firm* efficiency gains associated with production in sizeable quantities in a single plant. External economies are those that come from producing in large quantities in a single place. They are traditionally classified as localization economies arising from *within-industry* economic interactions, and urbanization economies arising from *between-industry* interactions. External economies are synonymous with "agglomeration economies," which comprise the benefits of proximity to other producers of the same commodity or service, and diversity, or being close to other producers of a wide range of commodities and services. This paper deals mainly with production-related scale economies.[3]

Internal economies arise from the size of a plant's ability to exploit fixed costs better. For example, a larger steel mill can get volume discounts from suppliers—implying fixed costs of transport and trade—and reap the benefits of division of labor within the firm. But the steelmaker's location near other firms is not the main source of gains from internal scale economies. *Localization economies* arise from the presence of a large number of firms in the same industry in the same place. For example, textile firms cluster to share a large pool of suitably skilled and unskilled labor and other intermediate input suppliers. *Urbanization economies* arise from the presence of a large number of different industries in the same place. For example, a hedge fund can benefit from locating near business schools, banking and other financial service providers, management consulting companies, legal and accounting firms, as well as manufacturers. This diversity breeds new ideas and innovation.

As summarized in Duranton and Puga (2004), agglomeration economies depend not just on size (that is, on how big a city or industry is). The reasons for producers to gain from proximity to others lie in the sharing of capital inputs, information, and labor; improving the matches between production requirements and types of land, labor, and intermediate inputs; and learning about new techniques and products. A plant in an isolated location can benefit from internal scale economies, but unless it is situated in an area of density it cannot enjoy the benefits associated with localization or urbanization economies. Cities bring together large pools of skilled labor and suppliers of specialized intermediate inputs and by so doing enhance employer–employee and buyer–seller matches. Input sharing is an important channel for agglomeration economies. Averaging across industries, a firm's relocation from a less dense location (of fewer than 500 neighboring employees in the same industry) to a denser location (of 10,000–25,000) results in a 3 percent increase in purchased input intensity (see, for example, Holmes and Stevens 2002).

This paper does not discuss such scale diseconomies as congestion, crime, noise, flooding, and pollution because the empirical literature on the subject is thin. While there is ample theoretical discussion (see, for example, Tolley 1974; Henderson 1975, 1977; Arnott 1979; Sasaki 1998), there is little empirical evidence of such relationships. In a seminal paper, Mills and Ferranti (1971) rebut the conventional statement that pollution distorts city size. They argue that urban population is not the only factor that affects pollution, and increased recycling, treatment of waste, and improved efficiency of urban transportation systems can significantly reduce pollution. Henderson (1974b) and Tolley and Crihfield (1987) show that policy measures that reduce diseconomies will enhance the welfare of population and profit of production, and, as a result, cities will become even bigger.

A large city provides a large labor pool to firms and a dependable supply of specialized labor, providing more job options to workers and reducing the volatility of labor demand. Smaller specialized cities expose workers to greater industry-specific shocks but provide favorable match-specific advantages. In both cases the concentration of economic activity lowers the search costs between firms and workers, which results in fewer unfilled vacancies, lower risk of job loss, and shorter duration of unemployment. Costa and Kahn (2000) document that because of better matching married couples with university education are increasingly found in large cities, up from 32 percent in 1940 to 50 percent in 1990 in the US. The matching mechanisms depend on information transmission among people (see Bikhchandani, Hirshleifer, and Welch 1998 for a survey of the literature on social learning). Cities make it easier for producers to find inputs and for consumers to experiment and discover new suppliers.

Learning mechanisms are another important explanation for agglomeration in cities. Marshall (1890) emphasizes how cities favor the diffusion of innovations and ideas, while Jacobs (1969) stresses how the environment offered by cities improves the prospects for generating new ideas. Lucas (1988) also suggests that learning encompasses, not just cutting-edge technology, but incremental mundane knowledge (knowing how, knowing who, and so on) through intended and unintended communications, and that cities are the best place for knowledge transmission. Knowledge spillovers are difficult to measure because they can seldom be traced through transactions. With patent citations, however, it is possible to identify a paper trail for some, although by no means all, knowledge spillovers. Jaffe, Trajtenberg, and Henderson (1993) find that patent citations are spatially concentrated, with citations between them 5 to 10 times more likely to come from the same standard metropolitan statistical areas in the United States as originator patents. Jaffe (1986) also finds evidence for local R&D spillovers among U.S. firms where the number of patents per dollar of R&D spending is higher for firms located in areas with above average R&D spending.

Another strand of research focuses on workers as the primary vehicles of knowledge. This implies that economies with substantial labor mobility across industries will exhibit a greater spread of ideas and growth. Rauch (1993) shows that wages are higher where average education levels are high, since workers will be more productive and employers will be willing to pay high wages in competing for them. Moretti (2004a, 2004b) finds a positive effect of the presence of college graduates on a city's wages. Charlot and Duranton (2003) use survey data to show that workplace communication is more extensive in urban areas and that this communication affects wages positively.

Internal Scale Economies

Internal increasing returns to scale are well documented in manufacturing, based on cost and value-added data, engineering estimates, and trade data and markups. Table 1 summarizes the stylized findings in the literature. Internal scale economies range from negligible among light industries to high among heavy and high-tech industries. Based on engineering estimates of cost savings and the minimum efficiency scale, increasing returns are found in transport equipment, chemicals, machinery, engineering, and paper and printing. Internal scale economies are negligible in rubber and plastics, leather goods, footwear, clothing, and textiles. Estimates based on cost and value-added data at different scales of output—from cross-section, time series, and panel data of firms—point to similar findings.

Likewise, increasing output brings about disproportionate savings in the European car, truck, and consumer durables industries as well as in the

Table 1. Internal Scale Economies Are Low in Light Industries and High in Heavy Industries

Findings of returns to scale	*Data source*	*Studies*
Low: apparel, leather, footwear, wood products *High*: machinery, pharmaceuticals, instruments, iron and steel, petroleum and coal products	Trade data	Antweiler and Trefler (2002)
Low: leather goods, clothing, wood prods, textiles *High*: Vehicles, transport equipments, chemicals, printing and publishing	Engineering estimates of costs and changes in minimum efficiency scale	Junius (1997) and Prateen (1988)
Low: footwear, apparel, food products, leather *High*: tobacco, pharmaceuticals, computing machinery, railroad equipment	Markups for manufacturing in 14 OECD countries	Junius (1997) and Martins, Scarpetta, and Pilat (1996)
Low: apparel, leather products, textiles *High*: electric, gas, and sanitary services; motor vehicles and equipment, chemicals, tobacco	Markups for two-digit sectors in the US, 1953–84 (24 sectors)	Roeger (1995)
Low: textiles, milk products, lumber mills, fish oil and meal products *High*: metal, transport equipment, cement products	Production function estimates for 1963 in Norway (27 industries)	Griliches and Ringstad (1971)
Low: clothing, knitting, leather, textiles *High*: petroleum, basic and fabricated metal, transport equipment	Cost and profit data in 167 four-digit industries for 1970 in Canada	Baldwin and Gorecki (1986)
Low: textiles, apparel, lumber, wood *High*: paper products, chemicals, fabricated metals, stone, clay, glass	Dynamic cost functions for 450 U.S. industries, 1958–89	Paul and Siegel (1999)
Low: clothing, knitting, leather, textiles *High*: petroleum, basic and fabricated metal, transport equipment	Labor productivity estimates for 1965–70 for 90 four-digit industries in Canada	Gupta (1983) and Fuss and Gupta (1981)
Low: apparel, wood products *High*: other chemicals, food products, printing and publishing	Firm-level production functions for 6,665 plants in Chile, 1979–86	Levinsohn and Petrin (1999)

manufacturing sectors in the United States, Norway, and Chile. Based on trade data, a third of all goods-producing industries have increasing returns to scale. Manufacturing industries with the highest plant-level economies and industry-level externalities are petroleum and coal products (1.40), petroleum refinery (1.19), pharmaceuticals (1.31), machinery (1.11), and iron and steel (1.15). Industries with constant returns are footwear, leather, textiles, apparel, and furniture and fixtures. Similar findings can be found with markups. Because increasing

returns to scale confer some market power on firms, markups of price over marginal cost can be a proxy for plant-level scale economies. Markups for U.S. industries range from 15 percent in apparel to over 200 percent in electricity, gas, and sanitary services.

External Scale Economies

The literature distinguishes between localization and urbanization economies. Urbanization economies come from industrial diversity which foster the exchange of ideas and technology, and produce greater innovation and growth. They tend to be more important for more sophisticated activities requiring learning and new or high-tech activities. Spatial concentration of a diversity of people and products reduces knowledge production costs because information transmission, competition, spying, imitation, learning, innovation, and commercialization of new ideas are easier. Feldman and Audretsch (1999) document that in the United States a staggering 96 percent of innovations are in metropolitan areas.

Localization economies arise from input sharing and competition within the industry and come from geographically concentrated groups of firms linked by the technologies they use, the skills they require, the markets they serve, and the products and services they provide. Competitive pressures that force firms in the same sector to innovate or fail also lead to productivity growth. Conditions tend to be competitive when upstream and downstream firms and associated institutions in a particular industry (say, electronic machinery or petrochemicals) cluster.

The extent of agglomeration economies is influenced by their geographical scope. Density of economic activity and the distance between economic agents influence the productivity gains of scale economies. Table 2 summarizes some stylized findings in the literature. Evidence from Brazil and the United States indicates that doubling the distance to dense metropolitan centers reduces productivity by 15 percent; and doubling the distance from 280 to 550 kilometers reduces profits by 6 percent. The concept of distance can be generalized, in this context, from distance in physical space to distance in industrial space. For example, spillovers between industries are more likely if the industries in question share related scientific spaces and are, therefore, closer in industrial space (Feldman and Audretsch 1999, pp. 409–29). Furthermore, the extent to which distance attenuates agglomeration economies differs for different types of agglomeration. For example, knowledge spillovers that rely on face-to-face communication will decay more quickly with distance than the home market effect (Venables 2006, pp. 61–85).

In contrast to distance, density raises productivity for plants and workers. A large literature on agglomeration economies suggests that doubling city size will

Table 2. External Economies Increase with Density and Decline with Distance

Findings	*Data source*	*Studies*
External economies increase with economic density . . .		
Doubling economic density increases productivity by 6 percent	1988 data on output per worker in U.S. states	Ciccone and Hall (1996)
Doubling employment density increases productivity by 4.5–5.0 percent	Nonagricultural value added per worker in NUTS regions, 1980s	Ciccone (2002)
A one-standard deviation increase in own-industry employment share raises employment growth by 16–31 percent	1970–19: five manufacturing industries in 224 US metro areas	Henderson, Kuncoro, and Turner (1995)
A 10 percent increase in local own-industry employment raises plant output by 0.6 to 0.8 percent for fixed inputs	Korean city-industry data for 1983, 1989, 1991–93	Henderson, Lee, and Lee (2001)
Own-county effect on plant productivity significant but no effect from neighboring county	Plant-level data on productivity, 1972–92 in 742 US counties	Henderson (2003b)
. . . ***and decline with distance***		
Increasing distance from the city center by 1 percent leads to a 0.13 percent decline in productivity	1980 data for 356 new manufacturing firms in Brazil	Hansen (1990)
Doubling the distance to a regional center lowers profits by 6 percent	Firms in auto-part and agricultural machinery (Brazil and the US)	Henderson (1994)
Doubling travel time to a city center reduces productivity by 15 percent	Data for eight industries in Brazil	Sveikauskas, Townroe, and Hansen (1985)
Effects of own-industry employment on new plant births attenuate within the first five 1-mile concentric rings	12 million U.S. establishments from Dun & Bradstreet, Marketplace database	Rosenthal and Strange (2003)

increase productivity by 3–8 percent (see, for example, Shefer 1973; Mera 1973; Segal 1976; Kawashima 1975; Sveikauskas 1975; Moomaw 1981, 1983; Bartlesman, Caballero, and Lyons 1994). Density of activity allows more refined specialization and a wider variety of intermediate inputs. In addition to static agglomeration economies, there are also dynamic agglomeration economies. For example, doubling the density of economic activity in European regions may increase Total Factor Productivity growth by 0.42 percentage points a year (Angeriz, McCombie, and Roberts 2008). While evidence of external economies comes primarily from developed countries (Carlton 1983; Wheeler and Mody 1992; Carlino 1979; Hay 1979), there is also evidence of external economies in developing countries. A survey of 12,400 firms in the manufacturing sector in 120 cities in China points to the higher productivity of firms in more populous cities (World Bank 2006). In the Republic of Korea, Henderson, Shalizi, and Venables (2001) find that a plant in a city with 1,000 workers could, without altering its input mix, increase output by 20–25 percent simply by relocating to

a city that has 15,000 workers in the same industry. Kuncoro (2008), focusing on four broad industry groups in Indonesia, finds that agglomeration economies vary over time and that localization, rather than urbanization, economies tend to dominate.

Despite the availability of land, almost all recent development in the United States has been less than one kilometer from earlier developments (Burchfield and others 2006). Even today, only about 2 percent of the land area of the United States is built-up or paved. This extreme clustering of firms and workers in cities can be explained only by agglomeration economies. As countries develop, their economies become more knowledge-based and service-oriented. Their spatial concentration of activity also rises as knowledge spillovers which require close proximity become more important.

As countries develop and manufacturing and service activities become more important, firms cram in closer to harness agglomeration economies. In the United States (Ellison and Glaeser 1997), the United Kingdom (Devereux, Griffith, and Simpson 2004), and France (Maurel and Sedillot 1999) about 75–95 percent of industry is localized. For instance, in the United States more than a third of the total production of aerospace engines is concentrated in three cities: Hartford with about 18 percent of total employment, and Cincinnati and Phoenix with another 18 percent together (Rosenthal and Strange 2004). Using continuous space without considering administrative boundaries and based on concentration of plants, more than half of the United Kingdom's 122 four-digit industries are localized and only 24 percent are dispersed (Duranton and Overman 2002).

Spatial clustering is more pronounced with high-skill and high-technology industries (electronic computing machinery, process control instruments, semiconductors, pharmaceuticals) than light industries (Kim 1995; Henderson, Lee, and Lee 2001). This is consistent with the documented findings of higher scale effects in heavier industries. High-skill and high-tech industries have more capital-intensive production technology. They are also likely to benefit more from the various mechanisms that generate external economies, as discussed earlier.

Services are even more spatially concentrated than manufacturing for two reasons. First, they tend to use less land per employee. Banks, insurance companies, hospitals, and schools can operate comfortably in high-rise buildings that economize on land and allow for very high density. Second, because of external economies, business services have even greater potential for agglomeration, as firms serve one another: every bank needs advertising, and every advertising firm needs a bank account. The potential for codependence and agglomeration is thus intrinsic to services (Feldman 1994; Dekle and Eaton 1999; Feldman and Audretsch 1999).

Services are among the most agglomerated activities in the United States (Fafchamps and Desmet 2000; Kolko 2007). With the rising prominence of services economic activity will become more concentrated. Jobs in the United States became more spatially concentrated between 1972 and 1992, driven primarily by the rising localization of service activities in larger cities as small and medium size counties lost jobs to the more urban areas (Carlino and Chatterjee 2001; Desmet and Fafchamps 2006). As communication costs fall, services become more tradable, allowing providers to take advantage of narrower specialization and agglomeration economies.

An Urban Hierarchy

The relative importance of localization or urbanization economies differs across the urban hierarchy of cities, depending on a product's life cycle, an industry's tenure in the city, and the city's amenities, natural resources, current industrial profile, historical industrial environment, and local culture. An urban hierarchy exhibits some stylized patterns. Larger metropolises tend to be more diversified and service-oriented. They innovate, invent, breed new firms, and expel mature industries (Calem and Carlino 1991; Glaeser and others 1992; Black and Henderson 1999; Fujita and Ishii 1999; Feldman and Audretsch 1999; Combes 2000; Duranton and Puga 2001, 2004; Desmet and Fafchamps 2006). Smaller cities tend to be industrially specialized. They produce or manufacture standardized products and receive relocated industries from diversified cities (Moomaw 1981, 1983; Nakamura 1985; Sveikauskas, Gowdy, and Funk 1988; Henderson, Kuncoro, and Turner 1995; Henderson 1997a, 1997b, 2003a; Kolko 1999; Glaeser and Mare 2001; Rosenthal and Strange 2001, 2003). And the relative city-size distribution and industrial concentration in cities tend to be stable over time. The stylized observation in most countries is an urban hierarchy of few metropolises and many smaller cities with complementary economic functions.[4]

Improved infrastructure and falling transport costs have encouraged standardized manufacturing production to move out of high-rent centers to smaller cities. Production in large cities focuses on services, research and development, and nonstandardized manufacturing (Glaeser and others 1995; Fafchamps and Desmet 2000). Glaeser and Kahn (2001) document that while it is common to find that manufacturing activities de-concentrate from city centers to their suburbs, services do not. The relocation of manufacturing to suburbs has also been documented in developing countries, for example Colombia, Indonesia, Korea, and Thailand (Henderson, Kuncoro, and Nasution 1996).

Even after controlling for natural comparative advantage, externalities are still important in explaining the patterns of specialization and diversity among cities.

Table 3 summarizes the relative importance of localization and urbanization economies. Intraindustry localization economies have the largest impact after 3–4 years, and eventually peter out after 5–6 years. Interindustry urbanization economies persist much longer. In fairly mature metropolises, competition and diversity help industrial growth. This greater importance of across-industry knowledge spillovers in metropolises seems consistent with the urbanization economies of between-sector innovation. The production of nontraditional items is more concentrated in diverse cities, while standardized traditional goods are located in smaller specialized cities. Similar patterns are found in the United States

Table 3. Localization and Urbanization Economies Vary by City Size and Profile, and by Industry Life Cycle

Findings	*Data source*	*Studies*
Localization economies are more important for heavy industries; urbanization economies are more important for light industries	Data for two-digit manufacturing industries in Japan	Nakamura (1985)
Localization economies become less important, giving way to urbanization economies as cities expand in size	Cross-sectional data for the US and Brazil	Henderson (1986)
Labor pooling is more important in newer and expanding markets while knowledge spillovers and specialized asset-sharing are dominant in mature markets	Firm employment data for four U.S. metro areas and two-digit industries	Hammond and Von Hagen (1994)
For mature capital goods industries, localization effects are significant. For new high-tech industries, both localization and urbanization effects are present	Panel data of 742 urban counties for 1970–87	Henderson, Kuncoro, and Turner (1995)
For all industries localization and urbanization effects are important. For traditional industries most effects die out after four years. For high-tech industries, the effects persist longer	Five traditional and three new high-tech industries in 224 metropolitan areas between 1970 and 1987	Henderson (1997a)
Historical industrial environments of cities matter. In fairly mature cities, urbanization economies encourage industrial growth	Data for the largest industries 1956–87 in 170 U.S. cities	Glaeser and others (1992)
In high-tech industries, a one-standard-deviation increase in industrial diversity raises productivity by 60 percent, but diversity has no effect on light industries	Korean city-industry data for years 1983, 1989, 1991–93	Henderson, Lee, and Lee (2001)

(Henderson 1997b), Japan (Fujita and Ishii 1999), and Korea (Kolko 1999; Henderson, Lee, and Lee 2001).

While new firms often start in diverse cities, they move to specialized ones after they mature (Feldman and Audretsch 1999). Of all new plants in France, for example, 84 percent were created in cities with above-median diversity. Some 72 percent of firm relocations are from a diverse city to a specialized one. In the United States almost all product innovations take place in metropolitan areas. Own-industry specialization in a city has a negative effect on innovative output. However, industrial diversity and city size both have a positive effect on innovative output. Trial plants are based in large cities in Japan, but mass production plants are in small cities or rural areas. Young firms need a period of experimentation. In the early learning phase, diversified cities act as "nurseries" for firms to try out a variety of processes. Once a firm identifies its ideal process, it can begin mass production in specialized cities, where all firms share similar processes or specializations (Fujita and Ishii 1999; Duranton and Puga 2001).

The different economic functions served by cities can also be seen in the clustering of headquarters from different sectors and concentrations of business services in a few large cities, while production plants from each sector congregate in smaller specialized cities. Duranton and Puga (2005) document that in 1950 there was little difference across U.S. cities in their proportions of managers and production workers. Although the largest cities already housed more managers, there was no clear ranking by city size. By 1980, however, the differences across cities had increased substantially, and a clear ranking by size had emerged. Larger cities had become specialized in management and information-intensive activity, which benefit from face-to-face contacts, and smaller cities had become specialized in production. This pattern became even more marked over the following decade.

As discussed above, at the top of this hierarchy are a few large cities serving as incubators for new industries which benefit most from *urbanization economies*. Below these metropolises are medium-sized cities, serving as regional hubs of transportation, finance, and commerce, as well as regional centers of advanced public health, higher education, and culture. They are typically more specialized, focusing on manufacturing. In these cities, *localization economies* tend to dominate. Finally, towns are linked to cities and connected to a mass of rural areas at the bottom of the hierarchy. They are facilitators of *internal scale economies* for mills and factories and market centers for farm products and rural output. Towns draw sustenance from surrounding rural areas and their prosperity also spills over to villages through nonfarm employment opportunities.[5]

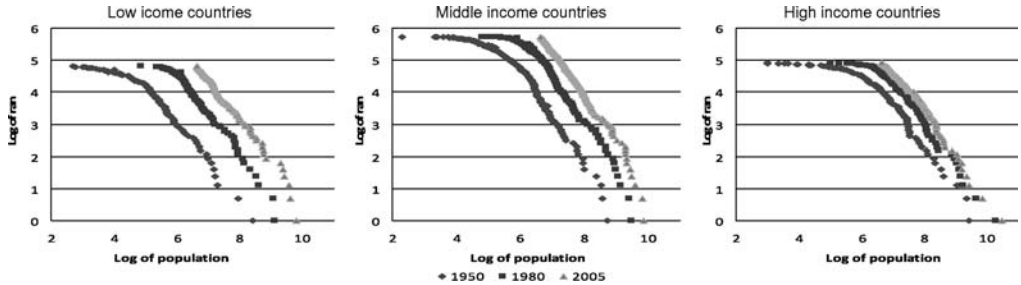
Externalities mean that history matters. Two otherwise identical enterprises in the same city could benefit differently from the local agglomeration depending on how long each has been in the city. Similarly, two otherwise identical cities would

offer different types of external economies depending on their history (Glaeser and others 1992; Henderson, Kuncoro, and Turner 1995). Such intangibles include the local stock of knowledge relevant for an industry or a labor force with specific acquired skills. The influences of history and specialization are consistent with the remarkably stable relative city-size distribution and the industrial concentration in specific cities over time even as incomes and populations grow. This pattern of stable relative size distribution is robust across income levels (figure 2).

This relationship is known as Zipf 's law. As Eeckhout (2004) notes, as early as 1682 Alexandre Le Maître observed a systematic pattern in the size of cities in France. Black and Henderson (2003) show that the relative sizes of U.S. cities have been stable over the last century. Similarly, Eaton and Eckstein (1997) and Dobkins and Ioannides (2001) find a pattern of stability in France, Japan, and the United States. They observe that the relative populations of the top 40 urban areas in France (1876–1990) and Japan (1925–85) remained essentially unchanged.

There is also persistence in the industrial concentration in specific cities (Henderson 1997b). Kim (1995) shows a high (0.64) correlation coefficient of regional localization for two-digit industries in the United States between 1860 and 1987 at the state level. Dumais, Ellison, and Glaeser (2002) also find that, for most industries, agglomeration patterns were strikingly stable over the period 1972–92. Likewise, Henderson (2003a) finds stable specialization patterns over 30 years of nine three-digit industries. Among mature industries the persistence in employment patterns across cities is high and the convergence in individual industry employment across cities is slow (Dunne, Roberts, and Samuelson 1989a, 1989b; Herzog and Schlottmann 1991; Davis and Haltiwanger 1992). Brezis and Krugman (1997) document that cities which underwent major sectoral overhauls were the exceptions. The norm has been the "lock-in" of industrial

Figure 2. Urban Hierarchies Remain Stable over Time and across Incomes



Source: United Nations (2006).

structure in specific cities, which can be explained by localization economies. These cities can better compete and, over time, retain plants and employment in that industry. A larger scale of own-industry activity historically means that firms in that locality today will operate more productively with greater accumulated knowledge about technology, sources of supply of different quality inputs, and local culture and its effect on the legal, business, and institutional climate. Henderson, Kuncoro, and Turner (1995) find that these localization advantages are relevant for traditional manufacturing industries which explain the longevity of particular clusters in certain locations, for example the world-class cutlery cluster in Solingen, Germany, since 1348 (Van der Linde 2003).

The evolving urban hierarchy is an enduring feature of economic development. Settlements of different sizes are linked through complementary economic functions.

Seen another way, places at different positions within the urban hierarchy are at different stages of urbanization. The types of scale economies that dominate in these places will depend on their stage of urbanization. Areas of *incipient* urbanization, with urban shares of about 25 percent, are predominantly agricultural or resource based and need policies to encourage internal economies of scale. Areas of *intermediate* urbanization, where urban population shares are about 50 percent, have seen economic alliances strengthen within and between urban areas. Many firms and plants in the same sector conglomerate to take advantage of shared inputs and knowledge spillovers. In such areas, the promotion of localization economies and congestion control is the highest priority. In areas of *advanced* urbanization, with urban shares of about 75 percent, productivity and consumption benefits arise from urbanization economies associated with the diversity and intensity of economic activity.

Implications for Policy

Towns, cities, and metropolises provide farms, firms, and families with the benefits of proximity, but the geographical compactness of activity also produces congestion, pollution, and social tension that can offset these benefits. Millions in Mumbai live in slums, with little or no access to basic amenities. High levels of crime are an accepted feature of urban living in Latin America. Workers in Jakarta, Kinshasa, Lagos, and Manila spend on average 75 minutes commuting to work. Even in better-managed cities like Seoul and Shanghai, downtown traffic averages 8 kilometers an hour (World Bank 2002). Crime, squalor, and congestion are the diseconomies of agglomeration which can erode the benefits of rising density.

Restricting the growth of cities is not the answer. While there is little evidence that agglomeration economies of megacities have been exhausted (Alonso 1971;

Sheshinski 1973; Kanemoto, Ohkawara, and Suzuki 1996; Au and Henderson 2006), the appropriate policies involve a balancing act between controlling negative externalities and steps to exploit the positive externalities from agglomeration. Most countries undergoing rapid urbanization have not matched urban growth with a commensurate expansion of transport infrastructure. While traffic in developing countries is increasing with per capita income along a path similar to that of the richer countries (World Bank 2002; McCrae 2006), road capacity has not increased enough. Successful cities react to growing congestion with spatially connective infrastructure. But preceding most such infrastructure investments in all successful cities are robust and versatile land market institutions. Land use and building regulations must adapt to shifts in population and production as urbanization advances. They must enable the integration of private and public use of land to provide space for transport infrastructure.

The congestion, grime, and crime-related costs that accompany and undermine rising concentration will, however, differ through the various stages of urbanization. Policies should be calibrated and sequenced accordingly. Towns require mainly the application of common institutions. Cities require the addition of spatially connective policies. And metropolises require all this and spatially targeted efforts such as slum development programs. The success of a policy initiative is predicated on the implementation of the ones introduced before it.

For areas of incipient urbanization, the policy goal in towns and villages includes such *spatially blind institutions* as flexible and efficient working of land markets, especially enforcement of property rights and provision of security, sanitation, streets, schools, and other basic services. For areas of intermediate urbanization, such spatially neutral efforts must be augmented by steps to ease congestion so that the benefits of rising density are more widely shared. *Spatially connective infrastructure* includes mass transit systems, roads and railways, and traffic demand management. Connections among rapidly urbanizing areas facilitate the relocation of production and people to places which supply most efficiently the desired internal or external economies. Finally, for metropolises or areas of advanced urbanization, policy must also address division within urban settlements, as reflected in the residential, as well as the economic and social, segregation of the poor in slums. *Spatially targeted interventions* such as slum upgrading and subsidized housing often become necessary for helping to facilitate urbanization economies.

Consistent with the implied prioritization, economic history and analysis of contemporary experience signal the need for a bedrock of spatially blind policies for equipping every place with the foundation of urbanization, while allowing markets to select the locations at which agglomeration happens. The United Kingdom in the nineteenth century is illustrative. The Reform Act of 1832 and the Municipal Corporations Act of 1835 allowed municipal governments to take

over privately owned sewerage, water, and gas systems to ensure universal provision of such basic amenities.[6] By the 1880s British land market institutions were mature enough to begin unifying patchy systems, separating sewage and drainage systems from the water systems, and extending the reach of basic services to poor areas (Land Enquiry Commission 1914; Hargan 2007). This facilitated rapid urban growth in the following decades (Offer 1981, p. 291).[7]

New York City provides another example of how versatile land market institutions allow a rapidly urbanizing city to accommodate complex public transit networks. Building a new or expanded transport network requires the purchase of contiguous plots of land, and holdouts can extract huge rents or thwart the project. The United States had a well-defined system of property rights by the mid-nineteenth century. As New York's transport networks expanded, its institutions could adapt. The 1916 Zoning Resolution was amended numerous times over the following decades as market needs changed (Dunlap 1992). Successful cities have flexible zoning laws to allow higher value users to bid for the valuable land. The interaction between such institutions and connective infrastructure enabled the density of Manhattan, the Bronx, Brooklyn, and Queens to increase from 230 people per square kilometer in 1820 to more than 5,000 in 1900 and about 12,000 today.[8]

Flexible land market institutions, the universal provision of basic social services, and investments in connective infrastructure help to make spatially targeted interventions such as slum clearance programs more likely to succeed. Hong Kong is an example (Bristow 1984; Adams and Hastings 2001; Cullinane 2002). Consistent with a tradition of minimal government intervention, Hong Kong's government contracted urban redevelopment to an organization led by private sector interests. Over five decades, slums were gradually eliminated. Similar success in a sustainable elimination of slums because of strong aspatial institutions coupled with connective infrastructure can be found in the United States (Chandler 1992; von Hoffman 1996; Hall 2002; Abreu 2008), Sweden (Swedish Council for Building Research 1990; Nesslein 2003; Borgegård and Kemeny 2004; Hall and Vidén 2005; Abreu 2008), and Costa Rica (Hall 1984; Trackman, Fisher, and Salas 1999; Ruster and Imparato 2003; World Bank 2007; Abreu 2008).

While an area's stage of urbanization (for example town, city, or metropolis) determines policy priorities, a nation's urban share can be a good indicator of the overall complexity of its urbanization challenges. Different parts of a country urbanize at different speeds. In the same way, more urbanized countries have more of their people in high-density areas and less urbanized countries have most people in rural areas. In the simplest case, one area may characterize an entire country, as in Singapore. For larger countries, more refined disaggregation can help determine priorities at different levels of government.

In a famous article Tinbergen (1952) proposed that one policy instrument is needed to address each policy objective. Urbanization's challenges increase with the stage of urbanization, so the required number of policy instruments increases as well. Fortunately for developing nations, the capacity of markets and governments grows as they urbanize. But these policies must be introduced in the right sequence. The institutional foundations of an inclusive urbanization have to be instituted early in the development process. Investments in infrastructure should come next. When implemented in settings where the institutions related to land markets and social services function reasonably well, and where transport infrastructure is adequate, targeted interventions aimed at reducing slums and squalor can be successful.

Notes

Indermit S. Gill and Chor-Ching Goh are, respectively, the Chief Economist and Senior Economist in Europe and Central Asia Region of the World Bank; email address: cgoh@worldbank.org. The authors would like to thank Sarah Boeckmann, Jean-Jacques Dethier, Vernon Henderson, John McCombie, Diego Puga, Mark Roberts, three anonymous referees, and the editor of this journal for helpful suggestions.

1. There were, of course, some exceptions, most notably in the work of Young (1928) and Kaldor (1972). These economists did not provide the technical solutions associated with the modeling of increasing returns to scale.

2. Warsh (2006) provides an entertaining and informative account of intellectual progress in dealing with increasing returns to scale. The advances are based on the special features of ideas, highlighted elegantly in Romer (1986, 1989). By adding knowledge explicitly to formulations of economic growth, economists are able to recognize the centrality of ideas and the importance of increasing returns, but this also requires recognizing the proliferation of imperfect competition.

3. Discussion of agglomeration economies dates as far back as Smith's consideration of specialization and the division of labor (Smith 1979); Marshall's information spillovers, searching and matching processes, and input sharing (Marshall 1890); and more recently interindustry supply linkages (Chinitz 1961); learning-by-doing (Arrow 1962); and cross-fertilization of ideas and innovation (Jacobs 1969). There are also consumption externalities from agglomeration, which are understudied. They are not covered here (but see Glaeser, Kolko, and Saiz 2001; Sinai and Waldfogel 2004).

4. This was recognized in geography as long ago as and.

5. Capello and Camagni (2000) argue that the real issue is not the size but the functional characteristics of a city, which depends on its position within the urban system. While negative externalities start to erode scale economies after a certain city size, urban development also generates conditions leading to structural readjustments which will create new economic advantages and innovative measures which will address diseconomies. These structural adjustments may be sectoral transformations toward higher-order functions or increases in external linkages with other cities. They provide econometric evidence of these processes among 58 Italian cities that confirm the integrative nature of an urban network.

6. Based on the historical collection of transport pamphlets the London School of Economics library has converted into digital format and made available for public use. See Roberts (2008).

7. More policy examples can be found in Chapter 7 of the *World Development Report 2009: Reshaping Economic Geography* (World Bank 2008).

8. Population and area figures are available at the US Census Bureau website.

References

The word *processed* describes informally reproduced works that may not be commonly available through libraries.

Abreu, Maria. 2008. "Effectively Dealing with Slums." Unpublished manuscript as background paper for the *World Development Report 2009 Reshaping Economic Geography.*

Adams, David, and E.M. Hastings. 2001. "Urban Renewal in Hong Kong: Transition From Development Corporation to Renewal Authority." *Land Use Policy* 18(3):245–58.

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth Through Creative Destruction." *Econometrica* 60(2):323–51.

_____. 2005. "Growth with Quality-Improving Innovations: An Integrated Framework." In Philippe Aghion, and Steven Durlauf, eds., *Handbook of Economic Growth*, Volume 1A. Amsterdam, Netherlands: North-Holland.

Alonso, William. 1971. "The Economics of Urban Size." *Regional Science Association Papers* 26: 67–83.

Angeriz, Alberto, John McCombie, and Mark Roberts. 2008. "Estimates of Returns to Scale and Spatial Spillovers for EU Regional Manufacturing." *International Regional Science Review* 31(62): 62–87.

Antweiler, Werner, and Daniel Trefler. 2002. "Increasing Returns and All That: A View from Trade." *American Economic Review* 92(1):93–119.

Arnott, Richard J. 1979. "Optimal City Size in a Spatial Economy." *Journal of Urban Economics* 6: 65–89.

Arrow, Kenneth J. 1962. "Economic Welfare and the Allocation of Resources for Invention." In Richard R. Nelson, ed., *The Rate and Direction of Inventive Activity.* Princeton, NJ: Princeton University Press.

Au, C.C., and Henderson, V. 2006. "Are Chinese Cities Too Small?" *Review of Economic Studies* 73: 549–76.

Baldwin, John R., and Paul K. Gorecki 1986. *Role of Scale in Canada–U.S. Productivity Differences in the Manufacturing Sector 1970–1979*. Toronto, Canada: University of Toronto Press.

Baldwin, Richard E., and Philippe Martin. 2004. "Agglomeration and Regional Growth". CEPR Discussion Paper 3960.

Bartlesman, Eric J., Ricardo J. Caballero, and Richard K. Lyons 1994. "Customer and Supplier-Driven Externalities." *American Economic Review* 84(4):1075–84.

Bikhchandani, Suchil, David Hirshleifer, and Ivo Welch. 1998. "Learning from the Behavior of Others: Conformity, Fads, and Information Cascades." *Journal of Economic Perspectives* 12(3): 151–70.

Black, Duncan, and J. Vernon Henderson 1999. "A Theory of Urban Growth." *Journal of Political Economy* 107 (2): 252–84.

_____ 2003. "Urban Evolution in the USA." *Journal of Economic Geography* 3(4):343–73.

Borgegård, Lars-Erik, and Jim Kemeny. 2004. "Sweden: High-Rise Housing for a Low-Density Country." In R. Turkington, R. van Kempen, and E. Wassenberg, eds., *High-Rise Housing in Europe: Current Trends and Future Prospects*. Delft, Netherlands: Delft University Press.

Brezis, Elise S., and Paul Krugman. 1997. "Technology and the Life Cycle of Cities." *Journal of Economic Growth* 2(4):369–83.

Bristow, R. 1984. *Land-Use Planning in Hong Kong: History, Policies, and Procedures*. Hong Kong: Oxford University Press.

Burchfield, Marcy, Henry G. Overman, Diego Puga, and Matthew Turner. 2006. "Causes of Sprawl: A Portrait from Space." *Quarterly Journal of Economics* 121(2):587–633.

Calem, Paul S., and Gerald A. Carlino 1991. "Urban Agglomeration Economies in the Presence of Technical Change." *Journal of Urban Economics* 29(1):82–95.

Capello, Roberta, and Roberto Camagni. 2000. "Beyond Optimal City Size: An Evaluation of Alternative Urban Growth Patterns." *Urban Studies* 37(9):1479–96.

Carlino, Gerald A. 1979. "Increasing Returns to Scale in Metropolitan Manufacturing." *Journal of Regional Science* 19(3): 369–72.

Carlino, Gerald, and Satyajit Chatterjee. 2001. "Employment Deconcentration: A New Perspective on America's Postwar Urban Evolution." Philadelphia, PA: Federal Reserve Bank of Philadelphia Working Paper Series 01-4.

Carlton, Dennis W. 1983. "The Location and Employment Choices of New Firms: An Econometric Model with Discrete and Continuous Endogenous Variables." *Review of Economics and Statistics* 65(3):440–9.

Chandler, Mittie Olion. 1992. "Public Housing Desegregation: What Are the Options?" *Housing Policy Debate* 3(2):509–34.

Charlot, Sylvie, and Gilles Duranton. 2003. "Communication Externalities in Cities." London, UK: Centre for Economic Policy Research Working Paper Series 4048.

Chinitz, Benjamin J. 1961. "Contrasts in Agglomeration: New York and Pittsburgh." *American Economic Review* 51(2):279–89.

Christaller, Walter. 1933. *Central Places in Southern Germany.* London, UK: Prentice Hall.

Ciccone, Antonio. 2002. "Agglomeration Effects in Europe." *European Economic Review* 46(2): 213–27.

Ciccone, Antonio, and Robert E. Hall 1996. "Productivity and the Density of Economic Activity." *American Economic Review* 86(1):54–70.

Combes, Pierre-Philippe. 2000. "Economic Structure and Local Growth: France 1984–1993." *Journal of Urban Economics* 47(3):329–55.

Costa, Dora L., and Matthew E. Kahn 2000. "Power Couples." *Quarterly Journal of Economics* 115(4): 1287–315.

Cullinane, Sharon. 2002. "The Relationship between Car Ownership and Public Transport Provision: A Case Study of Hong Kong." *Transport Policy* 9(1):29–39.

Davis, Steven J., and John C. Haltiwanger 1991. "Gross Job Creation, Gross Job Destruction, and Employment Reallocation." Chicago, IL: Federal Reserve Bank of Chicago Working Papers 91–5.

_____. 1992. "Gross Job Creation, Gross Job Destruction, and Employment Reallocation." *Quarterly Journal of Economics* 107(3): 819–63.

Dekle, Robert, and Jonathan Eaton. 1999. "Agglomeration and Land Rents: Evidence from the Prefectures." *Journal of Urban Economics* 46(2):200–14.

Desmet, Klaus, and Marcel Fafchamps. 2006. "Employment Concentration Across U.S. Counties." *Regional Science and Urban Economics* 36(4):482–509

Devereux, Michael P., Rachel Griffith, and Helen Simpson. 2004. "The Geographic Distribution of Production Activity in the U.K." *Regional Science and Urban Economics* 34(5):533–64.

Diamond, P.A., and J.A. Mirrlees 1971. "Optimal Taxation and Public Production; I: Production Efficiency" and "II: Tax Rules." *American Economic Review* 61:8–27 and 261–78.

_____. 1973. "Aggregate Productions and Consumption Externalities," *Quarterly Journal of Economics*, LXXXVI: 1–24.

Dixit, Avinash. 1973. "The Optimum Factory Town." *Bell Journal of Economics and Management Science* 4(2):637–51.

Dixit, Avinash, and Joseph E. Stiglitz 1977. "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67(3):297–308.

Dobkins, Linda H., and Yannis M. Ioannides 2001. "Spatial Interactions Among US Cities, 1900–1990." *Regional Science and Urban Economics* 31(6):701–32.

Dumais, Guy, Glenn Ellison, and Edward L. Glaeser 2002. "Geographic Concentration as a Dynamic Process." *Review of Economics and Statistics* 84(2): 193–204.

Dunlap, David W. 1992. "Some Land-Use History Highlights." *New York Times*, April 12.

Dunne, Timothy, Mark J. Roberts, and Larry Samuelson. 1989a. "Plant Turnover and Gross Employment Flows in the US Manufacturing Sector." *Journal of Labor Economics* 7(1):48–71.

Dunne, Timothy, Mark J Roberts, and Larry Samuelson. 1989b. "The Growth and Failure of US Manufacturing Plants." *Quarterly Journal of Economics* 104(4):671–98.

Duranton, Gilles. 2007. "Urban Evolutions: The Fast, the Slow, and the Still." *American Economic Review* 97(1):197–221.

Duranton, Gilles, and Henry G. Overman 2002. "Testing for Localisation Using Micro-Geographic Data." London, UK: Centre for Economic Performance Discussion Paper 540.

Duranton, Gilles, and Diego Puga. 2001 "Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of Products." *American Economic Review* 91(5): 1454–77.

_____. 2004. "Micro-Foundations of Urban Agglomeration Economies." In J. Vernon Henderson, and Jacque Thisse, eds., *Handbook of Urban and Regional Economies*, Volume 4. Amsterdam: North Holland.

_____. 2005 . "From Sectoral to Functional Urban Specialization." *Journal of Urban Economics* 57(2): 343–70.

Easterly, William, and Ross Levine. 2001. "What Have We Learned from a Decade of Empirical Research on Growth? It's Not Factor Accumulation: Stylized Facts and Growth Models." *World Bank Economic Review* 15(2):177–219.

Eaton, Jonathan, and Zvi Eckstein. 1997. "Cities and Growth: Evidence from France and Japan." *Regional Science and Urban Economics* 27(4–5):443–74.

Eeckhout, Jan. 2004. "Gibrat's Law for (All) Cities." *American Economic Review* 94(5):1429–51.

Ellison, Glenn, and Edward L. Glaeser 1997. "Geographic Concentration in US Manufacturing Industries: A Dartboard Approach." *Journal of Political Economy* 105(5):889–927.

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

Fafchamps, Marcel, and Klaus Desmet. 2000. "The Changing Spatial Distribution of Economic Activity Across US Counties." Oxford, UK: Economic Series Working Paper 43.

Feldman, Maryann P. 1994. *The Geography of Innovation*. Boston, MA: Kluwer.

Feldman, Maryann P., and David B. Audretsch 1999. "Innovation in Cities: Science-Based Diversity, Specialization and Localized Competition." *European Economic Review* 43(2):409–29.

Fujita, Masahisa, and Ryoichi Ishii. 1999. "Global Location Behavior and Organizational Dynamics of Japanese Electronics Firms and Their Impact on Regional Economies." In Alfred D. Chandler Jr., Peter Hagstrom, and Orjan Solvell, eds., *The Dynamic Firm: The Role of Technology, Strategy, Organization, and Regions*. Oxford, U.K.: Oxford University Press.

Fujita, Masahisa, Paul Krugman, and Anthony J. Venables 1999. *The Spatial Economy. Cities, Regions, and International Trade*. Cambridge, MA: The MIT Press.

Fuss, Melvyn A., and Vinod K. Gupta 1981. "A Cost Function Approach to the Estimation of Minimum Efficient Scale, Returns to Scale, and Suboptimal Capacity : With an Application to Canadian Manufacturing." *European Economic Review* 15(2):123–35.

Gill, Indermit, and Homi Kharas. 2007. *An East Asia Renaissance: Ideas for Economic Growth*. Washington, DC: The World Bank.

Glaeser, Edward L., José A. Scheinkman, and Andrei Schleifer. 1995. "Economic Growth in a Cross-Section of Cities." *Journal of Monetary Economics* 36(1): 117–43.

Glaeser, Edward L., and Matthew E. Kahn 2001. "Decentralized Employment and the Transformation of the American City." Boston, MA: Harvard Institute of Economic Research Working Paper Series 1912.

Glaeser, Edward L., and David C. Mare 2001. "Cities and Skills." *Journal of Labor Economics* 19(2): 316–42.

Glaeser, Edward L., Jed Kolko, and Albert Saiz. 2001. "Consumer City." *Journal of Economic Geography* 1(1):27–50.

Glaeser, Edward L., Hedi D. Kallal, Jose A. Scheinkman, and Andrei Shleifer. 1992. "Growth in Cities." *Journal of Political Economy* 100(6):1126–52.

Griliches, Zvi, and Vidar Ringstad. 1971. *Economies of Scale and the Form of the Production Function: An Econometric Study of Norwegian Manufacturing Establishment Data*. Amsterdam, Netherlands: North-Holland.

Grossman, Gene M., and Elhanan Helpman. 1991. "Quality Ladders in the Theory of Growth." *Review of Economic Studies* 58(1):43–61.

_____. 1995. "Comparative Advantage and Long Run Growth." In J.P. Neary, ed., *International Trade*. Aldershot, UK: Edward Elgar.

Gupta, Vinod K. 1983. "Labor Productivity, Establishment Size, and Scale Economies." *Southern Economic Journal* 49(3):853–9.

Hall, Carolyn. 1984. "Regional Inequalities in Well-Being in Costa Rica." *Geographical Review* 74(1): 48–62.

Hall, Peter. 2002. *Cities of Tomorrow*. Oxford, UK: Blackwell Publishing.

Hall, Thomas, and Sonja Vidén. 2005. "The Million Homes Programme: A Review of the Great Swedish Planning Model." *Planning Perspectives* 20(3):301–28.

Hammond, George William, and Jürgen von Hagen. 1994. "Industrial Localization: An Empirical Test for Marshallian Localization Economies." Working Paper 917, Centre for Economic Policy Research, London.

Hansen, Eric R. 1990. "Agglomeration Economics and Industrial Decentralization: The Wage-Productivity Trade-Offs." *Journal of Urban Economics* 28(2):140–59.

Hargan, Jim. 2007. "Prince Charles's Poundbury." *British Heritage*. May.

Hay, Donald Andrew. 1979. "The Location of Industry in a Developing Country: The Case of Brazil." *Oxford Economic Papers* 31(1):93–120.

Helpman, Elhanan, and Paul Krugman. 1987. *Market Structure and Foreign Trade*. Cambridge, MA: MIT Press.

Henderson, J. Vernon. 1974a. "The Sizes and Types of Cities." *American Economic Review* 64(4): 640–56.

_____. 1974b. "Optimum City Size: The External Diseconomy Question." *Journal of Political Economy* 82(2):373–88.

_____. 1975. "Congestion and Optimum City Size." *Journal of Urban Economics* 2:48–62.

_____, 1977. *Economic Theories and Cities*, New York: Academic Press.

_____. 1986. "Efficiency of Resources Usage and City Size." *Journal of Urban Economics* 19(1): 47–70.

_____. 1994. "Where Does an Industry Locate?" *Journal of Urban Economics* 35(1):83–104.

_____. 1997a. "Externalities and Industrial Development." *Journal of Urban Economics* 42(3): 449–70.

_____. 1997b. "Medium Size Cities." *Regional Science and Urban Economics* 27(6):583–612.

_____. 2003a. "Marshall's Scale Economies." *Journal of Urban Economics* 53(1):1–28.

_____. 2003b. "The Urbanization Process and Economic Growth: The So-What Question." *Journal of Economic Growth* 8(1):47–71.

Henderson, J., Vernon Ari Kuncoro, and Damhuri Nasution. 1996. "The Dynamics of Jabotabek Development." *Bulletin of Indonesian Economic Studies* 32(1):71–95.

Henderson, J., Vernon Ari Kuncoro, and Matthew Turner. 1995. "Industrial Development in Cities." *Journal of Political Economy* 103(5):1067–85.

Henderson, J., Vernon Todd Lee, and Yung Joon Lee. 2001. "Scale Externalities in Korea." *Journal of Urban Economics* 49(3):479–501.

Henderson, J. Vernon, Zmarak Shalizi, and Anthony J. Venables 2001. "Geography and Development." *Journal of Economic Geography* 1(1):81–105.

Herzog, Henry W., and Alan M. Schlottmann 1991. *Industry Location and Public Policy*. Knoxville, TN: University of Tennessee.

Hoffman A., von. 1996. "High Ambitions: The Past and Future of American Low-Income Housing Policy." *Housing Policy Debate* 7(3):423–46.

Holmes, Thomas J., and John J. Stevens 2002. "Geographic Concentration and Establishment Scale." *Review of Economics and Statistics* 84(4):682–91.

Jacobs, Jane. 1969. *The Economy of Cities*. New York: Vintage.

Jaffe, Adam B. 1986. "Technological Opportunity and Spillover of R&D: Evidence from Firms' Patents, Profits, and Market Value." *American Economic Review* 76(5): 984–1001

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic Localization of Knowledge Spillovers as Evidence by Patent Citations." *Quarterly Journal of Economics* 108(3): 577–98.

Junius, Karsten. 1997. "Economies of Scale: A Survey of Empirical Literature." Kiel, Germany: Kiel Institute of World Economics Working Paper Series 813.

Kaldor, Nicholas. 1972. "The Irrelevance of Equilibrium Economics." *Economic Journal* 82(328): 1237–55.

Kanemoto, Yoshitsugu, Toru Ohkawara, and Tsutomu Suzuki. 1996. "Agglomeration Economies and a Test for Optimal City Sizes in Japan." *Journal of the Japanese and International Economies*, 10(4):379–98.

Kawashima, Tatsuhiko. 1975. "Urban Agglomeration Economies in Manufacturing Industries." *Papers in Regional Science* 34(1):157–75.

Kim, Sukkoo. 1995. "Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in US Regional Manufacturing Structure, 1860–1987." *Quarterly Journal of Economics* 110(4):881–908.

Kim, Sukkoo, and Robert A. Margo 2004. "Historical Perspectives on U.S. Economic Geography." In J. Vernon Henderson, and Jacque-Francois Thisse, eds., *Handbook of Regional and Urban Economics, Volume 4: Cities and Geography*. Amsterdam: North-Holland.

Kolko, Jed. 1999. "Can I get Some Service Here? Information Technology, Service Industries, and the Future of Cities." PhD thesis. Harvard University.

_____. 2007. "Agglomeration and Co-Agglomeration of Services Industries." Public Policy Institute of California, San Francisco. Processed.

Krugman, Paul R. 1980. "Scale Economies, Product Differentiation, and the Pattern of Trade." *American Economic Review* 70(5):950–9.

_____ 1981. "Intraindustry Specialization and the Gains from Trade." *Journal of Political Economy* 89(5):959–73.

_____ 1991. "Increasing Returns to Scale and Economic Geography." *Journal of Political Economy* 99(3):483–99.

Kuncoro, Ari. 2008. "Spatial Agglomeration, Site Productivity and Local Governance: Indonesian Experience from 1980 to 2003." In Yukon Huang, and Alessandro Magnoli Bocchi, eds., *Reshaping Economic Geography in East Asia*. Washington, DC: World Bank; EAP Companion Volume to the WDR 2009.

Land Enquiry Commission. 1914. *The Land*, Vol. II. London: Hodder and Stoughton.

Levinsohn, James, and Amil Petrin. 1999. "When Industries Become More Productive, Do Firms?" NBER Working Paper 6893, National Bureau of Economic Research, Cambridge, MA.

Lösch, Auguste. 1940. *The Economics of Location*. New Haven, CT: Yale University Press.

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

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

Martins, Joaquim Oliveira, Stefano Scarpetta, and Dirk Pilat. 1996. "Mark-up Ratios in Manufacturing Industries: Estimates for 14 OECD Countries." Working Paper 162, Organisation for Economic Co-operation and Development, Paris.

Maurel, Françoise, and Beatrice Sedillot. 1999. "A Measure of the Geographic Concentration in French Manufacturing Industries." *Regional Science and Urban Economics* 29(5):575–604.

McCrae, Ian. 2006. "Uncertainties in Road Transport Emissions in Developing Countries." HM Treasury, London.

Mera, Koichi. 1973. "Regional Production Function and Social Overhead Capital: An Analysis of Japanese Case." *Regional Science and Urban Economics* 3(2):157–86.

Mills, Edwin S. 1967. "An Aggregative Model of Resource Allocation in a Metropolitan Area." *American Economic Review*, 57(2):197–210.

Mills, Edwin S., and David M. Ferranti 1971. "Market Choices and Optimal City Size." *American Economic Review* 61(2):340–345.

Moomaw, Ronald L. 1981. "Productivity and City Size: A Critique of the Evidence." *Quarterly Journal of Economics* 96(4):675–88.

_____. 1983. "Is Population Scale a Worthless Surrogate for Business Agglomeration Economies?" *Regional Science and Urban Economics* 13(4):525–45.

Moretti, Enrico. 2004a. "Estimating the Social Return to Higher Education: Evidence from Longitudinal and Repeated Cross-Sectional Data." *Journal of Econometrics*121(1–2):175–212.

_____. 2004b. "Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions." *American Economic Review* 94(3):656–90.

Nakamura, Ryohei. 1985. "Agglomeration Economies in Urban Manufacturing Industries: A Case of Japanese Cities." *Journal of Urban Economics* 17(1):108–24.

Nesslein, Thomas S. 2003. "Markets versus Planning: An Assessment of the Swedish Housing Model in the Post-War Period." *Urban Studies* 40(7):1259–82.

Offer, A. 1981. *Property and Politics, 1870–1914*. Cambridge, UK: Cambridge University Press.

Paul, Catherine, J. Morrison, and Donald S. Siegel 1999. "Scale Economies and Industry Agglomeration Externalities: A Dynamic Cost Function Approach." *American Economic Review* 89(1):272–90.

Prateen, Cliff. 1988. "Survey of the Economies of Scale." Economic Papers 67, Commission of the European Communities, Brussels.

Rauch, James E. 1993. "Productivity Gains from Geographic Concentration of Human Capital: Evidence from the Cities." *Journal of Urban Economics* 34(3):380–400.

Roberts, Mark. 2008. "Congestion and Spatially Connective Infrastructure: The Case of London in the 19th and early 20th Century." Unpublished manuscript as background paper for the *World Development Report 2009 Reshaping Economic Geography*.

Roeger, Werner. 1995. "Can Imperfect Competition Explain the Difference between Primal and Dual Productivity Measures? Estimates for U.S. Manufacturing." *Journal of Political Economy* 103(2): 316–30.

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

_____. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98(5): S71–102.

Romer, Paul M., Rosenthal, Stuart S., and William C. Strange 2001. "The Determinants of Agglomeration." *Journal of Urban Economics* 50(2):191–229.

_____. 2003. "Geography, Industrial Organization, and Agglomeration." *Review of Economics and Statistics* 85(2):377–93.

_____. 2004. "Evidence on the Nature and Sources of Agglomeration Economies." In J. Vernon Henderson, and Jacque Thisse, eds., *Handbook of Regional and Urban Economics*, Volume 4. Amsterdam:North-Holland: Elsevier.

Rossi-Hansberg, Esteban, and Mark L.J. Wright 2007. "Urban Structure and Growth." *Review of Economic Studies* 74(2):597–624.

Ruster, Jeff, and Ivo Imperato. 2003. *Slum Upgrading and Participation: Lessons from Latin America*. Washington, DC: World Bank.

Santos, Georgina, and Blake Shaffer. 2004. "Preliminary Results of the London Congestion Charging Scheme." *Public Works Management* 9:164–81.

Sasaki, Komei. 1998. "Optimal Urban Growth Controls." *Regional Science and Urban Economics* 28: 475–96.

Segal, David. 1976. "Are there Returns to Scale in City Size?" *Review of Economics and Statistics* 58(3):339–50.

Shefer, Daniel. 1973. "Localization Economies in SMSAs: A Production Function Analysis." *Journal of Regional Science* 13(1):55–64.

Sheshinski, Eytan. 1973. "Congestion and the Optimum City Size." *American Economic Review* 63(2):61–6.

Sinai, Todd, and Joel Waldfogel. 2004. "Geography and the Internet: Is the Internet a Substitute or Complement for Cities?" *Journal of Urban Economics* 56(1):1–24.

Smith, Adam. 1979. *An Inquiry into the Nature and Causes of the Wealth of Nations, 1939 Canaan Edition*. London: Penguin Books.

Spence, Michael, 1976. "Product Selection, Fixed Costs, and Monopolistic Competition." *Review of Economic Studies* 43(2):217–35.

Sveikauskas, Leo A. 1975. "The Productivity of Cities." *Quarterly Journal of Economics* 89(3): 393–413.

Sveikauskas, Leo A., John Gowdy, and Michael Funk. 1988. "Urban Productivity: City Size or Industry Size." *Journal of Regional Science* 28(2):185–202.

Sveikauskas, Leo A., Peter M. Townroe, and Eric R. Hansen 1985. "Intraregional Productivity Difference in Sao Paulo State Manufacturing Plants." *Weltwirtschaftliches Archiv* 121(4):722–40.

Swedish Council for Building Research. 1990. *Housing Research and Design in Sweden*. Stockholm: Swedish Council for Building Research.

Tinbergen, J. 1952. *On the Theory of Economic Policy*. Amsterdam: North-Holland.

Trackman, B., W. Fisher, and L. Salas 1999. *The Reform of Property Registration Systems in Costa Rica: A Status Report*. Cambridge, MA: Law School Harvard University.

Tolley, George. 1974. "The Welfare Economics of City Bigness." *Journal of Urban Economics* 1: 324–45.

Tolley, George, and John Crihfield. 1987. "City Size and Place as Policy Issues." In E.S. Mills, ed., *Handbook of Regional and Urban Economics* Amsterdam: Elsevier North Holland :1285–311.

United Nations. 2006. "World Urbanization Prospects: The 2005 Revision". New York, United Nations, Department of Economic and Social Affairs, Population Division.

Van der Linde, Claas. 2003. "The Demography of Clusters: Findings from the Cluster Meta-Study." In J. Brocker, D. Dohse, and R. Soltwedel, eds., *Innovation Clusters and Interregional Competition*. Berlin, Germany: Springer.

Venables, Anthony J. 2006. "Shifts in Economic Geography and Their Causes." *Federal Reserve Bank of Kansas City Economic Review* Q IV:61–85.

Warsh, David. 2006. *Knowledge and the Wealth of Nations: A Story of Economic Discovery*. New York: W.W. Norton.

Wheeler, David, and Ashoka Mody. 1992. "International Investment Location Decisions: The Case for U.S. Firms." *Journal of International Economics* 33(1–2):57–76.

World Bank. 2002. *Cities on the Move: A World Bank Urban Transport Strategy Review*. Washington, DC: World Bank.

_____. 2006. *Governance, Investment Climate, and Harmonious Society: Competitiveness Enhancements for 120 Cities in China*. Washington, DC: World Bank.

_____. 2007. *World Development Indicators 2007*. Washington, DC: World Bank.

_____. 2008. *World Development Report 2009 Reshaping Economic Geography*. Washington, DC: World Bank.

Young, Allwyn. 1928. "Increasing Returns and Economic Progress." *Economic Journal* 38(152): 527–42.

Estimation of Water Demand in Developing Countries: An Overview

Céline Nauges • Dale Whittington

A better understanding of household water use in developing countries is necessary to manage and expand water systems more effectively. Several meta-analyzes have examined the determinants of household water demand in industrialized countries, but little effort has been made to synthesize the growing body of literature evaluating household water demand in developing countries. This article reviews what is known and what is missing from that literature thus far. Analysis of demand for water in developing countries is complicated by abundant evidence that, contrary to what is observed in most developed countries, households in developing countries have access to, and may use more than one of several types of, water sources. The authors describe the different modeling strategies that researchers have adopted to estimate water demand in developing countries and discuss issues related to data collection. The findings from the literature on the main determinants of water demand in these countries suggest that, despite heterogeneity in the places and time periods studied, most estimates of own-price elasticity of water from private connections are in the range from -0.3 *to* -0.6*, close to what is usually reported for industrialized countries. The empirical findings on decisions relating to household water sources are much less robust and should be a high priority for future research.* JEL codes: D12, O13, Q25, Q56

This article reviews what is known and what is missing from the growing body of literature on household water-demand functions in developing countries. We also discuss the challenges researchers face in carrying out studies of household water demand in the constrained data environment of developing countries, and how these can be overcome.

Studies of residential water demand in industrialized countries have mainly concerned measurement of price and income elasticities. In these countries

The World Bank Research Observer

doi;10.1093/wbro/lkp016 Advance Access publication November 11, 2009 25:263–294

almost all households have a connection to the piped water network. Tap water, generally of good quality, is the primary source for all water uses. These characteristics permit a relatively straightforward estimation of the household water-demand function. The chief methodological issue that has been extensively discussed in this literature is the nonlinearity of the pricing scheme, which may cause an endogeneity bias at the estimation stage.

Analyzes of household water demand in developing countries first appeared in the work of White, Bradley, and White (1972), Katzman (1977), and Hubbell (1977), but they remain limited even today. One reason for this lack of attention is that such analyzes are more difficult to do. This is mainly because conditions surrounding water access often vary across households, and this variability makes it almost impossible to base a comprehensive analysis of household water demand on secondary data from the water utility. Households often rely on a variety of water sources, including piped and nonpiped sources with different characteristics and levels of services (price, distance to the source, quality, reliability, and so on). For many households in developing countries water is a heterogeneous good, which is not usually the case in industrialized countries (Mu, Whittington, and Briscoe 1990). Obtaining water from nontap sources outside the house involves collection costs that need to be taken into account to assess household behavior accurately.

Researchers have employed four principal strategies to obtain the information needed to investigate the behavior of household water demand in developing countries. First, well-designed household surveys can be used to complement existing data from public (and private) utilities.[1] Second, households can be asked questions about how they would behave in hypothetical water-use situations (for example, Whittington and others 1990; The World Bank Water Demand Research Team 1993; Whittington and others 2002). Third, researchers can look to secondary markets such as housing to draw inferences about how households value improved water services (for example, North and Griffin 1993; Daniere 1994; and, for a review, Komives 2003). Fourth, experimental methods (including randomized controlled trials) can be used to test how households behave in response to different water supply interventions (Kremer and others 2007, 2008).

This paper reviews the literature that uses data from utilities and household surveys to estimate household water-demand functions, not papers that investigate the behavior of water demand based on stated preference techniques, revealed preference techniques, or experimental methods. We begin with an overview of three large groups of households in developing countries and discuss why water planners need somewhat different information about the behavior of household water demand to address the policy challenges each household group poses. We then provide a brief overview of the literature on the estimation of water-demand functions in industrialized countries because research based on data from developing countries has been informed by findings from this work.

Methodologies developed to correct for price endogeneity under nonlinear pricing have in particular been applied in recent studies of household water-demand functions in developing countries.

Next we describe the different modeling strategies that researchers have adopted to estimate water-demand functions in developing countries, and discuss issues related to data collection. We then review the findings from the literature on the main determinants of water-demand functions in developing countries: water price, cost of water collection, quality of water service, and household socio-economic characteristics. In our conclusions we discuss the policy implications of the findings from this literature and indicate directions for future research. In the Appendix we offer some recommendations for the design of household surveys that collect data for estimating water demand functions.

Background

Broadly speaking, there are three large groups of households in developing countries today, each with its own distinct set of water and sanitation challenges. First, there are hundreds of millions of households living in the medium and large cities of China, India, Southeast Asia, and Latin America with monthly incomes of US$150–400. Most of these households can now afford municipal piped-water services in their homes or will soon be able to do so. For many of these households, full sewerage collection and treatment may remain financially out of reach for some time, but rising incomes will increase demand for the services of a modern piped-water supply and put pressure on governments to ensure that better services are provided. The challenge for water supply managers serving this first group of households is to raise the financing necessary to pay for the capital-intensive investments needed to expand system capacity and improve water quality and service reliability (Whittington and others 2009). An understanding of how the quantity of water used by households is affected by tariff structures and other factors is needed to help guide public pricing policies, that is to design tariff structures that will both raise funds for financing system improvements and better balance the economic value of water to households with the rising costs of supply.

The second large group of households live in the expanding slums of cities through the developing world and typically have incomes of less than US$150 per month. Many of these households currently lack in-house piped connections and the income to obtain them. In densely crowded slums there are often large positive externalities associated with improved sanitation. Because such sanitation is crucial for public health, improvements in water supply must compete with sanitation investments for limited public subsidies. Here the challenge is to design

tariffs and subsidies so that the basic needs of all households can be met. At the same time the incomes of many of these households are also growing, and water planners should not design service options and tariffs that trap these slum households for long periods with only intermediate water and sanitation services. For this second group, water planners need a better understanding of both the factors that determine the water source choices made by households and the quantity of water used, so that piped services can be offered to the minority of households that can afford them—and other households can be served by cheaper, more basic, levels of service.

The third large group of households live in the rural areas of Sub-Saharan Africa and South Asia on less than US$1 per person per day. For the majority of these households, in-house piped water and sanitation services are prohibitively expensive and will remain out of reach for the foreseeable future. The design of rural water supply projects and programs to reach this third group of households has a long history of failure (Therkildsen 1988). Hundreds of millions of dollars have been spent by donors on projects that households do not want and that are subsequently abandoned. Regardless of the type of technology utilized by donors, systems were not repaired and fell into disuse. Cost recovery was minimal and revenues were often insufficient to pay for even basic operation and maintenance, much less capital costs. Communities did not have a sense of ownership in their water projects, and households were not satisfied with the type of services that donors and national governments provided.

Over the past two decades a global consensus has gradually emerged among national governments and donors about what has been learned from this failure and how best to design rural water-supply programs to serve households in such communities (Whittington and others 2009). Most sector professionals would now agree that a well-designed rural water-supply program should include the following:

1. Involve households in the choice of both technology (service level) and institutional and governance arrangements;
2. Give women a larger role in decision-making;
3. Require households to pay all of the operation and maintenance costs of providing water services and at least some of the capital costs;
4. Transfer ownership of the facilities to the community;
5. Involve households in the design of cost recovery systems and tariffs to be charged.

The role of higher level government (for example national, state, province), perhaps assisted by donors, is to decide the following:

1. The eligibility rules (that is which communities are eligible to participate in the program);

2. The feasible technological options to offer to communities;
3. The cost-sharing rules (how much will government pay, how much will communities pay);
4. The protocol for transferring ownership of facilities to the communities;
5. The government's financing of the program (grants versus loans);
6. How best to provide communities with information about the program.

In order for governments and donors to make informed decisions about the design of these program rules for rural water-supply programs, they need better information in particular about source choices made by households, that is the factors that determine whether or not households will decide to use the public taps and community handpumps that are the typical service options provided by rural water supply programs. For this third group of households, the interconnection between sanitation and water investments is less critical than for households living in urban slums. In rural areas, the negative externalities associated with poor sanitation often can be more effectively addressed by behavioral change than by infrastructure investments.

We acknowledge that there is considerable heterogeneity among households in each of these three groups. Nevertheless, we believe this simple typology is helpful because it illustrates that the information on the behavior of household water demand that is needed for policy decisions is somewhat different for the three groups. For households with piped connections living in the nonslum parts of medium and large cities, water planners need to know how household water use from piped connections responds to changes in tariffs, given that some households may rely on multiple water sources. For poorer households living in slum areas, information on how households with piped connections respond to changes in tariffs is still important, though source choices made by households themselves assume greater policy relevance because the decision by households to connect to the piped distribution system cannot be taken for granted. Finally, for poor households in rural areas that cannot afford a connection to a piped distribution system, water planners primarily need information about the determinants of the source choices made by the households, not the quantity of water used.

Estimation of Household Water-Demand Functions in Industrialized Countries

Research based on data from developing countries has been significantly informed by the findings from the literature on the estimation of water-demand functions in industrialized countries.

Literature

This literature includes many empirical papers, starting with the work of Gottlieb (1963) and Howe and Linaweaver (1967). Studies have been made in a large number of countries, including Australia (Grafton and Ward 2008), Canada (Kulshreshtha 1996), Denmark (Hansen 1996), France (Nauges and Thomas 2000), Spain (Martínez-Espiñeira 2002), Sweden (Höglund 1999), and especially the United States (Foster and Beattie 1979; Agthe and Billings 1980; Chicoine, Deller, and Ramamurthy 1986; Nieswiadomy and Molina 1989; Hewitt and Hanemann 1995; Pint 1999; Renwick and Green 2000). For comprehensive reviews of this literature see Espey, Espey, and Shaw (1997), Hanemann (1998), Arbués-Gracia, García-Valiñas, and Martínez-Espiñeira (2003), and Dalhuisen and others (2003).

Modeling Strategies

In almost all studies performed in industrialized countries, the residential water-demand function is specified as a single equation linking (tap) water use (the dependent variable) to water price and a vector of demand shifters (household socioeconomic characteristics, housing features, climatologic variables, and so on) to control for heterogeneity of preferences and other variables affecting water demand.[2] A popular functional form is the double log, which yields direct estimates of elasticities but constrains the elasticity to be constant. There are few discussions on the choice of functional form, except by Griffin and Chang (1991), who advocate more flexible forms such as the generalized Cobb-Douglas, and Gaudin, Griffin, and Sickles (2001), who discuss the tradeoff between simplicity and parsimony of parameters.

This single-equation modeling strategy implicitly assumes that there is no substitute available for water. The only exception is Hansen (1996), who considers water and energy prices in the demand function for water. Water quality and the reliability of the water supply service are generally not included in the single-equation model as controls because there is little variation in terms of service quality across households on the same distribution system. The focus instead has been on the estimation of price elasticity and the measurement of the impact of socioeconomic characteristics (mainly income) on the quantity of water used.

The main methodological issues relate to the choice of marginal or average price and to price endogeneity when households face a nonlinear pricing scheme (for example increasing or decreasing block-pricing tariff structures). Although economic theory suggests the use of marginal price (the price of the last cubic meter), average price (computed as total bill divided by total consumption) has often been preferred. Authors who use average price argue that households are rarely well informed about

the tariff structure used by their local water utility and are thus more likely to react to adjustments in average price than in marginal price. Estimation of the residential water-demand function when the pricing scheme is nonlinear has been the focus of numerous articles, including Agthe and others (1986); Deller, Chicoine, and Ramamurthy (1986); Nieswiadomy and Molina (1989); Hewitt and Hanemann (1995); Olmstead, Hanemann, and Stavins (2007).

Data

In studies of household water-demand functions in industrialized countries, data for the model estimation typically come from water utility records. An important advantage of relying on these records is that panel data on each household's water use are usually available. A disadvantage is that water utilities typically maintain little socioeconomic or demographic information on the households they serve. There is also little variation in potentially important covariates, such as the tariff structure itself and water quality and reliability.

Results

Most studies find that household water demand is both price and income inelastic. Espey, Espey, and Shaw (1997) report an average own-price elasticity of -0.51 from industrialized countries. Income elasticity has often been estimated in the range $0.1-0.4$ (see Arbués-Gracia, García-Valiñas, and Martínez-Espiñeira 2003). Other household characteristics (size and composition), housing characteristics (principal versus secondary residence; size of the lawn or garden, if any; stock of water-using appliances), and weather data are commonly acknowledged as determinants of water use in industrialized countries.

Estimation of Household Water-Demand Functions in Developing Countries: Modeling Strategies

Analysis of demand for water in developing countries is complicated by abundant evidence that, contrary to what is observed in most developed countries, households in all three groups described above have access to and may use more than one of several types of water sources, such as in-house tap connections, public or private wells, public or (someone else's) private taps, water vendors or resellers, tank trucks, water provided by neighbors, rainwater collection, or water collected from rivers, streams, or lakes. That some households utilize more than one source may indicate that their use of a particular convenient source is rationed (implying

that additional water must be taken from an alternative source); or that it is relatively cheap to take some water but not all from a particular source (for example the household may have limited capacity to haul cheap water from a given source and may prefer to obtain the rest more expensively from another source); or that waters from different sources are used for different purposes (drinking, bathing, cleaning, and so on). The choice determined upon, as well as the conditions of access, can vary significantly across households. In the formal parts of large cities, piped networks are typically common, but many people may not be connected, for a variety of reasons, and even those that are connected may use a variety of other water sources. In urban slums residents sometimes have access to a connection to a piped network but often exploit a wide variety of water sources. In poorer rural areas, piped distribution networks with private connections are the exception.

Three basic approaches to estimating household water-demand functions in developing countries can be seen in the literature:

1. *Estimation of (unconditional) demand for water coming from one particular source.* When households rely on a unique source or when water comes primarily from one source, a demand equation for water from that particular source can be estimated from data on the subsample of households using that source. For example, Rizaiza (1991) estimates separately water-demand equations for households with a private connection and for households supplied with tankers in the four major cities (with populations between 700,000 and 4 million) of the western region of Saudi Arabia (namely Jeddah, Makkah, Madinah, and Taif). Crane (1994) specifies separate demand equations for a sample of households in Jakarta (population around 8 million), Indonesia, that were supplied by water vendors, and for households relying on public taps (hydrants). David and Inocencio (1998) use data from Metro Manila (population around 11 million), the Philippines, to estimate separate demand equations for households supplied by water vendors and for households with a private connection. Rietveld, Rouwendal, and Zwart (2000) and Basani, Isham, and Reilly (2008) estimate the water-demand function for households with a piped connection in, respectively, Salatiga (a medium-sized city of about 150,000 inhabitants in Central Java, Indonesia) and seven provincial towns in Cambodia (all between 400,000 and 1 million inhabitants).
2. *Discrete analysis of source choice.* In some cases (Crane 1994; David and Inocencio 1998) dummy variables are introduced in single demand equations to control for possible use of additional sources. The estimation of (single) source-specific demand equations provides insight into the sensitivity of water use to the price of water from that particular source. However, this approach does not allow the analyst to measure cross-price elasticities in the case of

households that combine water from different sources. A system of water-demand equations is a better specification in this case, because it allows the analyst to identify substitutability and complementarity relationships between sources (Cheesman, Bennett, and Son 2008; Nauges and van den Berg 2009).

3. *A combination of the source choice model and a model of water use conditional upon source choice*. Several papers have studied household choice of water source, either as a primary focus (Mu, Whittington, and Briscoe 1990; Madanat and Humplick 1993; Hindman Persson 2002; Briand, Nauges, and Travers forthcoming) or in combination with estimations of conditional water-demand models (Larson, Minten, and Razafindralambo 2006; Nauges and Strand 2007; Basani, Isham, and Reilly 2008; Cheesman, Bennett, and Son 2008; Nauges and van den Berg 2009). Most of these studies were conducted in urban areas, very often in medium or large cities. Authors generally agree that source attributes (for example price, distance to the source, quality, and reliability) and household characteristics (income, education, size, and composition) should both enter the source choice model. Whereas source attributes account for heterogeneity in water from different sources, household characteristics account for differences in personal taste, opportunity cost of time, and perception of health benefits from improved water.[3]

The most frequent specifications for source choice models are the probit model and the multinomial logit (MNL) model. The probit model has been used when the household choice being modeled is whether or not to acquire a private connection (Larson, Minten, and Razafindralambo 2006; Basani, Isham, and Reilly 2008; Nauges and van den Berg 2009). The MNL model has proved useful for describing either the primary source of water chosen by households (Mu, Whittington, and Briscoe 1990; Nauges and Strand 2007) or the water source that is chosen for a specific use such as drinking, bathing, or cooking (Madanat and Humplick 1993; Hindman Persson 2002).[4] The MNL model considers choices between exclusive alternatives and relies on the assumption of independence from irrelevant alternatives. However, for modeling household choice of water sources so as to allow for a combination of sources, the multivariate probit or nested logit models should be the preferred alternative. In the multivariate probit setting, households are assumed to make several decisions, each between two alternatives. Briand, Nauges, and Travers (forthcoming), in a study of households living in 11 formal but poor districts in Dakar, Senegal (population around 1 million), estimated a bivariate probit model to describe household decisions to rely on a private connection, public standpipes, or both. The nested logit specification can be seen as a two- (or more) level choice problem (for more details on these models see Greene 2003, chapter 21; for recent approaches that may be useful in the present context see Bhat 2005).[5]

Models describing household choice of water source have recently been combined with conditional models of water demand. The simultaneity between choice of water source and choice of quantity was first acknowledged by Whittington, Briscoe, and Mu (1987), who argued that a complete set of water-demand relationships should include models of both water source choice and the quantity of water demanded. If both factors are not taken into account, the simultaneity in both decisions could lead to biased estimates of the demand parameters. In particular, if some unobserved variables affect both the choice of water source(s) and the quantity of water used, estimated parameters could suffer from a selection bias (Heckman 1979).

This issue has been discussed by several authors, and a two-step Heckman procedure for correcting selection bias has been applied by, among others, (i) Larson, Minten, and Razafindralambo (2006) on data from Fianarantsoa, Madagascar (population around 100,000); (ii) Nauges and Strand (2007) on data from Central American cities (Santa Ana, Sonsonate, and San Miguel in El Salvador, population between 65,000 and 200,000; and Tegucigalpa in Honduras, population of about 900,000); (iii) Basani, Isham, and Reilly (2008) on data from seven provincial towns in Cambodia; and (iv) Cheesman, Bennett, and Son (2008) on data from Buon Ma Thuot, in the Central Highlands of Vietnam (population around 135,000). Selectivity correction terms are computed from the estimation of the discrete choice models described above and added linearly to the demand equations. Statistical significance of these correction terms indicates the presence of a selectivity bias.[6]

These estimates of the household water-demand function have never been used to derive welfare measures, except by Cheesman, Bennett, and Son (2008), who derive the effects of quantity restrictions on the surpluses of Vietnamese households.[7] They find that consumer surplus losses from reduced total monthly household municipal water supplies are more pronounced among households that use only municipal water than among those that combine municipal water with well water. This is as expected, because the former group of households has a more inelastic own-price demand and a lack of substitution opportunities.

Welfare analysis following changes in the conditions of water supply for households in developing countries remains a difficult question, in particular when piped water is charged according to a block-pricing scheme and when households that are currently without a connection become able to connect to the piped network. In cases where block pricing is used, consistent estimation of water demand and calculation of the change in consumption following a change in price become computationally challenging (for details, see Olmstead, Hanemann, and Stavins 2007). The problem is that it is difficult for the researcher to assess demand for piped water among households that do not yet have a connection to the piped network. The assumption that such households, after being connected,

will behave the same as those that already are connected is likely to be too strong in most cases: there is evidence that a household's own characteristics drive the choice of both access to specific water sources and the quantity of water used.

The determinants of how total water consumption is allocated among different uses (drinking, cooking, bathing, and so on) is a question that has not yet been studied, so far as we know. This question is likely to be more relevant for developing countries, because water from different sources may be used for different purposes.

Estimation of Household Water-Demand Functions in Developing Countries: Data Issues

Analysts attempting to estimate household water-demand functions in developing countries face at least four difficult challenges when assembling data. First, households that are connected to piped-water networks may nevertheless have unmetered connections: thus no household-level data on the quantity of water used is available from the water utility. In such situations households themselves usually have little idea how much water they use, and direct interviews with them will be of no use in determining any exact or approximate quantity. In such a situation the main options open to the analyst are to install meters (which may change behavior), to monitor (directly watch) household use of water over some interval of time, or to ask the household to keep a detailed water-use diary.

Second, even when households have metered connections, the meter readings are often unreliable. Many piped-water systems in developing countries do not provide 24-hour service. When water service in a piped distribution system is intermittent, the water pressure fluctuates. Meters typically will not provide accurate readings because air intermittently enters the pipes, such that the meter may register water as passing through when in fact it is only air. Also, because water prices are so low in many places, and because corruption is common (Davis 2003), water utilities have little incentive to keep meters in good working order; nor are they replaced on a timely basis. The end result is that in many cases no one knows how much water a household is using—not the utility, not the household, and certainly not the researcher.

Third, when an analyst wants to model source choices made by households that have multiple potential sources, the model requires data not only on the water sources chosen but also on the sources *not* chosen. For example, a household's decision to purchase water from a vendor will depend not only on the price charged but also on how far household members would have to walk to fetch the water from, say, a well. The analyst would need to know the distance to the well

even if the household bought all of its water from a vendor. However, standardized household surveys that include questions about a household's water source generally ask the respondent only about the sources the household uses, not the attributes of the sources not chosen. Thus models of water source choices made by households, and discrete–continuous models, almost always require specially designed household surveys, even when utility records are available. Even the specially designed household surveys may need to be supplemented with additional data collection activities, because households may not be able to provide quantitative information on some attribute of the sources not chosen (for example distance from the dwelling to the source).

Fourth, because information on the quantity of water used is often not available (even from a utility) or is of poor quality, researchers have typically relied on cross-sectional surveys of households in the community under study. It is possible to use cross-sectional data in regression models to determine associations between the source chosen (and the quantity of water used) and covariates such as household income, housing type, education levels of household members, and the collection costs of water. It is often difficult, however, confidently to ascribe a causal relationship of the independent variables (the covariates) to the dependent variables (source chosen, quantity of water used) on the basis of analysis of cross-sectional data. Many of the independent variables are arguably endogenous, and good instruments for these are rarely available. For variation in key independent variables over time intervals, time series data are generally required.

Nevertheless, most researchers seeking to estimate household water-demand functions in developing countries have used data from cross-sectional household surveys. Occasional attempts have been made to escape the cross-sectional dilemma. For example, Cheesman, Bennett, and Son (2008) built an "artificial panel" data set by combining revealed and stated preference data. Diakite, Semenov, and Thomas (2009) use utility data for 156 small towns (all above 3000 inhabitants) in Côte d'Ivoire over the years 1998–2002.

In addition to these four data problems, researchers encounter challenges associated with measuring specific variables within the model specifications. We discuss some of these below.

Water Price

When data are obtained from one-time household surveys conducted in a single city or village, there may be little or even no cross-sectional variation in policy-relevant variables such as connection costs, tariff, and levels of service. In fact, Larson, Minten, and Razafindralambo (2006) exclude the price of water altogether from their analysis of water demand in Fianarantsoa, Madagascar, because all surveyed households had the same price schedule. One may attempt

to overcome this problem by combining revealed and stated preference data, that is, by asking respondents how much water they would use at different (hypothetical) prices (Acharya and Barbier 2002; Cheesman, Bennett, and Son 2008). However, respondents simply may not know how much water they would use if the prices for water proposed by the researcher are outside their experience.

For water from nonpiped sources, contingencies vary across places and sources. Water may be distributed free of charge, or perhaps it is charged at a fixed price per liter. If the surveyed households obtain water from various nontap sources, some cross-sectional variation will most likely be observed in the data. Because data on price (and consumption) for households relying on nonpiped sources are usually based on self-reported information (households are usually asked to report the number of buckets that they collect everyday), there is room for substantial error in measurement.

Costs of Water Collection

Even if water is available from a source away from home free of charge, its collection involves time to go to the source, to wait at the source (queuing), and time to haul the water back home. One may choose to convert collection time into collection costs using an assumed value of time. However, the value of time may differ widely across households depending on who is responsible for collecting water, and even within a specific household over time of day or day of week. In localities lacking formal labor markets or with high unemployment, estimating an average value of time for a study population is largely guesswork. Many analysts thus do not attempt to convert the time cost of water collection into a pecuniary collection cost. For example, Larson, Minten, and Razafindralambo (2006) consider round-trip walking time to water source and waiting time at the source. David and Inocencio (1998), on a sample from Metro Manila in the Philippines, use distance from source in meters as an explanatory variable in their demand model. Strand and Walker (2005) consider hauling time per unit of water consumed.

Whittington, Mu, and Roche (1990) are among the only authors to provide some empirical evidence about the pecuniary cost of collecting water from nontap sources. Using data from Ukunda, a small market town in Kenya, they develop two approaches, based on discrete choice theory, for estimating the value of time spent collecting water. Their results indicate that the value of time for households relying on nontap sources (kiosks, vendors, or open wells in the village) was at least 50 percent of the market wage rate and likely to approach that for unskilled labor for some households. These estimates were higher than the ones recommended by the Inter-American Development Bank at the time, which were that time savings should be valued at 50 percent of the market wage rate for unskilled labor (Whittington, Mu, and Roche 1990). Nevertheless, this small

study for a single community in Kenya cannot be easily generalized to other locations. Nauges and Strand (2007), on household data from Santa Ana, Sonsonate, and San Miguel in El Salvador, and Tegucigalpa in Honduras, have conducted the only study where hauling time is translated into a corresponding pecuniary time cost. They use the average hourly wage in the individual household as the shadow cost of time but acknowledge that even this approximation may overestimate actual costs if the hauling is performed by a child. Mu, Whittington, and Briscoe (1990) note that in places where queuing time varies significantly over the course of the day, collection time could be determined endogenously.

Quality of Water Service

Because water quality and reliability may vary from one source to another, such variables should be included in household water-demand functions for developing countries (as well as in models describing source choice). These include opinion variables about the taste, smell, and color of the water (at all available sources), the hours of water availability, and potential pressure problems (for piped water). These data are typically provided by households themselves and may be subject to misreporting. Variables measuring household opinion (or perception) about water quality should also be used with caution, because they may introduce endogeneity into the demand model. For example, households that suffered from water-related diseases in the past may be more inclined than other, healthy households to believe that water is unsafe and may therefore exhibit different behavior regarding water use (Nauges and van den Berg 2006). Also, quality perceptions may be correlated with income and education, implying collinearity issues (Whitehead 2006). To avoid such biases, one could develop an average of opinion (on water quality) for households living in the same neighborhood, or relying on the same water source, if the average could be computed without considering the opinion of the individual household (Briand, Nauges, and Travers forthcoming).

Socioeconomic Characteristics of Households

Household surveys often gather a large amount of information on socioeconomic and demographic characteristics, such as size and composition (by sex and age) of the household, the education level, occupation, and earnings of each member, as well as data on living conditions (structural materials, conditions of access to various services such as electricity, schooling, doctors, and so on). Income is one important variable in the study of water demand that may be difficult to gather in some places. Whittington, Mu, and Roche (1990) used several variables as income (or wealth) proxies, including the construction of the respondent's house

(whether the house was painted, whether the roof was straw or tin, whether the floor of the house was dirt or concrete). Basani, Isham, and Reilly (2008) use household expenditures as a proxy for income, arguing that in surveys households are more likely to understate their incomes than to overstate their expenditures. Another possible proxy approach would be to state wealth via a linear index of asset ownership indicators, using principal components analysis to derive weights (Filmer and Pritchett 2001).

Household Water Demand in Developing countries: Results

The studies reviewed in this article have used data from various regions in the world—Central America (El Salvador, Guatemala, Honduras, Nicaragua, Panama, Venezuela), Africa (Kenya, Madagascar), and Asia (Cambodia, Indonesia, the Philippines, Saudi Arabia, Sri Lanka, Vietnam)—and cover a 20-year time span (the earliest survey dates back to 1985; the most recent was conducted in 2006). Tables 1 and 2 summarize, respectively, the main characteristics of studies that model water source choice and water demand. They both include the author(s) of the research, number of households surveyed, study areas, time periods (when available), types of water access available to surveyed households, econometric approach used for model estimation, and main estimation results. With the exception of the research conducted in Ukunda, Kenya, these studies were conducted in medium- to large-sized cities in developing countries.

Water Consumption

The case studies described in the discussion throughout this article illustrate the heterogeneity of conditions for access to water across developing countries. The average water use by households with piped connections varies across places: 72 liters per capita per day (lpcd) in a group of seven provincial towns in Cambodia (Basani, Isham, and Reilly 2008), 88 lpcd in Fianarantsoa, Madagascar (Larson, Minten, and Razafindralambo 2006), 120 lpcd in Buon Ma Thuot, Vietnam (Cheesman, Bennett, and Son 2008), 130 lpcd in Salatiga City, Indonesia (Rietveld, Rouwendal, and Zwart 2000), and 135 lpcd in urban areas of medium cities from three districts in Southwest Sri Lanka, namely Gampaha, Kalutara, and Galle (Nauges and van den Berg 2009).[8]

Households without a piped connection have lower water consumption in general, with important differences depending on the source on which they rely. Households with a private well usually have a higher consumption level than households relying on public sources. In Santa Ana, Sonsonate, and San Miguel (El Salvador) and Tegucigalpa (Honduras), nonconnected households relying on

Table 1. Household Choice of Water Source: An Overview

Reference	*Household N, study region, period*	*Type of water access*	*Decision variable*	*Choice set*	*Model specification and estimation method*	*Significant explanatory variables*
Mu, Whittington, and Briscoe (1990)	69 hh from Ukunda (Kenya), 1986	Kiosks Water vendors Open wells Hand pumps	Choice of primary water source	Assumed to be exogenous: all hh face the same choice set	MNL ML	Collection time Price of water Number of women in the hh
Madanat and Humplick (1993)	900 hh from Faisalabad (Pakistan), 1992	Private connection Public piped water Motor/hand pumps Water vendors Wells/ponds/canals	Choice of usage-specific water source (for drinking, bathing, washing, etc.)	Not all the alternatives are available to all hh Focus here on hh that can connect to the water system	Two-level decision model Sequential ML estimation	Education level Presence of a storage tank Piped-water pressure level
Hindman Persson (2002)	769 hh in Cebu (Philippines)	Piped in house Pump in house or yard Rainwater Public pump or piped water Open well Surface water Purchased water	Choice of drinking water source	Suggests that the set of available sources is determined by choice of living areas	Suggests using a nested conditional logit but uses an MNL instead because of too few observations ML	Annual labor income Walking time to source
Larson, Minten, and Razafindralambo (2006)	547 hh from Fianarantsoa (Madagascar), 2000	Private connection Collecting hh	Decision to get or not to get a private connection	Assumed to be exogenous: all hh face the same choice set	Probit model ML	Education level Income

Nauges and Strand (2007)	553 nontap hh from three cities in El Salvador and 826 nontap hh from Tegucigalpa (Honduras), 1995–97	Private or public well Someone else's private tap Public tap Trucks Rivers/lakes	Choice of primary nonpiped water source	Assumed to be exogenous: all hh face the same choice set	MNL ML	Income Size of the property Access to electricity Household size Interviewee reads and writes
Basani, Isham, and Reilly (2008)	782 hh from seven provincial towns in Cambodia	Private connection Rivers/streams Tanks Wells Vendors	Decision to get or not to get a private connection	Assumed to be exogenous: all hh face the same choice set	Probit model ML	Connection fee Expenditure (as a proxy for income) Ethnic group
Briand, Nauges, and Travers (forthcoming)	301 hh from Dakar (Senegal), 2005	Private connection Public standposts	Decision to use a private connection and/or a public standpost	Assumed to be exogenous: all hh face the same choice set	Bivariate probit ML	Distance to standpost Household's head is a widow Interviewee reads and writes Average opinion on piped water reliability Average opinion about service at the standpost Renter/owner status
Nauges and van den Berg (2009)	1,800 hh from Sri Lanka, 2003	Private connection Public taps Public and private wells Neighbors Surface water	Decision to get or not to get a private connection	Assumed to be exogenous: all hh face the same choice set	Probit model ML	Income Education of hh's head Access to other sources Taste and reliability of water from other sources

Notes: hh: household; MNL: multinomial logit model; ML: maximum likelihood approach.

Table 2. Estimation of Water Demand in Developing Countries: An Overview

Reference	*Households N, study region, period*	*Type of water access*	*Demand model specification and estimation method*	*Dependent variable*	*Significant explanatory variables and estimated elasticities*
Mu, Whittington, and Briscoe (1990)	69 hh from Ukunda (Kenya), 1986	Kiosks Water vendors Open wells Hand pumps	Single demand equation with dummy to control for type of water access OLS method	Water used per capita per day	Collection time (−) Income (+)
Rizaiza (1991)	563 hh from four major cities in Saudi Arabia, 1985	Private connection Tankers	Separate demand equations for hh with a private connection and hh supplied with tankers OLS method	Annual water demand per household	Price elasticity ranging from −0.40 (for tankers water) to −0.78 (for piped water) Family size (+) Income elasticity in the range 0.09–0.20 Average temperature (+) Dummy for garden in the property (+)
Crane (1994)	291 hh from Jakarta (Indonesia), 1991	Piped system Water vendors Public hydrants Household resellers Neighbors with in-house connection	Separate demand equations for hh supplied by vendors and hh relying on hydrants plus dummy to control for use of extra sources OLS method	Household monthly water use	Price elasticity ranging from −0.48 (for vended water) to −0.60 (for hydrant water) Time per purchase (− for vended water) Capacity of water reservoir (+ for vended water)
David and Inocencio (1998)	506 hh from Manila (Philippines), 1995	Piped system Tubewell pumps Water vendors	Separate demand equations for hh supplied by vendors and for hh with a private connection 2SLS estimation for correction of price endogeneity	Household monthly water use	Price elasticity estimated at −2.1 for vended water Income elasticity estimated at 0.3 for vended water

Rietveld, Rouwendal, and Zwart (2000)	951 hh from Salatiga (Indonesia), 1994	Private connection Neighbors Community water terminal Wells Rivers	Single demand equation for water from a private connection Discrete–continuous approach of Burtless and Hausman (1978); ML method	Household monthly water use	Price elasticity −1.2 Income elasticity 0.05 Household size (+) Use of extra sources (−)
Strand and Walker (2005)	About 3,700 hh from 17 cities in Central America, 1995–98	Private connection Public taps Trucks Wells Rivers/lakes	Separate demand equations for hh with a private connection and for nontap hh 2SLS estimation for tap water equation and OLS estimation for nontap water equation	Household monthly water use	Price elasticity in the range from −0.3 (for hh with a private connection) to −0.1 (for nontap hh) Income elasticity less than 0.1 Household size (+) Hauling time (− for nontap water)
Larson, Minten, and Razafindralambo (2006)	547 hh from Fianarantsoa (Madagascar), 2000	Private connection Public taps Wells Natural sources	Separate demand equations for collecting hh and hh with private connections Two-step Heckman approach to control for use of a private connection	Household monthly water use	Household size (+) Roundtrip collection time (− for collecting hh)
Nauges and Strand (2007)	553 nontap hh from three cities in El Salvador and 826 nontap hh from Tegucigalpa (Honduras), 1995–97	Private or public well Someone else's private tap Public tap Trucks Rivers/lakes	Single demand equation for nontap water, allowing for elasticities to water cost varying with type of water access Two-step Heckman approach to control for choice of primary nontap source	Water use per capita per month	Total water cost (price + hauling cost) elasticity in the range from −0.4 to −0.7 Income elasticity in the range 0.2–0.3 Household size (−)

Continued

Table 2. *Continued*

Reference	*Households N, study region, period*	*Type of water access*	*Demand model specification and estimation method*	*Dependent variable*	*Significant explanatory variables and estimated elasticities*
Basani, Isham, and Reilly (2008)	782 hh from seven provincial towns in Cambodia	Private connection Rivers/streams Tanks Wells Vendors	Single demand equation for connected hh Two-step Heckman approach to control for use of a private connection	Household monthly water use	Price elasticity in the range from −0.5 to −0.4 (connected hh) Expenditure elasticity in the range 0.2–0.7 (connected hh)
Cheesman, Bennett, Son (2008)	166 hh from Buon Ma Thuot (Vietnam), 2006	Private connection Private wells Vendors	Separate estimation for hh using a private connection only (single equation) and hh combining water from a private connection and well water (system) Two-step Heckman approach to control for use of well water	Household monthly water use	Price elasticity for piped water estimated at −0.06 for hh using a private connection only and at −0.53 for hh using a private connection and well water Income elasticity 0.14 Household size (+) Use of a storage tank (+)
Nauges and van den Berg (2009)	1,800 hh from Sri Lanka, 2003	Private connection Public taps Public and private wells Neighbors Surface water	Separate systems of demand equations for piped and nonpiped hh Two-step Heckman approach to control for use of a private connection; Tobit model for censored observations	Water use per capita per month (for piped hh) or per day (for nonpiped hh)	Price elasticity in the range from −0.15 to −0.37 for piped hh Collection time (− for nonpiped water) Income elasticity 0.14 for piped hh and 0.20 for nonpiped hh Use of a storage tank (+) Hours of piped water availability (+ for piped water) Household size (− for nonpiped hh)

Notes: hh: household; OLS: ordinary least-squares; 2SLS: two-stage least-squares; ML: maximum likelihood approach.

public taps outside the home consume on average 25 lpcd, whereas those relying on a private well consume on average 110 lpcd (Nauges and Strand 2007). In Jakarta (Indonesia) nonconnected households that buy water from resellers purchase on average 27 lpcd, whereas those that buy water from vendors purchase 15 lpcd on average (Crane 1994).

Water Price

Despite heterogeneity in the places and time periods studied, authors seem to agree on the inelasticity of water demand in developing countries, with most estimates for households with a private connection in the range from −0.3 to −0.6. Espey, Espey, and Shaw (1997) report an average own-price elasticity of −0.51 from industrialized countries, suggesting that own-price elasticity for households in developed countries and for those in developing countries is in the same range. Only two studies from developing countries find evidence of an elastic water demand: David and Inocencio (1998) use data from Metro Manila, the Philippines, to estimate price elasticity for vended water at −2.1; Rietveld, Rouwendal, and Zwart (2000) use data from Jakarta, Indonesia, to estimate price elasticity for piped water at −1.2. Interestingly, Rietveld, Rouwendal, and Zwart (2000) are the only authors to use the discrete–continuous model first proposed by Burtless and Hausman (1978), which was first used for estimating household water demand by Hewitt and Hanemann (1995) in a study in Texas that yielded a price elasticity of −1.6, a figure above (in absolute value) most elasticities that had been estimated in developed countries.[9] In our opinion, the price elasticity estimated by David and Inocencio (1998) should be regarded with some caution, as alternative estimation techniques used on the same data (by the same authors) seem to provide very different price elasticities.

Nauges and van den Berg (2009) on data from three districts in Southwest Sri Lanka (Gampaha, Kalutara, and Galle) and Cheesman, Bennett, and Son (2008) on data from Buon Ma Thuot, Vietnam, estimate systems of water demand for households with private connections that also consume water from nonpiped sources. Both studies show that piped water and nonpiped water are used as substitutes and that households that rely solely on piped water are less sensitive to price changes than connected households that complement their piped-water consumption with water from a private well.[10]

Costs of Water Collection

Collection time and distance to the source are found to be significant drivers of household choice of water source(s) (Mu, Whittington, and Briscoe 1990, using data from Ukunda, Kenya; Hindman Persson 2002, using data from metropolitan

Cebu, the Philippines; Briand, Nauges, and Travers forthcoming, using data from Dakar, Senegal). They also have a significant negative effect on the quantity of water collected from nontap sources (Mu, Whittington, and Briscoe 1990; Strand and Walker 2005; Larson, Minten, and Razafindralambo 2006; Nauges and Strand 2007; Nauges and van den Berg 2009). With data from Santa Ana, Sonsonate, and San Miguel (El Salvador) and Tegucigalpa (Honduras), Nauges and Strand (2007) estimate elasticity to price and hauling cost to be in the range from −0.4 to −0.7.

Quality of Water Service

Choice of water source is found to be driven by piped-water pressure level (Madanat and Humplick 1993) and by opinions about the taste and reliability of water (Briand, Nauges, and Travers forthcoming; Nauges and van den Berg 2009). If service from a piped connection is available for longer hours, water use by connected households increases (Nauges and van den Berg 2009). However, the magnitude of the effect is found to be quite small: an extra hour of piped-water availability would increase per capita consumption of households in Sri Lanka (districts of Gampaha, Kalutara, and Galle) by 2 percent on average. Variables measuring household opinion about water quality are not found to be significant in household water-demand functions in general.

In response to deficiencies in the water supply system, households may invest in coping strategies; that is, they may incur fixed costs in the form of investments in alternate supply sources, storage facilities, or both (Pattanayak and others 2005). For example, a household may buy a storage tank in order to mitigate problems with reliability and pressure that may be associated with private house connections, or, if the household relies on well water, pumping equipment may be purchased.

A demand equation that controls for household use of a water storage tank or for tank capacity is featured in analyzes by Crane (1994), Cheesman, Bennett, and Son (2008), and Nauges and van den Berg (2009). Crane (1994) notes that the use of a storage tank (and its capacity) could be endogenously determined in the demand model, as the investment decision regarding the tank (and its capacity) was certainly codetermined with the expected need for water. Endogeneity may not be present if the investment decision was made a long time before the actual (observed) water purchase. Using data for urban households from three districts in Southwest Sri Lanka (Gampaha, Kalutara, and Galle), Nauges and van den Berg (2009) estimate that a storage tank in the house increases per capita (piped) consumption by 13 percent on average.

Household Socioeconomic Characteristics

Income (or expenditure) and education level (or the ability of the head of household to read and write) have been found to be positively associated with household choice of improved water source (Madanat and Humplick 1993; Hindman Persson 2002; Larson, Minten, and Razafindralambo 2006; Nauges and Strand 2007; Basani, Isham, and Reilly 2008; Nauges and van den Berg 2009; Briand, Nauges, and Travers forthcoming). Mu, Whittington, and Briscoe (1990) and Briand, Nauges, and Travers (forthcoming), using data from Ukunda (Kenya) and Dakar (Senegal), respectively, find evidence that household composition affects choice of water source. In Ukunda; households with more women were less likely to purchase from vendors (and more likely to rely on water from wells and kiosks), presumably because more people are available in the household unit to carry water. In Dakar, the probability that households used water from the piped system increased if the head of household was a widow.

In studies estimating water demand, income elasticity (or expenditure elasticity) is found to be quite low, most often in the range 0.1–0.3. Household size is found to be significant in most cases. When the dependent variable is total household consumption, larger households are found to have greater water use. When the dependent variable is per capita consumption, scale effects are confirmed: per capita consumption decreases with the number of members in the household. Using data from Buon Ma Thuot (Vietnam), Cheesman, Bennett, and Son (2008) found that doubling the number of permanent residents in the household increased household consumption from a piped network by approximately 50 percent.

Concluding Remarks

Our review of the emerging literature on household water-demand functions in developing countries suggests that estimates of own-price elasticity for water from private connections is in the range from −0.3 to −0.6 and that income elasticity is typically in the range 0.1–0.3, both close to what is usually reported for industrialized countries. These findings have three important implications for water utility managers in developing countries. First, in medium and large cities with significant numbers of middle-income households, tariff increases probably can be used to increase revenues in order to raise some of the funds needed to finance system improvements and expansion. Second, although demand for water from private connections is inelastic, tariff increases will induce a reduction in the quantity of water demanded, and thus can be an important component of a water-demand management program. Third, although the estimates of income

elasticities are relatively small, in countries that are experiencing high rates of economic growth, water utility managers should anticipate powerful upward pressures on household water demand from increases in income. In locations where the marginal costs of water supply are increasing, this reinforces the need to use tariff increases to manage demand better.

In contrast, the literature on household water source choice, especially in rural areas, is still in its infancy, and in our judgment the empirical findings are much less robust. The explanatory variables suggested by economic theory are, in fact, associated with household water source choices and are often statistically significant and have the expected signs. However, the magnitude of the parameter estimates seems to us quite location specific, and so the policy implications are less clear. We speculate that further research will show that in many circumstances water source choices made by households will be quite sensitive to changes in prices of water from different sources and household incomes, in contrast to the findings from the literature on the quantity of water demanded by households with private connections living in medium to large cities. Programs designed to recover operation and maintenance costs, and some capital costs, thus may have significant effects on the use of new water infrastructures by households, especially in rural areas, a conclusion reached by Kremer and Miguel (2007) for households in their study villages in rural Kenya. This suggests that better demand information on source choices by households is needed in many rural areas and urban slums for designing subsidies and tariffs. This should be a high-priority research area.

Because so many people in developing countries lack improved water supply services, public health officials and many donors have little patience with economists' arguments about the optimal allocation of investment funds and the need to subject water supply projects to cost–benefit analysis. It seems obvious that people are dying from diseases that could be largely eliminated by improved water and sanitation services. Improvements are thus needed urgently, and if subsidies are necessary, so be it. If the water policy discourse is framed in this manner, information about how customers respond to different service options and pricing schemes is not likely to be a high priority to decision-makers.

Because many water utilities in developing countries have few incentives to undertake careful economic appraisals of investment projects, or to price the delivery of water to their customers in order to recover costs or meet an economic efficiency objective, water utility managers have not placed a high priority on obtaining better information on the water-demand behavior of households. Few water utilities in developing countries are financially self-sufficient; most receive capital, and in many cases even operating subsidies, from governments and donors. Their focus is naturally on the providers of these subsidies.

However, there are reasons why this situation may soon change, and the findings from the literature on household water-demand functions may become more policy relevant. At the macrolevel, economic growth and the increased hydrologic variability brought about by climate change are placing new pressures on the water sources used by all three groups of households described in this paper. As variability in the raw water supplies increases, providing reliable supply to households becomes more expensive. Governments throughout the world are also facing increasing challenges over allocation of water resources among different users. Both climate change and intersectoral competition for water make demand management increasingly important, and thus reinforce the need for a better understanding of household demand for improved water services in different circumstances.

There have, however, been exceptions to policymakers' general neglect of this literature on household demand for improved water services. First, there is an increasing recognition that existing water and sanitation tariff structures are not achieving their stated equity objectives, and that subsidies to the water sector are not reaching the poor (Komives and others 2005; Boland and Whittington 2000). This has led to a new willingness on the part of some utilities to experiment with different water tariff structures, which leads naturally to a consideration of how consumers will respond to changing prices and incomes.

Second, there is a growing appreciation among water utility managers that water pricing decisions regarding public taps and private connections need to be coordinated. This has been due in part to the findings from the literature reviewed in this paper. In some cases demand studies have suggested that water from public taps can be provided free because this policy will not affect demand for water from private connections (World Bank Water Demand Research Team 1993). In other locations this is not the case, and information on household demand is needed to avoid the serious policy mistakes that can arise from pursuing independent, uncoordinated pricing strategies (Whittington, Davis, and McClelland 1998).

Third, in many cities in developing countries, water utility managers are increasingly recognizing the competition they face from water vendors (Whittington, Lauria, and Mu 1991). Utility managers that are losing sales and market share to water vendors may wonder what attributes of the services of water vendors households prefer, and what it would take to get households to connect to the water utility's distribution system. This financial interest in increased revenues leads utility managers to the water-demand literature reviewed in this paper.

Fourth, numerous international organizations, including the Gates Foundation, have recently focussed their attention on the need to improve the quality of drinking water provided to households in developing countries. There has been

renewed interest in point-of-use technologies to treat water in the home. This has also raised questions about household demand for improved drinking water quality; how much households value different attributes of service, and the costs and benefits of different point-of-use technologies (Whittington and others 2009). Again, these policy issues are generating new interest in the water-demand literature in developing countries.

Two important questions about household water-demand behavior in developing countries remain unanswered, or simply cannot be addressed with the existing data. First, existing data do not permit measurement of how household water use would respond to the establishment of dual networks (one for drinking and cooking water, the other for uses that do not require high-quality water). Analyzes of household allocation of water among various uses could be a first step. Second, welfare analysis following changes in the conditions of water supply for households in developing countries remains a difficult question, in particular when piped water is charged according to a block-pricing scheme and when scenarios involve the connection of currently nonconnected households to the piped network.

Appendix: Data Collection Issues

Our overview in this paper of empirical issues has shown that careful analysis of household water demand in developing countries requires gathering a great deal of information from each household surveyed. The prudent researcher will bear in mind the following seven issues when designing a household survey on water use:

1. Surveys should ideally be made in more than a single city or village, in order to acquire data with cross-sectional variation regarding conditions of water services, in particular price, connection fee, and quality and reliability of services.
2. In most cases only data on sources that are actually used by the surveyed household are available. Ideally one should identify the complete set of sources available to the household (whether used or not) and gather information on the time to walk from home to any off-site source(s) used or not used, the waiting (queuing) time at the source(s), price of the water, possible rationing or constraints (opening hours, limited availability), and quality of the water from each source (whether used or not). These considerations are a prerequisite for consistent estimation of household choice of water sources.
3. For households relying on nonpiped sources, information on the persons in charge of collecting the water should be gathered, so that appropriate wage rates can be applied for estimating the shadow price of hauling.

4. At the time of the survey, interviewers should test each household about its knowledge of consumption and water expenditure during the last piped water billing period and of the pricing scheme.
5. It may be important to control for demand seasonality, because demand (in total and for water by source) may vary over the course of a year.
6. For planning it may also be important to control for number of permanent and nonpermanent household members.
7. To determine whether water infrastructure (storage tank, pumping equipment) is endogenous, that is, whether current household water usage might be linked to the acquisition of new infrastructure, installation dates can be recorded to serve as a measure of how recently these were purchased.

Notes

Céline Nauges (corresponding author) is a senior researcher at the National Institute for Research in Agriculture (INRA) and at the Toulouse School of Economics, France; email address: celine.nauges@toulouse.inra.fr. Dale Whittington is a professor in the departments of Environmental Sciences and Engineering, and City and Regional Planning, at the University of North Carolina at Chapel Hill (USA), and is also a professor at the Manchester Business School, University of Manchester (UK); email address: Dale_Whittington@unc.edu. The authors gratefully acknowledge comments from the editor and three anonymous referees, Jeremy Cheesman, Sylvestre Gaudin, and participants at the seminar organized by the Crawford School of Economics and Government (at The Australian National University, Canberra).

1. Most analyzes made in industrialized countries have been based on aggregate consumption data provided by water utilities (usually from billing records).

2. When working on data from industrialized countries, authors commonly assume that the water-demand function derives from the maximization of a household's utility, subject to a budget constraint, under the assumption that water is a homogeneous good that has no direct substitute or complement. In developing countries, the underlying theoretical model is described slightly differently: water demand is usually assumed to derive from a model in which the household is considered a joint production and consumption unit (for a description of such demand models see Behrman and Deolalikar 1988). In such models, the demand for water can be regarded as a derived input demand in the production of household health (because water consumption may have health consequences). As a result, health enters a household's utility, along with consumption goods, leisure time, and other household characteristics such as education. This preference function is then maximized subject to a time–income constraint and a set of production functions. For related discussions see Acharya and Barbier (2002) and Larson, Minten, and Razafindralambo (2006).

3. Kremer and others (2007) use data on household water source choices (in a travel cost model) to estimate a revealed preference measure of household valuation of the water quality gains generated by spring protection in rural Kenya.

4. Hindman Persson (2002) estimates a probit model but points out that a nested conditional logit model would be better suited. Madanat and Humplick (1993) estimate a two-level sequential choice model to distinguish between the decision to obtain a private connection and the choice of nontap sources.

5. Analysis of source choice may be complicated insofar as the entire set of sources available to the household is not known to the econometrician. Hindman Persson (2002) assumes that each household's location within the city determines its set of available sources.

6. For computation of correction terms when a probit is used in the first estimation stage, see Heckman (1979); for computation when an MNL is used, see Lee (1983) and Dubin and McFadden (1984).

7. Cheesman, Bennett, and Son (2008) employed a combination of revealed and stated preference techniques. These authors asked households how much water they would use if its price changed. They find that the own-price elasticity for household water use is extremely low (−0.059). This estimate needs to be interpreted carefully in the context of the local water situation in their study area in Vietnam. The local water utility was only charging about US$0.15 per cubic meter. It is thus not surprising that households would say that they would not change the amount of water they would use if the price doubled because the volumetric rate would still be very cheap. It seems to us implausible that the own-price elasticity is −0.059 throughout the range of hypothetical prices offered to respondents. Such extremely inelastic demand might be plausible at much lower levels of water use per capita. However, per capita water use in the study area was relatively high. A typical Vietnamese household in the sample without a private well was using about 16 cubic meters per month—roughly equivalent to household water use in many European cities. To see how odd these results are, consider the estimates of gross surplus losses. A typical Vietnamese household in the sample was paying a water bill of about US$2.25. Cheesman, Bennett, and Son's welfare calculations suggest that this household would be willing to pay about $8 to avoid having a supply restriction imposed of 3.5 cubic meters per month (from 16 to 12.5 cubic meters per month). In other words, the household would be indifferent with regard to paying US$2.25 for 12.5 cubic meters per month or paying $10.25 for 16 cubic meters per month. We doubt that these Vietnamese households actually place such a high value on modest supply restrictions given that their income is quite modest.

8. The European average is 150 lpcd; see European Environment Agency at http://themes.eea.europa.eu/Specific_media/water/indicators/WQ02e%2C2003.1001.

9. Meta-analyzes conducted on data from industrialized countries provide mixed evidence on the effect of functional form and estimation method on the level of price elasticities. Espey, Espey, and Shaw (1997), in a meta-analysis of 124 price elasticity estimates generated between 1963 and 1993, find no significant effect either of the functional form (linear versus log-linear) or of the estimation method (OLS versus others). Dalhuisen and others (2003), who extended Espey, Espey, and Shaw's database up to the year 2000, found that discrete–continuous choice (DCC) models (see Hewitt and Hanemann 1995; Rietveld, Rouwendal, and Zwart 2000) produced price elasticities that were significantly higher (in absolute value) than elasticities obtained using other approaches. However, this finding is weakened by the recent study of Olmstead, Hanemann, and Stavins (2007): using a DCC model on household data from 11 urban areas in the United States and Canada they find a moderate price elasticity of −0.33. The number of studies using the DCC model is too small to be able to draw any definite conclusion on the link between functional form, estimation method, and price elasticity.

10. Nauges and van den Berg (2009) use the approach introduced by Shonkwiler and Yen (1999) to control for censoring of observations in a system of simultaneous equations. This is because not all piped households complement their tap water consumption with water collected from a private well.

References

Acharya, G., and E. Barbier. 2002. "Using Domestic Water Analysis to Value Groundwater Recharge in the Hadejia-Jama'are Floodplain, Northern Nigeria." *American Journal of Agricultural Economics* 84:415–26.

Agthe, D.E., and R.B. Billings. 1980. "Dynamic Models of Residential Water Demand." *Water Resources Research* 16(3):476–80.

Agthe, D.E., R.B. Billings, J.L. Dobra, and K. Raffiee. 1986. "A Simultaneous Equation Demand Model for Block Rates." *Water Resources Research* 22(1):1–4.

Arbués-Gracia, F., M.A. García-Valiñas, and R. Martínez-Espiñeira. 2003. "Estimation of Residential Water Demand: A State of the Art Review." *Journal of Socio-Economics* 32(1):81–102.

Basani, M., J. Isham, and B. Reilly. 2008. "The Determinants of Water Connection and Water Consumption: Empirical Evidence from a Cambodian Household Survey." *World Development* 36(5):953–68.

Behrman, J.R., and A.B. Deolalikar. 1988. "Health and Nutrition." In J. Behrman, and T.N. Srinivasan eds., *Handbook of Development Economics*. Amsterdam: Elsevier Science, Vol. 1, pp. 633–711.

Bhat, C.R. 2005. "A Multiple Discrete–Continuous Extreme Value Model: Formulation and Application to Discretionary Time-use Decisions." *Transportation Research* 39(8):679–707.

Boland, J., and D. Whittington. 2000. "The Political Economy of Increasing Block Water Tariffs in Developing Countries." In Ariel Dinar ed., *The Political Economy of Water Pricing Reforms*. Oxford: Oxford University Press, pp. 215–36.

Briand, A., C. Nauges, and M. Travers. Forthcoming "Choix d'approvisionnement en eau des ménages de Dakar: Une étude économétrique à partir de données d'enquête." *Revue d'Economie du Développement*.

Burtless, G., and J. Hausman. 1978. "The Effect of Taxation on Labour Supply: Evaluating the Gary Income Maintenance Experiment." *Journal of Political Economy* 86(1):103–30.

Cheesman, J., J. Bennett, and T.V.H. Son. 2008. "Estimating Household Water Demand Using Revealed And Contingent Behaviors: Evidence from Viet Nam." *Water Resources Research* 44, W11428, doi:10.1029/2007WR006265.

Chicoine, D.L., S.C. Deller, and G. Ramamurthy. 1986. "Water Demand Estimation under Block Rate Pricing: A Simultaneous Equation Approach." *Water Resources Research* 22(6):859–63.

Crane, R. 1994. "Water Markets, Market Reform and the Urban Poor: Results from Jakarta, Indonesia." *World Development* 22(1):71–83.

Dalhuisen, J.M., R. Florax, H. De Groot, and P. Nijkamp. 2003. "Price and Income Elasticities of Residential Water Demand: A Meta-Analysis." *Land Economics* 79(2):292–308.

Daniere, A. 1994. "Estimating Willingness to Pay for Housing Attributes: An Application to Cairo and Manila." *Regional Science and Urban Economics* 24:577–99.

David, C.C., and A.B. Inocencio. 1998. "Understanding Household Demand for Water: The Metro Manila Case." Research Report, EEPSEA, Economy and Environment Program for South East Asia (http://web.idrc.ca/en/ev-8441-201-1-DO_TOPIC.html).

Davis, J. 2003. "Corruption in Public Services: Experience from South Asia's Water and Sanitation Sector." *World Development* 32(1):53–71.

Deller, S.C., D.L. Chicoine, and G. Ramamurthy. 1986. "Instrumental Variables Approach to Rural Water Service Demand." *Southern Economic Journal* 53(2):333–46.

Diakite, D., A. Semenov, and A. Thomas. 2009. "A Proposal for Social Pricing of Water Supply in Côte d'Ivoire." *Journal of Development Economics* 88(2):258–68.

Dubin, J.A., and D.L. McFadden. 1984. "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption." *Econometrica* 52:345–62.

Espey, M., J. Espey, and W.D. Shaw. 1997. "Price Elasticity of Residential Demand for Water: A Meta-analysis." *Water Resources Research* 33(6):1369–74.

Filmer, D., and L.H. Pritchett. 2001. "Estimating Wealth Effects without Expenditure Data—or Tears: An Application to Educational Enrollments in States of India." *Demography* 38(1):115–32.

Foster, H.S.J., and B.R. Beattie. 1979. "Urban Residential Demand for Water in the United States." *Land Economics* 55(1):43–58.

Gaudin, S., R.C. Griffin, and R. Sickles. 2001. "Demand Specification for Municipal Water Management: Evaluation of the Stone-Geary Form." *Land Economics* 77(3):399–422.

Gottlieb, M. 1963. "Urban Domestic Demand for Water in the United States." *Land Economics* 39(2): 204–10.

Grafton, R.Q., and M. Ward. 2008. "Prices versus Rationing: Marshallian Surplus and Mandatory Water Restrictions." *The Economic Record* 84:57–65.

Greene, W.H. 2003. *Econometric Analysis*, 5th edn. Upper Saddle River, NJ: Prentice Hall.

Griffin, R.C., and C. Chang. 1991. "Seasonality in Community Water Demand." *Western Journal of Agricultural Economics* 16(2):207–17.

Hanemann, W.M. 1998. "Determinants of Urban Water Use." In Duane Baumann, John Boland, and Michael Hanemann eds., *Urban Water Demand Management and Planning*. New York: McGraw Hill, pp. 31–75.

Hansen, L.G. 1996. "Water and Energy Price Impacts on Residential Water Demand in Copenhagen." *Land Economics* 72(1):66–79.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47:153–61.

Hewitt, J.A., and W.M. Hanemann. 1995. "A Discrete/Continuous Approach to Residential Water Demand under Block Rate Pricing." *Land Economics* 71:173–92.

Hindman Persson, T.H. 2002. "Household Choice of Drinking-water Source in the Philippines." *Asian Economic Journal* 16(4):303–16.

Höglund, L. 1999. "Household Demand for Water in Sweden, with Implications of a Potential Tax on Water Use." *Water Resources Research* 35(12):3853–63.

Howe, C.W., and F.P. Linaweaver. 1967. "The Impact of Price on Residential Water Demand and Its Relationship to System Design and Price Structure." *Water Resources Research* 3(1):13–32.

Hubbell, L.K. 1977. "The Residential Demand for Water and Sewerage Service in Developing Countries: A Case Study of Nairobi." Urban Regional Report 77-14, Development Economics Department, World Bank, Washington, D.C.

Katzman, M. 1977. "Income and Price Elasticities of Demand for Water in Developing countries." *Water Resources Bulletin* 13:47–55.

Komives, K. 2003. "Infrastructure, Property Values, and Housing Choice: An Application of Property Value Models in the Developing Country Context." Ph.D. diss. University of North Carolina–Chapel Hill, Department of City and Regional Planning.

Komives, K., V. Foster, J. Halpern, and Q. Wodon. 2005. *Water, Electricity, and the Poor: Who Benefits from Utility Subsidies*. Washington, D.C.: World Bank.

Kremer, M., and E. Miguel. 2007. "The Illusion of Sustainability." *Quarterly Journal of Economics* 122(3):1007–65.

Kremer, M., J. Leino, E. Miguel, and A. Zwane. 2007. "Spring Cleaning: A Randomized Evaluation of Source Water Quality Improvement." Working Paper. UC Berkeley. (http://www.cepr.org/MEETS/WKCN/7/770/papers/Miguel.pdf).

Kremer, M., C. Null, E. Miguel, and A. Zwane. 2008. "Diffusion of Chlorine Drinking Water Treatment in Kenya." UC Berkeley, Working Paper. (http://are.berkeley.edu/~clair/Clair_Null_2nd_paper.pdf).

Kulshreshtha, S.N. 1996. "Residential Water Demand in Saskatchewan Communities: Role Played by Block Pricing System in Water Conservation." *Canadian Water Resources Journal* 21(2): 139–55.

Larson, B., B. Minten, and R. Razafindralambo. 2006. "Unravelling the Linkages between the Millennium Development Goals for Poverty, Education, Access to Water and Household Water Use in Developing countries: Evidence from Madagascar." *Journal of Development Studies* 42(1): 22–40.

Lee, L.F. 1983. "Generalized Econometric Models with Selectivity." *Econometrica* 51:507–12.

Madanat, S., and F. Humplick. 1993. "A Model of Household Choice of Water Supply Systems in Developing countries." *Water Resources Research* 29:1353–8.

Martínez-Espiñeira, R. 2002. "Residential Water Demand in the Northwest of Spain." *Environmental and Resource Economics* 21(2):161–87.

Mu, X., D. Whittington, and J. Briscoe. 1990. "Modeling Village Water Demand Behavior: A Discrete Choice Approach." *Water Resources Research* 26(4):521–9.

Nauges, C., and J. Strand. 2007. "Estimation of Non-tap Water Demand in Central American Cities." *Resource and Energy Economics* 29:165–82.

Nauges, C., and A. Thomas. 2000. "Privately-operated Water Utilities, Municipal Price Negotiation, and Estimation of Residential Water Demand: The Case of France." *Land Economics* 76(1):68–85.

Nauges, C., and C. van den Berg. 2006. "Perception of Health Risk and Averting Behavior: An Analysis of Household Water Consumption in Southwest Sri Lanka." Working Paper 08.09.253. LERNA, Toulouse, France.

_____. 2009. "Demand for Piped and Non-piped Water Supply Services: Evidence from Southwest Sri Lanka." *Environmental and Resource Economics* 42(4):535–49.

Nieswiadomy, M.L., and D.J. Molina. 1989. "Comparing Residential Water Demand Estimates under Decreasing and Increasing Block Rates Using Household Data." *Land Economics* 65(3):280–9.

North, J.H., and C.C. Griffin. 1993. "Water Source as a Housing Characteristic: Hedonic Property Valuation and Willingness to Pay for Water." *Water Resources Research* 29:1923–9.

Olmstead, S.M., W.M. Hanemann, and R.N. Stavins. 2007. "Water Demand under Alternative Price Structures." *Journal of Environmental and Economics Management* 54:181–98.

Pattanayak, S., J.C. Yang, D. Whittington, and B. Kumar. 2005. "Coping with Unreliable Public Water Supplies: Averting Expenditures by Households in Kathmandu, Nepal." *Water Resources Research* 41, W02012, doi:10.1029/2003WR002443.

Pint, E. 1999. "Household Responses to Increased Water Rates during the California Drought." *Land Economics* 75(2):246–66.

Renwick, M.E., and R. Green. 2000. "Do Residential Water Demand Side Management Policies Measure Up? An Analysis of Eight California Water Agencies." *Journal of Environmental Economics and Management* 40(1):37–55.

Rietveld, P., J. Rouwendal, and B. Zwart. 2000. "Block Rate Pricing of Water in Indonesia: An Analysis of Welfare Effects." *Bulletin of Indonesian Economic Studies* 36(3):73–92.

Rizaiza, O. 1991. "Residential Water Usage: A Case Study of the Major Cities of the Western Region of Saudi Arabia." *Water Resources Research* 27(5):667–71.

Shonkwiler, J.S., and S.Y. Yen. 1999. "Two-Step Estimation of a Censored System of Equations." *American Journal of Agricultural Economics* 81:972–82.

Strand, J., and I. Walker. 2005. "Water Markets and Demand in Central American Cities." *Environment and Development Economics* 10(3):313–35.

Therkildsen, O. (1988). *Watering White Elephants: Lessons from Donor-Funded Planning and Implementation of Rural Water Supplies in Tanzania*. Uppsala: Scandinavian Institute of African Studies.

White, G., D. Bradley, and A. White. 1972. *Drawers of Water: Domestic Water Use in East Africa*. Chicago: University of Chicago Press.

Whitehead, J.C. 2006. "Improving Willingness to Pay Estimates for Quality Improvements through Joint Estimation with Quality Perceptions." *Southern Economic Journal* 73(1):100–11.

Whittington, D., J. Briscoe, and X. Mu. 1987. "Willingness to Pay for Water in Rural Areas: Methodological Approaches and an Application in Haiti." Field Report 213, September. Water and Sanitation for Health Project, U.S. Agency for International Development, Washington, D.C.

Whittington, D., J. Davis, and E. McClelland. 1998. "Implementing a Demand-driven Approach to Community Water Supply Planning: A Case Study of Lugazi, Uganda." *Water International* 23: 134–45.

Whittington, D., X. Mu, and R. Roche. 1990. "Calculating the Value of Time Spent Collecting Water: Some Estimates for Ukunda, Kenya." *World Development* 18(2):226–80.

Whittington, D., D.T. Lauria, and X. Mu. 1991. "A Study of Water Vending and Willingness to Pay for Water in Onitsha, Nigeria." *World Development* 19 (2/3):179–98.

Whittington, D., J. Briscoe, X. Mu, and W. Barron. 1990. "Estimating the Willingness to Pay for Water Services in Developing countries: A Case Study of the Use of Contingent Valuation Surveys in Southern Haiti." *Economic Development and Cultural Change* 38(2):293–311.

Whittington, D., S.K. Pattanayak, Y. Jui-Chen, and K.C. Bal Kumar. 2002. "Household Demand for Improved Piped Water Services: Evidence from Kathmandu, Nepal." *Water Policy* 4(6):531–56.

Whittington, D., J. Davis, L. Prokopy, K. Komives, R. Thorsten, H. Lukacs, W. Wakeman, and A. Bakalian. 2009. "How Well is the Demand-Driven, Community Management Model for Rural Water Supply Systems Doing? Evidence from Bolivia, Peru, and Ghana." *Water Policy* 11(6).

World Bank Water Demand Research Team. 1993. "The Demand for Water in Rural Areas: Determinants and Policy Implications." *The World Bank Research Observer* 8(1):47–70.

To Mitigate or to Adapt: Is that the Question? Observations on an Appropriate Response to the Climate Change Challenge to Development Strategies

Zmarak Shalizi • Franck Lecocq

*Climate change is a new and important challenge to development strategies. In light of the current literature a framework for assessing responses to this challenge is provided. The presence of climate change makes it necessary to at least review development strategies—even in apparently nonclimate-sensitive and nonpolluting sectors. There is a need for an integrated portfolio of actions ranging from avoiding emissions (mitigation) to coping with impacts (adaptation) and to consciously accepting residual damages. Proactive (*ex ante*) adaptation is critical, but subject to risks of regrets when the magnitude or location of damages is uncertain. Uncertainty on location favors nonsite-specific actions, or reactive (*ex post*) adaptation. However, some irreversible losses cannot be compensated for. Thus, mitigation might be in many cases the cheapest long-term solution to climate change problems and the most important to avoid thresholds that may trigger truly catastrophic consequences. To limit the risks that budget constraints prevent developing countries from financing reactive adaptation—especially since climate shocks might erode the fiscal base—"rainy-day funds" may have to be developed within countries and at the global level for transfer purposes. Finally, more research is required on the impacts of climate change, on modeling the interrelations between mitigation and adaptation, and on operationalizing the framework.* JEL codes: O10, Q54, Q56

Until recently, policymakers and development experts could at least assume that where there was water today, there would still be water in the future. Or that

The World Bank Research Observer

doi;10.1093/wbro/lkp012 Advance Access publication September 2, 2009 25:295–321

where there was a coastline suitable for a port, that coastline would still be there in the future. In other words, the geographical and physical foundations for development, and for the determination of competitive advantage, were treated as stable and reliable.

This presumption is no longer true, as climate change threatens to bring about important shifts in precipitation and weather patterns, sea levels, and water flows (IPCC 2007a), ratcheting up pressure on the land and on ecosystems (IPCC 2007b), thereby making previously stable parameters less stable. In fact the fingerprints of climate change are increasingly evident in changing weather patterns across the globe. From a developing country perspective, climate change is thus yet another important (and new) challenge within which development takes place. It may bring new opportunities, but also many constraints and risks. As such, development and growth will be taking place against a new headwind.

The key question for developing countries is the extent to which this new challenge—climate change—will modify the allocation of resources in development strategies. For a variety of reasons, the international debate on climate change is currently framed around "reducing emissions", that is, *mitigation*. Since we all emit greenhouse gases (GHG) into a shared atmosphere, protecting the quality of the atmosphere as a global public good requires global collective action. Developed countries have taken some (modest) emission reduction commitments, and developing countries are now under pressure to commit to emissions targets as well. Developing countries have resisted this call on the grounds that they have contributed little to the current concentrations of GHG in the atmosphere and that they cannot afford to increase the cost of development given their poverty and low standards of living. Prominent on their agenda, in fact, is "making their development strategies robust to climate change", that is, *adaptation*.

The present paper addresses the following questions within a framework for assessing responses to the challenge that climate change poses to development strategies. Is it worth revising development strategies for climate change? How can one think about the balance between mitigation and adaptation? And how might the balance in this portfolio be affected by the uncertainties about the extent and location of climate damages?

The paper is organized as follows. First, we provide a brief review of the literature on the risks that climate change poses to economic development, and the options available to respond. Second, we argue for *an integrated portfolio of actions*, including both mitigation of and adaptation to climate change (for both developed and developing countries)—because mitigation and adaptation are not entirely substitutable. We also review the literature on how to balance this portfolio given the uncertainties about climate damages. Third, we discuss why an integrated portfolio remains necessary despite criticisms that mitigation is a low priority for

development and that there is no public sector role in adaptation. Fourth, we argue for a broad review of development strategies. Fifth, we discuss some of the limitations of the current literature on the topic and we identify areas for further research.

Setting The Stage: Evaluating The Risks To Economic Development And Identifying The Options

The causal chain linking economic behavior today to economic consequences tomorrow *via climate change* can be summarized as follows:

Economic activities → GHG emissions → Atmospheric concentrations of GHG → Climate change → Impacts on physical and ecological systems → Impacts on economies and human welfare

Climate Change Threatens Economic Development

Climate change matters to policymaking only if its potential ultimate damages are expected to be significant, that is only if the expected laissez-faire or business-as-usual scenario *with* climate change differs markedly from the expected laissez-faire scenario *without* climate change. A converging set of data and projections suggest that such a significant difference is in fact likely (IPCC 2007a, 2007b). The ultimate damages of climate change arise from both gradual and rapid changes in climatic averages (mean precipitation, temperature, and so on), as well as increases in the variance, frequency, or magnitude of climatic shocks—all of which have potentially significant economic implications.

The estimates of the economic costs associated with climate change impacts focus mostly on gradual changes. Early figures in the mid-1990s estimated these costs at around 1.5–2.0 percent of World GDP in 2100 for temperature increases of 2–3 degrees Celsius by 2100 (see Pearce and others 1996 for a summary). New analysis in the early 2000s found generally lower costs (Mendelsohn and others 2000; Nordhaus and Boyer 2000; Tol 2002a, 2002b). But Stern (2007) has recently reported costs in the range of a 5–20 percent annual equivalent loss of World GDP from now up to 2200. The changing estimates reflect improvements in data and methods over time (for example, the inclusion of risks including catastrophes), evolution of scientific views about certain aspects of climate change (for example, mean temperature increase in 2100 for a given emissions path), and differences in the choice of key parameters (for example, the value of the pure rate of time preference). Despite the dispersion of the results, there is an emerging consensus that climate change will have a net negative impact on

developing countries, that the aggregate impacts of climate change on economic growth can be significant in individual countries (Lecocq and Shalizi 2007b), and that the impact on specific resources can be high, generating additional tensions where resource availability is already an issue (see, for example, Kundzewicz and others 2007, for freshwater).

The development literature shows that climatic shocks (not necessarily due to climate change) have already had large impacts on economic growth in many countries, such as Madagascar, Bangladesh, or Central American countries (IMF 2003), and play a significant role in explaining cross-country economic differences (Easterly and others 1993; Collier and Dehn 2001). This suggests that tomorrow's climate-change-induced climatic shocks, which are likely to be larger and more frequent than today's, may affect economic growth further, within the same countries, their neighbors, as well as in others with similar characteristics (Martin and Bargawi 2004). Though there is no empirical or theoretical consensus on the key mechanisms through which climate shocks have such large impacts on growth, a number of factors—such as the size of climate-sensitive sectors (for example, agriculture or tourism), the indirect impacts on nonclimate-sensitive sectors, rigidities in factor allocation and in price adjustments (Hallegatte, Hourcade, and Dumas 2007), and the strength of institutions and the cohesiveness of society (Rodrick 1999)—appear to play an important role.

An increase in the frequency and magnitude of shocks due to climate change would also increase the chance of countries falling into "poverty traps" or would reduce their chances of getting out of them (as a result of path-dependent multiple equilibria combined with stochastic shocks; Azariadis and Stachurski 2004). Similarly, increasing returns to agglomeration (Fujita, Krugman, and Venables 1999) can magnify the national or global consequences for economic growth of localized impacts of climate change on key localities (such as major[1] or coastal[2] cities that constitute national engines of growth) (Huq and others 2007).

The analysis above thus suggests that climate change is an important enough risk to development that it warrants a response at the national and international level. The rest of the paper aims at providing some views on what that response should be.

The Response Options: Mitigation, Proactive Adaptation, Reactive Adaptation, and Accepting Residual Damages

To limit the impacts of climate change on economies, countries can *mitigate* emissions or *adapt* to climate change consequences. Mitigation consists of reducing emissions (or removing GHG out of the atmosphere) at the beginning of the chain to minimize climate change in the first place. By contrast, adaptation consists of responding to climate change impacts at the end of the chain.

For example, shifting from coal- to gas-fired power plants (thereby reducing GHG emissions per kilowatt-hour produced), developing renewable energy, or reducing deforestation and associated emissions of carbon dioxide are mitigation actions. In the literature, "mitigation" also encompasses "carbon sequestration" (either biological via photosynthesis or physical through carbon capture and storage), though sequestration does *not* avoid emissions but removes carbon from the atmosphere (that is, it reduces *net* emissions and not *gross* emissions). Relocating people and capital away from new flood-prone areas, shifting to crops that are more resistant to drought, or responding to and rehabilitating areas following natural disasters are all examples of adaptation actions.

In addition, following Smit and others (2000), two forms of adaptation are distinguished. *Reactive* adaptation focuses on *coping ex post* with the adverse impacts of climate change, when they occur. *Proactive* (or anticipative) adaptation, on the other hand, focuses on *lowering the costs of coping ex ante*. Proactive adaptation encompasses measures taken in advance to limit the ultimate damages of climate change or to reduce the extent of reactive adaptation required when climate change impacts materialize. For example, evacuating people from a flood-hit area is reactive adaptation, while modifying zoning laws on coasts in anticipation of stronger sea surges is proactive adaptation. Even though proactive adaptation and mitigation are both *ex ante* actions, proactive adaptation only reduces the cost of *ex post* adaptation but not the need for it, because it does not reduce emissions as mitigation would.

As noted by Fankhauser, Smith, and Tol (1998), the distinction between proactive adaptation and reactive adaptation is intuitively clear but difficult to delineate with precision in a dynamic setting. For example, after the heat wave of August 2003 in France, which is estimated to have caused in excess of 11,000 deaths over the historical average, the French government prepared a Heat Wave National Action Plan[3] that includes among other things the creation of a national alert system, a strong effort for prevention and information, and a clearer division of tasks among public agencies. This plan was adopted both in reaction to the 2003 heat wave itself (that is, reactive adaptation) and in anticipation of future repeat events (that is, proactive adaptation)—as such, it is a typical example of a "co-evolution" of problems and responses in a dynamic setting.

However, the distinction between proactive and reactive adaptation is important from a policy point of view because the rationale for the two actions is very different. Proactive adaptation (like mitigation) uses resources *now* to prevent possible crises in the future, while reactive adaptation uses resources to cope with events *at the time they occur. The crux of the problem is that, in practice, behavioral changes and policy decisions are often easier to implement once a crisis has occurred than in anticipation of a crisis.* Yet from an economic point of view, examples such as the 2004 Indian Ocean Tsunami or the 2005 hurricane Katrina suggest that the costs of preventive action—for example, installing early warning systems or fixing

dikes—are often lower than the costs of deferred action,[4] even when appropriately discounted, thereby making proactive adaptation preferable (Athukorala and Resosudarmo 2005; Burby 2006).

Finally, the term "ultimate damages" is used for damages that would be incurred *in the absence of any policies*—even if some private adaptation is implemented—and "residual ultimate damages" is used for those damages that remain *after all mitigation and adaptation expenditures have been incurred*, because they are technically irreversible (for example, lost species) or economically irreversible (that is, reversibility may be feasible technically, but is considered too costly, for example, the full restoration of the Everglades ecosystem or the Aral Sea).

There are thus four main options only for reducing the ultimate damages of climate change: mitigation (*ex ante*), proactive adaptation (*ex ante*), reactive adaptation (*ex post*), and accepting residual damages.

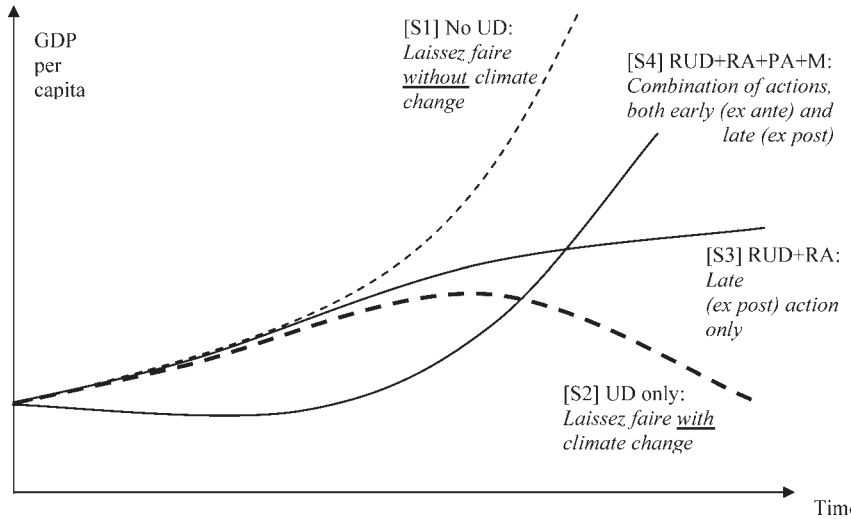
The Appropriate Counterfactual for Evaluating Climate Change Options: No Action and Full Ultimate Damages

Before choosing among these options, an important methodological point about the appropriate counterfactual must be made. Most assessments of mitigation policies have used as the baseline for ranking options a business-as-usual or laissez-faire growth (scenario S1 in figure 1) in the *absence* of climate change. Yet the uncertainty about the likelihood of human-induced global warming has been essentially resolved by now. In fact, from many different directions scientific evidence suggests that some degree of climate change is already occurring (IPCC 2007a). Thus, a laissez-faire scenario without climate change (S1) no longer describes any real-world situation. The appropriate counterfactual scenario (to determine the *real opportunity costs*) is now one in which no action whatsoever is taken against climate change and that, as a result, the full set of associated damages are incurred on the whole portfolio of assets (S2).

Against the first counterfactual (S1), any policy action against climate change, or any combination thereof, has a *net cost*. In other words, it looks as if any policy action was making the economy worse off. (This is not the case, of course, because even *no* action, as in S2, leads to a net cost relative to S1.) However, relative to the laissez-faire scenario in the presence of climate change (S2), it will be seen that a combination of policy actions (S3) or (S4) might bring *net benefits*.

Worrying about the appropriate counterfactual may seem trivial *if* the ranking of policy options does not change. However, the message conveyed will be different depending on how the problem is *framed*: the net benefits of mitigation action will be reported as a *positive* with respect to a counterfactual *with* climate change, and as a *negative* with respect to counterfactual *without* climate change (Mohr 1995). At a more fundamental level, the set of mitigation options available in the

Figure 1. Illustrative Growth over Time in the Laissez-faire Cases *without* Climate Change and *with* Climate Change; and in Two Policy Scenarios without and with Anticipative Action



Note: UD, ultimate damages; RUD, residual ultimate damages; RA, reactive adaptation; PA, proactive adaptation; M, mitigation.

S1, Laissez-faire Scenario without Climate Change; S2, Laissez-faire Scenario with Climate Change; S3, Policy Scenario with Reactive Adaptation Only; S4, Policy Scenario with Reactive Adaptation, Proactive Adaptation, and Mitigation.

presence of climate change is also likely to be more restricted than the set of options available in the absence of climate change because the efficiency of mitigation actions and the intensity of climate change are interdependent. This *can* affect the ranking of policy options (and targets).[5] For example, if climate change were not already occurring, investing in hydropower instead of using fossil fuels to generate energy could be a very appropriate mitigation measure in countries with glacial melt, such as Bolivia or Afghanistan. With climate change, however, the glaciers will initially melt faster than historic patterns show—thereby generating potentially higher volumes of water and requiring larger or taller dams. But eventually, once glaciers have disappeared, there will be insufficient water and some or all of the investments could be stranded or wasted, especially if the timeline is compressed due to an acceleration in climate change.[6]

An Integrated Portfolio Of Actions Is Needed To Minimize The Climate Bill

The decision problem faced by the international community is to choose the best combination of options *to minimize the global climate bill*, that is, to minimize the

sum of the costs of mitigation, proactive adaptation, reactive adaptation, *and* the acceptance of residual ultimate damages (the latter depending on the levels of mitigation, proactive adaptation, and reactive adaptation) through appropriate incentives and transfer mechanisms.

The decision problem is different at the national level because individual countries, whether small or large emitters, have little direct control over total world emissions. This has two critical implications. *First, they have to set domestic proactive and reactive adaptation levels given other countries' mitigation decisions. Second, their mitigation decisions, for the most part, make sense primarily in the context of global action*. Here, however, large and small emitters differ. For countries that are large emitters (such as China,[7] India, Brazil, Indonesia, Mexico, and a few others among developing countries), domestic mitigation decisions can have direct measurable implications for domestic damages, not just indirect via global collective action. For countries that are small emitters, domestic mitigation decisions will still matter, but only in certain circumstances; for example, if the use of cleaner fuels is also cheaper in the long run or if domestic commitment to mitigation action facilitates global collective action.

In this section we discuss how options to address climate change might be balanced, primarily at the international level, but also with reference to the national level.

Putting the Horse Before the Cart: Deriving the Need for Mitigation from the Inability to Adapt

Since climate change emerged as a major international issue some 20 years ago, the debate has focused for the most part on mitigation. As a result, the major pieces of international law that currently address climate change—the UN Framework Convention on Climate Change (1992), the Kyoto Protocol (1997), and the EU Emissions Trading Scheme (2003)—all focus primarily on mitigation, recognizing common but differentiated responsibilities.[8]

In this context, the debate over the participation of developing countries in the future climate regime narrows down to two questions: when will developing countries take on mitigation commitments? And how stringent will these commitments be? Negotiating on this basis has proved very difficult, and the controversy over how and when they should join the mitigation effort is far from resolved to date. Some countries, chief among them the United States, have consistently argued that large emitters among developing countries should take on commitments rapidly. Most developing countries, on the other hand, have typically been reluctant to even discuss this possibility (Hourcade 2003) and have called for more attention to adaptation.

The poorest countries emphasized adaptation early on, on the grounds that they would not be able to contribute much to mitigation, but would suffer from the costs of adapting to a changing climate to which they had not contributed (and many would not contribute) significantly.[9] In addition, most developing countries (like many developed countries) were concerned that mitigation would adversely affect economic growth and the ability to develop (Heller and Shukla 2003). In fact, although the IPCC (2007c) reports relatively modest global costs of mitigation, the range of modeling results is large. The impact on growth, in the case of developed countries, is expected to operate through increased capital and operating costs in emitting sectors (for the same output) and premature retirement of existing capital stock. In the case of developing countries, the impact on growth is expected to come through the higher cost of modernization, if it takes place in the context of expensive rather than cheap energy—thus making it more difficult to close the per capita income gap with industrial countries. Finally, competitiveness issues arise in both developed and developing countries if individual countries try to take mitigation actions unilaterally outside a collective action framework.

We argue that there are good reasons to *treat adaptation as the primary rather than the secondary concern* when addressing the climate change challenge within countries, as well as globally. First, as noted above, some countries have essentially very limited mitigation opportunities, but *all* face adaptation needs. Second, in practice, there have been delays in coming to an effective agreement on mitigation, and we expect that these delays will continue. *If, globally, no (or insufficient) action is being undertaken to mitigate, then implicitly one is behaving as if adaptation were cheaper.*[10] Third, the early impacts of climate change are being observed (in part a consequence of the delayed action of the past 20 years) and there is already a need for adaptation to deal with these initial stages.

A response to climate change based solely on reactive adaptation, however, is very unlikely to minimize the total climate bill for two key reasons. First, even if it is technically and economically feasible to cope with some impacts of climate change, it is not necessarily cheaper to do so than to engage in *ex ante* actions (proactive adaptation or mitigation). For example, though it might be technically and economically feasible to evacuate coastal cities, building dikes *ex ante* might prove cheaper, and reducing emissions to limit sea surges even more so. Much of the reluctance to move more effectively on mitigation is predicated on the assumption that adaptation costs will either be low or occur faraway in the future, when many countries will be better off and able to cope with the consequences of climate change (Schelling 1995). Yet it is unclear whether in fact the costs of adaptation will be lower than the costs of mitigation, and how they will be distributed over space and time, because information on the costs of adaptation is still limited (Adger and others 2007). Closing this information gap is a critical challenge facing analysts and modelers.

Second, and more importantly, *mitigation and adaptation are not perfect substitutes for each other*. If some losses are irreversible, then proactive adaptation (or reactive adaptation, for that matter) cannot restore them (for example, the loss of polar species, glacial ecosystems, or submerged coastal cities): large *residual* damages remain. Mitigation, on the other hand, can avoid those irreversible losses, provided it is undertaken early enough. In addition, mitigation is the only game in town to avoid potentially catastrophic consequences of climate change (such as a shift in thermohaline circulation).

Thus, the likely inability to adapt fully at low cost requires some degree of mitigation action. On the other hand, *a policy response relying on mitigation only is also very unlikely to minimize the total climate bill*—if only because some future changes in the climate are already locked in, leaving no option but to adapt to those consequences.

Combining Options is Preferable to Picking One: The Case for an Integrated Portofolio of Actions

The previous discussion thus suggests that *an integrated portfolio of actions that encompasses simultaneously some mitigation actions, some proactive adaptation actions, some reactive adaptation actions, and explicit acceptance of residual damages will be superior to any individual type of action alone in minimizing the total climate bill.*[11] Working backwards from the ability to adapt to climate change one gets the following priorities:

- Where the ultimate damages are likely to be low or inconsequential, the whole problem can be ignored—but it has been argued above that this is not likely to be the case globally in the long run (though it could be relevant in the short run for some locations).
- Where ultimate damages are expected to be of a magnitude or a type that one can cope with at low cost, reactive adaptation will dominate. However, even then one has to identify *irreversibilities* that might be incurred, and consciously make a decision that the residual damages associated with these irreversibilities are acceptable.
- Where the ultimate damages associated with climate change will generate vulnerabilities that cannot easily be coped with, or will generate irreversibilities that cannot be accepted, preventive *ex ante* action—proactive adaptation or mitigation—is necessary. The balance between proactive adaptation and mitigation will depend on the structure of uncertainties and risks (which is discussed below).

Because of the *inertia* in the climate system, there will always be a lag between *ex ante* actions and their effects, so one needs to schedule *ex ante* actions (whether mitigation or proactive adaptation) well in advance. This implies that a portfolio

of action is needed *at any given moment in time*. For example, all three types of action to deal with climate change damages are needed simultaneously *now*: some damage is occurring already and requires reactive adaptation *now*; other damage is unavoidable in the near to medium term. The cost of coping with this damage can be reduced by proactive adaptation *now*, while other damage may occur further in the future, which cannot be coped with cost-effectively even with proactive adaptation, thus requiring mitigation *now*. Some *ex ante* actions (such as proactive adaptation) can be taken unilaterally, but others (primarily mitigation) will require collective action.

The balance between actions, however, will need to adapt *over time*. First, because of the time lag between action and consequences, windows of opportunities to avoid or reduce particular categories of damages are closing continuously. In addition, climate change is likely to be nonlinear. The speed with which temperature will increase is not known, but the presumption is that it could accelerate in the absence of mitigation due to positive feedbacks between emissions and temperature (Friedlingstein and others 2003). Also, the extent of additional damages that will accompany each supplementary degree above current levels is uncertain, but again the presumption is that the damage increment will be larger with each additional degree (IPCC 2007b). Thus, minimizing the cost of climate change in the presence of uncertainty and nonlinearities must be treated within a portfolio of actions that is *capable of adapting. It should be negotiated and planned as such*.

The discussion above has tried to clarify, from a practical point of view, why a portfolio of action that combines proactive and reactive adaptation with mitigation is desirable and possibly unavoidable. In fact, mitigation and adaptation are so tightly linked that they need to be thought through and addressed jointly and simultaneously. Since the need for adaptation usually depends on the level of mitigation, and since the level of mitigation depends on the ability to adapt, what should be the balance between the two? Within the adaptation portfolio, what should be the balance between proactive and reactive adaptation? The first question is especially relevant for negotiations at the global level, whereas the second is also relevant at the national level, especially for countries with limited ability to contribute to mitigation.

The Interactions Between Mitigation and Adaptation Reinforce the Need for an Integrated Portfolio of Actions

Building on Kane and Shogren (2000), Lecocq and Shalizi (2007a) analyze the optimal balance between mitigation, proactive adaptation, reactive adaptation, and residual ultimate damages in a partial equilibrium, dynamic optimization model. The analytical resolution of the model confirms many standard results in

economic analysis of mitigation policies—notably that the marginal costs of abatement must be equal to the discounted sum of the marginal damages of emissions in all sectors/regions over all future periods. Also, adaptation—whether proactive, reactive, or a combination of the two—in specific sectors, regions, and periods should be financed up to the point where the last dollar spent is matched exactly by the discounted value of the avoided damages (in the *future* for proactive adaptation, vs *now* for reactive adaptation).

The analytical resolution of the model also confirms that the interactions among mitigation, proactive adaptation, and reactive adaptation determine the optimal levels of these three components in the model (Shibata and Winrich 1983). If mitigation and adaptation, whether proactive or reactive, are independent, then their optimal levels can be determined separately. In particular, the optimal level of adaptation would *not* depend on the success or failure of mitigation policies. So observed delays in implementing mitigation measures would not have any consequences for adaptation actions. If, on the other hand, mitigation and adaptation are interdependent—that is if the marginal benefits of mitigation in terms of avoided damages depend on the level of adaptation—then *the optimal level of mitigation and the optimal level of adaptation cannot be determined independently*. Then, a commitment to more adaptation would require less stringent mitigation targets. By the same token observed delays in implementing mitigation measures would require a different amount of (and probably more) proactive adaptation actions.

In fact, mitigation and adaptation are often interdependent (Klein and others 2007). First, some activities simultaneously influence both mitigation and adaptation—sometimes reinforcing and sometimes offsetting each other's effects. For example, planting trees can have a cooling effect on surrounding areas by providing additional moisture (adaptation) while removing carbon dioxide from the atmosphere—here adaptation reinforces mitigation. By contrast, developing air conditioning to cope with warming, or desalinization to cope with warming-induced water stress (both adaptation measures) may result in higher energy demand and GHG emissions if electricity is generated by fossil fuels—here adaptation undermines mitigation.

Second, the *efficiency* of adaptation measures often depends on the *level* of mitigation, and vice versa. For example, in the absence of mitigation, sea-level rise in some areas may be such that no sea wall can possibly protect the coastline. Neighborhoods and even cities may have to be relocated, at very high cost, with limited possibility for reducing the residual damages because the submerged physical capital is lost. With sufficient mitigation, on the other hand, the sea level may not rise as much and sea walls would be effective. It might then be more cost-effective to invest in proactive adaptation in the form of sea walls than to relocate the city. In this example, mitigation and proactive adaptation are

complements, but they can also be substitutes. For example, high levels of mitigation that limit temperature increase may allow some threatened ecosystems to survive and thus make it unnecessary to adopt costly proactive protection measures.

Similarly, proactive adaptation and reactive adaptation are often not independent, and their optimal levels cannot be determined separately. For example, modifying zoning plans *ex ante* to account for increased risks of floods reduces the need for costly evacuation and sheltering of victims in response to a flood *ex post* (substitution). On the other hand, investing in advance in upgrading the capabilities of emergency response teams (proactive adaptation), through better training and equipment, will enable them to react more efficiently when a disaster occurs (reactive adaptation)—hence *they complement each other*.

The interdependence between mitigation and adaptation has four implications. First, it reinforces the importance of improving our knowledge about the costs and benefits of adaptation—currently underdeveloped relative to mitigation. Second, it suggests that introducing adaptation into numerical models that assess the costs and benefits of climate policies is very important—not as an add-on, but as a potentially important factor in shaping mitigation decisions. Third, from a policy perspective, it suggests that *mitigation policies and adaptation policies should be negotiated jointly and not separately as is essentially the case today at the international level, and that they should be developed and planned jointly at the national level*. Fourth, the interdependence between mitigation and adaptation is an additional argument in favor of an integrated portfolio of actions.

Balancing the Portfolio of Actions under Uncertainty

In frameworks where only mitigation is discussed, the shape of the *aggregate* damage function—that is, the timing and size of the damage—is the most important uncertainty for setting the optimal level of mitigation (Ambrosi and others 2003). However, when adaptation is introduced, uncertainty on the *distribution of damages across space* (regions) also becomes an issue, since the benefits of adaptation are sector and site-specific.

This has implications for the optimal division of resources among mitigation, proactive adaptation, and reactive adaptation. First, the more uncertain the location of an impact, the more cost-effective mitigation becomes relative to proactive adaptation[12] (Lecocq and Shalizi 2007a). Second, in contrast to the mitigation vs proactive adaptation balance, where uncertainty favors earlier action, in the case of the proactive vs reactive balance, uncertainty favors reactive adaptation over proactive adaptation—to the extent that proactive adaptation and reactive adaptation are substitutes (but not if they are complements). This is

because *proactive adaptation measures, and particularly those that consist of building or strengthening fixed, long-lived capital stock, have a higher chance of being misdirected* when there is uncertainty about the location of impacts. This uncertainty is resolved once climate change events occur, hence the greater efficiency of reactive adaptation. In other words, with uncertainty about location, the costs of making mistakes—that is, of adapting in sectors/regions that finally will suffer less than expected—begin to erode the expected benefits of proactive adaptation. *Mitigation, on the other hand, reduces all damages regardless of the region/sector, and is thus unaffected by uncertainty over the spatial distribution of impacts.* (However, it remains affected by the uncertainty over the timing and magnitude of impacts.)

Lecocq and Shalizi (2007a) draw from this analysis a typology of situations with different implications for the balance of the portfolio of actions under uncertainty:

- If uncertain impacts are expected to occur with high confidence in known locations, then targeted (site-specific) proactive adaptation has the highest chance of remaining cost-effective[13] (even if it involves producing fixed long-lived capital stock), because the risk of misdirecting investment toward the wrong region/sector is low. For example, one might invest in water management infrastructure to reduce tensions on water resources that are likely to emerge in regions already under high water stress.
- If uncertain impacts are expected to occur with high confidence, but whose precise location remains uncertain, then nontargeted (nonsite-specific) proactive adaptation measures may still remain cost-effective if they cover enough sectors or regions. Examples of nontargeted proactive adaptation measures include setting up country-wide disaster response and management capabilities, or developing appropriate insurance markets.
- If uncertain impacts are expected to occur with low confidence, and location remains uncertain, then, depending on the scale of the impact, mitigation, or reactive adaptation is more likely to be cost-effective relative to proactive adaptation.

An Integrated Portfolio Makes Sense Even Within A Broader Perspective

The discussion above is framed in a partial equilibrium approach focused primarily on climate change. However, given the magnitude of the problem and the fact that resources are limited, it is necessary to look at climate change in the broader context of other challenges to development.

Views on the Appropriateness of Incorporating Mitigation into Development Strategies from a Broader Perspective

When looking at the problem from a broader perspective, the idea that early anticipative action (be it mitigation or proactive adaptation) is necessary has been challenged by a number of authors. Three lines of criticism stand out.

The first line of criticism is associated with Schelling (1995, 2006). He argues that the main beneficiaries of mitigation will be developing countries, since they account for the largest part of the Earth's surface and the greatest proportion of population exposed to climate variability. They are also the most vulnerable—with limited ability to cope, given their current level of development. He further argues that instead of industrialized countries putting a lot of funds into mitigating themselves, or putting pressure on developing countries to mitigate, they would be better off transferring equivalent resources *directly* to developing countries *today*—in order to facilitate more rapid growth in those countries and increase their ability to cope with climate change consequences when they arrive—rather than *indirectly* through avoided costs in the *future*. The argument rests on the assumption that future generations will be wealthier and technologically more capable than today's (that is, there is less need for *inter*generational transfers), so that there is a premium on helping the poor today (that is, there is more need for *intra*generational transfers now). This is an important and valid point, but, as noted above, it is predicated on the assumption that adaptation will be cheap relative to mitigation, which has not yet been demonstrated. In addition, *the core weakness of this position is that it understates potential economic and technical irreversibilities*. Mainly encouraging growth and development in the hope that it will increase adaptation capabilities in the future does not address the fact that adaptation and mitigation are not perfect substitutes, since adaptation cannot meaningfully and cost-effectively address many types of species extinction, ecological destruction, or other potentially catastrophic events.

The second line of criticism is associated with the Copenhagen Consensus (Lomborg 2004),[14] which notes that there are many immediate and important risks and challenges facing developing countries that dominate the actions for addressing climate change. In other words, policies that address many of these other problems have a higher cost–benefit ratio than policies that address climate change. This is also an important and valid argument. But it does not exclude the need for mitigation, as even the analysis of climate policies on which the Copenhagen Consensus is based (Yohe and others 2008) has a positive cost–benefit ratio, despite the fact that it does not fully include uncertainty or catastrophes. In addition, *this analysis does not take into account the fact that because of the decade-long lag structures between action and consequences in the climate change arena, earlier actions that avoid bad lock-ins and favor good lock-ins can dominate later actions*.

A third line of criticism comes from technological optimists who argue that exogenous technological change will allow us to reduce emissions drastically (see for example, the lowest-emissions business-as-usual scenarios reported in IPCC 2007c) or to find a geo-engineering solution to climate change (even if the latter is still recognized to be costly financially). As a result, they argue that there is less need for costly anticipative actions now. While this argument is enticing and there is ample evidence that exogenous technological change is a major driver of growth and development, *there is much less certainty that the necessary technologies will automatically emerge in time* to forestall the negative consequences of climate change (that is, without early changes in incentives and institutions to stimulate research into and diffusion of desired technologies). Since insurance markets cannot address systemic global risk, there is a need for a global insurance policy in the form of actions now on incentives and institutions to stimulate R&D into desired technologies and to ensure that they are brought on line in time. Otherwise the world, and particularly developing countries, could be confronted with the need for draconian adjustments and potentially serious social conflict if the necessary technologies do not automatically emerge in time. The geo-engineering options also carry the risk of uncertain consequences associated with large-scale interventions on the Earth's climate in a web of relations not yet fully understood.

A key issue ignored in the lines of criticism above is the role of learning by doing—both in terms of speeding up the generation of information that resolves uncertainty and in terms of speeding up the rate at which the cost of action (for example, developing viable alternative technologies) is lowered over time. Arrow and Fischer (1974) and Henry (1974) have noted that, in the presence of uncertainty and inertia, increasing information that might resolve uncertainty in the future (at least partially) increases the cost-effectiveness of courses of action that leave options open (that is, there is an option value to retaining flexibility). Translating this approach to climate change, Ha-Duong, Grubb, and Hourcade (1997) show that risk-averse actions (in their model, more mitigation now, which can be revised upward or downward as new information on climate change materializes) often dominate risk-neutral approaches. The mitigation actions contemplated here would be separate from, and in addition to, actions that generate information to resolve uncertainty (that is, more research on climate change mechanisms and impacts). Several authors have also argued that in the presence of "learning by doing", early action can expedite the move along the technology cost curve to lower the cost side of the cost–benefit calculations (for example, Grubb, Carraro, and Schellnhuber 2006).

Another fundamental problem not taken into account by the three lines of criticism above is the limited ability of cost–benefit analysis and standard discounting to handle the large uncertainty on catastrophic events with low probability or unpredictable systemic effects. In the presence of "unreckonable risks"

(Chomitz 2007), or of uncertainty in the sense of Knight (1921) (that is, without ex ante knowledge of the probability of occurrence of the different states of nature), Weitzman (2007) argues that one must act as if the chance of an extreme event is significant. He even argues that this issue can dominate discount-rate debates. Such arguments will favor larger investments on mitigation to avoid crossing "catastrophe generating" thresholds.[15]

To sum up, the three lines of criticism of mitigation action (in both developed and developing countries) raise important and legitimate points. However, they do not address the whole gamut of issues associated with climate change, nor do they use methodologies fully adequate to the nature of the problem. Thus, even when looked at from a broader perspective, mitigation will remain an important component of any integrated portfolio of actions to address climate change—whether in developed or developing countries.

Sharing Responsibilities: The Role of Adaptation in the Public Sector Response

Though a portfolio of action among mitigation, proactive adaptation, reactive adaptation, and accepting residual damages might be necessary for society, a question remains as to whether adaptation should be part of the portfolio of *public* actions. Since mitigation reduces all climate-related risks—both known and unknown—everywhere, it is a global public good requiring collective action (by all nations at the international level, and by all subnational actors at the national level). By contrast, as already noted above, adaptation reduces specific classes of risks, often in specific locations. Thus, adaptation can be site-specific (for example, land-use planning), risk-specific (for example, R&D on heat-tolerant crops), or both (for example, hardening of local and regional infrastructure). As such, adaptation provides a private good (for example, a more resistant building benefiting its inhabitants only), a club good (for example, a mutual insurance fund), or a local public good (for example, a dike).

Economic theory suggests that such goods should be self-supplied by the individuals, firms, or local communities that benefit from them at the subnational level and not by national governments or public agencies. Similarly, from an international point of view, economic theory suggests that adaptation measures that benefit individual countries should be self-financed by the countries themselves and not by the international community.

The rationale for public provision of resources for adaptation at the national or international level is thus less obvious than the rationale for public provision of mitigation. However, public intervention may still be justified—at the national level (in relation to the subnational level, nonpublic entities, or both) and at the international level (in relation to the national level, global civil society entities, or both)—for standard well-known economic reasons. These latter include imperfect

information, barriers to collective action (at the subnational level within a country, or at the national level relative to the international community), moral hazard or free rider problems, externalities within and across countries, network or public good aspects of high fixed-cost national and international assets, and budget constraints and the ability of the poor to pay (see Lecocq and Shalizi 2007a for a discussion). There are also a wide range of instruments that can be used, ranging from indirect actions (such as information provision or standard setting) to direct actions (such as taxes on ill-adapted assets or direct provision of adaptation resources and institutions). There are also multiple ways in which the international community may support adaptation at the country level on top of what individual countries are doing.

Further empirical work is required to determine how much adaptation is required, how much private agents, developing country governments, and the international community can afford, and *whether the existing framework and level of international funds for adaptation is capable of meeting the needs*.[16] However, the cost–benefit criterion applies to government, as well as private, actions. Thus, the government should only support *proactive* adaptation measures to the extent that the benefits to society outweigh the public costs of implementation.

One advantage of anticipative action, including proactive adaptation, is that, if properly planned, expenditures can more easily be *spread out* over time, whereas reactive adaptation may require large expenditures in short periods of time. In the words of Chomitz (2007), "smoothly adaptive" expenditure or investment strategies may be preferable to "lumpy" expenditure or investment strategies in the face of "inexorable calamities" when timing uncertainty is taken into consideration.

Relying on reactive adaptation runs another risk. Public resources are rarely stable over long periods of time, especially in developing countries. Both sudden and prolonged climatic shocks can erode the country's fiscal base. Thus, the risks of climate impacts and low availability of public resources are at least partly correlated and must be addressed in advance. Setting up rainy-day funds (Sobel and Holcombe 1996; Lecocq and Shalizi 2007a) may be an appropriate solution. Such funds could still be cost-effective even with low returns, so long as the risk of not being able to react adequately is high because of budget constraints. At the global level, the rainy-day fund is a form of self-insurance whose usefulness is highest when contributions cumulate in the medium term. At the national level, however, resources might be insufficient relative to the size of the impacts. Therefore, financing of reactive adaptation may have to be split between a national rainy-day fund and transfers from abroad.

However, even when there is uncertainty about the location of damages, the rainy-day fund may complement, but not necessarily replace, proactive adaptation. More research is required to determine fully the conditions under which rainy-day funds are effective, notably, taking into account the uncertainty as to

when damages might occur, and taking into account that proactive adaptation typically reduces damages during more than one period. More empirical research on the returns to these funds, and their contingency to institutional structures in developing country contexts, is also necessary.

The Need For A Broad Review Of Development Strategies

The emerging risks associated with climate change make it necessary that all actors, public and private, at least review their development strategies, policies, and projects. The discussion above provides a qualitative framework for such a review.

This review should be conducted for all investment projects that take place within the country, be they funded by corporations, communities, households, or individuals—and not only for investment programs and projects funded by the government. Similarly, the design of long-lasting institutional arrangements may have to be revised to take climate change into account. For example, when water run-offs are expected to diminish, it is all the more important for long-term water rights' arrangements to include strong provisions for resolving tensions (Miller, Rhodes, and MacDonnell 1997).

The review should also encompass all sectors, not just climate-sensitive or GHG-emitting ones. In fact, current *adaptation* literature focuses mostly on a limited number of sectors (notably agriculture) (Adger and others 2007) and on developing countries (Dasgupta and others 2007; Mendelsohn and others 2007). This is understandable given the large share of agriculture in the GDP of many developing countries and the sensitivity of this sector to the vagaries of the climate. However, because of growing interindustry linkages as development progresses many more sectors will exhibit sensitivity to climate, and attention will have to expand to these other sectors as well—such as various infrastructure sectors (roads for rural markets and global trade, changing the engineering design of infrastructure in areas where glaciers are disappearing, hardening buildings and infrastructure in coastal areas prone to storms and storm surges, and so on) and emerging alternatives to agriculture (such as tourism). Finally, the indirect effects of climate change on nonclimate-sensitive sectors, via, for example factor mobility or markets, may also require that adaptation measures be taken there.

The *mitigation* literature, though much larger, also tends to concentrate on a narrow range of sectors, namely energy supply (volume and mix of fuel). However, there also is a need to review development strategies in other sectors (Sathaye and others 2007). For example, demand management policies in energy-intensive sectors can be very effective in reducing the long-term trajectory

of energy consumption and emissions. In fact, it may be easier to build more compact cities with more balanced multimodal transport systems (lower inefficient use of energy in transportation) or to construct buildings with better insulation and energy efficiency (lower energy demand through better construction) than to risk lock-ins and be left with "retrofitting" long-lasting capital as the only mitigation option. *This is especially important since developing countries are undergoing massive urbanization and will be installing a large part of their long-lived capital stock in the next 15–30 years: addressing mitigation opportunities in these other sectors is critical to avoid potential lock-ins.*

One might object to the need for such a review on the grounds that in many countries adaptation to *current* climate variability is already part of development strategies. Yet this does not mean that these strategies are also adapted to *future* climate variability associated with climate change. "Win–win" opportunities, in which improving adaptation to current climate variability is aligned with adaptation to future climate variability, may well exist (Smit and others 2000), but careful examination is still warranted. For example, a key development goal for a small, very poor country with a high share of GDP in agriculture might be to improve smallholders' agricultural productivity and their integration into agroprocessing. To meet this goal, the high vulnerability of smallholders to the currently observed range of weather-related shocks must be reduced through irrigation projects, improved management of key watersheds, and other agriculture development programs that include weather-risk mitigation. It would seem at first glance that many countries' current development strategies already target vulnerable rural communities with the objective of reducing the impacts of weather-related risks. However, these projects and programs may *not* be sufficient to cope with *increased variability* in climate, or with *sustained climate patterns for which there is no precedent*—such as 100-year floods or multiyear droughts occurring much more frequently. And they might even be a waste of resources if climate variability increases so much in the future that outmigration[17] of the local population, shifting the domestic economy towards other, less climate-sensitive sectors, or both, become the only viable solution.

Finally, though the discussion has focused on developing countries, it must be clear that the challenge of adapting development strategies to climate change is a global one. And unlike other challenges, it is new for all countries involved, not just developing ones.

Conclusions And Recommendations

Improving people's quality of life, and not just standard of living, is a major goal of most societies globally. This involves transforming institutions to manage

a broad portfolio of assets: not just physical and human capital, but also social and environmental assets, as well as knowledge and technology (World Bank 2002). Development strategies that transform institutions and policies to move from a low and narrow asset base to a high and broad asset base operate within a set of constraints including, among other things, geography, endowments in natural resources, climate, history, culture, and economic environment. Climate change is a new challenge creating headwinds for development.

This paper has provided a roadmap for assessing the consequences of climate change for development strategies. *The presence of climate change makes it necessary to at least broadly review development strategies*, regardless of the sources of funding (foreign, domestic, or both). Though the mitigation debate focuses primarily on the energy sector, and though the adaptation debate focuses primarily on climate-sensitive sectors such as agriculture, input–output relations in multisector models highlight the importance of indirect effects of climate change on the rest of the economy, hence the importance of reviewing development strategies *even in apparently nonclimate-sensitive and nonGHG-emitting sectors*. In particular, it is critical to review projects and programs that involve long-lived, fixed capital stock and to adjust investment strategies—notably in countries undergoing major urbanization. Similarly, the design of long-lasting institutional arrangements will have to be revised to take climate change risks into account.

An argument has been made that because climate change poses both imminent and long-run impacts, and because the policy responses to climate change risks are interdependent, *there is a need for an integrated portfolio of actions spanning a spectrum from avoidance of climate change to explicitly accepting the residual damages generated by climate change*. In this portfolio, proactive (*ex ante*) adaptation is critical, but also subject to risks of regrets in cases of uncertainty about the location of damages. In particular, uncertainty as to location favors either nonsite-specific actions (such as strengthening the ability to react and manage disasters), or reactive (*ex post*) adaptation. Although adaptation often provides private benefits, it should not be left entirely to private agents: *there is a strong rationale for public intervention for adaptation both at the national and international levels*—for example, when there are spillovers, such as conflicts. To limit the risks that budget constraints might prevent developing countries from financing reactive adaptation—especially since climate shocks often erode the fiscal base—*rainy-day funds may have to be developed within countries*, and also at the global level for transfer purposes.

However, the effectiveness of proactive or reactive adaptation is limited as some losses are technically or economically irreversible, such as for biodiversity or culturally valued sites and monuments. As a result, some level of mitigation might be, in many cases, the cheapest option for addressing long-term climate change. It is also the only option for avoiding thresholds that generate truly catastrophic

consequences. In fact, *it is highly unlikely that the amount of effort on any one type of action in the portfolio will be zero in any country over the next few decades*. However, the balance between the different components of the portfolio is likely to evolve over time. It should be negotiated and planned as such.

Finally, *since mitigation and adaptation are interdependent, mitigation policies and adaptation policies should be negotiated jointly at the international level and not separately*, as is essentially the case today, and they should be designed and implemented jointly at the national level. Adopting such a package could in turn increase the probability that a global treaty can be devised that fairly addresses the needs and capabilities of the diverse constituencies.

A caveat must be noted here on the methodology used in this analysis. Optimization tools are very powerful for conceptually determining an optimal portfolio of actions. In practice, however, it may be very difficult operationally to define such a portfolio without additional information and data. It may even be *irrelevant* in some cases to worry about trade-offs at the margin, if current actions are *suboptimal* in aggregate, and if one needs to move forward on multiple fronts simultaneously (because one is far inside the production frontier, rather than being on the frontier). Weitzman's criticism (2007) of consumption smoothing cost–benefit analysis in the presence of potential catastrophic events also applies to our framework. However, the key message remains: resources can be misallocated if one just funds activities because they might hypothetically address climate change. Some effort has to be made to construct a portfolio of actions that recognizes trade-offs and the fact that one faces simultaneously different damages with different lag structures between actions and their benefits.

The paper has also identified big gaps in the literature. First, more *disaggregated* information about likely damages—in terms of magnitude, location, and timing—is necessary to get a better quantitative sense of the optimal balance among mitigation, reactive adaptation, proactive adaptation, and residual ultimate damages. In particular, more research is required on path dependency (lock-ins) and poverty traps. But as noted above, there is already a lot of scope for action with the information currently available. Second, introducing adaptation in numerical models that assess the costs and benefits of climate policies is very important—not as an add-on, but as a potentially important factor in shaping mitigation decisions. Third, further empirical work is needed to determine how much adaptation is required; how much private agents, developing country governments, and the international community can afford; and whether the existing framework and level of international funds for adaptation is capable of meeting the needs. Fourth, more work is required on whether the costs of adaptation can *meaningfully* be separated from normal development expenditures, in order to determine the extra resources required.[18] Finally, more research is needed to determine the conditions under which rainy-day funds are effective.

Finally, a priority for future research should be on how to *operationalize* this framework—that is, determining how development strategies should be modified and what should be the balance of actions in an integrated portfolio of actions—at the country or regional level. This requires in particular further analysis of the relationships between the national and international levels. If, in the presence of uncertainty, mitigation is indeed more cost-effective than adaptation, then one needs to understand how the need for collective action on mitigation can be strengthened at the international negotiation level, including by developing country negotiators. Second, since the extent of mitigation is for the most part exogenous for individual country policymakers, one needs to explore to what extent optimal country-level adaptation strategies depend on this exogenous parameter. In conclusion, on the one hand, the first part of the paper argues that development strategies will in fact look different in the presence of climate change than in its absence, because of the need to incorporate a portfolio of actions to minimize the climate bill. On the other hand, the second part of the paper argues that climate change negotiations are less likely to succeed if they do not explicitly address development issues[19] and the need for an integrated portfolio of actions.

Funding

Franck Lecocq acknowledges support from the World Bank Knowledge for Change program (7137924).

Notes

Zmarak Shalizi (corresponding author) is an independent scholar, formerly Senior Research Manager for Infrastructure and Environment at the World Bank; his email address is zmarakshalizi@yahoo.com. Franck Lecocq is an economist at the AgroParisTech, Engref and INRA, UMR 356 Économie Forestière, F-54000 Nancy, France; his email address is franck.lecocq@nancy-engref.inra.fr. The authors thank three anonymous reviewers, Ariel Dinar, Jean-Charles Hourcade, Eric Langlais, David Wheeler, and participants at workshops in Paris, Capetown and Goteborg for useful comments on earlier versions of this paper.

1. Large water-stressed inland cities such as Beijing, Delhi, Kabul, and Tehran.
2. Cheap sea transportation favors export industries in coastal cities. Yet the latter are more vulnerable to climate-change-induced sea-level rise, hurricanes, and associated sea surges.
3. http://www.sante-jeunesse-sports.gouv.fr/IMG//pdf/PNC-2008.pdf, last accessed July 7, 2009.
4. Lump sum transfers of resources to victims of climate change are not included in reactive adaptation, because they do not affect the size or efficiency of the economy.
5. Currently, most studies rank competing concentration targets by adding the costs of mitigation and the ultimate damages associated with each target, the former and the latter being established independently.
6. Hydropower potential may increase in northern latitudes as a result of climate change, but may decrease in temperate/Mediterranean regions (Lehner, Czisch, and Vassolo 2005) and in tropical areas such as the Andes (Bradley and others 2006).

7. Now estimated to be the largest in the world.

8. In fact, a Least Developed Countries Fund under the UNFCCC supports, among other things, the preparation of National Adaptation Plans of Actions. The Kyoto Protocol also establishes an Adaptation Fund supported by shares of the proceeds of the sales of Certified Emission Reductions under the Clean Development Mechanism. But overall, resources available for adaptation remain limited (Tompkins and Hultman 2007). Another reason for focusing on mitigation is the availability of a common metric for most actions. A comparable metric is not yet available to evaluate and compare the efficacy of adaptation actions.

9. For a history of climate negotiations, see Grubb, Vrolijk, and Brack (1999).

10. Or, cynically, as if the costs were to fall conveniently on the weakest members of the society within countries or globally. (Although failure to take action could also be explained by the difficulty of collective action between sovereign nations without an external enforcement mechanism; see also endnote 19.)

11. Few studies attempt to estimate numerically the optimal balance between mitigation and adaptation. Bosello (2004) and de Bruin, Dellink and Tol (2007) both find that introducing reactive adaptation in global optimization models significantly reduces the total climate bill and that the optimal portfolio of actions include both mitigation and adaptation. However, as of spring 2009, their models include neither proactive adaptation nor uncertainty.

12. Even here, there is a caveat. The uncertainty associated with mitigation investments, for example, in hydropower or wind farms increases once climate change is already underway, because past patterns and locations of water or wind flows are no longer reliable.

13. Callaway (2004) makes a similar point that irreversible investment for adaptation will be undertaken when it becomes clear that the climatic events they are aimed at adapting to are not random, but part of climate change.

14. See also http://www.copenhagenconsensus.com, last accessed July 7, 2009.

15. Heal (2008) shows that neither low values for the pure rate of time preference nor catastrophes are necessary to generate the need for high levels of mitigation action early.

16. Compare Note 8.

17. Historically, outmigration was a natural adaptive response to climate events such as prolonged droughts. However, it often led to conflicts with settled or nomadic populations in the regions to which the eco-refugees moved.

18. Politically, there is an imperative to separate mitigation, adaptation, and development finance. But this raises complex analytic problems. In light of the argument made earlier in this paper, "normal" development in the absence of climate change can *no longer* be meaningfully defined, because the pattern of development expenditures going forward will already be different (resulting in a higher or lower level of expenditures) in the presence of climate change in order to handle a continuous and changing set of adaptation needs brought about by climate change.

19. Large inequalities lead to a large divergence of interests, which in turn *undermine* the ability to resolve *collective action problems* (World Bank 2002). This is why climate change negotiations will not generate robust solutions until development issues are resolved.

References

Adger, W.N., S. Agrawala, M.M.Q. Mirza, C. Conde, K. O'Brien, J. Pulhin, R. Pulwarty *et al.* 2007. "Assessment of Adaptation Practices, Options, Constraints and Capacity." In IPCC (2007b).

Ambrosi, P., J.-C. Hourcade, S. Hallegatte, F. Lecocq, P. Dumas, and M. Ha-Duong. 2003. "Optimal Control Models and Elicitation of Attitudes Towards Climate Damages." *Environmental Modeling and Assessment* 54:133–47.

Arrow, K.J., and A.C. Fischer. 1974. "Environmental Preservation, Uncertainty and Irreversibility." *Quarterly Journal of Economics* 88(2):312–9.

Athukorala, P., and B.P. Resosudarmo. 2005. "The Indian Ocean Tsunami: Economic Impact, Disaster Management, and Lessons." *Asian Economic Papers* 4(1):1–39.

Azariadis, C., and J. Stachurski. 2004. "Poverty Traps." Research Paper 913. University of Melbourne.

Bosello, F. 2004. *Timing and Size of Adaptation, Mitigation and R&D Investments in Climate Policy*. Venice: Fondazione Eni Enrico Mattei.

Bradley, R.S., M. Vuille, H.F. Diaz, and W. Vergara. 2006. "Threat to Water Supplies in the Tropical Andes." *Science* 312:1755–6.

de Bruin, K.C., R.B. Dellink, and R.S.J. Tol. 2007. "AD-DICE: An Implementation of Adaptation in the DICE Model." Working Paper FNU-126. Hamburg University and Center for Marine and Atmospheric Science.

Burby, R.J. 2006. "Hurricane Katrina and the Paradoxes of Government Disaster Policy: Bringing About Wise Governmental Decisions for Hazardous Areas." *Annals of the American Academy of Political and Social Sciences* (604):171–91.

Callaway, J.M. 2004. "The Benefits and Costs of Adapting to Climate Variability and Change." In Corfee-Morlot, J., and S. Agrawala eds., *The Benefits of Climate Change Policies: Analytical and Framework Issues*. Paris: OECD.

Chomitz, K.J. 2007. "Inexorable Catastrophes and Unreckonable Risks: A Framework for Thinking about Climate Adaptation." Mimeo.

Collier, P., and J. Dehn. 2001. "Aid, Shocks, and Growth." Policy Research Working Paper 2688. World Bank, Washington, DC.

Dasgupta, S., B. Laplante, C. Meisner, D. Wheeler, and J. Yan. 2007. "The Impact of Sea Level Rise on Developing Countries: A Comparative Analysis." Policy Research Working Paper 4136. World Bank, Washington, DC.

Easterly, W., M. Kremer, L. Pritchett, and L. Summers. 1993. "Good Policy or Good Luck? Country Growth Performance and Temporary Shocks." *Journal of Monetary Economics* 32(3):459–83.

Fankhauser, S., J.B. Smith, and R.S.J. Tol. 1998. "Weathering Climate Change: Some Simple Rules to Guide Adaptation Decisions." *Ecological Economics* 30:67–78.

Friedlingstein, P., J.-L. Dufresne, P.M. Cox, and P. Rayner. 2003. "How Positive is the Feedback between Climate Change and the Carbon Cycle?" *Tellus B* 55(2):692–700.

Fujita, M., P. Krugman, and A.J. Venables. 1999. *The Spatial Economy*. Cambridge, MA: MIT Press.

Grubb, M., C. Carraro, and J. Schellnhuber, eds. 2006. "Endogenous Technological Change and the Economics of Atmospheric Stabilisation." *The Energy Journal* 27(April; special issue).

Grubb, M., C. Vrolijk, and D. Brack. 1999. *The Kyoto Protocol, a Guide and an Assessment*. London: Royal Institute of International Affairs.

Ha-Duong, M., M. Grubb, and J.-C. Hourcade. 1997. "Influence of Socioeconomic Inertia and Uncertainty on Optimal CO2-emission Abatement." *Nature* 390:270–3.

Hallegatte, S., J.-C. Hourcade, and P. Dumas. 2007. "Why Economic Dynamics Matter in Assessing Climate Change Damages: Illustration on Extreme Events." *Ecological Economics* 62(2):330–40.

Heal, G. 2008. "Climate Economics: A Meta-Review and Some Suggestions." NBER Working Papers 13927.

Heller, T.C., and P.R. Shukla. 2003. "Development and Climate: Engaging Developing Countries." In *Beyond Kyoto: Advancing the International Effort against Climate Change*. Washington, DC: Pew Center on Global Climate Change.

Henry, C. 1974. "Investment Decisions under Uncertainty: The 'Irreversibility Effect'." *American Economic Review* 64(6):1006–12.

Hourcade, J.-C. 2003. "Dans le labyrinthe de verre. La négociation sur l'effet de serre." *Critique Internationale* (15):143–59.

Huq, S., S. Kovats, H. Reid, and D. Satterthwaite. 2007. "Editorial: Reducing Risks to Cities from Disasters and Climate Change." *Environment & Urbanization* 19(1):3–15.

IMF (International Monetary Fund). 2003. *Fund Assistance for Countries Facing Exogenous Shocks*. Washington, DC: Policy Development and Review Department.

IPCC (Intergovernmental Panel on Climate Change). 2007a. "Climate Change 2007: The Physical Science Basis." Contribution of Working Group I to the Fourth Assessment Report of the IPCC (Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller, eds.) Cambridge, UK: Cambridge University Press.

———. 2007b. Climate Change 2007: "Impacts, Adaptation and Vulnerability." Contribution of Working Group II to the Fourth Assessment Report of the IPCC (Parry, M.L., O.F. Canziani, J.P. Palutikof, and P.J. van der Linden, and C.E. Hanson, eds.) Cambridge, UK: Cambridge University Press.

———. 2007c. Climate Change 2007: "Mitigation." Contribution of Working Group III to the Fourth Assessment Report of the IPCC (Metz, B., O.R. Davidson, P.R. Bosch, R. Dave, and L.A. Meyer, eds.) Cambridge, UK: Cambridge University Press.

Kane, S., and J.F. Shogren. 2000. "Linking Adaptation and Mitigation in Climate Change Policy." *Climatic Change* 45:75–102.

Klein, R.J.T., S. Huq, F. Denton, T.E. Downing, R.G. Richels, J.B. Robinson, and F.L. Toth. 2007. "Inter-relationships between Adaptation and Mitigation." In IPCC (2007b).

Knight, F.H.. 1921. *Risk, Uncertainty and Profit*. Boston, MA: Houghton Mifflin.

Kundzewicz, Z.W., L.J. Mata, N.W. Arnell, P. Döll, P. Kabat, B. Jiménez, K.A. Miller *et al.* 2007. "Freshwater Resources and their Management." In IPCC (2007b).

Lecocq, F., and Z. Shalizi. 2007a. "Balancing Expenditures on Mitigation of and Adaptation to Climate Change—An Exploration of Issues Relevant to Developing Countries." Policy Research Working Paper 4299. World Bank, Washington, DC.

———. 2007b. "How Might Climate Change Affect Economic Growth in Developing Countries? A Review of the Growth Literature with a Climate Lens." Policy Research Working Paper 4315. World Bank, Washington, DC.

Lehner, B., G. Czisch, and S. Vassolo. 2005. "The Impact of Global Change on the Hydropower Potential of Europe: A Model-based Analysis." *Energy Policy* 33:839–55.

Lomborg, B. (ed.) 2004. *Global Crisis, Global Solutions: World-Leading Economists Confront World-Threatening Problems*. Cambridge, UK: Cambridge University Press.

Martin, M., and H. Bargawi. 2004. "Protecting Africa Against Shocks." Background paper prepared for the Commission for Africa. Development Finance International.

Mendelsohn, R., A. Basist, P. Kurukulasuriya, and A. Dinar. 2007. "Climate and Rural Income." *Climatic Change* 81:101–18.

Mendelsohn, R., W. Morrison, M.E. Schlesinger, and N.G. Andronova. 2000. "Country-Specific Market Impacts of Climate Change." *Climatic Change* 45:553–69.

Miller, K.A., S.L. Rhodes, and L.J. MacDonnell. 1997. "Water Allocation in a Changing Climate: Institutions and Adaptation." *Climatic Change* 35:157–77.

Mohr, E. 1995. "Greenhouse Policy Persuasion: Towards a Positive Theory of Discounting the Climate Future." *Ecological Economics* 15(3):235–45.

Nordhaus, W.D., and J.G. Boyer. 2000. *Warming the World: The Economics of the Greenhouse Effect*. Cambridge, MA: MIT Press.

Pearce, D.W., W.R. Cline, A.N. Achanta, S. Fankhauser, R.K. Pachauri, R.S.J. Tol, and P. Vellinga. 1996. "The Social Costs of Climate Change: Greenhouse Damage and Benefits of Control." In Bruce, J.P., H. Lee, and E.F. Haites eds., *Climate Change 1995: Economic and Social Dimensions of Climate Change*. Cambridge, UK: Cambridge University Press.

Rodrick, D.. 1999. "Where Did All the Growth Go? External Shocks, Social Conflict, and Growth Collapses." *Journal of Economic Growth* 4(4):385–412.

Sathaye, J., A. Najam, C. Cocklin, T.C. Heller, F. Lecocq, J. Llanes-Regueiro, J. Pan *et al.* 2007. "Sustainable Development and Mitigation." In IPCC (2007c).

Schelling, T.C. 1995. "Intergenerational Discounting." *Energy Policy* 23(4/5):395–401.

_____. 2006. "Global Warming: Intellectual History and Strategic Choices." Fourth Annual Hans Lansberg Memorial Lecture, Resources for the Future.

Shibata, H., and S.J. Winrich. 1983. "Control of Pollution When the Offended Defend Themselves." *Economica* 50:425–37.

Smit, B., I. Burton, R.J.T. Klein, and J. Wandel. 2000. "An Anatomy of Adaptation to Climate Change and Variability." *Climatic Change* 45:223–51.

Sobel, R.S., and R.G. Holcombe. 1996. "The Impact of State Rainy-day Funds in Easing State Fiscal Crises During the 1990–1991 Recession." *Public Budgeting and Finance* 16(3):28–48.

Stern, N. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge, UK: Cambridge University Press.

Tol, R.S.J. 2002a. "Estimates of the Damage Costs of Climate Change, Part I. Benchmark Estimates." *Environmental and Resource Economics* 21(1):47–73.

_____. 2002b. "Estimates of the Damage Costs of Climate Change, Part II. Dynamic Estimates." *Environmental and Resource Economics* 21(2):135–60.

Tompkins, E.L., and N.E. Hultman. 2007. "Funding Adaptation to Climate Change: Are the Emerging Institutions for Financing Adaptation already too Inflexible?" Working Paper. Oxford University Centre for the Environment.

Weitzman, M.L. 2007. "On Modeling and Interpreting the Economics of Catastrophic Climate Change." AEI-Brookings Joint Center Working Paper 07-11.

World Bank. 2002. *World Development Report 2003: Sustainable Development in a Dynamic World: Transforming Institutions, Growth, and Quality of Life*. Washington, DC: World Bank.

Yohe, G.W., R.S.J. Tol, R.G. Richels, and G.J. Blanford. 2008. *Copenhagen Consensus 2008 Challenge Paper: Global Warming*. Frederiksberg: Copenhagen Consensus Center.

Nuclear Power and Sustainable Energy Policy: Promises and Perils

Ioannis N. Kessides

The author examines the challenges and opportunities of nuclear power in meeting the projected large absolute increase in energy demand, especially electricity, throughout the industrialized and developing world, while helping to mitigate the threat of climate change. A significant global nuclear power deployment would engender serious risks related to proliferation, safety, and waste disposal. Unlike renewable sources of energy, nuclear power is an unforgiving technology because human lapses and errors can have ecological and social impacts that are catastrophic and irreversible. However, according to some analysts, advances in the design of nuclear reactors may have reduced their associated risks and improved their performance. Moreover, while a variety of renewable energy sources (hydro, wind, modern biomass, solar) will play important roles in the transition to a low-carbon economy, some analysts perceive that nuclear power is the only proven technology for generating electricity that is both largely carbon-free, not location specific (as with wind, hydro and solar), and amenable to significant scaling up. Thus given the projections of threats from climate change, and if the considerable strain experienced by world energy markets in recent years is a harbinger of things to come, then there is a rationale for examining the pros and cons of nuclear power as a supply option within low-carbon strategies. It should be noted that despite the emerging centrality of climate change and security of supply in the energy policy debate, nuclear power is still viewed with a great deal of skepticism and in fact continues to elicit considerable opposition. Indeed the views on nuclear power in the context of sustainable energy policy are highly divergent. A thorough evaluation of all aspects of the issue is warranted. JEL codes: L52, L54, L94, L97, L98

The world is facing an enormous energy challenge. Despite continuous declines in energy intensities, population growth and rising incomes in developing economies are stimulating substantial global energy demand. An adequate, secure, clean, and

The World Bank Research Observer

doi;10.1093/wbro/lkp010 Advance Access publication July 30, 2009 25:323–362

competitively priced supply of energy is vital for sustainable development. Without it, the modernization, growth, and international competitiveness of developing economies will suffer. In meeting this energy demand, however, the world is faced with two major dilemmas: climate change and security of energy supplies.

Climate change is rapidly becoming the defining environmental, economic, and political challenge of our era. It poses a significant threat to global ecosystems and it could have devastating impacts on such climate-sensitive sectors as agriculture, forestry, public health, water supplies and coastal areas. Climate change could undermine the stock of natural capital and inflict serious damage to human systems, and thus reduce the overall productivity of the capital needed for socioeconomic development. These adverse impacts will be most striking in developing countries because of their greater dependence on natural resources and their limited mitigative and adaptive capacities. Ultimately, climate change could compound existing poverty and undermine sustainable development.

According to the International Energy Agency (IEA), world primary energy demand in the Reference Scenario (in which government policies are assumed to remain unchanged from mid-2008) is projected to grow by 45 percent between 2006 and 2030. Nearly 90 percent of the increased energy demand will come from non-Organisation for Economic Co-operation and Development (OECD) countries, led by China and India, between them accounting for more than half the total increase. Global electricity demand is projected to nearly double from 15,665 terawatt-hours in 2006 to 28,140 terawatt-hours in 2030, with nearly 80 percent of the increase in non-OECD countries (IEA 2008b).[1] To meet these needs, the world's electricity-generating capacity will have to increase from about 4,343 gigawatts in 2006 to 7,484 gigawatts in 2030. In 2006, coal-fired generation accounted for 41 percent of world electricity supply; in 2030, its share is projected to rise to 44 percent. Coal is the most carbon intensive of the fossil fuels: a 1 gigawatt coal-fired plant emits approximately 10 million tons of carbon dioxide a year. Without specific policies to limit greenhouse gas (GHG) emissions, energy-related carbon dioxide emissions would rise from 28 to 41 billion tons per year during the same period, with 97 percent of the increase in non-OECD countries, with the share of coal rising from 42 to 46 percent—a scenario which in the face of increasing concerns about climate change could be deemed environmentally unsustainable (IEA 2008b).

As evidence mounts on the threats of climate change, pressures for curtailing carbon dioxide emissions from coal-fired electricity generation are likely to escalate sharply. This gives rise to one of the central challenges in global energy policy: in the context of a carbon-constrained world, with coal and to a lesser extent natural gas being limited in their future growth by policy decisions to limit carbon dioxide emissions, what sources will provide the estimated additional 3,140 gigawatts of new electricity generation capacity that it is estimated the

world will need by 2030? Since the bulk of that additional capacity will be required in the developing world, whether and how this challenge is to be met is a dilemma of unprecedented complexity and magnitude for sustainable development.

A number of energy sources and technological options exist—though with major environmental, social, and economic tradeoffs. Immediate reductions in GHG emissions can be achieved through lowering demand by increasing the efficiency with which electricity is used and by expanding the deployment of existing renewable generation technologies. However, it is argued that these benign measures will not be enough to make ends meet (Pacala and Socolow 2004; Frois 2005). Even restraining the increase in per capita energy consumption in the face of growing energy demand in developing countries will require very substantial improvements in energy efficiency. Hydropower is cost effective in a number of locations, but utilization of potential new sites is likely to be limited given that these sites are often less accessible and precious for environmental and social reasons. A major expansion of biomass fuels would require vast land areas for cultivation, in competition with increasing food production and the preservation of natural ecosystems. And the cost characteristics of solar photovoltaics (PV) are still highly unfavorable (EC 2007), except in off-grid locations where the costs of alternatives are even higher. Although there is considerable interest in concentrating solar power (CSP), it is only commercially semi-mature and involves costs and location constraints on the delivery from source to demand.

The most promising renewable technology for the near to medium term is seen by many to be wind power, which is already near commercial viability and is achieving high penetration rates in some countries (for example, Denmark, Germany, and Spain).[2] When combined with hydro storage—and where a wide area grid can even out local fluctuations in wind strength, as in the United Kingdom and Europe—problems of intermittency can be handled up to appreciable shares in total generation. While many developing countries have a substantial wind resource—Central America, Chile, Brazil, Pakistan, Mexico, Mongolia—in others wind resources are less satisfactory and would require substantial complementary investments in transmission and reserve capacity.[3]

Fossil fuels can be decarbonized, but the underlying technology is in the early stages of development and is clouded by large uncertainties regarding the speed of implementation and the ultimate feasibility of large-scale application of carbon capture and storage (CCS). That should change in the near future as commercial scale pilot installations are under consideration in a number of countries, and experience of social acceptability of on-shore transmission of the carbon dioxide is gathered. More to the point, CCS is doubly capital intensive, both in terms of the extra equipment to handle the carbon dioxide and the lowered efficiency of the overall plant. Current cost projections suggest CCS would require a carbon dioxide

price of U.S. $50/ton to be commercial in the United States (and higher where coal costs and quality are lower, as in South Africa, China, and India), making it costlier than some other low-carbon options (MIT 2007). That said, provided that CCS can be technically proven, it will be needed on a large scale, and will need either a continuing high carbon price to ensure that it is used, or a mandated position in the merit order and close monitoring of carbon dioxide emissions. There is the telling point that attempts to discourage the consumption of fossil fuel does not remove the temptation to extract that fuel in the future (and might increase it by lowering the price). The obvious way to ensure that the carbon content of that fuel does not enter the atmosphere is to separate it and bury it—hence the ultimate need for widespread CCS.[4]

The energy environment for the 21st century is opaque and uncertain—especially in developing countries. Future energy prices are particularly hard to forecast with any precision, and even more so is the future of carbon pricing and climate change agreements. Providing sufficient energy to meet the requirements of a growing world population with rising living standards will be a challenge. Doing it without substantially exacerbating the risks of climate change will be even more challenging. It is likely to require a significant shift in the historic pattern of fossil-fuel use and a transformation of the global energy system. Technological development, however it unfolds, will almost certainly play a critical role in minimizing the costs of various GHG stabilization constraints to the global economy. The most recent research suggests that there is no obvious "silver bullet": the solution will comprise a variety of technologies on both the supply and demand side of the energy system (Richels and others 2007). In the face of significant technological and market risks and uncertainties, prudence calls for technological diversification.

A broad portfolio of low-carbon technologies and energy sources needs to be investigated and developed as part of a general strategy to confront the growing energy problems of both industrial and developing countries (EC 2007). Fuel switching, price-induced conservation and end-use efficiency, and expanded deployment of existing renewable technologies can provide significant opportunities for addressing carbon dioxide emissions and energy security in the shorter term, albeit at a potentially significant price tag depending on the speeds of deployment (IPCC 2007a).[5] Over the longer term, it is very likely that other technologies will need to play a large role in managing the costs of transition to a low-carbon economy. Two such technologies that are considered likely by many to do some of the heavy lifting in the future are CCS and nuclear power. Major technological shifts like advanced nuclear power may require a long transition as learning-by-doing accumulates and markets expand. While nuclear power has low running costs and hence, once built, is attractive to continue to run, CCS attached to coal-fired generation raises running costs and without a strong

carbon price signal it will be subject to bypass. Renewables continue to fall in price, and technological change, particularly for solar power, may make them commercially competitive at the kinds of carbon prices that reflect the likely social cost of climate change damage (Stern 2007). They also share with nuclear power the attraction of low variable costs and hence a continuing incentive for use once built.

This paper focuses mainly on nuclear power because amongst the low-carbon technologies it is the most controversial and susceptible to instinctive rejection. My main objective is to review the issues that are relevant for the analysis of the pros and cons of nuclear energy. I am not arguing that other low or zero carbon technologies are inferior, only that a balanced portfolio approach that minimizes not just the expected cost but also the risk of excessive costs, will likely include nuclear power.

Electricity Demand and Alternative Sources of Electricity Supply

The life-cycle GHG emissions per unit of electricity from nuclear power are much lower than from either coal or gas generated power and comparable to those from solar and other renewables. Although nuclear power is a well-established technology for generating electricity, it has long been considered unattractive by many environmental groups and ordinary citizens. These unfavorable attitudes emanate from concerns about the potential hazards of reactor meltdowns (and their potentially catastrophic ecological and social impacts), unresolved issues related to nuclear waste disposal, and potential problems with diversion and proliferation of fissile material. Still, there is a revival of worldwide interest in nuclear power—a result of rapidly rising and volatile fossil fuel prices, concerns about the security of energy (particularly gas) supplies, and global climate change.

Proponents argue that, in relation to the objectives of mitigating the threat of climate change, resource efficiency, and supply security, nuclear power performs very well. Nuclear power is seen as: (a) a well-established technology for generating electricity that has life-cycle GHG emissions as low as the best renewables such as wind; (b) amenable to significant scaling-up and thus can provide large amounts of power; and (c) using a natural resource (uranium) which is found in abundance (2–3 parts per million, ppm) in the earth's crust and, with advanced technologies, enough fuel could be provided to meet the world's electricity needs for several centuries (IAEA 2006; WNA 2008).

Skeptics claim that nuclear power is a costly, complex source of energy (Thomas 2005; Greenpeace 2007). It involves the use of highly toxic materials

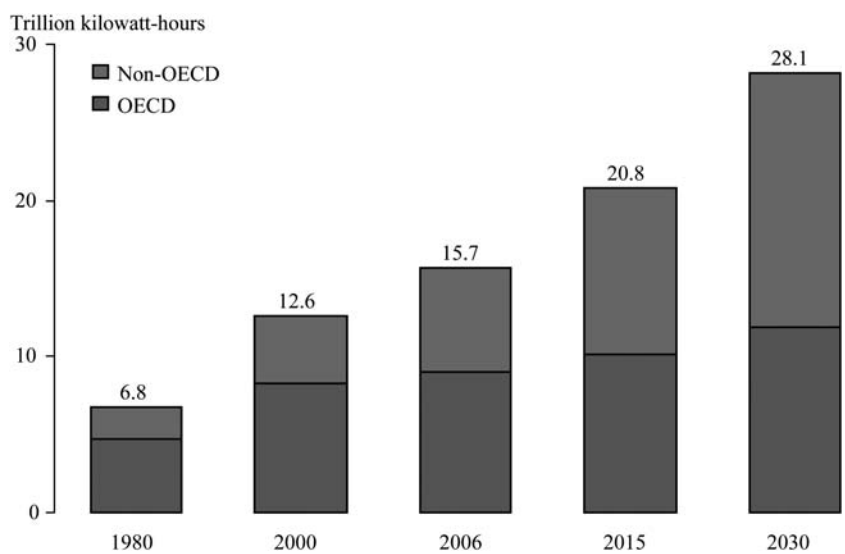
that must be kept secure from attack or theft; and a demonstrated viable technology for the permanent disposal or reprocessing of spent nuclear fuel does not yet exist. Moreover, even in a carbon-constrained world, nuclear power may be less economically attractive than many renewable options in favorable locations, as well as a host of decentralized energy efficiency and distributed generation technologies (Lovins 2005).

Public opposition to nuclear power facilities and, in particular, their association with nuclear weapons is entirely understandable (UCS 2007). For nuclear power to gain greater public acceptance, become a significant option for mitigating GHG emissions, and meet growing needs for electricity supply, four critical challenges must be overcome: safety, waste, proliferation, and costs (MIT 2003).

Projected Growth in Global Energy Demand

Over the next several decades global energy use is expected to rise substantially. Approximately one-half of the projected increase in world energy consumption in the IEO2008 (EIA 2008) reference case is attributed to electricity demand. Indeed, net electricity consumption is projected to nearly double between 2006 and 2030 (figure 1).[6] Over three-quarters of the reference forecast increase in electricity demand is expected to come from non-OECD countries, where

Figure 1. World Final Electricity Consumption between 1980 and 2030 (trillion kilowatt-hours)



Source: IEA (2008b).

electricity consumption is projected to increase by 3.8 percent a year during 2006–2030 (compared with 1.1 percent in OECD countries), fueled in part by the doubling of urban populations over this period (IEA 2008b, EIA 2008).

Fossil Fuel Supply Issues

Fossil fuels (oil, coal, and natural gas) supply most of the world's energy. Under business-as-usual scenarios, oil—the most convenient and multipurpose of these fuels—will continue to be the main source of energy, specifically for the transport sector, up to 2030. For electricity generation, coal and natural gas will remain the primary energy sources, with coal possibly increasing its market share from 41 to 44 percent (IEA 2008b) or 46 percent (EIA 2008).

Based on current consumption, proven economic global reserves of coal will last another 143 years, natural gas 63 years, and oil 43 years (EIA 2008). Probable reserves for oil and natural gas are substantially higher—by some estimates twice as great. Extracting these reserves will eventually become more difficult and costly, putting upward pressure on prices—though how quickly remains a matter of debate. Regardless of how rapidly prices rise, however, their volatility is likely to continue or increase. Indeed, the considerable strain experienced by world oil markets in recent years could be a harbinger of things to come. This recent market stress was due to the tight balance between supply and demand, a legacy of a period of under-investment during the previous decades of low prices, declining production from non-OPEC fields, and rising resource nationalism. The natural gas reserves needed to meet the growing demand are costly to access, constrained by the capacity of pipelines and liquefied natural gas (LNG) tankers, and often needing routes through politically contested regions.[7]

There may be abundant unconventional fossil-fuel alternatives to oil and natural gas; and there is a consensus that undiscovered global coal reserves are huge. Oil sands and shale represent very significant potential fossil resources. In addition to being plentiful sources for power generation, these resources can be transformed into gaseous and liquid fuels using known technologies at manageable costs (Bartis and others 2008). The problem is that all of these options are much more carbon dioxide intensive than conventional oil and gas alternatives, and even with CCS their carbon dioxide intensity is no lower than conventional fossil alternatives. Thus, they carry an even heavier GHG burden than conventional fuels. They also involve a range of local environmental problems from fuel extraction. The availability of these secure, economically feasible, yet environmentally unattractive options lies at the heart of the dilemma facing energy policymakers today.

The Challenging Transition to New Energy Sources

Market systems exhibit remarkable adaptability and flexibility, and market forces can play a crucial role in conserving scarce energy resources—directing them to their most valuable uses. The same price signals that balance supply and demand in the short run will signal profitable opportunities for, and help unlock, new approaches to energy production and use in the long run, provided they are allowed to respond to demand and supply—but too many developing countries respond to price shocks by capping domestic prices. In 2007, the IEA estimates that the 20 largest non-OECD countries paid energy subsidies of U.S. $310 billion. The world's economies will change in response to rising oil and natural gas prices. Shifts in the structure of economic activity and productivity improvements will reduce the energy intensity of those economies.[8]

Throughout history there has been a tendency to underestimate grossly the potential for technological developments that dampen the predicted doom from natural or human-made calamities. Technologies for energy supply (from oil exploration and recovery to solar PV), distribution, and use will likely improve significantly due to the improved incentives to engage in the relevant research. And if history is any guide, technology will help replace oil and natural gas use with cheaper alternatives before supplies run out as their prices rise.[9] But this transition to new energy sources will take time. And, unlike other such transitions in the past, it will be severely constrained by what is quickly becoming one of today's defining economic and political challenges—climate change due to GHG emissions.

Pressures Curtailing Carbon Dioxide Emissions from Coal-fired Electricity Generation

Approximately two-thirds of the world's electricity is generated using fossil fuels.[10] In 2006, fossil-fuel power plants emitted 11.4 gigatonnes of carbon dioxide accounting for 41 percent of the world total. In the IEA Reference Scenario, the carbon dioxide emissions of electricity will reach 16 gigatonnes in 2020 and 18 gigatonnes in 2030. Its share of world total carbon dioxide emissions will increase to 44 percent in 2020 and 45 percent in 2030. Electricity generation is projected to contribute half of the projected increase in energy-related carbon dioxide emissions to 2030 (IEA 2008b). Per unit of output, coal-fired electricity generation produces nearly twice as much carbon dioxide as does natural gas. In 2005 coal overtook oil as the leading contributor to global energy-related carbon dioxide emissions (EIA 2008).

The cost of heat energy (in importing countries) from coal has varied by a factor of six between 2000 and 2008—from U.S. $33/ton to U.S. $200/ton in north-west Europe, or from U.S. $1.4 to U.S. $8 per million British thermal units

(BTUs) (IEA 2008c). Coal within exporting countries is often considerably cheaper, particularly if it is of lower quality and far from ports, for example in Northern China and central India or near the main coalfield belt in Southern Africa. Gas prices have been similarly volatile over this period and also vary widely across countries, as it is costly to ship over long distances. U.S. gas prices for power production have varied from less than U.S. $3/million BTU in 2002 to nearly U.S. $12 in June 2008 before falling back to U.S. $5 in January 2009, while oil has been if anything even more volatile (and considerably more expensive for power production). Thus coal is generally the economic fuel of choice for baseload electricity generation—especially in countries with large deposits. Coal is not nearly as vulnerable to supply disruptions and price shocks as oil and natural gas, given the low costs of storage and the fact that most coal consumed is produced domestically. Thus, under normal circumstances, in countries with domestic coal reserves increased reliance on coal would be the obvious response to looming shortages of oil and natural gas and risks of interruptions in their supply. But converting all the coal in the ground into energy would generate enormous carbon dioxide emissions, with dangerous consequences for the earth's climate.

After years of intense debate, the question of whether climate change can be attributed to anthropogenic factors is almost resolved (IPCC 2007b). Extensive data show that anthropogenic emissions over the past century have already caused a significant rise in global temperatures, relative to temperatures in their absence. Including 2006, six of the seven warmest years on record have occurred since 2001 and the 10 warmest years have occurred since 1995 (Princeton Environmental Institute 2007). Thus prudence dictates reversing the current course of rising energy-related carbon dioxide emissions. As evidence accumulates on the extent of human augmentation of the natural greenhouse effect, demands to reduce the carbon dioxide emissions from coal will increase considerably. In the IEA Reference Scenario, total coal-fired electricity generation is projected to increase from 7,756 terawatt-hours in 2006 to 14,600 terawatt-hours by 2030. In the 550 Policy Scenario, it reaches 9,700 terawatt-hours in 2030, 33 percent less than the Reference Scenario. In the 450 Policy Scenario, the 1,382 gigawatts of conventional coal-fired plants in use today are, to a considerable extent, replaced with CCS coal plants and other low-carbon technologies, and conventional coal generation is one-third of the Reference Scenario (IEA 2008b).

Potential Impacts of the Curtailing of Carbon Dioxide Emissions on Developing Countries

Electricity is essential for producing almost all goods and services and so is vital for economic development. Reliable electricity services have become more

important because businesses and households use electric and electronic devices to perform an enormous range of tasks, basic and advanced. Thus adequate, reliable, and low-cost electricity is essential for modernization, growth, and international competitiveness in developing countries—and delivering it is among the most urgent challenges they face.

Today there is a huge global energy imbalance. Some 1.6 billion people lack access to electricity. Fortunately, efforts to curtail GHG emissions should not conflict with the aim of connecting the unconnected, as the following back-of-the-envelope calculation suggests. To supply an extra 1.6 billion people or 320 million households with an average household consumption of 1,200 kilowatt-hours/year will take 384 terawatt-hours/year or 2 percent of current consumption. For industrial countries, meeting the energy challenge of this century is central to maintaining high living standards; given the political will, low-carbon options are affordable. For developing countries, the energy and climate change challenges put at stake the very sustainability of their economic development.

The Hurdles of Decarbonization and Sequestration Technologies

Carbon sequestration offers a possibility for reducing the carbon dioxide emissions associated with fossil fuel use, based on a two-step process of CCS. Proponents argue that there are no major technical obstacles to sequestering carbon dioxide in deep geologic formations or in the ocean (MIT 2007).[11] Several pilot-scale plants have been or are being built (in the 10–25 megawatt range, compared to the 300–500 megawatt scale for commercial deployment), and there is growing experience of storing carbon dioxide in off-shore reservoirs such as in the Sleipner field. Various European union (EU) countries such as the United Kingdom are committed to larger scale trials in the near future, and the UK Climate Change Committee has argued that no coal-fired generation in the United Kingdom can operate without CCS post-2020, so that new-build coal-fired plants will need to be designed capture-ready. Until experience of commercial scale trials are available, the details on how and when this technology will be deployed must remain somewhat speculative. What is clear is that its required use will be massive, with hundreds of gigatons of carbon being stored over the course of this century and beyond (Dooley and others 2002).

The storage available in geologic and ocean reservoirs is probably sufficient to make CCS a viable mitigation option. But there are unresolved issues related to public perceptions about the long-term stability and environmental safety of such storage. Rapid escape of large amounts of carbon dioxide from geologic formations could pose a serious risk of asphyxiation to humans and animals, and cause substantial mechanical damage due to the huge amounts of compressive energy that

would be stored in a typical reservoir.[12] Against that, natural gas has been stored under pressure in large reservoirs for geologic periods of time and clearly does not raise the same concerns. Gradual leaks may pose little danger, but they could still defeat the mitigation goals of sequestration.[13] Additional concerns include acidification of groundwater from carbon dioxide.

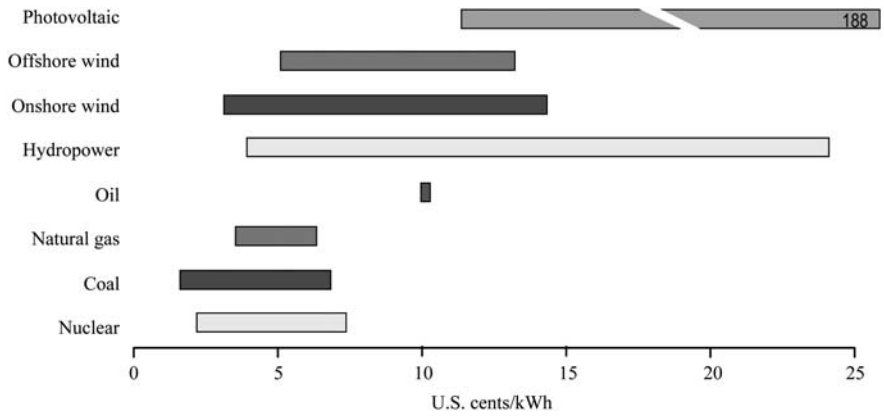
Retrofitting existing coal-fired plants to capture GHGs is very costly, although the costs can be reduced by designing the plant to be "capture-ready." Plants using "clean coal" technologies, such as integrated (coal) gasification combined cycle (IGCC) and other pre- or post-combustion plants with CCS that also remove sulfur dioxide and nitrogen oxide emissions, are therefore more likely to be built from scratch. These new technologies, including sequestration, would raise the cost of generating electricity from coal by about half. Nevertheless, in a range of low-emission scenarios, CCS is predicted to play an important part in addressing the problem. The IEA scenarios contemplate actions to stabilize carbon dioxide concentrations at either 550 ppm (ppm of carbon dioxide concentration in the atmosphere at which there is still a considerable risk of dangerous global warming) or 450 ppm (which lowers that risk considerably, but which is significantly more challenging). In both cases electricity is the key sector to decarbonize, as it requires the least change to behavior or other technology—low-carbon or "green" electricity is identical to electricity from any other source. Even under the 550 ppm scenario, the power sector will need to construct between 2021 and 2030 130 gigawatts of plants with CCS, 127 gigawatts of new nuclear plants, 450 gigawatts of wind turbines, and 250 gigawatts of hydropower. Thus in the decade to 2030 CCS would contribute about the same carbon saving as new nuclear power (IEA 2008b).

The Promises and Limitations of Renewable Energy Sources

Renewable energy sources are indigenous and abundant—and their use could significantly increase the long-term security of energy supplies and reduce GHG emissions. Governments have recognized that for the less mature technologies (solar, off-shore wind, wave, and tidal stream) current costs are both high (figure 2) and a poor guide to what is ultimately achievable.[14] Thus, they are actively supporting such technologies, through carbon prices, feed-in tariffs, and renewables obligations, and are beginning to address problems of intermittency through better grid design and management.

In 2006 renewables-based electricity generation amounted to 3,740 terawatt-hours or 18 percent of the total output. In the IEA Reference Scenario, it is projected to increase to 4,970 terawatt-hours in 2015 (overtaking gas as the world's second-largest source of electricity) and over 7,700 terawatt-hours by 2030, or 23 percent of the total. In the 550 Policy Scenario, renewables will play

Figure 2. Levelized Costs Associated with New Construction for Different Electricity Generation Technologies (U.S. cents per kilowatt-hour)



Notes: The levelized cost of a project is equivalent to the constant dollar price of electricity that would be required over the life of the plant to cover all operating expenses, interest and repayment obligations on project debt, and taxes plus an acceptable return to equity investors over the economic life of the project. The levelized costs for different electricity-generating technologies in this figure are based on estimates from seven recent studies: DGEMP (2003), MIT (2003), Ayres and others (2004), METI (2004), RAE (2004), UofC (2004), and IEA/NEA (2005). The very broad cost ranges and the consequent overlap among alternative technologies are due to different assumptions across the studies regarding the weighted average cost of capital, plant construction time and economic life, load factor, fuel prices, and other pertinent economic and technical parameters. The rather precise estimate of oil reflects a particular assumption of that price, and while gas prices might move with oil prices, those of other technologies would be less tied to the oil price. The MIT study assumes that the debt–equity ratio and the nominal cost of equity are different for nuclear, coal, and natural gas plants and adopts the following discount rates: 11.5 percent for nuclear and 9.6 percent for coal and natural gas plants; similarly, the UofC assumes that the cost of capital is 12.5 percent for nuclear and 9.5 percent for coal and natural gas. The other studies employ a common financing model and the same discount rate for the different generating technologies: 7.5 percent in the RAE study, 10 percent in the CETI study, 8 percent in the DGEMP study, and 5 percent and 10 percent in the OECD/NEA/IEA study. The price of natural gas is assumed to be $3.50/MBTU and $3.39MBTU in the MIT and UofC studies, €3.30/MBTU by the DGEMP, £2.18/GJ in the Royal Academy of Engineers, and Can$6.47/Mcf in Ayres and others. All these figures are subject to considerable reservation, as the cost of capacity also depends on materials prices, financial conditions, and the allocation of risk, which can be affected by regulatory regimes.

Source: IAEA (2006).

a major role, supplying more than 30 percent of total electricity by 2030; they achieve even faster deployment in the 450 Policy Scenario, reaching 12,000 terawatt-hours or 40 percent of the total electricity generation in 2030 (IEA 2008b).

Hydroelectric power is the world's largest source of renewable electricity, with considerable unexploited potential. Less than a fifth of technically exploitable hydropower has been used to date (WEC 2007).[15] Much of the potential for new

development is in Asia, Latin America, Africa, and the former Soviet Union. In the IEA Reference Scenario, global hydropower capacity reaches around 1,400 gigawatts in 2030 compared to its current level of approximately 900 gigawatts; and hydropower generation is anticipated to increase by 60 percent from 3,035 terawatt-hours in 2006 to 4,810 terawatt-hours in 2030, when it would contribute 14.5 percent of the world total—40 percent more than nuclear energy, compared to 9 percent more in 2006. In the 550 and 450 Policy Scenarios, hydropower capacity increases further to 1,600 gigawatts and 1,922 gigawatts respectively by 2030 (IEA 2008b).

Hydropower has a very low GHG footprint, unless one counts the emissions from entrapped biomass, and it can give rise to important synergies with other renewable energies. Indeed, a natural synergy exists between hydroelectric generation and wind and solar power. Storage hydro and pumped storage can provide the firming capacity that is needed to smooth out the fluctuations and fill the supply gaps associated with these intermittent renewable resources. By enhancing the flexibility of the power system, the presence of quickly dispatchable hydropower could facilitate higher penetrations by intermittent renewable technologies without compromising system quality and security. Synergies might also develop with geothermal, marine power, and bio-generation, especially as these technologies undergo expanded deployment.[16]

Although hydropower is nonpolluting, it can have serious environmental impacts. Hydroelectric facilities, unless carefully located and managed, can disrupt natural river flows with adverse impacts on the health of important, instream ecosystems; flood riverside lands, destroying riparian and upland habitats and force the relocation of populations; cause irregular water releases, disturbing variations in seasonal flows that trigger natural growth and reproduction cycles in many species; and impede the natural flow of sediments, degrading downstream habitats. These environmental and social impacts could limit the scale of hydropower development relative to its physical potential. However, there are mitigation strategies that could play an increasingly important role in the future. Also, the more than 170 gigawatts of hydro capacity currently under construction is frequently part of multipurpose developments providing such important benefits as irrigation water, industrial and drinking water supply, flood control, and improved navigation (Bartle 2002). Moreover, as the anxiety about climate change increases, the perceived value of the various environmental amenities that may be adversely impacted by the expansion of hydro is likely to change.

Among nonhydroelectric renewable electricity generation sources, *wind power* is the world's fastest growing—increasing by 25 percent a year over the past few years. Global wind power capacity rose from about 6 gigawatts in 1996 to 74 gigawatts in 2006 and 94 gigawatts in 2007. In the IEA Reference Scenario,

global wind power capacity is projected to increase to 271 gigawatts in 2015 and 551 gigawatts by 2030; and global wind power output is projected to increase from 130 terawatt-hours (less than 1 percent of the total) in 2006 to more than 660 terawatt-hours (2.7 percent of the total) in 2015 and 1,490 terawatt-hours (4.5 percent of the total) in 2030. In the 550 Policy Scenario, wind supplies 1,947 terawatt-hours, or over 6 percent of the total, by 1930 (IEA 2008b).

Wind power is inexhaustible (as long as the sun shines on the earth unevenly).[17] Over the past 20 years its costs have dropped by more than 80 percent.[18] Thus utility-scale wind systems are rapidly becoming cost competitive relative to conventional fossil fuel generation plants, certainly at carbon prices that would be needed to deliver 550 ppm stabilization. The scale of wind power needed to make a material difference in stabilizing atmospheric carbon dioxide is considerable—the 550 gigawatts needed in the IEA Reference Scenario by 2030 represents 183,000 three-megawatt wind turbines (offshore turbines are likely to be larger, some onshore ones are smaller). Between 2025 and 2030 some 16,000 megawatts would be needed each year, requiring roughly 3,000–5,000 wind turbines (depending on size) or approximately 15 a day. Mackay's (2008) estimates indicate that in windy Britain in the best locations the average output might be around 16 gigawatt-hours per year per sq. km. Thus, to produce the 2,000 terawatt-hours per year in the 550 Policy Scenario, the land take would be 125,000 sq. km or 35 percent of the area of Germany, although offshore wind farms (projected to supply one-quarter of the total) would lessen the pressure on alternative land uses. Moreover, the intermittency of wind generation presents major challenges for integrating it with the electricity grid on a large scale. Deployment beyond 20 percent of grid capacity would require some combination of significant spinning reserve capacity, adding to generation costs; much greater transmission capacity to link areas with different simultaneous wind strengths; or low-cost, large-scale energy storage that does not yet exist.[19] Thus both the area of suitable land (and accessible offshore locations) and the intermittency of wind power will ultimately limit its penetration—although penetrations as high as 30 percent can be achieved depending on local and national grid strength.

The *solar* resource is ultimately the driver of all renewable energy, and is potentially enormous, providing only that it can be tapped efficiently and cost effectively.[20] The efficiency with which solar insolation can be transformed into energy varies widely, from less than 2 percent for biomass (before energy conversion losses) to a high of perhaps 20 percent for solar PV technology.[21] Although PV technology is advancing rapidly, it remains expensive relative to other electricity generation technologies, and the manufacturing of PV panels consumes significant energy. Moreover, PV technology suffers from diurnal variability and is least available when most needed (at least in temperate climates): in winter and after dark. In comparison with wind, its predictability requires less sophisticated

system management to accommodate and where central heating loads are high there is a natural complementarity, but otherwise, even if apparently cost competitive, there may be hidden costs in raising reserve requirements to deal with winter peaks.

MacKay (2008) provides useful data to estimate the potential contribution that PV might make in various countries. Thus equipping all Britain's south-facing roofs with 20 percent efficient (still high by current standards) PV would generate approximately 100 terawatt-hours per year or 30 percent of current electricity consumption. Covering 1 percent of Britain's land area with 10 percent efficient PV panels (which is still twice the output per area of existing solar parks like that in Bavaria) would produce about 200 terawatt-hours per year. More ambitious plans, such as those of DESERTEC (see www.desertec.org) that would cover North Africa with CSP linked to Europe by high voltage DC power lines, provide estimates of the "economic potential" at 620,000 terawatt-hours per year with coastal potential at 6,000 terawatt-hours per year. The latter requires some 300 high voltage DC interconnectors, with an average power density of 15 megawatts per sq. km, and therefore 45,000 sq. km, or 8 percent of the area of Germany or 2 percent of the area of Algeria.[22] The obvious problem is the cost and noncoincidence of supply and demand in time and space.

Geothermal plants can operate around the clock and so do not face the intermittency problems of wind power and solar PV systems. But their development is limited to certain volcanic areas, although optimists consider that drilling down to the magma could tap huge supplies (see http://www.magma-power.com) enough in the (geologically favored) U.S. for 500–5,000 years' energy supply. Mackay estimates that Britain might have access to minable hot rocks that could supply 20 terawatts per year, less than 10 percent of electricity demand. Thus geothermal resources are unlikely to become a major contributor to the world's energy supply, although they can make a sizeable contribution in specific locations.[23]

By contrast, *biomass* (agricultural and forestry residues, energy crops, and wastes) offers a huge, continuously replenishing energy source. Substituting biomass for coal in electricity generation can reduce emissions of carbon dioxide and other GHGs.[24] The main impediment to expanding the use of biomass in electricity generation is the availability and cost of feedstock. Increased demands on land and water risk raising the cost of food and further damaging vulnerable populations. Mackay (2008) estimates that best European practice can generate 0.5 megawatts per sq. km, which, if burned at 40 percent efficiency to generate electricity, would produce 1.6 gigawatt-hours per year per sq. km, or less than 2 terawatt-hours per thousand sq. km. If 10 percent of Germany were devoted to biomass at this efficiency, it would produce less than 60 terawatt-hours or 10 percent of their electricity consumption.

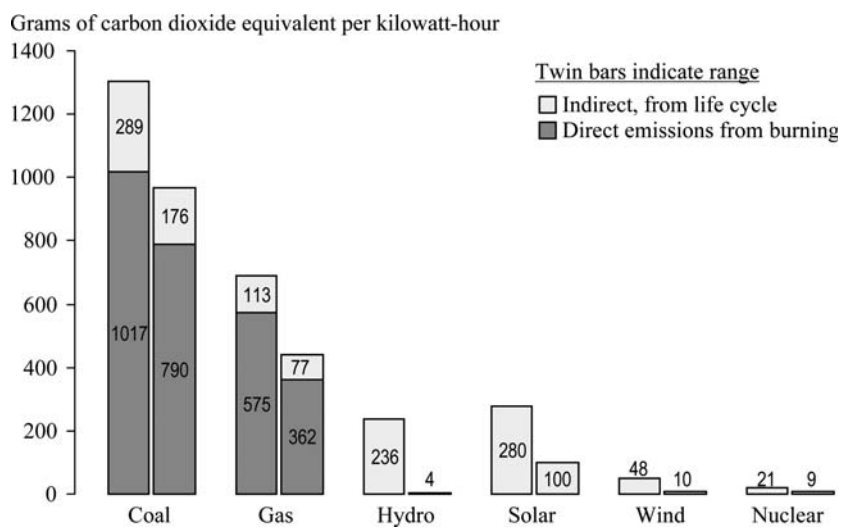
Nuclear Power: Promises

Several factors are driving the global resurgence of interest in nuclear power:

- A desire to make energy supplies more secure by diversifying fuel sources and reducing dependence on fuel imports, and to stabilize energy prices by reducing reliance on fossil fuels.
- Pressures to reduce air pollution: nuclear power does not produce the airborne pollutants that fossil fuels do.
- The increasing urgency of mitigating climate change by reducing GHG emissions: especially carbon dioxide (figure 3).

Proponents of nuclear power highlight the fact that, together with hydro and wind power, it is a currently proven producer of electricity that does not emit GHGs and can be scaled up quickly enough to replace baseload[25] fossil fuel electricity without the major environmental,[26] land use, or technological constraints impinging upon hydro, biomass, wind power, or solar—and so tackle the challenges of energy supply security and climate change.[27] At the current rate of use, with reactors operating on a once-through fuel cycle,[28] the world's present measured resources of uranium (4.7 million tons) will last about 70 years. The introduction of fast breeder reactors[29] and the recycling of plutonium from

Figure 3. GHG Emissions from Different Electricity Generation Technologies (grams of carbon dioxide equivalent per kilowatt-hour)



Source: IAEA (2000).

reprocessed spent fuel would increase the energy potential of today's uranium reserves by 50-fold or more (NEA 1998; IAEA 2006; UIC 2007a).

Energy security concerns were the main factor motivating the nuclear power build-up following the energy crises of the 1970s, especially in countries with limited energy resources. Such concerns have become even more pressing in recent years due to heightened geopolitical tensions.

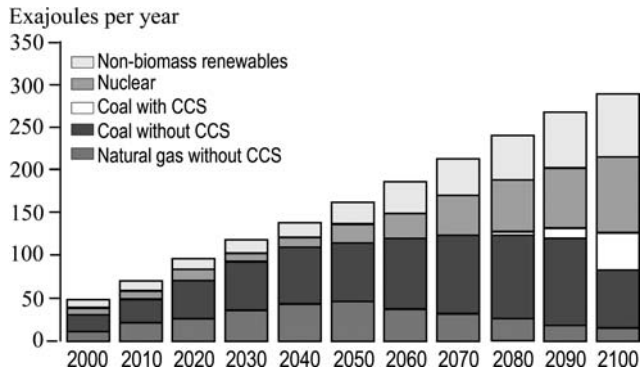
The world energy markets have experienced considerable strain in recent years. Robust economic growth, especially in some large developing countries, has caused a significant structural shift in global demand for fossil fuels and a tightening balance between supply and demand. Indeed, the recent rise in and volatility of fossil fuel prices reflect a tightening balance of supply and demand, while Europe's concerns over gas security of supply have been reawakened by Russia's failure to secure satisfactory transit agreements through their near-eastern neighbors, such as the Ukraine. Certainly the new EU accession states have placed considerable emphasis on reducing their import dependence on Russia, and many view nuclear power as part of that solution.

Reducing the Costs of GHG Stabilization Constraints

In recent years, a variety of models has been used to estimate the impacts on the gross world product of stabilizing GHG concentrations in the earth's atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system.[30] An important aspect of this modeling has been to assess the potential contributions that advanced technologies could make in achieving various atmospheric stabilization targets and minimizing the corresponding abatement costs.

While these studies employ different modeling approaches, technology representations and climate policies, most of them seem to indicate that nuclear power has the technical and economic potential to play a significant role in emissions mitigation.[31] This is true even at relatively modest mitigation targets (for example, carbon dioxide concentrations of 750 ppm; figure 4). In the IEA Reference Scenario, world nuclear capacity is projected to increase from 368 gigawatts in 2006 to 433 gigawatts in 2030 and nuclear generation to rise from 2,793 terawatt-hours in 2006 to 3,458 terawatt-hours during the same period—although its share of total electricity generation falls from 15 percent in 2006 to 10 percent in 2030. In the 450 Policy Scenario, nuclear capacity reaches 619 gigawatts and nuclear generation rises to 4,000 terawatt-hours, or 14 percent of the total, by 2030. And in the 450 Policy Scenario, nuclear power capacity reaches 680 gigawatts and supplies over 5,200 terawatt-hours, or 18 percent of total electricity generation, by 2030. Thus, as we move to more

Figure 4. Power Generation Shares in the 750 ppm (Carbon Dioxide Concentration) Scenario: MERGE Model (exajoules per year)



Note: One exajoule (EJ) equals 277,778 terawatt-hours.
Source: Clarke and others (2007).

stringent constraints on GHG concentrations, nuclear power makes a comparable contribution to CCS, hydro, and other renewables (IEA 2008b).

The costs of stabilization increase substantially if nuclear power is excluded from the supply options. Difiglio and Gielen (2004) used the IEA ETP model to analyze the average and marginal costs of emission reductions. They estimated the worldwide cost of not having nuclear power while achieving 550 ppm to be almost U.S. $300 billion per year. In the IEA ACT Map Scenario,[32] making nuclear power more available results in a U.S. $9 per ton of carbon dioxide reduction in marginal costs and moderates the estimated average increase in electricity prices during 2030–50 by lowering that average by 11 percent (IEA 2008b).

The Potential Benefits of "Keeping the Nuclear Option Open"

Previous work on the optimal degree of generating diversity has identified two principal macroeconomic benefits of fuel diversification and technology-mix: (i) nonfossil fuel technologies reduce fossil price risk and help avoid costly economic losses; and (ii) a diverse system is intrinsically more robust to supply shocks and thus diversification benefits security of supply (Stirling 2001; Awerbuch and Berger 2003). In the face of the current disturbing trends in climate change caused by the anthropogenic emission of carbon dioxide and other GHGs, diversification into generating technologies that do not emit such gases will have the added climate-change mitigation benefit.

Nuclear generation costs are fairly insensitive to oil, gas, and carbon prices.[33] Nuclear power could, therefore, offer a hedge to an electric utility against the uncertainty and volatility and risk of oil, gas, and carbon prices. This hedging and the flexibility to choose between nuclear power and other generating technologies, as new information emerges about fossil-fuel supply conditions and evidence accumulates on global warming, creates an option value for nuclear power (Graber and Rothwell 2006; Rothwell 2006; Rothwell 2007). This hedging value cannot be adequately taken into account in the context of the standard levelized life-cycle cost methodology. It requires a dynamic framework to capture fully the value of the flexibility of waiting for more information on the supply conditions of oil and gas and the policy toward carbon (Roques and others 2006). However, any evaluation of the diversification of nuclear power has to take explicitly into account the countervailing inflexibility arising from the fact that the nuclear fuel cycle produces dangerous radioactive waste that "essentially lasts forever."

Facilitating the Transition to a Low-oil Transport Economy

In recent years interest in a transition from oil-based liquid fuels to some low-carbon alternative, as the energy source for transportation, has grown markedly because of increased concerns about the volatility of oil prices, energy security, and climate change. Several candidates have been proposed, from bio-fuels in the near term, through hybrid electric vehicles in the medium run, to hydrogen as the energy vector used in fuel cells in some more distant vision (ANL 2003).

While the wide diffusion of hydrogen-based transportation would by itself reduce carbon dioxide emissions emanating from oil-based fuels, the net gain from such technological transformation depends critically on the production process of hydrogen fuels. Currently, most hydrogen is manufactured with an economical steam-reforming process from methane or natural gas. A key disadvantage of this process is that it releases large amounts of carbon dioxide, so the environmental advantage of the hydrogen fuel it produces is limited (Herring 2004). Using electrolysis to split water into hydrogen and oxygen is the simplest (although not the most economical) way to produce hydrogen, and it releases no carbon dioxide into the atmosphere. However, if the electricity needed to produce hydrogen electrolytically is generated using fossil fuels, again the environmental advantage is nullified. It is argued that nuclear energy could be part of the solution by generating the large quantities of electricity needed to facilitate the transition to a hydrogen-based transportation system (Baurac 2004).[34] It seems more likely that transport solutions will be derived directly from electricity (via batteries or in rail) than indirectly via hydrogen, at least in the next 20–40 years, again strengthening the case for expanding the supply of low- or zero-carbon electricity.

Nuclear Power: Perils

Two accidents have indelibly marked the history of nuclear power, leaving impressions in the public mind that, many years later, still affect reactions to this form of energy: the 1979 accident at the Three Mile Island nuclear power plant (in Pennsylvania, the United States) and the 1986 accident at Chernobyl (in Ukraine). The environmental and health effects of the Chernobyl accident were far more severe than those from Three Mile Island. Those differences confirmed the critical importance of inherent safety features—especially a strong containment building enclosing the reactor's primary system. Although about half of the reactor core melted at Three Mile Island, the released radionuclides were mostly deposited on the inside surface of the plant or were dissolved in condensing steam. The containment building prevented any significant release of radioactive material. Except for some early Soviet-designed systems, most nuclear power plants currently operating have such containment buildings.

After the Three Mile Island accident, significant improvements were made to both operating and new reactors, making them much less vulnerable to accidents—whether due to equipment malfunctions or human error. The industry has developed many new reactor designs. To improve safety, the general design trend has been to add passive features that ensure responses to equipment malfunctions and other mishaps are founded on basic physical laws such as gravity and natural circulation—rather than engineered systems (such as safety pumps) or backup systems that rely on power availability or human intervention (Ahearne 2006; USNRC 2007). In addition, there has been extensive global cooperation to improve the safety of Soviet-designed plants. So far, only two major accidents have occurred in more than 12,000 reactor-years of commercial operations in 32 countries. It is claimed that the newest plants would suffer no more than one incident of severe core damage for every 100,000 reactor-years of operations—and this without environmental release of radioactive material (UIC 2007b).

The Chernobyl accident was unquestionably severe. Approximately 6 percent of the radioactive content of the reactor core was released into the atmosphere. The accident led to 31 short-term deaths and an additional 106 individuals experienced serious radiation effects. Around 200,000 workers who were engaged in cleanup activities during 1986–87, received, on average, exposures double the yearly permitted occupational exposure and about the same level as individuals in high radon areas of Europe. The number of cleanup workers ultimately rose to 600,000, but most of the additional individuals received limited exposure. Some 116,000 inhabitants were evacuated from the 30 km exclusion zone around the plant site, but the majority (95 percent) of them received on average less exposure than the cleanup workers. The more than 400,000

inhabitants of areas classified as strict control zones received significantly less radiation exposure. The IAEA and World Health Organization estimated that out of the total of 1,116,000 individuals identified above, some 3,500 to 4,000 could eventually die from radiation-induced cancers, mainly late in life (Chernobyl Forum 2005; IAEA 2006). However, other groups have disputed these estimates as being too low (Greenpeace 2006).

Still, it is useful to compare the effects on human life of various electricity generation technologies per unit of electricity produced. A 1998 study by the Paul Scherrer Institut (commissioned by the Swiss Federal Office of Energy) of 4,290 energy-related accidents found that for each terawatt-year of production, hydropower caused 883 deaths, coal 342, natural gas 85, and nuclear power 8 (figure 5).

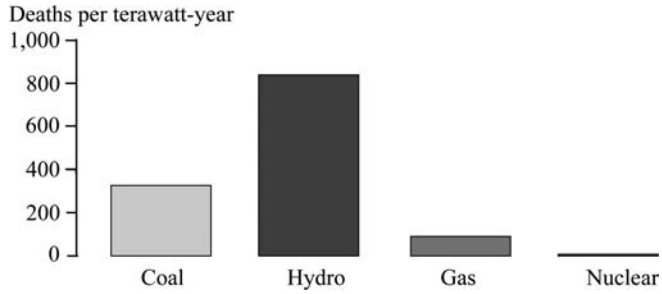
Nuclear power plants are also vulnerable to land-based assaults, deliberate aircraft crashes, and other terrorist acts. Such attacks could lead to the dispersal of large amounts of radioactive fission products, which could pose a direct radiation hazard, contaminate soil and vegetation, and be ingested by humans and animals.

The Challenges of Nuclear Waste Disposal

The nuclear fuel cycle produces a variety of radioactive waste, including low- and intermediate-level waste, transuranic waste, and spent fuel and high-level waste.[35] The spent fuel and high-level waste create by far the most serious problems and so dominate the debate.

For political, social, technical, and economic reasons, the issue of nuclear waste disposal is extremely complex. Today, more than 40 years after the first

Figure 5. Energy-related Fatalities of Different Technologies (deaths per terawatt-year)



Source: WNA (2009).

commercial nuclear plant became operational, no country has yet succeeded in disposing of high-level waste—the longest lived, most highly radioactive, and most technologically challenging of the waste streams generated by the nuclear industry (MIT 2003). Because such waste poses danger to current and future generations, the public is understandably apprehensive. Indeed, public attitudes toward nuclear energy are strongly influenced by concerns about waste disposal.[36] Thus the spent fuel or high-level waste management could and should be a significant consideration in the discussion of the potential for expanded nuclear deployment. Current research is examining the use of accelerator-driven systems to reduce the volume and radioactive toxicity of nuclear waste (Brolly and Vertes 2004; Gokhaleb, Deokatteya, and Kumar 2006). Because of the potential future uses of materials in spent nuclear fuel, disposal programs strive to ensure the retrievability of such waste, and research and development programs are aimed at achieving this goal.

The challenges of nuclear waste disposal remain a potent impediment to the expansion of nuclear power around the world. Some technical solutions for the safe storage of spent fuel have been proposed. Deep geologic formations are the preferred technical solution, and most countries have selected or are selecting sites as permanent repositories, with appropriate studies and permits. But institutional obstacles remain formidable, and the public's aversion to having a repository in its "backyard" continues. Citizens of countries planning to install or expand nuclear power units can be expected to ask questions about plans for safe, sustainable storage and disposal of nuclear waste.

Many countries have opted to reprocess spent fuel, not so much to extend the resource as to reduce and simplify radioactive waste management. In 2006 the U.S. administration announced that it would move toward that option. Some developing countries that seem poised to accelerate the expansion of nuclear power are apparently intending to store waste first onsite for a fairly long time—and expecting to ship spent fuel to its country of origin or to countries willing to accept it for a fee. Sending spent fuel back to the supplier for storage or other handling fits well into a strategy of nonproliferation, as it removes the potential for spent fuel to be used to produce nuclear weapons. However, such decisions may shift with changes of political regimes.

Presently, the global public remains deeply skeptical about nuclear waste disposal. There is considerable justification for the expressed skepticism. Although experimental and pilot facilities have been built, there are no operating high-level waste repositories and all countries have encountered significant difficulties with their waste management programs. Implementation of geologic repositories has proven a highly demanding task that has placed considerable stress on operating, regulatory, and political institutions.

The Proliferation Risks of Expanded Nuclear Deployment

All civilian nuclear power plants and their associated fuel cycles employ dual-use technology that can serve both military and nonmilitary purposes. Two points of the nuclear power cycle in particular form sensitive links between civilian uses and weapons applications: uranium enrichment and spent-fuel reprocessing.

The extent to which nuclear power will prove an acceptable and enduring option for meeting the future energy requirements in many regions of the world will depend in part upon the ability of the international community to minimize the associated proliferation risks (APS 2005). The common fear is that such an expansion will increase the risk that weapons-usable fissile materials, facilities, technology, or expertise might be diverted or stolen and thus make it easier for countries to acquire technology as a precursor to developing nuclear weapons capability or for terrorist groups to obtain nuclear materials. This risk could be further compounded by the likelihood that plutonium-fueled breeder reactors will be widely used to stretch uranium resources under expanded nuclear power deployment. Proliferation could prove to be the Achilles heel of a global nuclear energy expansion.

Technical (Intrinsic) Barriers to Proliferation

Over the past three decades several attempts have been made to develop alternative nuclear technologies and cycles with greater resistance to proliferation. These efforts have focused on:

- Advanced reactor designs, new fuels, or both which allow very high burnup and produce less plutonium than current reactors (such as, for example, the pebble-bed high-temperature, gas-cooled reactor).
- Breeder or particle-accelerator-driven reactors that collocate sensitive activities and processes with the reactor, and do not separate the plutonium from other actinides.

Analyses of various reactor cycles have shown that all have some potential for diversion—that is, there is no proliferation-proof nuclear power cycle (APS 2005). Reactor concepts that do not require refueling (have 15–20 years of core life), especially under a hub-and-spoke architecture, could enhance proliferation resistance (Wade 2005). Small innovative reactors (SIRs) have special attributes that make them more proliferation resistant than the larger, conventional nuclear reactors. These attributes include infrequent refueling, restricted access to nuclear fuel, and elimination of the host country needs or rationale to construct facilities that could be diverted from civilian to military purposes and ultimately used for clandestine production of nuclear material (Greenspan and Brown 2001).

Another advantage claimed for the SIRs is that they can be constructed more rapidly and track actual capacity needs, especially in developing countries, more closely. Thus some observers believe that there is significant scope for improving intrinsic barriers to proliferation through high-burnup fuels (including uranium and thorium), nonfertile fuels, closed fuel cycles, high-temperature gas-cooled reactors, and SIRs (Baker Institute 2001; INL 2002).

Complementary Institutional (Extrinsic) Measures

Over the past 35 years, IAEA's safeguards system under the Nuclear NonProliferation Treaty has proven fairly effective in restraining the diversion of fuel-cycle materials and facilities from civilian to military uses.[37] Indeed, the adoption of institutional measures to mitigate proliferation risks has played a key, if not dominant, anti-proliferation role. Consequently, there is an emerging consensus that most of the progress made toward improving proliferation resistance can be attributed to the increased authority accorded to the IAEA to detect clandestine facilities and undeclared operations within declared facilities. This includes the statutory authority provided through the "Additional Protocol" to the existing agreements governing IAEA's safeguards system (CGSR 2000).[38]

Under a robust global nuclear power expansion program, there will be increasing pressures on countries to reprocess and recycle. While reprocessing and recycling have advantages in terms of resource utilization and spent fuel disposal, they will require strong process safeguards against misuse, diversion, or theft. Reactor-grade plutonium is weapons usable, whether by unsophisticated proliferators or by states seeking nuclear weapons capability. Thus, the primary challenge is to account and control adequately for weapons-usable material during normal operations of the nuclear energy system; and to monitor, detect, and prevent process modification or facilities diversion to produce or acquire such material (USDoE 1997; MIT 2003).

International or Multinational Energy Parks

One potential way of mitigating the proliferation risks of expanded nuclear deployment might be through the adoption of hub-and-spoke configurations that restrict all sensitive activities (such as isotope separation of uranium or reprocessing of spent fuel) to large, international/regional energy parks which would export fuel, hydrogen, and even small (40–50 megawatts) sealed reactors to client states (Feiveson 2001; Kursunoglu and Mintz 2001). At the end of their core life (say 15–20 years) the reactors would be returned to the central park unopened. Thus during the years of operation there would be no refueling and consequently the client countries would need no fuel fabrication facilities and management capabilities. To the extent that such modular reactors would operate almost

autonomously, the hub-and-spoke architecture could reduce substantially the rationale and opportunities for countries to develop nuclear research laboratories and train technical specialists and scientists whose know-how could later be diverted to weapons activities (Feiveson and others 2008).

Although international energy parks and the hub-and-spoke nuclear architecture are technically feasible, they could prove politically very difficult to implement. Countries might view these arrangements as encroaching upon their energy independence. Moreover, the hub-and-spoke system would require the spoke countries to accept restrictions on their nuclear activities that would not be similarly imposed on the countries hosting the nuclear parks. Inevitably, such restriction would be viewed as being discriminatory, unless all countries (including the advanced industrial ones) were willing to accept a high degree of international control over their nuclear energy programs.

The Economics of Nuclear Power

In a deregulated global electricity marketplace, economics will be a key consideration in future decisions to build new nuclear plants. Thus assessing the forward-looking cost elements of nuclear power and the uncertainties underlying those cost estimates is key to evaluating its potential role in balancing the electricity supply and demand over the next several decades and mitigating the threat of climate change. Even if countries decide that the challenge of decarbonizing the electricity sector requires more state control, economics will continue to be important, although the perceived costs of risk might then be somewhat lower.

One of the fundamental problems underlying the debate on the potential role of nuclear power in meeting the future global energy needs relates to the continuing lack of consensus on what will be the costs of new nuclear-generating plants—and this is only likely to be resolved with accumulating information about the full costs of new nuclear build (Rothwell 1992).

Past trends would lead observers to assume a modest nuclear power component in the future global energy mix: the simple fact is that no new nuclear plants have been built in the United States and Europe for more than two decades. Skeptics point out that by the early 2000s, nuclear plants ended up achieving less than 10 percent of the capacity and 1 percent of the new orders (all from countries with centrally planned energy systems) that were forecast a quarter of a century before. They claim that this fact demonstrates nuclear power has not been commercially attractive (Lovins and Lovins 2001).

The history of construction of nuclear plants is replete with huge cost overruns. The specter of increasing construction times (table 1), especially from the late 1980s until 2000, and the consequent high construction costs was a major

Table 1. Construction Time of Nuclear Power Plants Worldwide

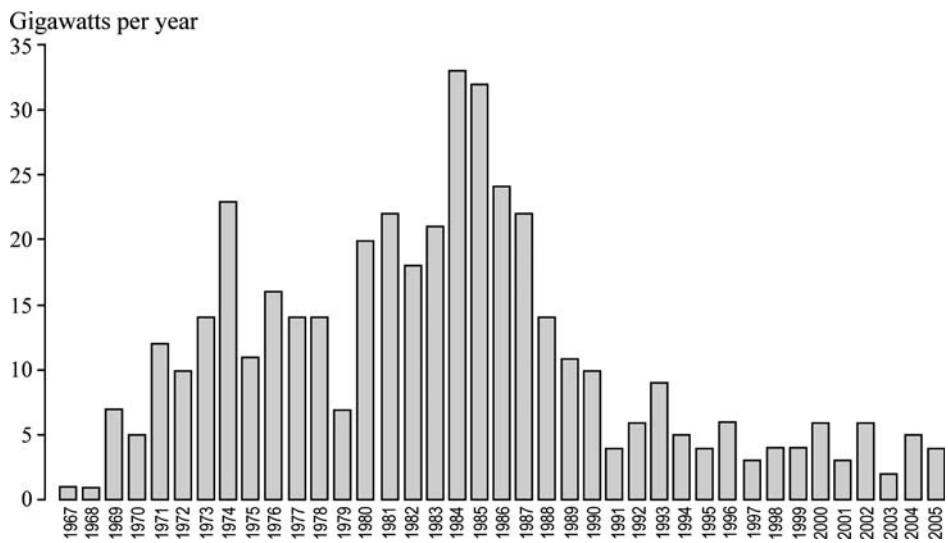
Period of reference	*Number of reactors*	*Average construction time (months)*
1965–70	48	60
1971–76	112	66
1977–82	109	80
1983–88	151	98
1995–2000	28	116
2001–05	18	82

Source: Clerici (2006).

cause of the lack of commercial interest in investment in new nuclear plants. This is evidenced by the significant decline in construction of new nuclear plants—from a peak of over 30 gigawatts of new capacity per year in the 1980s to an average of 4 gigawatts per year during the last decade (figure 6).

Part of this can be explained by the sharp fall in the real cost of relevant fuel and the dramatic improvement in gas turbine technology that undermined the apparent cost advantage of nuclear power compared to oil-fired generation in the aftermath of the 1970 oil shocks, while the escalation of construction costs had much to do with the growing safety concerns after Three Mile Island (and later Chernobyl).

Figure 6. Installation of New Nuclear Capacity (gigawatts per year)



Source: Greenpeace (2007).

Critics of nuclear energy point out that the promises made by some in the nuclear industry in the mid-1950s for cheap power ("power too cheap to meter") have seldom been kept (Greenpeace 2007). The nuclear industry is claiming that it has learned from its past mistakes. Joskow (2006a) argues that in recent years nonfuel operation and maintenance costs have fallen significantly, plant capacity factors have increased dramatically, and safety has improved considerably as well. Moreover, it is argued that improved big-project management techniques and new plant designs hold considerable promise for lower and more predictable construction times and costs, particularly for successor projects to the first of a kind.

The economics of nuclear power depend heavily on the long-term real discount rate, the cost of alternative fuels, the cost of carbon, and the cost of capital (including the time cost of the construction period). All of these factors had changed by early 2008, altering the economics of nuclear power in recent years. Specifically:

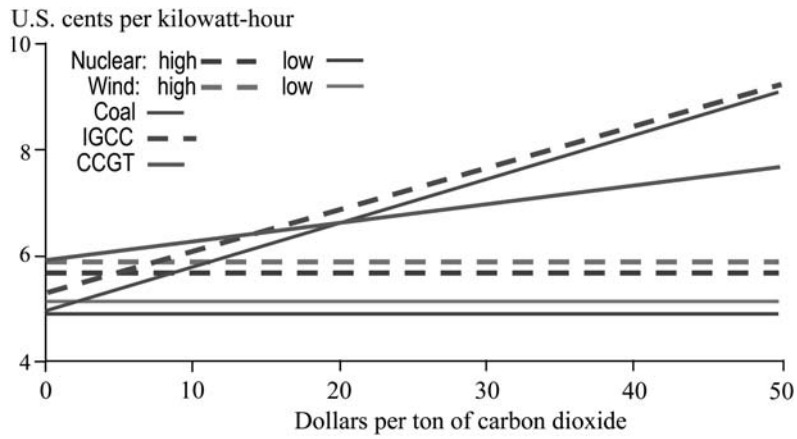
- The weighted average cost of capital had fallen, particularly for long-dated indexed instruments, where the demand by pension funds exceeds supply, driving down the longer-term rates.
- Oil prices, and to a considerable extent natural gas prices, had risen sharply and were not forecast to return to the previous benchmark of U.S. $25 per barrel. Gas (preferably pipeline, but even LNG) has been the preferred fuel for many new power projects, and these have become substantially more costly.
- Carbon prices in the EU had risen to €20 per ton of carbon dioxide (U.S. $90 per ton of carbon) partly in response to the rising price of natural gas, which increases demand for coal-fired generation and hence increases the demand for EU Emission Allowances, driving up the price of carbon.
- Standardized reactor design and streamlined licensing procedures held out the prospect of reduced construction time and costs, longer operating life, simpler operations and maintenance (O&M), and improved productivity of nuclear plants (in terms of capacity factors, operating costs, and safety indicators).

Since then the world's most serious financial crisis since the Great Depression has changed the financial landscape and put equity finance for new investment in a perilous state. Energy prices have fallen, the return to savers has collapsed, but the cost of borrowing has remained stubbornly high as credit risk has dramatically increased and projections of future GDP growth have been slashed, all developments that would appear inimical to capital intensive electricity investment, particularly for nuclear power. There are several possible reactions to this combination of circumstances. Major industrialized economies might reasonably argue that stimulating investment demand to match the need of the private sector to increase savings is a better solution to a recession than stimulating consumption and thereby increasing public and private net debt (that is, debt not offset by valuable

assets). Given that the real cost of borrowing to the public sector is low, that the demand for low-carbon energy is as pressing as ever, and the desirability of stimulating investment is therefore high, there is a case for redoubling efforts to increase investment in the low-carbon future. If this cannot be done readily with the current financial and energy market structures, other mechanisms may be needed. One potentially attractive possibility would be a public financial institution that would issue electricity bonds to the public, paying for example the value of 1,000 kilowatt-hours of retail electricity (suitably defined) per bond. These would have the attractions of indexed public debt (possibly greater, as they would provide a hedge against a particular and important price risk, that of volatile electricity prices). As such they could be issued at a low real interest rate and would provide an ideal form of finance for a reliable electricity generator, particularly one not dependent on the prices of fuel and carbon—such as renewables and nuclear power.[39]

The relative costs of nuclear power, coal, and combined cycle gas turbine (CCGT) generation have been extensively analyzed in several recent studies. There is a consensus that in deregulated markets, nuclear power is not now cost competitive with coal and natural gas without an adequate carbon price and suitable financial/contractual arrangements (Finon and Roques 2008). However, plausible reductions by industry in capital cost, operation and maintenance costs, and construction time could reduce the gap. Carbon emission credits, if enacted by government, can give nuclear power a cost advantage (MIT 2003; see figure 7).[40]

Figure 7. Impact of Carbon Value on Electricity-Generating Costs



Note: CCGT stands for combined cycle gas turbine and IGCC for integrated (coal) gasification combined cycle.
Source: Heller (2007).

While recent studies point to improved economics of nuclear power, their analysis highlights considerable uncertainty surrounding the economic appraisal of potential nuclear investments, mainly because of disparate forecasts of construction cost and fossil fuel and carbon prices. Part of the disparity relates to the date of forecast, as nuclear plant construction costs, even more than that of other generation plant costs, escalated very rapidly relative to other prices in the period 2005–08. With recent falls in steel and other prices, this may well change again. Other sources of uncertainty relate to the ever-changing information about the costs and delays of constructing the EPR Olkiluoto 3 plant in Finland.[41] There is an urgent need to model this uncertainty and assess the economic viability of nuclear power under a range of scenarios reflecting different assumptions relating to plant construction costs, costs of uranium and alternative fossil fuels and their real rates of escalation over time, and carbon prices or taxes.

Components of Nuclear Power Costs and Levelized Costs of Alternative Baseload Generation Technologies

The costs of nuclear power comprise four major components.[42]

- Capital or construction costs: The costs incurred during the planning, preparation, and construction of a new nuclear power station.
- O&M: It relates to administration, management, support and upkeep of a power station (labor, material and supplies, spares, insurance, security, planned maintenance and contractor services, licensing and regulatory fees, and corporate overhead costs).
- Fuel costs: It reflects the cost of fuel for the power station.
- Back-end costs: The costs related to the decommissioning of the plant at the end of its operating life and the long-term management and disposal of radioactive waste.

Much of the uncertainty surrounding nuclear power's future costs relates to construction cost—the most important component (roughly two-thirds) of total generating costs. The industry has a notoriously poor historical record on construction cost estimation, realization, and time to build. Indeed, the construction of most nuclear plants around the world has been plagued with delays and substantial cost overruns. In the United States, for example, the final costs of plants that commenced commercial operations in the late 1970s were in some cases several times greater than their initial cost estimates (table 2). Costs escalated dramatically as the industry scaled up its building rate and the size of plants. There has also been a global trend in increasing construction times (table 1). Prior research has identified several explanatory factors for the observed uncertainty and miscalculation or escalation of construction costs: (i) incorrect

Table 2. Historical U.S. Construction Cost Experience

Construction started	*Estimated overnight cost (kWe)*	*Actual overnight cost (kWe)*	*Percentage over*
1966–67	$560	$1,170	209
1968–69	$679	$2,000	294
1970–71	$760	$2,650	348
1972–73	$1,117	$3,555	318
1974–75	$1,156	$4,410	381
1976–77	$1,493	$4,008	269

Note: The overnight construction cost is the hypothetical cost of a generating plant if it could be built instantly ("overnight"): it does not reflect inflation, the costs of construction financing, or the length of time that it takes to build the plant and associated cash flows. There has been no new nuclear construction in the United States after the 1979 accident at Three Mile Island.

Source: Joskow (2006b).

understanding of economies of scale—early cost projections tended to ignore the potential diseconomies of scale due to the increased complexity and greater management requirements of larger nuclear plants; (ii) design flaws that necessitated costly redesign and caused significant construction delays which at a time of high interest rates substantially increased the cost of build; (iii) an unwieldy licensing process and burdensome and ever-increasing regulatory requirements, often changing in mid-course, leading to regulatory turbulence and construction delays; (iv) nonuniform designs which inhibited the exploitation of economies of volume and further compounded the complexity of the licensing process (Zimmerman 1982; Cantor and Hewlett 1988).

The nuclear industry has put forward very optimistic construction cost estimates based on simpler plant designs (including innovative reactors incorporating passive safety systems) and shorter build times facilitated by improved construction methods, standardization of reactors, learning-by-doing, and a more predictable licensing process. During the past 15 years there have been few new nuclear plants constructed around the world, and most of those have been in Asia, mostly in countries with low-wage costs, such as China and Korea. Thus, there is very limited actual cost experience to verify the industry's projections. Still, the standardization of France's and Japan's reactor programs, and the evolutionary design innovations in South Korea's nuclear plants, seem to indicate significant construction and operating learning effects.[43] And most recent studies assume construction cost improvements relative to historic levels (table 2) that closely match optimistic but plausible forecasts. Nevertheless, it is argued by some that the industry's quoted construction cost projections (based on largely non-transparent engineering cost calculations) for the future nuclear build should be viewed with some skepticism and be subjected to very vigorous analysis (Joskow 2006b).

Economies of Scale and the Economics of Smaller Sized Nuclear Reactors

Most of the nuclear reactors currently in operation are medium- to large-scale plants sized at 500–1,500 megawatts, utilizing tested technologies. The first generation nuclear power plants had a capacity of about 400–500 megawatts. However, because of the general belief that nuclear power operations are characterized by significant economies of scale at the plant level, there was a definite trend toward larger units. By the mid-1960s, the industry scaled up to about 800 megawatts and, before those units were completed, new ones with capacities of over 1,300 megawatts were planned and constructed (Cantor and Hewlett 1988).

What type of nuclear energy is likely to emerge under the expanded global deployment scenario? The power grids in many developing countries that could consider nuclear power are not large enough to support deployment of very large units. Power systems science dictates that no single unit should be larger than about 15 percent or at most 20 percent of system demand. This implies that the 1,000 megawatts size unit, the smallest of the three reactor types currently being actively promoted in the international market, cannot be considered in systems with a peak demand lower than about 5,000 megawatts. These considerations suggest that there may be considerable scope for small size reactors that would permit a more incremental investment than the large units of the past and provide a better match to the limited grid capacity of many developing countries (Schock, Brown, and Smith 2001).[44]

It is argued that smaller reactors could offer significant advantages in terms of overall simplicity, modularity and speed of build, passive safety features, proliferation resistance, and reduced financial risk. However, the other side of the coin is that such small plants may not be economic because they will not be able to exploit important economies of scale. And in any case, most of the cost estimates for small nuclear reactors have not yet been tested by actual construction and operation. Only time and experience will verify the claims made by vendors and the industry.

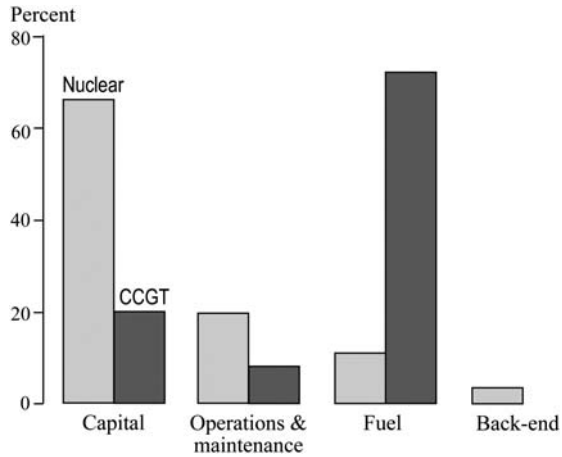
The Nexus between Electricity Market Structure and the Commercial Attractiveness of Nuclear Power

Until recently, no new nuclear power station had been commissioned in a liberalized electricity industry. In deregulated, competitive electricity markets, it is power-company investors and not ratepayers who must bear the bulk of the financial risk of new generating capacity. While some of the risks associated with the future value of electricity can be shifted to consumers and marketers with forward contracts, power-company investors will continue to bear some of the market risk and all of the construction cost, operating cost, performance, and

residual regulatory risk (MIT 2003; Joskow 2006a). In such a market environment, investors will naturally tend to favor less capital intensive and shorter construction lead-time investments. For nuclear power, construction accounts for most of the costs, whereas for gas-fired generation fuel is the largest component (figure 8).

Moreover, nuclear power lead times are, for engineering and licensing reasons, much longer than those of alternative baseload technologies—especially CCGT but also coal-fired generation. Thus, the redistribution of risk among different stakeholders is likely to make nuclear generation unattractive for private investors, even when its levelized costs are similar to those of the alternative technologies.[45] Taylor (2007) argues that the risk profile of merchant nuclear power with high fixed costs, typically fixed fuel-cycle costs and back-end costs, facing a volatile electricity price, requires a low debt–equity ratio as it is effectively more highly leveraged than conventional plants and that it has taken the collapse of British Energy to demonstrate this fact to the financial sector. Finon and Roques (2008) argue that the choice of contractual and financial arrangements for new nuclear build in liberalized markets will need to change, and that a consumers' consortium with credit-worthy participants (the Finnish model for Olkiluoto 3) is the most favorable arrangement. The model of indexed electricity bonds mentioned above represents a way of delivering the same advantages as the Finnish model (which depended on large industrial customers with long time horizons) and deserves investigation. The vertically integrated utility model that allows risks to

Figure 8. Cost Profile of Nuclear Power and Gas-fired Generation



Note: O&M, operations and maintenance.
Source: DTI (2007).

be shifted on to final consumers also has obvious attractions in lowering the effective cost of capital which is such a large element in the cost of nuclear power.

Summary

Increasing concerns related to energy supply security and widespread perceptions about the urgency of mitigating climate change are generating significant challenges in the global energy policy framework. Because of the shortage of low-cost substitutes for high carbon-emitting technologies, supplying the world's energy needs while stabilizing GHG emissions to prevent dangerous climate change is likely to prove a great challenge. If this challenge is to be met in a cost-effective way and not be disruptive to economic growth, it will almost certainly require contributions from a combination of existing, improved or transitional, and advanced technologies on both the supply and demand side of the energy system to (i) increase efficiency in electricity use; (ii) expand the deployment of renewable technologies; (iii) effect CCS; and (iv) expand nuclear power. However, many of these technologies confront substantial hurdles and would need to meet a complex array of conditions before they could be successfully implemented.

Within this context of the urgent need to develop a full arsenal of low-carbon energy technologies, nuclear power is viewed by some with a great deal of skepticism and in fact it continues to elicit considerable opposition from various electorates. Since the views on nuclear power remain highly divergent, the debate about its re-emergence requires a fresh, careful, and objective assessment of its costs (risks) and benefits. The consideration of nuclear power as a significant low-carbon supply option requires addressing four critical issues: cost, safety, waste, and proliferation. There is an urgent need for further research to assess what progress has been made on addressing these challenges and the implications for costs and benefits, so as to inform the debate on the viability of nuclear power as an important ingredient in the world's clean energy development in the coming decade.

Notes

Ioannis N. Kessides is a Lead Economist in the Development Research Group of the World Bank; his email address is ikessides@worldbank.org. The author is indebted to David Newbery, Gershon Feder, Daryl Field, Masami Kojima, Nataliya Kulichenko, Jessica Lin, Laszlo Lovei, Gordon McGerron, Lucio Monari, Mudit Nairan, William Nuttal, Burton Richter, Fabien Roques, Geoffrey Rothwell, Jamal Saghir, Gary Stuggins, Michael Toman, and three anonymous referees for their valuable comments and suggestions. The views expressed in this paper are solely the responsibility of the author and should not be interpreted as reflecting those of the World Bank, its executive directors, or any of its member countries.

1. Similarly, according to the Energy Information Administration of the U.S. government, under the baseline "business as usual" scenario, world energy consumption is projected to grow by 50 percent and net electricity consumption will nearly double over the 2005–2030 period (EIA 2008).

2. It should be noted, however, that these high penetration rates are often the result of direct or indirect subsidies.

3. The reserve capacity requirement becomes onerous only when wind penetration exceeds a certain threshold level which in turn depends on the extent to which the grid is integrated (both within the country and also regionally).

4. Optimists might hope for other means of extracting carbon dioxide from the atmosphere, perhaps by burying biomass, but it is normally cheaper to extract gases where their concentration is highest, for example in the flu gases.

5. The Fourth Assessment Report of the Intergovernmental Panel on Climate Change highlights the following mitigation technologies and practices as being currently commercially available: improved supply and distribution efficiency; nuclear power; renewable heat and power (hydropower, solar, wind, geothermal, and bioenergy); and early applications of CCS. And CCS for gas, biomass and coal-fired electricity-generating facilities, advanced nuclear power, and advanced renewable energy (tidal and wave energy, concentrating solar, and solar PV) are considered likely to play an important role and to be commercialized before 2030 (IPCC 2007a).

6. Energy demand is sensitive to relative prices. The long-run demand growth projections cited above depend on specific assumptions about relative prices. In a more careful and complete treatment of global energy demand, these assumptions would have to be spelled out more clearly. However, such a detailed demand analysis is outside the scope of this paper.

7. This constraint is probably not binding at present.

8. Energy intensity (energy use per dollar of GDP) in OECD countries fell from 0.31 tons of oil equivalent per $1,000 in 1973 to 0.20 tons in 2003 (measured in 2000 dollars; IAEA 2006).

9. According to remarks made by Alan Greenspan at the Economic Club of New York, May 20, 2005.

10. In 2006, coal-fired generation worldwide reached 7,756 terawatt-hours, over 40 percent of the total; gas-fired generation reached 3,807 terawatt-hours, 20 percent of the total; and generation based on oil products reached around 1,140 terawatt-hours, 6 percent of the total. In South Africa and Poland the share of coal-fired generation is above 90 percent; in China and Australia it is approximately 80 percent; in India it is more than two-thirds; and it is about one-half in the United States and Germany (IEA 2008a).

11. Although the MIT study is dated 2007, it has not obviously been superseded by any better cost estimates, and that is likely to remain the case until commercial trial data are available.

12. The compressive energy stored during the 30-year operating life of a carbon dioxide disposal project associated with a 1,000 megawatt coal-fired plant would be equivalent to the energy content of more than 1 megaton of dynamite.

13. Leaving fossil fuel in the ground maintains the risk that it will be burned and carbon dioxide released, whereas leaving carbon dioxide in the ground at most runs the risk that it will be released as it would have been if the fuel had been burned conventionally, and as such is not a cogent argument compared to the health risks of sudden release.

14. It should be noted that the estimates presented in figure 2 need to be revised because, from early 2004 until the beginning of 2008, strong global demand led to sharp increases in the prices of materials, basic inputs, and equipment in the energy sector. This escalation in the cost of basic materials and equipment, coupled with a rise in the cost of craft labor, led to significant increases in the construction costs of generating facilities. For example, during 2006–08, the energy projects financed by the World Bank experienced 30–50 percent increases above the original cost estimates (ESMAP 2008). However, during the past few months in the face of the worst global financial crisis since the Great Depression, the prices of materials and energy have reversed their recent trend and begun declining. Thus, the projections of future prices need to be revised. Moreover, the interplay of technological progress and changes in the level of demand relative to other structural characteristics

of the markets for equipment and other basic inputs of generating plants have differential impacts across the different energy submarkets. For example, the prices of land-based and offshore wind turbines have been increasing in the last 3 years due to the rising costs of materials and capacity constraints in the wind-power market. The prices of solar panels, on the other hand, have been declining and are poised for further sharp drops mainly due to rapid technological change and the exploitation of important scale economies.

15. The global technically exploitable hydropower capability is estimated to be 16,494 terawatt-hours (WEC 2008), although the economically exploitable capability might be considerably less.

16. Water resources are closely linked to climate change. Thus, the potential vulnerability of hydropower to global warming needs to be carefully examined (Iimi 2007).

17. Wind power potential in 2000 was estimated to be approximately 72 terawatts (Archer and Jacobson 2005).

18. See www.awea.org/faq/cost.html.

19. The United Kingdom is fortunate in having a low correlation of wind speed at modest distances, reducing the country-wide problem of intermittency to manageable levels even at penetration rates of 25 percent or more. In Europe, Danish and North German wind power can cause current surges through neighboring countries, but with better information exchange and system operation these can be managed considerably better than at present.

20. The sun supplies the earth with a staggering amount of energy—in 1 hour it delivers the amount of energy that humans use annually. This enormous power that the sun continuously delivers to earth, 1.2×10^5 terawatts, dwarfs every other source of energy—for example in 2006 global power-generation capacity was just over 4 terawatts (Crabtree and Lewis 2007).

21. Typical solar panels have an efficiency of about 10 percent and more expensive ones can perform at 20 percent. The efficiency of PV systems is limited by the underlying laws of physics to a best of 60 percent with perfect concentrating mirrors and 45 percent without mirrors. Achieving 30 percent efficiency with a mass produced device would be remarkable (MacKay 2008).

22. CSP has a lower power density per unit area than PV (Mackay 2008).

23. The same can be said of solar thermal power.

24. For example, biomass-integrated gasification combined-cycle-generating plants can reduce nitrogen oxide emissions by a factor of six relative to average pulverized coal-fired plants.

25. Baseload power is that part of electricity supply that is continuous and does not vary over a 24-hour period, and thus is approximately equivalent to the minimum daily load. It is normally the power source with the lowest variable cost. Wind has virtually zero variable cost and is intermittent, and so can displace other power when operating, but cannot be relied upon to meet firm demand. Run-of-river hydropower has zero variable cost and must be used or lost (like wind) but may be seasonal, while storage hydro with spare reservoir capacity may appear to have zero variable cost but can have a high opportunity cost as it can serve to provide peak power and to store excess wind power.

26. Assuming that problems of waste storage can be adequately addressed.

27. Wind power in favorable locations (with access to a wide area grid and storage hydro) can also be scaled up and provide base load power (at zero variable cost). Hydropower typically requires long-distance transmission but can be complementary to wind power that is often in similarly isolated locations.

28. When reactors operate in a "once-through" mode, the fuel is used once and then sent directly to disposal without further processing.

29. These are rectors that are configured to produce more fissile material than they consume, using fertile material such as depleted uranium in a blanket around the core.

30. These models include among others the Integrated Global Systems Model of the Massachusetts Institute of Technology; the Model for Evaluating the Regional and Global Effects (MERGE) of GHG reduction policies developed at Stanford University and the Electric Power Research Institute; and the MiniCAM Model of the Joint Global Change Research Institute.

31. Four radiative forcing stabilization scenarios have figured prominently in the climate change literature. These include limits on long-term radiative forcing (relative to pre-industrial times) of 3.4, 4.7, 5.8, and 6.7 watts per sq. m which would correspond roughly to carbon dioxide concentrations of 450 ppm, 550 ppm, 650 ppm, and 750 ppm.

32. In the ACT scenario, technologies that are currently available or are in an advanced state of development, bring global carbon dioxide emissions back to the current levels by 2050.

33. Capital costs, particularly steel and concrete, can increase with fossil energy costs. However, these costs once sunk have no further effect on running costs.

34. In the future, advanced gas-cooled, high-temperature reactors that operate at about 900 degrees Celsius could be used to produce hydrogen more efficiently through high-temperature thermochemical cracking of water.

35. Waste classification is based on the concentration of radioactivity it contains (Rahn and others 1984).

36. A recent survey in 25 EU countries found that 6 in 10 citizens acknowledge the benefits of using nuclear energy—for diversifying energy supplies, reducing dependence on oil, and cutting greenhouse emissions. Still, only 37 percent were in favor of using nuclear energy, while 55 percent were opposed. But if the issues related to radioactive waste disposal were resolved, 38 percent of those opposed to nuclear energy would change their opinions (Botella and others 2006).

37. The International Safeguards are a set of activities that the IAEA uses to verify that a country is adhering to international commitments not to use its nuclear program for nuclear weapons purposes. The safeguards system is based on regularly verifying the accuracy and completeness of a country's declarations to the IAEA concerning nuclear-related activities and seeking to assure that no undeclared nuclear materials or activities exist within the country. In total, presently more than 900 declared facilities in 71 countries are "safeguarded" and subject to inspection (APS 2005).

38. The Additional Protocol is a legal document granting the IAEA complementary inspection authority to that provided in underlying safeguards agreements. A principal aim is to enable the IAEA inspectorate to provide assurance about both declared and possible undeclared activities. Under the Protocol, the IAEA is granted expanded rights of access to information and sites, as well as additional authority to use the most advanced technologies during the verification process (www.iaea.org/Publications/Factsheets/English/sg_overview.html).

39. Commercial operators might find a final consumer appetite for bonds indexed to the price of electricity (a liquid substitute for a long-term contract)—provided of course the company issuing the debt were sound. Given the essential nature of electricity demand, and a reasonably well-defined cost base (compared to property backing subprime mortgages), it should be possible to reduce risks considerably.

40. The costs of decommissioning modern reactors are modest and would likely be a charge on the reactor to be managed by a decommissioning fund.

41. The prices of new power stations remain volatile—after a rapid escalation between 2000 and 2008 they now appear to be falling back, as the price of steel has fallen with the recession and so has demand for and hence pressure on the prices of plants.

42. The direct costs do not include those associated with accidents and proliferation.

43. Japan's last six reactors built between 1997 and 2005 varied between U.S. $2,700 per kilowatt and U.S. $3,700 per kilowatt and took between 50 and 95 months from start of construction to criticality. The last three reactors built in China cost between U.S. $1,800 per kilowatt and U.S. $2,700 per kilowatt with construction times falling to 50 months. Korea's last four reactors decreased steadily in cost from U.S. $2,200 per kilowatt to U.S. $1,600 per kilowatt with construction times falling from 62 months to 48 months (Matzie 2005). Matzie argues that Korea's performance has resulted from standardization, localization, and low labor costs.

44. The IAEA defines as "small" those reactors with power less than 300 megawatts.

45. Clearly these biases against nuclear are less pronounced in countries where the government continues to play a significant role under managed liberalization in the electricity sector.

References

Ahearne, J. 2006. "Advanced Nuclear Reactors: Their Use in Future Energy Supply." *Forum on Physics & Society of the American Physical Society* 35(2):2–6.

ANL (Argonne National Laboratory). 2003. "Nuclear Energy: Power for the Twenty-First Century." Argonne National Laboratory Report ANL-03/12.

APS (American Physical Society). 2005. "Nuclear Power and Proliferation Resistance: Securing Benefits, Limiting Risk." Report by the Nuclear Energy Study Group of the American Physical Society Panel on Public Affairs.

Archer, C., and M. Jacobson. 2005. "Evaluation of Global Wind Power." *Journal of Geophysical Research* 110, D12110, doi:10.1029/2004JD005462.

Awerbuch, S., and M. Berger. 2003. "Energy Security and Diversity in the EU: A Mean-Variance Portfolio Approach." International Energy Agency Research Paper, Paris.

Ayres, M., M. MacRae, and M. Stogran. 2004. *Levelized Unit Electricity Cost Comparison of Alternate Technologies for Baseload Generation in Ontario*. Calgary, Alberta: Canadian Energy Research Institute.

Baker Institute. 2001. "New Energy Technologies: A Policy Framework for Micronuclear Technology." Baker Institute Study 17, Rice University.

Bartis, J., F. Camm, and D. Ortiz. 2008. *Producing Liquid Fuels from Coal: Prospects and Policy Issues*. Santa Monica, CA: Rand Corporation.

Bartle, A. 2002. "Hydropower Potential and Development Activities." *Energy Policy* 30:1231–9.

Baurac, D. 2004. "Nuclear Plants may be Clean Hydrogen Source." *Logos* 22(1), Spring. (www.anl.gov/Media_Center/logos22-1/hydrogen.htm).

Botella, T., J. Coadou, and U. Blohm-Hieber. 2006. "European Citizens' Opinions Towards Radioactive Waste: An Updated Review," European Commission, Directorate General for Energy and Transport.

Brolly, A., and P. Vertes. 2004. "Concept of a Small-scale Electron Accelerator Driven System for Nuclear Waste Transmutation." *Annals of Nuclear Energy* 32(4):417–33.

Cantor, R., and J. Hewlett. 1988. "The Economics of Nuclear Power." *Resources and Energy* 10: 315–35.

CGSR (Center for Global Security Research). 2000. Report of the International Workshop on Technology Opportunities for Increasing the Proliferation Resistance of Global Civilian Nuclear Power Systems (TOPS).

Chernobyl Forum. 2005. *Chernobyl's Legacy: Health, Environmental and Socio-Economic Impacts*. (www.chernobyl.undp.org/english/docs/chernobyl.pdf).

Clarke, L., J. Edmonds, H. Jacoby, H. Pitcher, J. Reilly, and R. Richels. 2007. *Scenarios of Greenhouse Gas Emissions and Atmospheric Concentrations*. Sub-report 2.1A of Synthesis and Assessment Product 2.1 by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. Washington, D.C.: Department of Energy, Office of Biological & Environmental Research.

Clerici, A. 2006. *The Future of Nuclear Energy in Europe*. European Regional Study Group, World Energy Council. ABB Italy, June 13.

Crabtree, G., and N. Lewis. 2007. "Solar Energy Conversion." *Physics Today* (March):37–42.

Deutch, J., and E. Moniz. 2007. "A Future for Fossil Fuel." *The Wall Street Journal*, March 15.

DGEMP (General Directorate for Energy and Raw Materials). 2003. *Study of Reference Costs for Power Generation*. Paris, France: French Ministry of the Economy, Finance and Industry.

Difiglio, C., and D. Gielen. 2004. "What is the Cost of Not Having Nuclear Power or Carbon Capture and Storage While Still Stabilizing GHG Concentrations at Safe Levels?" International Energy Agency, Energy PMP.

Dooley, J., J. Edmonds, R. Dahowski, and M. Wise. 2002. *Modeling Carbon Capture and Storage Technologies in Energy and Economic Models*. Intergovernmental Panel on Climate Change Workshop on Carbon Dioxide Capture and Storage, Regina, Canada, November 18–21.

DTI (Department of Trade and Industry). 2007. *The Future of Nuclear Power in a Low Carbon Economy*. Consultation Document, London.

EC (European Commission). 2007. *The Sustainable Nuclear Energy Technology Platform*. Special Report EUR 22842.

EIA (Energy Information Administration). 2006. *International Energy Outlook*. Washington, D.C: EIA.

_____. 2008. *International Energy Outlook*. Washington, D.C.: EIA.

ESMAP (Energy Sector Management Assistance Program). 2008. *Study of Equipment Prices in the Energy Sector*. World Bank, Washington, D.C.

Feiveson, H. 2001. "The Search for Proliferation-Resistant Nuclear Power." *Journal of the Federation of American Scientists* 54(5):1–9.

Feiveson, H., A. Glaser, M. Miller, and L. Scheinman. 2008. "Can Future Nuclear Power Be Made Proliferation Resistant?" Center for International and Security Studies at Maryland.

Finon, D., and F. Roques. 2008. *Financing Arrangements and Industrial Organization for New Nuclear Build in Electricity Markets*. Cambridge: EPRG0826. (www.eprg.group.cam.ac.uk/wp-content/uploads/2008/11/eprg08262.pdf).

Frois, B. 2005. "Advances in Nuclear Energy." *Nuclear Physics* A 752:611c–22c.

Gokhaleb, P., S. Deokatteya, and V. Kumar. 2006. "Accelerator Driven Systems (ADS) for Energy Production and Waste Transmutation: International Trends." *Progress in Nuclear Energy* 48(2): 91–102.

Graber, R., and G. Rothwell. 2006. "Valuation and Optionality of Large Energy Industry Capital Investments." *Cost Engineering* 48(8):20–6.

Greenpeace International. 2006. "Chernobyl Death Toll Grossly Underestimated." (www.greenpeace.org/international/news/chernobyl-deaths-180406).

_____. 2007. *The Economics of Nuclear Power*. Report prepared for Greenpeace International. The Netherlands: Greenpeace International.

Greenspan, E., and N. Brown. 2001. "Small Innovative Reactor Designs: Useful Attributes and Status of Technology." The James A. Baker III Institute for Public Policy, Rice University.

Heller, T. 2007. *US-Japan Workshop on Climate Actions and Development Co-benefits*. Stanford Program on Energy and Sustainable Development.

Herring, S. 2004. "Hydrogen Production using Nuclear Energy." Almquist Lecture. Idaho: National Engineering and Environmental Laboratory.

IAEA (International Atomic Energy Agency). 2000. Bulletin 42/2.

_____. 2006. *Nuclear Power and Sustainable Development*. Vienna, Austria: IAEA.

IEA (International Energy Agency). 2008a. *World Energy Outlook*. Paris: OECD.

_____. 2008b. *Energy Technology Perspectives*. Paris: OECD.

_____. 2008c. *Energy Prices and Taxes. 4th Quarter 2008*. OECD/IEA.

IEA (International Energy Agency)/NEA (Nuclear Energy Agency). 2005. Projected Costs of Generating Electricity—2005 Update. Paris: OECD.

Iimi, A. 2007. "Estimating Global Climate Change Impacts on Hydropower Projects." Policy Research Working Paper 4344. World Bank, Washington, D.C.

INL (Idaho National Laboratory). 2002. *A Technology Roadmap for Generation IV Nuclear Energy Facilities*. GIF-002-00.

IPCC (Intergovernmental Panel on Climate Change). 2007a. *Fourth Assessment Report: Climate Change 2007*. Working Group III Report "Mitigation of Climate Change." Cambridge: Cambridge University Press.

_____. 2007b. *Impacts, Adaptation and Vulnerability*. Working Group II, Fourth Assessment Report, Geneva.

Joskow, P. 2006a. "The Future of Nuclear Power in the United States: Economic and Regulatory Challenges." 06–019, Center for Energy and Environmental Policy Research, MIT.

_____. 2006b. "Prospects for Nuclear Power: A U.S. Perspective." Lecture, University of Paris-Dauphine, May 19.

Kursunoglu, B., and S. Mintz. 2001. *Global Warming and Energy Policy*. New York: Plenum Publishers.

Lovins, A. 2005. "Mighty Mice." *Nuclear Engineering International*, 44–48. December 21.

Lovins, B., and L. Lovins. 2001. "The Nuclear Option Revisited." *The Los Angeles Times*, July 8.

MacKay, D. 2008. *Sustainable Energy—without the Hot Air*. Cambridge: UIT.

Matzie, R. 2005. "Building New Nuclear Plants to Cost and Schedule: An International Perspective." Presentation to the Nuclear Fission Seminar, London, September 29.

METI (Ministry of Economy, Trade and Industry). 2004. Investment Plans Submitted to METI in accordance to the Electric Regulation Law, March 31, Tokyo, Japan.

MIT (Massachusetts Institute of Technology). 2003. *The Future of Nuclear Power*. Cambridge, Mass: MIT.

_____. 2007. *The Future of Coal*. Cambridge, Mass: MIT.

NEA (Nuclear Energy Agency). 1998. "Nuclear Power and Climate Change." Paris: OECD.

Pacala, S., and R. Socolow. 2004. "Stabilization Wedges: Solving the Climate Problem for the Next 50 Years with Current Technologies." *Science* 305:968–72.

Paul Scherrer Institut. 1998. *Severe Accidents in the Energy Sector: Report on Externalities of Power Production*. Commissioned by the Swiss Federal Office of Energy. PSI Bericht 98-16. (gabe.web.psi.ch/pdfs/PSI_Report/ENSAD98.pdf).

Princeton Environmental Institute. 2007. *Stabilization Wedges: A Concept & Game*. Carbon Mitigation Initiative, Princeton University.

RAE (Royal Academy of Engineering). 2004. *The Cost of Generating Electricity*. London: RAE.

Rahn, E., A. Adamantiades, J. Kenton, and C. Braun. 1984. *A Guide to Nuclear Power Technology*. New York: John Wiley.

Richels, R., T. Rutherford, G. Blanford, and L. Clarke. 2007. "Managing the Transition to Climate Change." Working Paper 07-01, American Enterprise Institute-Brookings Joint Center for Regulatory Studies.

Roques, F., W. Nuttall, D. Newbery, R. de Neufville, and S. Connors. 2006. "Nuclear Power: a Hedge against Uncertain Gas and Carbon Prices." *Energy Journal* 27(4):1–23.

Rothwell, G. 1992. "Can Nuclear Power Compete?" *Regulation* (Winter):66–74.

_____. 2006. "A Real Options Approach to Evaluating New Nuclear Plants." *Energy Journal* 27(1): 37–53.

_____. 2007. "An Evaluation of the Real Option of Starting to Build a Nuclear Power Plant in Chile in 2020." Santiago, Chile: Centro de Estudios Publicos.

Schock, R., N. Brown, and C. Smith. 2001. "Nuclear Power, Small Reactor Technology, and the Role of Technical Innovation: An Assessment." James A Baker III Institute of Public Policy, Rice University.

Stern, N. 2007. *The Economics of Climate Change*. Cambridge: Cambridge University Press.

Stirling, A. 2001. "Science and Precaution in the Appraisal of Electricity Supply Options." *Journal of Hazardous Materials* 86:55–75.

Taylor, S. (2007) *Privatization and Financial Collapse in the Nuclear Industry; The Origins and Causes of the British Energy crisis of 2002*. Abingdon, UK: Routledge.

Thomas, S. 2005. "The Economics of Nuclear Power." Heinrich Boll Shifting, Nuclear Issues Paper 5.

UofC (University of Chicago). 2004. *The Economic Future of Nuclear Power*. Chicago: UofC.

UCS (Union of Concerned Scientists). 2007. "Nuclear Power and Global Warming." Position Paper.

UIC (Uranium Information Center). 2007a. "World Energy Needs and Nuclear Power." Nuclear Issues Briefing Paper 11. (www.uic.com.au/nip14.htm).

_____. 2007b. "Safety of Nuclear Power Reactors." Nuclear Issues Briefing Paper 14. (www.uic.com.au/nip14.htm).

USDoE (United States Department of Energy). 1997. *Nonproliferation and Arms Control Assessment of Weapons-Usable Fissile Material Storage and Excess Plutonium Disposition Alternatives.* Washington, DC: USDoE.

USNRC (United States Nuclear Regulatory Commission). 2007. "Backgrounder on New Nuclear Plant Designs." (www.nrc.gov/reading-rm/doc-collections/fact-sheets/new-nuc-plant-des-bg.html).

Wade, D. 2005. "The STAR Concept: a Hierarchical Hub-Spoke Nuclear Architecture Based on Long-Refueling-Interval Battery Reactors and Regional Fuel Cycle Centers." Annex 11 of Innovative Small and Medium Sized Reactors: Design Features, Safety Approaches and R&D Trends. Final Report of a Technical Meeting Held in Vienna, June 11–14. IAEA-TECDOC-1451, pp. 171–91 (May).

WEC (World Energy Council). 2007. *Survey of Energy Resources*. (www.worldenergy.org/documents/ser2007_final_online_version_1.pdf).

WNA (World Nuclear Association). 2008. "Supply of Uranium." Information Papers, June.

_____. 2009. "Energy Subsidies and External Costs." Information Papers, February.

Zimmerman, M. 1982. "Learning Effects and the Commercialization of New Energy Technologies." *Bell Journal of Economics* 13(2):297–310.

THE WORLD BANK

1818 H Street NW
Washington, DC 20433, USA
World Wide Web: http://www.worldbank.org/
E-mail: pinearchiveserver@worldbank.org

ISBN 978-0-19-509191-6
9 780199 591916