The Journal of Economic Perspectives

A journal of the American Economic Association

Winter 2013

The Journal of
Economic Perspectives

A journal of the American Economic Association

Editorial offices:
Journal of Economic Perspectives
American Economic Association Publications
2403 Sidney St., #260
Pittsburgh, PA 15203
email: jep@jepjournal.org

The *Journal of Economic Perspectives* gratefully acknowledges the support of Macalester College.
Registered in the US Patent and Trademark Office (®).

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, USA
Printed by R. R. Donnelley Company, Jefferson City, Missouri, 65109, USA

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Winter 2013, Vol. 27, No. 1. The *JEP* is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $20.00, $30.00, or $40.00 depending on income; for an additional $15.00, you can receive this journal in print. E-reader versions are free. For details and further information on the AEA go to http://www.vanderbilt.edu/AEA/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 27 • Number 1 • Winter 2013

Symposia

Articles

Features

Statement of Purpose

The *Journal of Economic Perspectives* attempts to fill a gap between the general interest press and most other academic economics journals. The journal aims to publish articles that will serve several goals: to synthesize and integrate lessons learned from active lines of economic research; to provide economic analysis of public policy issues; to encourage cross-fertilization of ideas among the fields of economics; to offer readers an accessible source for state-of-the-art economic thinking; to suggest directions for future research; to provide insights and readings for classroom use; and to address issues relating to the economics profession. Articles appearing in the journal are normally solicited by the editors and associate editors. Proposals for topics and authors should be directed to the journal office, at the address inside the front cover.

Policy on Data Availability

It is the policy of the *Journal of Economic Perspectives* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Details of the computations sufficient to permit replication must be provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the above requirements cannot be met.

Policy on Disclosure

Authors of articles appearing in the *Journal of Economic Perspectives* are expected to disclose any potential conflicts of interest that may arise from their consulting activities, financial interests, or other nonacademic activities.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 3–22

The Case Against Patents

Michele Boldrin and David K. Levine

The case against patents can be summarized briefly: there is no empirical evidence that they serve to increase innovation and productivity, unless productivity is identified with the number of patents awarded—which, as evidence shows, has no correlation with measured productivity. This disconnect is at the root of what is called the "patent puzzle": in spite of the enormous increase in the number of patents and in the strength of their legal protection, the US economy has seen neither a dramatic acceleration in the rate of technological progress nor a major increase in the levels of research and development expenditure.

Both theory and evidence suggest that while patents can have a partial equilibrium effect of improving incentives to invent, the general equilibrium effect on innovation can be negative. The historical and international evidence suggests that while weak patent systems may mildly increase innovation with limited side effects, strong patent systems retard innovation with many negative side effects. More generally, the initial eruption of innovations leading to the creation of a new industry—from chemicals to cars, from radio and television to personal computers and investment banking—is seldom, if ever, born out of patent protection and is instead the fruit of a competitive environment. It is only after the initial stage of rampant growth ends that mature industries turn toward the legal protection of patents, usually because their internal growth potential diminishes and they become more concentrated. These observations, supported by a steadily increasing body of evidence, are consistent with

■ *Michele Boldrin is Joseph Gibson Hoyt Distinguished University Professor of Economics and David K. Levine is John H. Biggs Distinguished University Professor of Economics, both at Washington University in St. Louis, Missouri. They are also both Research Fellows with the Federal Reserve Bank of St. Louis. Their email addresses are mboldrin@artsci.wustl.edu and david@dklevine.com.*

http://dx.doi.org/10.1257/jep.27.1.3. doi=10.1257/jep.27.1.3

theories of innovation emphasizing competition and first-mover advantage as the main drivers of innovation, and they directly contradict "Schumpeterian" theories postulating that government-granted monopolies are crucial to provide incentives for innovation. A properly designed patent system might serve to increase innovation at a certain time and place—and some patent systems, such as the late-nineteenth century German system allowing only process but not final product patents, have been associated with rapid innovation. Unfortunately, the political economy of government-operated patent systems indicates that such systems are susceptible to pressures that cause the ill effects of patents to grow over time. The political economy pressures tend to benefit those who own patents and are in a good position to lobby for stronger patent protection, but disadvantage current and future innovators as well as ultimate consumers. This explains why the political demand for stronger patent protection comes from old and stagnant industries and firms, not from new and innovative ones. Our preferred policy solution is to abolish patents entirely and to find other legislative instruments, less open to lobbying and rent seeking, to foster innovation when there is clear evidence that laissez-faire undersupplies it. However, if that policy change seems too large to swallow, we discuss in the conclusion a set of partial reforms that could be implemented as part of an incremental strategy of reducing the harm done by the patent system.

Do Patents Encourage Productivity Growth?

If there is to be any rationale for patent systems, with all their ancillary costs, it must be that they increase innovation and productivity. What is the evidence?

Simply eyeballing the big trends shows that patenting has exploded over the last decades. In 1983 in the United States, 59,715 patents were issued; by 2003, 189,597 patents were issued; and in 2010, 244,341 new patents were approved. In less than 30 years, the flow of patents more than quadrupled. By contrast, neither innovation nor research and development expenditure nor factor productivity have exhibited any particular upward trend. According to the Bureau of Labor Statistics, annual growth in total factor productivity in the decade 1970–1979 was about 1.2 percent, while in the decades 1990–1999 and 2000–2009 it has been a bit below 1 percent. Meanwhile, US research and development expenditure has been oscillating for more than three decades in a narrow band around 2.5 percent of GDP. The recent explosion of patents, in other words, has not brought about any additional surge in useful innovations and aggregate productivity. In new industries such as biotechnology and software—where innovation was already thriving in their absence—patents have been introduced without any positive impact on the rate of innovation. The software industry is an important case in point. In a dramatic example of judge-made law, software patents became possible for the first time in the early 1990s. Bessen and Meurer, in a large body of empirical work culminating in *Patent Failure* (2008), have studied the consequences of this experiment and have concluded that it damaged social welfare.

Academic studies have also typically failed to find much of a connection between patents and innovation. In Boldrin and Levine (2008b), we conducted a metastudy gathering the 24 studies (including three surveys of earlier empirical work) we could find in 2006 that examined whether introducing or strengthening patent protection leads to greater innovation. The executive summary states: "[T]hese studies find weak or no evidence that strengthening patent regimes increases innovation; they find evidence that strengthening the patent regime increases patenting! They also find evidence that, in countries with initially weak IP [intellectual property] regimes, strengthening IP increases the flow of foreign investment in sectors where patents are frequently used." Actually, the issue of promoting foreign direct investment, while a well-established empirical consequence of strengthening patent regimes, is entirely beside the point of this essay. There are a number of ways to strengthen a country's institutions and infrastructure in a way that would encourage foreign direct investment—and, in any case, foreign direct investment is not equivalent to innovation.

Our conclusion was in keeping with other studies that have addressed this question. Some studies have failed to find any connection even between changes in the strength of patent law and the amount of patenting, while others fail to find a connection between patents and some measure of innovation or productivity. For example, after failing to find a single study claiming that innovation increased as a consequence of the strengthening of US patent protection in the 1980s, Gallini (2002, p. 139) wrote in this journal: "Although it seems plausible that the strengthening of US patents may have contributed to the rise in patenting over the past decade and a half, the connection has proven difficult to verify." Similarly, Jaffe (2000) also examines many studies and concludes: "[D]espite the significance of the policy changes and the wide availability of detailed data relating to patenting, robust conclusions regarding the empirical consequences for technological innovations of changes in patent policy are few. There is widespread unease that the costs of stronger patent protection may exceed the benefits. Both theoretical and, to a lesser extent, empirical research suggest this possibility."[1]

[1] The study by Kanwar and Evanson (2001) illustrates some of the issues that arise in these kinds of studies. They have two five-year averages on 31 countries for the period 1981–1990. They find support for the idea that higher patent protection leads to higher research and development spending as a fraction of GDP. However, a different story seems equally plausible. Countries with a larger market can more easily pay the fixed costs of innovation. Indeed, one perspective is that their data essentially compares countries with relatively small economies, little intellectual property protection, and low R&D spending with countries with relatively larger economies, greater intellectual property protection, and higher R&D spending. For example, R&D spending as a fraction of GDP in their data ranges from a ten-year average of 0.2 percent in Jordan to 2.8 percent in Sweden. If we combine their data with GDP data from *The 1990 CIA World Fact Book* to take account of the size of the economy, increasing the strength of intellectual property protection from 0 to 1 to 2 on their five-point scale does increase R&D expenditure. But as intellectual property protection is increased further, the gains to R&D expenditure levels then falls. Even at the lower levels, we are probably observing primarily the effect of foreign direct investment: that is, among poor countries with near-zero intellectual property protection, increases bring in more foreign investment and in doing so directly raise R&D spending. In higher-income countries with larger economies, foreign investment is not an issue, and increases in intellectual property have little or no effect on innovation.

The Lerner (2002) study is especially notable because he examined all significant changes in patent law in all countries over the last 150 years. His conclusion: "Consider, for instance, policy changes that strengthen patent protection. Once overall trends in patenting are adjusted for, the changes in patents by residents of the country undertaking the policy change are negative, both in Great Britain and in the country itself. Subject to the caveats noted in the conclusion this evidence suggests that these policy changes did not spur innovation." This, in summary, is what is currently known as the "patent puzzle"—although as we will explain, it is substantially coherent with a theory of innovation that emphasizes the gains from competition and first-mover incentives, rather than benefits from the monopoly power of patents.

Evidence at the sectoral level of the US economy shows the same disconnect between patenting and productivity. In Boldrin, Correa, Levine, and Ornaghi (2011), we carried out a sequence of statistical tests and econometric estimations on two datasets: an original microeconomic dataset obtained by combining firm-level information obtained through Compustat, the National Bureau of Economic Research, and the Bureau of Labor Statistics and an enriched version of the dataset used by Aghion, Bloom, Blundell, Griffith, and Howitt (2005) in their study of industry-level mark-ups. Conclusions must of course be drawn with care from this kind of data because, across industries, the strength of competition, patenting, and productivity are simultaneously determined and intertwined with technological change. With that reservation appropriately noted, at the industry level there is, in general, no statistically significant correlation between measures of productivity (whether measured by labor or total factor productivity) and of patenting activity (whether measured by number of patents or citations of patents).

We then investigated the relationships between patents, competition, and productivity further. When we regressed measures of patents (or patent citations) on a measure of competition (as measured by the inverse of profitability) used by Aghion, Bloom, Blundell, Griffith, and Howitt (2005), we found a positive relationship that is remarkably robust to changes in industry classification, time period, and set of sampled industries. That is, patents were more common in competitive industries. We also studied the correlation between the same measure of competitive pressure and objective measures of labor productivity growth. In our preferred specification, we found that average annual growth of productivity in the sectors with the highest level of competition is up to 2 percent bigger than in the sectors with the lowest level of competition. These are strikingly large differences when cumulated over various decades, as it is the case in our dataset. This finding of a positive correlation between competition and productivity at the sectorial level replicated a pioneering, and unfortunately forgotten, pattern reported in Stigler (1956).

The accumulated findings of no positive relationship between patenting and productivity are not conclusive, and arguments have raged over the specific data used, whether to look for a structural break in the data, how the researcher seeks

to correct for endogeneity, and so on.[2] However, it is fair to say that the sector-level, national, and cross-national evidence fail to provide any clear empirical link from patents to innovation or to productivity. This lack of connection is consistent with the view that the use of patents either as a defensive or as a rent-seeking tool is more widespread than one might have predicted. In addition, the empirical evidence is consistent with the proposition that greater competition, not patents, is the main factor leading to innovation and greater productivity.

Theory and Practice of Patents and Innovation

There is little doubt that providing a monopoly as a reward for innovation increases the incentive to innovate. There is equally little doubt that granting a monopoly for any reason has the many ill consequences we associate with monopoly power—the most important and overlooked of which is the strong incentive of a government-granted monopolist to engage in further political rent seeking to preserve and expand its monopoly or, for those who do not yet have a monopoly, to try to obtain one. These effects are at least to some extent offsetting: while the positive impact of patents is the straightforward partial equilibrium effect of increasing the profits of the successful innovator to the monopolistic level, the negative one is the subtler general equilibrium effect of reducing everybody else's ability to compete while increasing for everyone the incentive to engage in socially wasteful lobbying efforts.

Downstream Innovation, Defensive Patenting, and Patent Trolls

In the long run, even the positive partial equilibrium effect of patents in providing an incentive for innovation may be more apparent than real: the existence of a large number of monopolies created by past patent grants reduces the incentives for current innovation because current innovators are subject to constant legal action and licensing demands from earlier patent holders. The downstream blocking effect of existing monopoly grants on incentives for future innovation

[2] For a sense of these controversies, Aghion, Bloom, Blundell, Griffith, and Howitt (2005) find an "inverted-U" relationship between the extent of competition, as measured by the inverse of mark-ups, and a measure of patenting activity, based on a dataset of US patents of UK firms. In other words, they find that the maximum innovative effort (as measured by patents) occurs at some "intermediate" position between a high and low level of competition. However, Hashmi (2011) reexamines the inverted-U relationship using data from publicly traded US manufacturing firms and finds a robust positive relationship between the inverse of markups and citation-weighted patents. Correa (2012) reexamines the same dataset of UK firms and shows that the prediction of an inverted-U is overturned when allowing for the possibility that innovations follow a "memory process," where the current probability of introducing a new innovation increases when a firm successfully innovated in the previous period. He also finds a structural break in the data in 1981, when the Court of Appeals for the Federal Circuit was established to hear appeals of patent cases. Overall, Correa finds a positive innovation–competition relationship for the memory industries before the 1982 reform, but no relationship between innovation and competition for those industries that he classifies as memory-less.

has greatly increased in recent decades because modern products are made up of so many different components. The recent—and largely successful—efforts of Microsoft to impose a licensing fee on the large and expanding Android phone market is but one case in point. With the exception of Motorola Mobility, all the handset manufacturers have agreed to the fee, and Motorola lost its first battle against the fee in spring 2012—fought not in court but in the more receptive domain of the US International Trade Commission (Investigation Number 337-TA-744, May 18, 2012). Microsoft is attempting to charge a licensing fee solely over a patent involving the scheduling of meetings—a rarely used feature of modern smartphones. The meeting-schedule feature is but one of many thousands of patented "ideas" used in a modern smartphone, and each owner of each patent potentially can charge a licensing fee. Hence, the main dynamic general equilibrium effect of a patent system is to subject future inventions to a gigantic hold-up problem: with many licenses to be purchased and uncertainty about the ultimate value of the new innovation, each patent holder, in raising the price of his "component," imposes an externality on other patent holders and so charges a higher than efficient licensing fee. In Boldrin and Levine (2005) and Llanes and Trento (2009), we and others have explored the theory; and many case studies involving patents (and other fractionated ownership problems) can be found in Heller (2008).

To understand more about the actual effect of patents in the real world, consider the recent purchase by Google of Motorola Mobility, primarily for its patent portfolio—not for the ideas and innovations in that portfolio. Few if any changes or improvements to Google's Android operating system will result from the ownership or study of these software patents. Google's purpose in obtaining this patent portfolio is purely defensive: it can be used to countersue Apple and Microsoft and blunt their legal attack on Google. These remarks apply to the vast bulk of patents: they do not represent useful innovation at all and are just weapons in an arms race. This is not news: the same message emerged decades ago from the Levin, Klevorick, Nelson, and Winter (1987) and Cohen, Nelson, and Walsh (2000) surveys of research and development managers.

One could argue that the costs of building up a patent portfolio to engage in this sort of defensive patenting are not too large: after all, it can cost as little as $15,000 to file a successful patent application, and filing applications on a larger scale might be cheaper. However, the acquisition of large patent portfolios by incumbents creates huge barriers to entry. In the smartphone market, for example, Apple is the market leader and Microsoft is unable to produce a product that appeals to consumers. Each are incumbent firms with a large patent portfolio. In this market, Google is the new entrant and innovator and, while wealthy, Google found itself lacking a large defensive patent portfolio. Hence we see both Apple and Microsoft attacking Google with patent litigations, generating hundreds of millions in wasteful legal costs and no social benefit whatsoever.

Despite the fact that patents are mostly used for arms races and that these, in turn, are driven by patent trolls, there does not yet exist convincing formal models of the ways in which this interaction can inhibit innovation. In a pure arms

race theory, if all firms get counterbalancing patent portfolios and all innovate, then they would all have innovated in the absence of patents—hence, patents do not encourage innovation. This follows because with counterbalancing patent portfolios, no firm can sue any other firm—exactly as would be the case in the absence of patents. Hence in this setting patents simply add a cost to innovation: if you wish to innovate, you must acquire an expensive patent portfolio to avoid trolls. On the other hand if a patentholder does not produce a marketable product and hence cannot be countersued—like Microsoft in the phone market or other patent trolls in other markets—then patents become a mechanism for sharing the profits without doing the work. In this scenario, not only do patents discourage innovation, but they are also a pure waste from a social standpoint.

Patents and Information Disclosure

Another widely cited benefit of patent systems—although not so much in the economics literature—is the notion that patents are a substitute for socially costly trade secrecy and improve communication about ideas. From a theoretical point of view, the notion that patents are a substitute for trade secrecy fails in the simplest model. If a secret can be kept for N years and a patent lasts M years, then an innovator will patent when $N < M$. In other words, ideas will be patented when it seems likely that the secret would have emerged before the patent expired and not patented if the secret can be kept. In practice, it is uncertain when the secret will leak out, but it can be shown that the basic intuition remains intact in the face of uncertainty (Boldrin and Levine 2004; Ponce 2007).[3]

It is also the case that the extent of practical "disclosure" in modern patents is as negligible as the skills of patent attorneys can make it. It is usually impossible to build a functioning device or software program from a modern patent application; this is made especially clear by the fact that some patented ideas do not and *cannot* work. For example, US Patent 6,025,810 was granted for moving information through the fifth dimension. While detailed studies of the usefulness of disclosure in patent applications are not available, companies typically instruct their engineers developing products to avoid studying existing patents so as to be spared subsequent claims of willful infringement, which raises the possibility of having to pay triple damages. According to sworn testimony by Google's chief of Android development during the legal battles between Oracle and Google (for example, Niccolai 2012), the engineers that developed Android were unaware of Apple (or other) patents, and so were unlikely to have been helped by them. The opinion of Brec (2008), a Microsoft developer, reflects that of many practitioners:

> [Microsoft policy is for developers to] never search, view, or speculate about patents. I was confused by this guidance till I wrote and reviewed one of my

[3] A more subtle point is that secrecy may bias the type of inventive activity away from innovations that are not easily kept secret to those that can be. In this symposium, Moser offers some of the historical evidence on this point.

own patents. The legal claims section—the only section that counts—was indecipherable by anyone but a patent attorney. Ignorance is bliss and strongly recommended when it comes to patents.

The related idea that patents somehow improve communication about ideas, thereby creating some positive externality—a notion key to the "public–private" partnership between governments and private research organizations in which the government funds the research and then gives the private organization a monopoly over what is developed in the course of research—is backed by neither theory nor evidence. It is impossible to study the history of innovation without recognizing that inventors and innovators exchange ideas as a matter of course and that secrecy occurs, when it occurs, typically in the final stages of an innovation process when some ambitious inventors hope to corner the market for a functioning device by patenting it. A good case in point is that of the Wright brothers, who made a modest improvement in existing flight technology that they kept secret until they could lock it down on patents, then used their patents both to monopolize the US market and to prevent further innovation for nearly 20 years (Shulman, 2003). The role that Marconi and his patent played in the development of the radio is altogether similar (Hong 2001), as are innumerable other stories. At the opposite extreme we have, again among many, the example of the Cornish steam engine discussed in Nuvolari (2004, 2006). Here engineers exchanged nonpatented ideas for decades in a collaborative effort to improve efficiency. The contemporary FLOSS (Free/Libre and Open Source Software) community is another successful example of how collaboration and exchange of ideas can thrive without the monopoly power granted by patents.

First-Mover Advantages and Incentives for Innovation

In most industries, the first-mover advantage and the competitive rents it induces are substantial without patents. The smartphone industry—laden as it is with patent litigation—is a case in point. Apple derived enormous profits in this market before it faced any substantial competition. The first iPhone was released on June 29, 2007. The first serious competitor, the HTC Dream (using the Android operating system) was released on October 22, 2008. By that time, over 5 million iPhones had been sold, and sales soared to over 25 million units during the subsequent year, while total sales of all Android-based phones were less than 7 million. In the tablet market, the iPad has no serious competitor as of late 2012 despite having been introduced on April 10, 2010. While it is hard to prove this delayed imitation also would have occurred in the complete absence of patents, intuition suggests—and our formal model in Boldrin and Levine (2004) predicts—that there is little reason to assert patent rights while the first-mover advantage is still active. Apple did not initially try to use patents to prevent the Android phones from coming into its market and the subsequent "patents' fight" has been taking place largely after 2010; these facts are consistent with a substantial first-mover advantage. How valuable for Apple was the delay in the Android phones entry? Largely because Apple kept its

first-mover advantage in spite of a large imitative entry in this market, the value of Apple stock—during a severe market downturn—rose by a factor of approximately five. While there may have been some delay in entry from the competition due to Apple's threat—since executed—of patent litigation, the fact is that similar but less-successful devices had been available for a number of years before Apple finally cracked the market.

Less anecdotal than the story of the iPhone is the survey of research and development managers in Cohen, Nelson, and Walsh (2000). Here, over 50 percent of managers indicate lead time (first-mover advantage) is important to earning a return on innovation; outside the pharmaceutical and medical instruments industry, less than 35 percent of managers indicate that patents are important.

To understand patents in practice, it is necessary to examine the lifecycle of industries (for example, Jovanovich and MacDonald 1994; Scherer 1990). Typically a new, hence innovative, industry begins with a competitive burst of entries through which very many innovators try hard to get their products to market. In these early stages, many firms bring different versions of the new product to the market (think of the American auto industry in the early twentieth century or the software industry in the 1980s and 1990s) while demand for the new product grows rapidly and the quality of products is rapidly improved. At this stage of the industry lifecycle, the price elasticity of demand is typically high; what is important is not to dominate the market, but rather to get your own products quickly to market and to reduce costs. From the perspective of competing firms, your cost-reducing innovation is good for me in the same way that my cost-reducing innovation is good for you—hence, let us all imitate each other and compete in the market.

As the industry matures, demand stabilizes and becomes much less price elastic; the scope for cost-reducing innovations decreases; the benefits of monopoly power grow; and the potential for additional product innovation shrinks. Typically there is a shakeout in which many firms either leave the industry or are bought out. The automobile industry is a classical historical example, but many readers will have a more vivid memory of the bursting of the dot-com bubble, which makes this point even more forcefully. At this stage of the industry lifecycle, rent seeking becomes important and patents are widely used to inhibit innovation, prevent entry, and encourage exit. If we look at patent litigation in practice—and as predicted by theories of first-mover competition (Boldrin and Levine 2004, among others)—it takes place when innovation is low. When an industry matures, innovation is no longer encouraged; instead, it is blocked by the ever-increasing appeal to patent protection on part of the insiders.

While patent litigation has increased, few patents are actively used. Patent litigation often involves dying firms that have accumulated huge stockpile of patents but are no longer able to produce marketable products and that are now suing new and innovative firms. For example, Texas Instruments was one of the first producers of microchips, and many in our generation remember the capabilities of their first TI calculator. But Texas Instruments was unable to make the transition to the personal computer revolution and became, for a while, the symbol of a dying company

trying to stay alive by suing the newcomers.[4] In more recent times, Microsoft—once the giant bestriding the software industry—has been unable to make the leap to portable devices such as telephones and tablet personal computers. Thus, Microsoft now uses patent litigation to try to claim a share of the profits Google generates in this market. Back in 1991, Bill Gates said: "If people had understood how patents would be granted when most of today's ideas were invented and had taken out patents, the industry would be at a complete standstill today . . . A future start-up with no patents of its own will be forced to pay whatever price the giants choose to impose." Today, Microsoft lobbies across Europe and Asia for the introduction of software patents, a prize it has already obtained in its home country.

The cost of litigating patents is not insubstantial either. Bessen and Meurer (2008) used stock market event studies to estimate the cost of patent litigation: they estimate that during the 1990s such costs rose substantially until, at the end of the period, they constituted nearly 14 percent of total research and development costs. A related but more difficult-to-quantify phenomenon is the rise of uncertainty caused by the legal system. A case in point is the NTP Inc. patents that were used to threaten the Blackberry network with a shutdown. In 2006, Research in Motion (RIM), the producer of Blackberry, agreed to pay $612.5 million to license the patent in question from NTP (Svensson 2006). The patent was later invalidated by the court—but RIM did not get its money back (Salmon 2012). Here, the behavior of a single judge cost RIM more than half a billion dollars. In this setting, it is no surprise that patent trolls hope to get rich quickly.

It is easier to list the main social welfare implications of the tradeoff between costs of legal monopoly and incentives to patent holders than it is to calculate their magnitudes. Still, the provisional evidence we have suggests that the net welfare effects of the current patent system could easily be negative. It is somewhat conventional to think of welfare losses from distortions as small, with the idea that welfare triangles due to monopoly power are small being the paradigmatic case in point. Unfortunately, monopolies have no incentive to avoid large social losses even when the private gains are small. Witness, for example, the fact that patented pharmaceutical products often sell for hundreds of times the marginal cost of production, as some astonishing pricing differences between the US and the European markets show. Most revealing is the empirical study of the Quinolones family of drugs (Chaudhuri, Goldberg, and Gia 2006). It measures the economic consequences of the introduction of pharmaceutical patents for this family of drugs and concludes that the consequence of patent protection to India will be nearly $300 million in welfare losses—while the gain to the pharmaceutical companies will be less than $20 million.[5]

[4] Texas Instruments is such an important source of litigation that empirical work on patent litigation usually uses a dummy variable for TI. Empirical studies of the importance of firms no longer doing business in an industry to litigation can be found in Bessen and Meurer (2005) and Hall and Ziedonis (2007).

[5] Although the focus of this paper is on patents rather than copyright, it is worth noting that most of the copyright wars revolve around measures to prevent piracy, empirically a relatively minor factor as far as profits of media corporations are concerned (see for example Sinha, Machado, and Sellman 2010; Danaher, Dhanasobhon, Smith, and Telang 2010; Sanchez 2012).

Pharmaceuticals

This brings us to the controversial issue of drug patents. The standard argument says: No patents, no drugs. The total cost of developing a new drug, including failures, is quickly approaching the $1 billion mark (DiMasi, Hansen, and Grabowski 2003). So how can anyone, faced with such a gigantic fixed cost and a microscopic marginal cost of reproduction, innovate without the protection of patents? But consider the following facts: Under current law, the chemical formula and the efficacy of the cure as established by clinical trials are made available to competitors essentially for free. About 80 percent of the initial fixed cost of drug development comes from Stage III clinical trials, a public good that legislation requires be *privately* produced. The downstream social cost of monopoly pricing of pharmaceutical products is highest for life-saving drugs, and the cost of monopoly pricing of other pharmaceutical products is also quite high. Given all this, various economists, such as Kremer and Williams (2009), have argued that *if* government intervention is indeed needed in this market, a system of prizes might be superior to the existing system of monopolies.

There are four things that should be born in mind in thinking about the role of patents in the pharmaceutical industry. First, patents are just one piece of a set of complicated regulations that include requirements for clinical testing and disclosure, along with grants of market exclusivity that function alongside patents. Second, it is widely believed that in the absence of legal protections, generics would hit the market side by side with the originals. This assumption is presumably based on the observation that when patents expire, generics enter immediately. However, this overlooks the fact that the generic manufacturers have had more than a decade to reverse-engineer the product, study the market, and set up production lines. Lanjouw's (1998) study of India prior to the recent introduction of pharmaceutical patents there indicates that it takes closer to four years to bring a product to market after the original is introduced—in other words, the first-mover advantage in pharmaceuticals is larger than is ordinarily imagined. Third, much development of pharmaceutical products is done outside the private sector; in Boldrin and Levine (2008b), we provide some details. Finally, the current system is not working well: as Grootendorst, Hollis, Levine, Pogge, and Edwards (2011) point out, the most notable current feature of pharmaceutical innovation is the huge "drought" in the development of new products.

With these four factors in mind, it is possible to make proposals for reforming the pharmaceutical industry along with the patent system. For example, we could either treat Stage II and III clinical trials as public goods (where the task would be financed by National Institutes of Health, who would accept bids from firms to carry out this work) or by allowing the commercialization of new drugs—at regulated prices equal to the economic costs of drugs—if they satisfy the Food and Drug Administration requirements for safety even if they do not yet satisfy the current (overly demanding) requisites for proving efficacy. In other words, pharmaceutical companies would be requested to sell new drugs at "economic cost" until efficacy is proved, but they could start selling at market prices after that. (It is ensuring

the efficacy—not the safety—of drugs that is most expensive, time-consuming, and difficult.) In this way, companies would face strong incentives to conduct or fund appropriate efficacy studies where they deem the potential market for such drugs to be large enough to bear the additional costs. The new policy could begin with drugs aimed at rare diseases, which, because of their small potential market, are not currently worth the costs of efficacy testing; without the new policy, they might never make it to market at all. If this new progressive approval approach works for rare diseases, it could be adopted across the board. Our broader point is that, rather than just ratcheting up patent protection, there are a number of moves we could make to reduce the risks and cost of developing new drugs.

The Political Economy of Patents

We do believe, along with many of our colleagues, that a patent system designed by impartial and disinterested economists and administered by wise and incorruptible civil servants could serve to encourage innovation. In such a system, very few patents would ever be awarded: only those for which convincing evidence existed that the fixed costs of innovation were truly very high, the costs of imitation were truly very low, and demand for the product was really highly inelastic. (The curious reader may check Boldrin and Levine, 2008a, for a more detailed explanation as to why these three conditions need to be satisfied to make a patent socially valuable). There is little dispute, among these same colleagues, that the patent system as it exists is very far from satisfying such requirements and it is, in fact, broken. To quote a proponent of patents, Shapiro (2007): "A growing chorus of scholars and practitioners are expressing concerns about the operation of the US patent system. While there is no doubt that the US economy remains highly innovative, and there is no doubt that the patent system taken as a whole plays an important role in spurring innovation, the general consensus is that the US patent system is out of balance and can be substantially improved." Actually, we believe the evidence is clear that the patent system taken as a whole does not play an important role in spurring innovation. But if a well-designed and well-administered patent system *could* serve the intended purpose, why not reform it instead of abolishing it?

To answer the question we need to investigate the political economy of patents: why has the political system resulted in the patent system we have? Our argument is that it cannot be otherwise: the "optimal" patent system that a benevolent economist–dictator would design and implement is not of this world. It is of course fine to recommend patent reform. But if political economy pressures make it impossible to accomplish that reform, or if they make it inevitable that the patent system will fail to meet its goals, then abolition—preferably by constitutional means as was the case in Switzerland and the Netherlands prior to the late nineteenth century—is the proper solution. This political economy logic brings us to advocate dismantlement of the patent system.

The political economy of patent protection is shaped by many players, but "consumers" are not prominent among them. On one side, the side of the potential patentees, there are individual inventors, corporate inventors, and patent trolls. Other players include the patent office, the patent lawyers who file and litigate patents, and the courts where the litigation takes place. The rules of the game are established by some combination of legislation, judicial action, and custom. But because patenting is a technical subject about which few voters know anything with clarity, interests of voters are not well represented. In many spheres of government regulation, this lack of representation for voters has often led to "regulatory capture"—as Stigler (1971) and other public choice theorists have argued—where regulators act in the interests of the regulated, not the broader public. Nowadays, if there is one "regulator" who is captured, it is the one in charge of regulating patents. To understand why, we need to understand the motivation and incentives of the relevant players.

Let us start with the US Patent Office and the infamous "one-click" patent #5960411 issued to Amazon in September 1999. According to 35 U.S.C. 103, the statute under which the Patent Office operates, to obtain a patent "the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been not obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains . . ." Now consider the patent in question, which claims, among other things, a monopoly over:

> 11. A method for ordering an item using a client system, the method comprising: displaying information identifying the item and displaying an indication of a single action that is to be performed to order the identified item; and in response to only the indicated single action being performed, sending to a server system a request to order the identified item whereby the item is ordered independently of a shopping cart model and the order is fulfilled to complete a purchase of the item.

The idea of taking a single action to accomplish a goal is hardly innovative, and applying the idea of taking a single action to making a purchase is obvious to anybody who has ever used a soft drink machine. Purchases were already being made over the Internet in 1999. It was thus clear that orders would be made by a credit card, and either the credit card information would be provided at the time of the transaction, or stored in advance by the retailer. Either way, the user must identify itself when the purchase is made. Those obvious steps are exactly what Amazon describes in its patent, albeit with a few flow charts thrown into the eleven-page patent application. But through the fog of those flow charts, it is relatively easy to see that the verbal description of the single-click procedure applies equally well to what happens on the Amazon site *and* to what happens in front of millions of vending machines every day. The Amazon patent was reexamined by the US Patent Office starting in May 2006. After a preliminary finding that, indeed, "obvious" means "obvious" even

at the Patent Office, the office then reversed itself and in October 2007, reaffirmed the Amazon patent, albeit limiting its scope slightly. So we cannot dismiss such an absurd patent as an aberration.

What lead the US Patent Office to interpret, essentially, the words "not obvious" as meaning "obvious"? The Patent Office is constantly under pressure from applicants and their lawyers to be more generous in issuing patents—that is, to adopt lower standards of obviousness and steeper standards for what is considered "prior art." The following statement by David Kappos (2010), director of the US Patent Office concerning the allowance rate—what fraction of patents are accepted—is revealing: "Overall in FY 2010, the allowance rate increased to 45.6%, compared to an allowance rate of 41.3% in FY 2009 . . . So, while we still have a lot of work to do, I think we are on the right path." Apparently, accepting a higher fraction of patents applications is defined as "the right path." Talk about "regulatory capture"!

Patent lawyers play a large role in the political economy of patents. According to Quinn (2011), who is a patent attorney, legal fees for filing a patent run upwards of $7,000 and roughly half are rejected. In 2010, according to the US Patent Office, 244,341 patents were issued, which would imply roughly $3 billion in legal fees per year. Obviously, patent attorneys as a group have a tremendous incentive to see that more patents are issued. This insight helps us understand better the role of the courts and their relatively recent reform. In 1982—lobbied by patent lawyers—Congress passed the Federal Courts Improvement Act, which moved federal patent appeals out of the regular court system to a special court system for dealing with patents. Naturally, many of the judges for this new court were chosen from the ranks of patent attorneys. For example, when a court voted, in a 1994 decision, to expand the scope of patents to software (In *re Kuriappan P. Alappat, Edward E. Averill and James G. Larsen 33 F.3d 1526* [July 29, 1994]), of the six judges who voted in favor, half had previously been patent attorneys, while of the two that voted against, neither had been. The referee of the patent game is biased both materially and ideologically. As Landes and Posner (2004, p. 26) write in their discussion of the political economy of patents: "That has been the experience with the Federal Circuit; it has defined its mission as promoting technological progress by enlarging patent rights."

Notice, too, that many patent lawsuits have a public goods aspect. Consider a case in which the plaintiff is asserting that its patent has been infringed. If the plaintiff wins the lawsuit, by confirming its monopoly position it appropriates all the benefits of winning the lawsuit. A victory by the defendant, by contrast, benefits partly itself, but also other firms that might be sued by the plaintiff for patent infringement as well as consumers who would have a more competitive market. Thus, the defendant receives only a slice of the overall benefits from winning the lawsuit, and will be willing to spend less on such lawsuits than it would if it were to receive all the benefits. This dynamic is nothing but the patent court version of the (already noted) fundamental asymmetry in the distribution of economic incentives that defines the foundations of the political economy of patent law.

Finally, political economy can be influenced by how standard terminology frames a problem. Landes and Posner (2004) point out that there is an "ideological"

argument in support of stronger patent rights: supporters of free markets tend to favor institutions of private property, and patents and copyright are intellectual "property." Hence, strengthening them is ideologically and politically consistent with the general principle that "private property is good for growth." But as we (Boldrin and Levine 2008b) and many others elsewhere have argued, patents are just a monopoly, not property.

Given this set of players and their incentives, the patent game moves naturally towards its equilibrium, as we have observed over time. Two centuries or so ago, patents were restricted in their areas of applicability and limited in both depth and duration over time; they were somewhat "reasonable," to the extent social gains and costs seemed balanced. But we have witnessed a steady process of enlargement and strengthening of patent laws. At each stage, the main driving force was the rent-seeking efforts of large, cash-rich companies unable to keep up with new and creative competitors. Patent lawyers, patent officials, and wannabe patent trolls usually acted as foot soldiers. While this political economy process is pretty straightforward in broad terms, we are still missing an empirical, quantitative analysis of the stakes involved and of the gains and losses accruing to both the active players and to the rest of society, from the general public to the innovators that never emerged due to preexisting patent barriers.

Perhaps surprisingly, despite the key importance of political economy in understanding why we have the patent system we have, economists have had relatively little to say on the subject. The few prominent papers that we know of on this subject typically build from analyses very similar to what we have presented here—but then shy away from drawing the logical conclusions.

For example, Landes and Posner (2004) recognize that patent laws are mostly designed by interest groups keen to increase their monopoly rents, not aggregate welfare, and that this drove the enormous growth in patent legislation and judiciary activity during the last 30 years. The more elaborate writing by Scherer (2009) on "The Political Economy of Patent Policy Reform in the United States" follows a similar approach. It focuses on the fact that "government emphasis on patent systems increased" while academic research was starting to become more and more aware that patents are playing a minor positive role, if any at all, in creating incentives for high R&D and in fostering productivity growth. After providing a concise and very well-informed historical survey of all major changes in US patent policies over the last century or so, Scherer (p. 195) wonders why the political system would increase patent protection so much in light of the fact "that the record of debates on the enabling bill contains no solid evidence that the change would in fact stimulate R&D, and that there is no evidence of an acceleration in company-financed R&D between the 27 years before the bill was enacted and the 18 years thereafter." He then extends the same argument to the international arena, paying particular attention to the case of pharmaceutical patents. While Scherer's language and arguments are strongly critical of current trends in patents, he does not seek to explain why an institution, such as the patent system, that was supposed to be theoretically sound would degenerate into something so socially damaging over same 30-year

period that academic researchers were realizing the institution's limitations and potential dangerousness.

In our view, even insightful writers such as Landes and Posner (2004) and Scherer (2009) seem unable to shake themselves free of the belief that patents are essential in fostering innovation and that any problems can be fixed with some tweaks to the patent system; they fail to seriously consider the possibility of intrinsic problems with the design of the institution itself. This belief in patents flies in the face of the structural realities: Marginal extensions of patents result in substantially higher per capita rents for the few holders of the right while marginally reducing the individual welfare of the much larger number of nonpatent holders. The rent of the monopolist is a lot higher than an individual consumer's deadweight loss, so the monopolist has an incentive to perpetuate the system while the individual consumer has no incentive to fight it. Those who possess a patent do not hold a "property right" in the conventional sense of that term, but they do hold a socially granted "monopoly" right, and will tend to leverage whatever initial rents their monopoly provides in order to increase their monopoly power until all potential rents are extracted (and, in all likelihood, also largely dissipated by the associated lobbying and transaction costs). This scenario helps explain how patents interact with the industry lifecycle—why patents are either ignored or scarcely used in new and competitive industries, while being highly valued and overused in mature and highly concentrated ones.

Conclusion

In 1958, the distinguished economist Fritz Machlup in testimony before Congress famously said: "If we did not have a patent system, it would be irresponsible, on the basis of our present knowledge of its economic consequences, to recommend instituting one. But since we have had a patent system for a long time, it would be irresponsible, on the basis of our present knowledge, to recommend abolishing it." A proposal to abolish patents may seem "pie in the sky." Certainly, many interim measures could be taken to mitigate the damage caused by the current system: for example, properly enforcing the standard that patents should only be granted for nonobvious insights; requiring genuine disclosure of working methods in patents (the opposite of certain recent "protectionist" proposals to institute secret patents); and allowing an "independent invention" defense against claims of patent infringement. But why use band-aids to staunch a major wound? Economists fought for decades—ultimately with considerable success—to reduce restrictions on international trade. A similar approach, albeit less slow, should be adopted to phase out patents. Because policy proposals are often better digested and metabolized in small bites, here is our list of small reforms that could be easily implemented.

1) Patents are time limited, which makes it relatively easy to phase them out by phasing in ever shorter patent durations. This conservative approach also

has the advantage that if reducing patent terms indeed has a measurable effect on innovation, the process can be reversed.

2) Stop the rising tide that, since the early 1980s, has extended the set of what can be patented and has shifted the legal and judicial balance substantially in favor of patent holders.
3) Because competition fosters productivity growth, antitrust and competition policies should seek to limit patents when they are hindering innovation. This policy may be of particular relevance for high-tech sectors, from software to bioengineering, to medical products and pharmaceuticals.
4) Current international trade negotiations that affect patents often occur as part of either the Agreement on Trade-Related Aspects of Intellectual Property Rights (*TRIPS*), which was signed in 1995 as part of the World Trade Organization negotiations, or as part of the World Intellectual Property Organization, an agency of the United Nations. The nature of these agreements and organizations is well indicated by the use of the propaganda term "intellectual property" in their titles. In both cases, these talks are often focused on how to prevent ideas from high-income countries from being used in low-income countries—what we would characterize as essentially a neomercantilist approach toward free trade in goods and ideas. We should be highly cautious about this agenda. Within a couple of decades, the "balance of trade in ideas" between the US and European economies and emerging economies in Asia might easily equalize or reverse. Engaging in "mercantilism of ideas" may seem favorable to certain large US firms now, but such rules may become costly to the US economy if they are applied to protect patents held in the future by producers in the now-developing Asian economies.
5) If the US economy is to have patents, we may want to start tailoring their length and breadth to different sectoral needs. Substantial empirical work needs to be done to implement this properly, although a vast legal literature is already pointing in this direction.
6) Patents should not be granted based only on technological insights, but should also take economic evidence into account. For example, if an invention is easy to copy or has a high fixed cost, then patent protection to provide an incentive for the inventor may be more suitable. Ultimately, patents should be awarded only when strictly needed on economic grounds, as spelled out earlier.
7) We advocate returning to the rule prior to the Bayh–Dole Act of 1980 according to which the results of federally subsidized research cannot lead to patents, but should be available to all market participants. This reform would be particularly useful for encouraging the dissemination of innovation and heightening competition in the pharmaceutical industry.
8) In several industries, notably pharmaceuticals, it would be useful to rethink all of the government policies that bear on incentives for invention. The broad point is that there are a number of ways to reduce the risks and cost of developing new drugs, rather than just trying to ratchet up patent protection.

In general, public policy should aim to decrease patent monopolies gradually but surely, and the ultimate goal should be the abolition of patents. After six decades of further study since Machlup's testimony in 1958 has failed to find evidence that patents promote the common good, it is surely time to reassess his conclusion that it would be irresponsible to abolish the patent system. The patent system arose as a way to limit the power of royalty to award monopolies to favored individuals; but now its primary effect is to encourage large but stagnant incumbent firms to block innovation and inhibit competition.

■ *We are grateful to the editors, the referees, and to Richard Stallman for a careful reading and comments.*

References

Aghion, Phillipe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. 2005. "Competition and Innovation: An Inverted-U Relationship." *Quarterly Journal of Economics* 120(2): 701–728.

Boldrin, Michelle, Juan Correa Allamand, David K. Levine, and Carmine Ornaghi. 2011, "Competition and Innovation." In *Cato Papers on Public Policy*, Vol. 1, edited by J. A. Miron, 109–172. Cato Institute.

Boldrin, Michele, and David K. Levine. 2004. "Rent Seeking and Innovation." *Journal of Monetary Economics* 51(1): 127–60.

Boldrin, Michele, and David K. Levine. 2005. "The Economics of Ideas and Intellectual Property." *Proceedings of the National Academy of Sciences* 102(4): 1252–56.

Boldrin, Michele, and David K. Levine. 2008a. "Perfectly Competitive Innovation." *Journal of Monetary Economics* 55(3): 435–53.

Boldrin, Michele, and David K. Levine. 2008b. *Against Intellectual Monopoly*. Cambridge University Press.

Bessen, James, and Michael J. Meurer. 2005. "The Patent Litigation Explosion." BU School of Law Working Paper 05–18, Boston University.

Bessen, James, and Michael J. Meurer. 2008. *Patent Failure: How Judges, Bureaucrats, and Lawyers Put Innovators at Risk*. Princeton University Press.

Brec, Eric. 2008. "NIHilism and Other Innovation Poison." MSDN Blogs, November 1. http://blogs.msdn.com/b/eric_brechner/archive/2008/11/01/nihilism-and-other-innovation-poison.aspx.

Chaudhuri, Shubham, Pinelopi K. Goldberg, and Panie Gia. 2006. "Estimating the Effects of Global Patent Protection in Pharmaceuticals: A Case Study of Quinolones in India." *American Economic Review* 96(5): 1477–1514.

Cohen, Wesely M., Richard R. Nelson, and John P. Walsh. 2000. "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)." NBER Working Paper 7552.

Correa, Juan A. 2012. "Innovation and Competition: An Unstable Relationship." *Journal of Applied Econometrics* 27(1): 160–66.

Danaher, Brett, Samita Dhanasobhon, Michael D. Smith, and Rahul Telang. 2010. "Converting Pirates without Cannibalizing Purchasers: The Impact of Digital Distribution on Physical Sales

and Internet Piracy." March 3. Available at SSRN: http://ssrn.com/abstract=1381827 or http://dx.doi.org/10.2139/ssrn.1381827.

DiMasi, Joseph A., Ronald W. Hansen, and Henry G. Grabowski. 2003. "The Price of Innovation: New Estimates of Drug Development Costs." *Journal of Health Economics* 22(2): 151–85.

Gallini, Nancy T. 2002. "The Economics of Patents: Lessons from Recent U.S. Patent Reform." *Journal of Economic Perspectives* 16(2): 131–154.

Gates, Bill. 1991. "Challenges and Strategy." Memo, Microsoft Corporation, May 16. http://www.std.com/obi/Bill.Gates/Challenges.and.Strategy.

Grootendorst, Paul, Aidan Hollis, David K. Levine, Thomas Pogge, and Aled M. Edwards. 2011. "New Approaches to Rewarding Pharmaceutical Innovation." *CMAJ: Canadian Medical Association Journal* 183(6): 681–85.

Hall, Bronwyn H., and Rosemarie Ham Ziedonis. 2007. "An Empirical Analysis of Patent Litigation in the Semiconductor Industry." http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.69.5271.

Hashmi, Aamir Rafique. 2011. "Competition and Innovation: The Inverted-U Relationship Revisited." Available at SSRN: http://ssrn.com/abstract=1762388 or http://dx.doi.org/10.2139/ssrn.1762388.

Heller, Michael. 2008. *The Gridlock Economy: How Too Much Ownership Wrecks Markets, Stops Innovation, and Costs Lives*. Basic Books.

Hong, Sungook. 2001. *Wireless: from Marconi's Black-Box to the Audion*. MIT University Press.

Jaffe, Adam B. 2000. "The U.S. Patent System in Transition: Policy Innovation and the Innovation Process." *Research Policy* 29(4–5): 531–57.

Jovanovic, Boyan, and Glenn M. MacDonald. 1994. "The Life Cycle of a Competitive Industry." *Journal of Political Economy* 102(2): 322–47.

Kanwar, Sunil, and Robert Evanson. 2001. "Does Intellectual Property Protection Spur Technological Change?" Levine's Working Paper Archive, no. 122247000000000455.

Kappos, David. 2010. "Reflections on the USPTO Dashboard." Director's Forum: David Kappo's Public Blog, October 13. http://www.uspto.gov/blog/director/entry/reflections_on_the_uspto_dashboard.

Kremer, Michael, and Heidi Williams. 2009. "Incentivizing Innovation: Adding to the Toolkit." In *Innovation Policy and the Economy*, Vol. 10, edited by Josh Lerner and Scott Stern, 1–17. Chicago University Press.

Landes, William M., and Richard A. Posner. 2004. *Political Economy of Intellectually Property Law*. Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.

Lanjouw, Jean O. 1998. "The Introduction of Pharmaceutical Product Patents in India: Heartless Exploitation of the Poor and Suffering?" NBER Working Paper 6366.

Lerner, Josh. 2002. "Patent Protection and Innovation over 150 Years." http://www.epip.eu/papers/20030424/epip/papers/cd/papers_speakers/Lerner_Paper_EPIP_210403.pdf (quotes are from this version). Abridged version published in 2002 as "150 Years of Patent Protection," *American Economic Review* 92(2): 221–25.

Lerner, Josh. 2009. "The Empirical Impact of Intellectual Property Rights on Innovation: Puzzles and Clues." *American Economic Review* 99(2): 343–48.

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*, no. 3, Special Issue on Microeconomics, pp. 783–820.

Llanes, Gastón, and Stefano Trento. 2009. "Patent Policy, Patent Pools, and the Accumulation of Claims in Sequential Innovation." *Economic Theory* 50(3): 703–25.

Machlup, Fritz. 1958. "An Economic Review of the Patent System." Study commissioned by the Senate Judiciary Subcommittee on Patents, Trademarks, and Copyrights, 85th Congress, 2nd session. http://mises.org/document/1182/.

Niccolai, James. 2012. "Android Developers Ignored Sun Patents, Google Exec Testifies." *Computerworld*, May 9.

Nuvolari, Alessandro. 2004. "Collective Invention during the British Industrial Revolution: The Case of the Cornish Pumping Engine." *Cambridge Journal of Economics* 28(3): 347–63.

Nuvolari, Alessandro. 2006. "The Making of Steam Power Technology: A Study of Technical Change during the British Industrial Revolution." *Journal of Economic History* 66(2): 472–76. Cambridge University Press.

Ponce, Carlos J. 2007. "More Secrecy . . . More Knowledge Disclosure? On Disclosure Outside of Patents." Levine's Working Paper Archive, no. 122247000000001600.

Quinn, Gene. 2011. "The Cost of Obtaining a Patent in the US." *IPWatchdog.com*, January 28. http://www.ipwatchdog.com/2011/01/28/the-cost-of-obtaining-patent/id=14668/.

Salmon, Felix. 2012. "Why Patent Trolls Don't Need Valid Patents." March 4. Reuters. http://blogs.reuters.com/felix-salmon/2012/03/04/why-patent-trolls-dont-need-valid-patents/.

Sanchez, Julian. 2012. "How Copyright Industries Con Congress." Posted at Cato@Liberty, http://www.cato-at-liberty.org/how-copyright-industries-con-congress/.

Scherer, F. M. 1990. *Industrial Market Structure and Economic Performance*. Houghton Mifflin Company.

Scherer, F. M. 2009. "The Political Economy of Patent Policy Reform in the United States." *Journal on Telecommunications and High Techonology Law* 7(2):167–216.

Shapiro, Cark. 2007. "Patent Reform: Aligning Reward and Contribution." NBER Working Paper 13141.

Shulman, Seth. 2003. *Unlocking the Sky: Glenn Hammond Curtiss and the Race to Invent the Airplane*. Harper Perennial.

Sinha, Rajiv K., Fernando S. Machado, and Collin Sellman. 2010. "Don't Think Twice, It's All Right: Music Piracy and Pricing in a DRM-Free Environment." *Journal of Marketing* 74(2): 40–54.

Stigler, George J. 1956. "Industrial Organization and Economic Progress." In *The State of the Social Sciences*, edited by Leonard D. White, 269–82. University of Chicago Press.

Stigler, George J. 1971. "The Theory of Economic Regulation." *Bell Journal of Economics and Management Science* 2(1): 3–21.

Svensson, Peter. 2006. "Long-Running Blackberrry Patent Battle Ends with $612.5 Million Settlement." *The Daily Reporter* (Associated Press), March 4. http://news.google.com/newspapers?nid=1907&dat=20060304&id=5bVGAAAAIBAJ&sjid=0_0MAAAAIBAJ&pg=4540,273762.

US Central Intelligence Agency. n.a. *The 1990 CIA World Fact Book*. Available from Project Gutenberg: https://ia700401.us.archive.org/32/items/the1990ciaworldf00014gut/world12.txt.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 23–44

Patents and Innovation: Evidence from Economic History

Petra Moser

What is the optimal system of intellectual property rights to encourage innovation? In the most basic theoretical models, patents pose a tradeoff between the social benefits from stronger incentives for invention and losses in consumer welfare as a result of monopoly pricing (Nordhaus 1969). But providing stronger patents for early generations of inventors may also weaken incentives to invest in research and development for later generations (for example, Scotchmer 1991 in this journal), so that the overall effects of stronger patents on innovation are difficult to predict. Negative incentive effects are particularly severe if the boundaries of intellectual property are poorly defined, so that later generations of inventors place themselves at risk of ruinous litigation. Litigation risks are exacerbated when incumbents build "thickets" of strategic patents that cover little innovative progress and instead serve as a legal weapon to protect incumbents' profits (Shapiro 2001; Hall and Ziedonis 2001). Recent patent wars over smart phones and tablet computers have moved these issues to the forefront of policy debates, but the underlying tensions are substantially more general. Empirical analyses that exploit a wealth of historical datasets and exogenous variation, when done carefully, can help to improve our understanding of these tensions and inform contemporary patent policy.

Empirical analyses of historical data have emphasized the role of patent laws in creating incentives to invent, promoting innovation, and encouraging economic growth (for example, Khan and Sokoloff 1993; Lamoreaux and Sokoloff 1999; Khan 2005). In the absence of economy-wide data on the quantity of innovations, patent

■ *Petra Moser is a Fellow at the Center for Advanced Studies in the Behavioral Sciences and Assistant Professor of Economics, Stanford University, Stanford, California, and Faculty Research Fellow, National Bureau of Economic Research, Cambridge, Massachusetts.*

http://dx.doi.org/10.1257/jep.27.1.23. doi=10.1257/jep.27.1.23

counts have become the standard measure of innovation (for example, Schmookler 1962, 1966; Sokoloff 1988; Moser and Voena 2012), fueled in part by the creation of National Bureau of Economic Research dataset of US patents and citations between 1976 and 2002 (Hall, Jaffe, and Trajtenberg 2001), and more recently by the availability of historical patent data since 1920 through a collaboration between the US Patent and Trademark Office and Google Patents.

Patent data may, however, fail to capture innovation that occurs *outside* of the patent system—for example, in countries without patent laws or in industries in which inventors rely on alternative mechanisms to protect their intellectual property. In fact, survey data for the late twentieth century indicate that commercial research and development labs in most industries deem alternative mechanisms, such as secrecy and lead-time (being the first firm to offer a new product) to be more effective than patents (Levin, Klevorick, Nelson, and Winter 1987; Cohen, Nelson, and Walsh 2000). Historical accounts also indicate that innovation often occurs independently of patents as a result of knowledge sharing (Allen 1983; Nuvolari 2004; Thomson 2009) or cultural attitudes that encourage risk taking (Landes 1969) and scientific experimentation (Mokyr 2009).

Historical events—including a series of prominent technology exhibitions that started with the 1851 Crystal Palace world's fair in London—have created rich archival records on innovation within *and outside* of the patent system, which offer opportunities to measure the share and the characteristics of innovations that occur outside of the patent system. Data on exhibits and prizes that international juries awarded to the most innovative exhibits make it possible to examine innovation in countries without patent laws, and thus to exploit a large amount of credibly exogenous variation in patent laws to investigate the effects of patent laws on innovation. Patent laws that were in force in the mid-nineteenth century had largely been adopted ad hoc according to idiosyncratic allegiances of national rulers (Penrose 1951, p. 13) and before interest groups from individual industries had learned to lobby for stronger patents. Scientific breakthroughs that reduced the effectiveness of alternative mechanisms to protect intellectual property created exogenous shifts towards patenting, which make it possible to examine the role that patents play, for example, in the diffusion of ideas. Historical events, such as the creation of the first patent pool in 1856 and the compulsory licensing of enemy-owned US patents as a result of World War I, create opportunities to examine the effects of policies that strengthen or weaken the monopoly power of patents.

To use historical evidence to guide patent policies today, one must carefully compare historical and modern institutions, political conditions, and changes in the technological characteristics of industries over time. Empirical evidence from economic history, however, can help to inform important policy questions that have proven difficult to answer with modern data. For example, does the existence of strong patent laws encourage innovation? What proportion of innovations is patented? Is this share constant across industries and over time? How does patenting affect the diffusion of knowledge? How effective are prominent mechanisms, such

as patent pools and compulsory licensing, that have been proposed to address problems with the patent system? [1]

Have Patent Laws Increased the Rate of Innovation?

In 1474, the Venetian Republic began to offer exclusive rights to inventors and entrepreneurs who had invented or brought new technologies to Venice. Intended to attract skilled artisans, the Republic's rudimentary patent system was copied by other European rulers to promote economic development and, more frequently, to reward political and financial support (David 1994, p. 134; Boldrin and Levin 2008, p. 43–44). In 1623, Britain's Statute of Monopolies transferred the right of granting monopolies from King James I to Parliament. North and Thomas (1973) argue that this shift, which replaced a royal prerogative to sell monopolies by a legal property rights in ideas, played a critical role in encouraging Britain's Industrial Revolution. The first article of the US Constitution instructed Congress to "promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." This provision established the foundation for the world's first modern patent system, which Khan and Sokoloff (1998, 2001) argue was instrumental in encouraging technological progress and economic growth in the United States.

Recent interpretations, however, contend that patents played no major role in encouraging technological development and economic growth during Britain's Industrial Revolution (Clark 2006; Mokyr 2009; Allen 2009). Mokyr (2009), for example, emphasizes the importance of a shift towards science-based experimentation during the Enlightenment in setting the stage for Europe's Industrial Revolution. Alternative accounts of US innovation have emphasized the importance of relative factor prices, and in particular, the high costs of labor relative to the abundance of natural resources, as an impetus for mechanization, and for the development of a specifically American system of manufacturing (Rothbarth 1946; Habbakuk 1962; Rosenberg 1963, 1969, 1972; Hounshell 1985).

Historical variation in patent laws in the nineteenth century—when some countries had not yet adopted patent laws while other abolished them for political reasons—offers unique opportunities to investigate the effects of patent laws on innovation. Switzerland, for example, had no patents until the country adopted a rudimentary patent system in 1888 and switched towards a full-fledged system in 1907 (Schiff 1971). Denmark provided limited patent protection for up to five years

[1] In addition to patents, innovation policy includes other types of intellectual property rights, such as copyrights, which protect books, music, and software. National governments have also begun to increasingly use prizes as an alternative mechanism to encourage innovation. More generally, the ability to attract high-skilled scientists and workers is likely to be a key factor in determining rates of innovation. Economic history also offers rich opportunities to explore the effectiveness of these alternative mechanisms (see for example Li, MacGarvie, and Moser 2012; Moser, Voena, and Waldinger 2011; Moser and Nicholas 2012).

in 1874, but waited until 1894 to enact an official patent law (Agnew 1874, p. 430; Boult 1895, p. 136). The Netherlands abolished its patent system in 1869 after a political victory of the free trade movement, which reflected a common view of patents as a form of protectionism and rejected them as a restriction on trade (Schiff 1971). Even for countries with patent laws, the strength of patents was far from uniform. In 1876, for example, patents in Denmark and Greece expired after five years, while patents in other countries lasted for a minimum of twelve years (Lerner 2000). Inventors around the world were also heavily dependent on domestic patent laws because patenting abroad was prohibitively expensive and—until the Paris Convention of 1883—national patent systems discriminated heavily against foreign patentees (Bilir, Moser, and Talis 2011).

Analyses of technologies that were exhibited at nineteenth-century world's fairs exploit such variation to examine differences in innovation for countries with and *without* patent laws. Exhibition catalogues, which guided visitors through the vast grounds of nineteenth- and early twentieth-century technology fairs, list all exhibits. Collecting these data and matching them with reports on prize-winning innovations, as well as with patent data and with geographic information, makes it possible to examine the number and the characteristics of innovations that occurred inside and outside of the patent system, which has been difficult to accomplish using patent counts as the standard indicator of innovation.

Exhibition data are available for the Crystal Palace Exhibition in London in 1851, the American Centennial Exhibition in Philadelphia in 1876, the World's Columbian Exhibition in Chicago in 1893, and the Panama-Pacific International Exposition in San Francisco in 1915. In 1851, the Crystal Palace, a 1,848-foot long greenhouse of cast iron and glass, was the largest enclosed space on earth; it housed 17,062 exhibitors from 40 countries. At a time when London had fewer than two million inhabitants, more than six million entry tickets were sold for the Crystal Palace. In 1876, visitors at the US Centennial Exhibition would have had to walk more than the distance of a marathon to see 30,864 exhibitors from 35 countries; almost ten million people visited the fair (Kroker 1975, p. 146). In 1893, the World's Columbian Exposition covered 717 acres of land and water in Jackson Park by Lake Michigan; it attracted 27.5 million visitors. In 1915, San Francisco's Marina and Presidio was converted to a fairground; it welcomed 30,000 exhibitors from 32 countries and 19 million visitors.

Analyses of the 1851 and 1876 exhibits reveal a perhaps surprising amount of high-quality innovations in countries without patent laws. In 1851, Switzerland and Denmark contributed 110 exhibits per million people, compared with a mean of 55 and a median of 36 per million people for all countries (Moser 2005). Swiss exhibits were also more likely to win prizes for exceptional novelty and usefulness. In 1851, 43 percent of Swiss exhibits won a prize, compared with a mean of 35 percent and a median of 33 percent for all countries. In 1876, Switzerland contributed 168 exhibits per million in population, compared with a mean of 87 and a median of 61 for all countries (Moser and Zimring 2012). The Netherlands—which had abolished patents in 1869—won more prizes per

exhibit than any other country, with 86 percent, compared with a mean of 46 and a median of 45 percent for all countries.

The world's fair data also indicate that only a small share of innovations were patented, calling into question the role of intellectual property rights in encouraging Britain's Industrial Revolution. In 1851, 11 percent of British exhibits were patented. These results are consistent with historical accounts, which emphasize the importance of cultural factors (Clark 2006; Mokyr 2009) as well as systems of collective invention without patents. For example, improvements in Cornish steam engines (Nuvolari 2004) and in blast furnaces in Cleveland's iron industry in the United Kingdom were shared freely within a system of collective invention (Allen 1983) in which patenting was rare.[2]

Data on prize-winning British exhibits help to shed light on the interaction between the quality of inventions and inventors' decision to use patents. Existing theoretical models indicate that firms may decide to keep important innovations secret because patents require disclosure, which is risky if patents are ineffective at blocking competitors from using a patented invention (Anton and Yao 2004; Horstmann, MacDonald, and Slivinski 1985). Exhibition data, however, indicate that high-quality innovations are slightly more likely to be patented: In 1851, 15 percent of British exhibits that won prizes for exceptional usefulness and quality were patented, compared with 11 percent of average-quality exhibits.

Exhibition data on the share of innovations without patents make it possible to examine how the characteristics of patent institutions influence inventors' use of patents. Khan and Sokoloff (1998, 2001, in this journal) have credited the design and low costs of patenting under the US system with encouraging technical progress and economic growth through the "democratization" of invention. In the mid nineteenth century, British inventors faced a drawn-out and expensive process, with exorbitant legal fees and bribes (MacLeod 1988, p. 76) in addition to official fees of $37,000 (in 2000 US dollars, Lerner 2000).[3] By comparison, US inventors could mail in their applications and paid only $618 in fees (in 2000 US dollars, Lerner 2000). Patenting rates, however, were only slightly higher for US compared with British exhibits—at 15 compared with 11 percent (Moser 2012, p. 54).

US courts have also always been more likely to uphold the patent rights of early generations of inventors, while British courts tended to be more anti-patent (Dutton 1984; Khan 2005). This pro-patent bias may, however, have *discouraged* US rates of innovation as early as the mid nineteenth century, anticipating problems with the current system (Bessen and Meurer 2008). In 1846, for example, the US Patent and Trademark Office issued patent 4,750 to Elias Howe for an *Improvement in Sewing Machines*. Howe's patent was broad enough to cover most commercially viable

[2] Inventions within systems of collective invention were predominantly incremental (or micro-, rather than macro-inventions, Mokyr 1990), which Landes (1969, p. 92) argues "were probably more important in the long run than the major inventions that have been remembered in history books."

[3] Reforms of the British and other European patent systems during the "Patent Controversy" (1855–1873) may have been triggered by the Crystal Palace exhibition and the unexpected quality of US innovations (Machlup and Penrose 1950; Rosenberg 1969, p. 2).

sewing machines at the time. Like a twenty-first century "patent troll," Howe used his patent to threaten litigation instead of commercializing his invention. In 1852, a District Court upheld Howe's patent, and he began to collect license fees of $25 per machine, roughly one-fifth the average price of a sewing machine (Lampe and Moser 2012b). Then other firms sued based on their own patents, and production came to a near halt in the 1851–1856 "sewing machine wars" (Bissell 1999, p. 84). By 1867, Howe had received $2 million in license fees (Parton 1867) roughly $27.8 million in 2011 dollars (converted using the GDP deflator, based on data from Officer and Williamson 2011).

Did the Creation of Plant Patents in 1930 Encourage Innovation?

Throughout the early twentieth century, living organisms such as livestock, bacteria, and plants could not be patented. After World War I, however, concerns about food security motivated the creation of intellectual property rights for plants that propagate asexually (through roots rather than seeds) in the US Plant Patent Act of 1930. Breeders of food crops had argued that, in the absence of effective alternative mechanisms, they were heavily dependent on patent rights to recover large development costs. The Stark Brothers Nursery, for example, had built a large cage, armed with a burglar alarm, to prevent competitors from stealing cuttings of the first Golden Delicious apple tree, as shown in Figure 1. By creating plant patents, Congress hoped to encourage domestic innovation and the development of a domestic US plant breeding industry.

Nearly half of all US plant patents between 1930 and 1970, however, were for roses, suggesting that the 1930 legislation may have missed its target of establishing food security (Moser and Rhode 2012, pp. 418–420). Anecdotal evidence indicates that the creation of plant patents may have facilitated the development of a research-based US rose breeding industry. Similar to pharmaceutical research and development today, it took up to twelve years to develop a new rose, and fewer than one in 1,000 seedlings typically proved commercially successful (Robb 1964, p. 389; Stewart 2007, p. 131). Once a new rose had been developed, it was easy for competitors to copy and propagate through cuttings, so that original breeders could not rely on secrecy or being first to recuperate their costs of research and development. Until World War II, US nurseries had depended on imported nursery stock from Europe, but in the 1940s, roughly a decade after the Plant Patent Act, commercial nurseries, which account for the majority of plant patents, began to build mass hybridization programs for roses.

Data on registrations of newly created roses between 1916 and 1970, as an alternative measure of innovation, however, suggest that the effect of plant patents was limited. Registration data suggest that US breeders created *fewer* new roses after 1931. Moreover, less than 20 percent of new rose varieties registered after 1930 were patented (Moser and Rhode 2012, pp. 429–434). In fact, information on lineage indicates that most roses that are commercially successful today descended from

Figure 1
A Cage that Stark Brothers Nursery Built around Its Golden Delicious Apple Tree

Source: Image from Rossman (1930, p. 395), reproduced in Moser and Rhode (2012, p. 415).
Note: The cage was built around the Stark Brother's Golden Delicious tree to prevent competitors from stealing shoots of the tree; it was equipped with an alarm.

the breeding efforts of public sector plant scientists that preceded the creation of plant patents. Furthermore, historical records suggest that the US rose industry received a boost when World War II cut off rose supplies from European competitors and US breeders began to produce their own nursery stock based on licensed European roses.

Patents, Secrecy, and the *Direction* of Technical Change

Exhibition data also indicate that the share of innovations that inventors chose to patent varied strongly across industries. For example, fewer than 5 percent of Britain's chemical exhibits in 1851, 10 percent of scientific instruments, and 8 percent of exhibits in food processing were patented, compared with 20 percent of manufacturing machinery (Moser 2012). Remarkably, US and British inventors appear to have relied on patents—and avoided patents—in the same industries despite vast differences between the British and the American patent system. Historical accounts suggest that variation in the effectiveness of secrecy, as an alternative

to patents, was instrumental in determining variation in the use of patents. Secrecy was an effective mechanism to protect mid-nineteenth-century improvements in chemicals because science had not yet evolved enough to allow competitors to reverse engineer them. Given the crude analytical tools of the time, valuable dyes such as indigo and madder red proved impervious to industrial espionage until the late nineteenth century (Haber 1958, p. 83). Secrecy was also effective in protecting improvements in the production of scientific instruments, such as the rectangular prisms of Swiss glassmaker T. Daguet of Soleure and the optical instruments of Danish makers (Berichterstattungs-Kommission, vol 1, 1853, pp. 813–19, 930–41). Watchmakers in the Swiss Valleé de Joux maintained tight secrecy surrounding an improved mechanism to measure minutes by agreeing not to take apprentices between 1823 and 1840 (Jaquet and Chapuis 1945, p. 165).

But if inventors' dependence on patents varies across industries, patent laws may influence the *direction* of technical change (Moser 2005): In countries without patent laws, inventors depend entirely on secrecy, lead time, and other alternatives to patents in protecting their intellectual property. As a result, investments in research and development may be most attractive in industries in which secrecy can effectively guarantee exclusive rights long enough to allow inventors to recoup their investments. In countries *with* patent laws, inventors can use legal protection to establish exclusivity in any industry, so factors other than the effectiveness of secrecy determine the direction of technical change.

Cross-country comparisons of exhibition data confirm that innovation in countries without patent laws focused on a narrow set of industries in which secrecy was effective. At the Crystal Palace, one-fourth of exhibits from countries without patent laws were scientific instruments, compared with one-seventh of exhibits from other countries (Moser 2005). Countries without patent laws also had larger shares of innovations in textiles, especially dyes, and in food processing.

In food processing, the history of margarine illustrates the effectiveness of secrecy relative to patents. The French chemist Mège Mouriès, for example, believed his invention to be protected by a patent, and disclosed the process of producing margarine from suet to two Dutch entrepreneurs, Jurgens and van den Bergh. Jurgens and van den Bergh began to manufacture margarine in 1871—two years after the Netherlands had abolished patent laws in response to a victory of the free-trade movement. After a falling out, van den Bergh kept his improvements secret, and Jurgens was unable to reverse engineer the superior taste of van den Bergh margarine (which allowed for its commercialization) until 1905 (Schiff 1971).

More generally, the share of Dutch innovations in food processing experienced a marked increase after the Netherlands abolished patents in 1869. In 1851, 11 percent of exhibits from the Netherlands were related to food processing. In 1876, 37 percent of Dutch exhibits, including a disproportionate amount of award-winners, originated from this industry (Moser 2005). Many other innovations in the field, including milk chocolate, baby foods, and ready-made soups, were made in Switzerland and the Netherlands when neither country offered patents (Schiff 1971, pp. 52–58).

Survey data from the late twentieth century indicate that the relative effectiveness of secrecy and patents continued to vary strongly across industries. For example, respondents from 634 American research and development labs in 1983 (Levin, Klevorick, Nelson, and Winter 1987) and from 1,478 American firms in 1994 (Cohen, Nelson, and Walsh 2000) report that secrecy is more effective than patents as a mechanism to protect intellectual property in most industries. Harhoff and Hoisl (2006) present comparable evidence for European countries. Only for pharmaceuticals and chemical inventions are patents consistently rated as an effective mechanism to protect intellectual property today. Compared with mid-nineteenth-century reports, which emphasize the effectiveness of secrecy to protect chemical inventions, these results indicate that the effectiveness of secrecy varies not only across industries, but also over time.

Scientific breakthroughs, which lowered the effectiveness of secrecy, may be one important factor that determines inventors' propensity to patent. In chemicals, for example, analytical advances such as August Kekulé's model of the benzene ring in 1865 and Dmitrii Mendeleev's publication of the periodic table in 1869, transformed chemical analysis in the second half of the nineteenth century. As a result of these advances, it became much riskier to protect chemicals through secrecy (Haber 1958, p. 81). At the same time, these analytical advances had no effects on innovations in machinery, which had always been easy to copy.

In Moser (2012), I exploit this differential shift to examine the effects of exogenous changes in the effectiveness of secrecy on inventors' propensity to patent. Difference-in-differences comparisons reveal a significant shift towards patenting in response to analytical advances: In 1851 and 1876, 0 and 5 percent of US chemical innovations were patented, respectively. In 1893 and 1915, 19 and 20 percent of US chemical innovations were patented, respectively. During the same time, patenting rates in manufacturing machinery—an industry in which secrecy was always ineffective—stayed roughly constant between 44 and 49 percent (Moser 2012, pp. 62–67). These results suggest that scientific breakthroughs, such as the publication of the periodic table in the nineteenth century or the decoding of the human genome today, may not only affect the speed of innovation but also increase inventors' dependency on patents.

Patent Laws and the *Diffusion* of Innovation

This science-driven shift towards patenting makes it possible to explore whether patent rights encourage the geographic diffusion of innovative activity, which in turn has important consequences for cumulative innovation and economic growth. Analyses of patent laws typically focus on incentive effects and have largely ignored diffusion, even though disclosure and teaching a new set of firms about the "mysteries" of more advanced technologies was an important goal of early patent systems (David 1994). In fact patents are often considered as a mechanism

to prevent rather than encourage the diffusion of patented ideas. As Abramovitz (1989, pp. 39–40) wrote:

> [T]here is a need to balance the potential private rewards of innovation, which are the incentive for private investment, against the social interest in spreading knowledge and encouraging its widespread and rapid commercial application. The first element calls for protecting the private investor in an exclusive right to exploit the new knowledge he has gained. The second calls for limiting that exclusive privilege to permit diffusion and to support the competitive investments of rivals.

Lamoreaux and Sokoloff (1999), however, link the increase in US patenting in the late nineteenth century with the emergence of professional patent agents, whose role was to facilitate the trade in patented ideas. The case of Mège Mouriès (the unfortunate inventor of margarine) suggests that inventors may be more willing to disclose technical information to competitors if they feel protected by a patent. In another example from early nineteenth-century England, the UK iron founder Robert Ransome began to advertise his plough-shares to all ironmongers in Norwich and 50 outlets in East Anglia after he received a patent in 1803 (MacLeod 1988, p. 100). By contrast, inventors have fiercely guarded knowledge from spreading to people outside their social network in the absence of intellectual property. For example, silk weavers in seventeenth-century Bologna hanged Ugolino Menzani for sharing the knowledge of a new silk twisting machine with Venetian weavers (Belfanti 2004, p. 581), and mechanics in the nineteenth-century Pennsylvania cotton industry relied on family relations to exchange technical knowledge (Wallace 1986, pp. 211–46).

In Moser (2011), I exploit the shift towards patenting in the nineteenth-century chemicals industry to explore whether patenting may, in fact encourage the diffusion of innovative activity: by creating intellectual property rights in ideas, patents may encourage inventors to disseminate knowledge of patented inventions, which in turn facilitates cumulative innovation and learning by doing.[4] A geographic analysis of exhibition data confirms that the shift towards patenting in chemicals was followed by a significant weakening in the geographic localization of inventive activity in chemicals. This decline in geographic concentration cannot be explained by changes in the localization of production; data from decennial census records for 1840 to 1920 indicate that the localization of chemical production remained relatively stable after 1876. Measuring changes in the diffusion of innovations by a geographic Herfindahl–Hirschmann index and using 1876 as a baseline, geographic concentration decreased by more than 70 percent for chemicals after 1876, compared with roughly 25 percent for manufacturing machinery. Difference-in-differences regressions, which compare changes after 1876 in the geographic

[4] See Scotchmer (1991) for a survey of the literature on cumulative innovation.

concentration of innovations in chemicals and manufacturing machinery, indicate that a 1 percent increase in the share of patented innovations was associated with a 1.3 percent decrease in localization.

Thus, the sum of the historical evidence from exhibition data, plant patents, and other sources indicates that patent laws may influence the direction of technological change and help to encourage the diffusion of knowledge, even though patent laws do not appear to be a necessary or sufficient condition for higher rates of innovation.

Mechanisms to Modify Patent Laws: Patent Pools

How can economic policy modify existing patent systems to make them more effective? A major problem with any patent system lies in the difficulty of defining the boundaries of the technology space that is covered by a patent. As a result, patent examiners may issue patents that cover overlapping areas of the technology space, such that two or more firms own blocking patents for the same technology. This in turn leads to infringement litigation, which impedes the production of new technologies and may discourage innovation.

Patent pools, which allow a group of firms to combine their patents, have emerged as a prominent mechanism to resolve blocking patents and prevent or resolve patent wars. In the 1990s, four pools formed in the information technology industry: the MPEG-2 pool, the 3G platform, and two DVD pools (Merges 2001). More recently, Google launched an open-source video format pool to counter MPEG LA's pool for the H.264 video coding standard, and MPEG LA has announced plans for a pool to cover kits for diagnostic genetic testing.

Although patent pools may weaken the intensity of competition, as they allow a group of firms to combine their individually held patents, regulators and courts have allowed pools, arguing, "In a case involving blocking patents, such an arrangement is the only reasonable method for making the invention available to the public" (*International Mfg. Co. v. Landon*, 336 F.2d 723, 729 [9th Cir. 1964]). Another argument in favor of pools is that, at least in theory, pools that combine complementary patents may reduce license fees for outside firms as they eliminate "n-marginalization," which occurs when firms that own patents for parts of a product charge license fees that are too high compared with the profit-maximizing fee for the complete product (Lerner and Tirole 2004; Shapiro 2001, p. 134).

This positive view of patent pools is consistent with the early history of a pool that formed in the US aircraft industry to encourage the production of planes during World War I. In 1917, patent litigation between the Orville and Wilbur Wright Company and their competitor, the Curtiss Company, had brought the US production of planes to a halt. A committee under Franklin Roosevelt, then Assistant Secretary of the Navy, recommended that Wright and Curtiss form a patent pool. After the pool had formed, US output of aircraft increased from 83 in 1916 to 11,950 in 1918 (Bittlingmayer 1988; Stubbs 2002). The aircraft pool remained in

effect until 1975, when the US Department of Justice decided to dissolve the pool, arguing that it had "lessened competition in research and development" (*Federal Register* 40(142), July 23, 1975, p. 30848). This decision exemplifies the tension between the potential benefits and costs of patent pools.

In theoretical models, the predicted effects of patent pools on innovation are ambiguous. The prospect of a pool may motivate firms to enter a race to patent the technologies that will form the pool; this race could be productive, or it may be socially wasteful if it encourages duplicative research and strategic patenting (Dequiedt and Versaevel 2012). The creation of a pool may also encourage investments in research and development by reducing litigation risks for members and thereby increasing expected profits from research and development (Shapiro 2001), but it may also lead pool members to cut their own investments in research and development because they hope to be able to free-ride on the investments of other members (Vaughn 1956, p. 67). Incentives to free-ride are particularly strong for pools that include "grant-back provisions," which require members to offer all new patents to the pool, and innovative members may abandon the pool to protect their patents (Aoki and Nagaoka 2004). Grant-back provisions may, however, also encourage innovation by reducing the potential for hold-up (Lerner, Strojwas, and Tirole 2007).

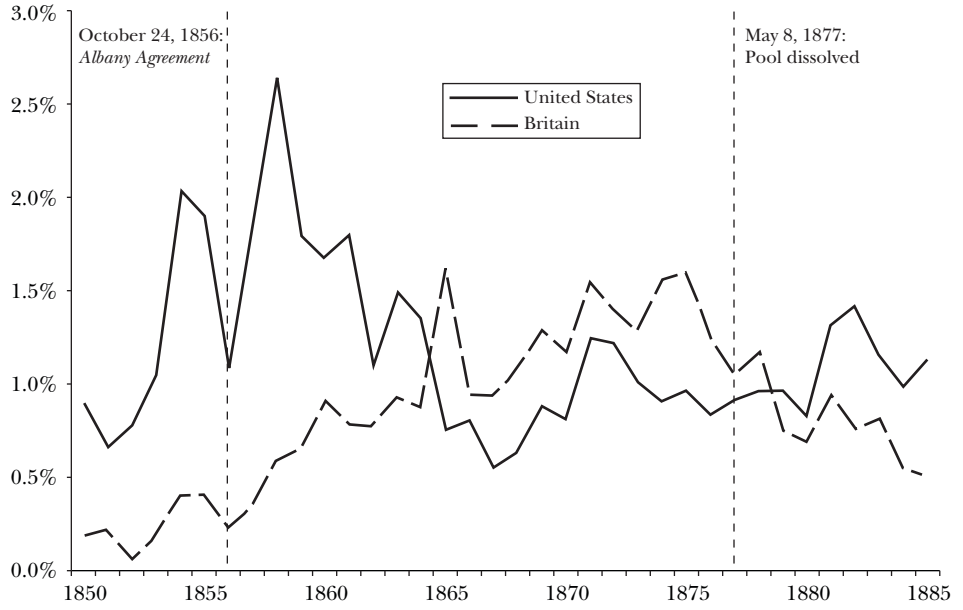
Empirical evidence on the effects of modern pools on innovation is limited so far. Qualitative evidence indicates that innovation increased in response to a pool for CDs, but declined in response to a pool for disk drives (Flamm 2012). In the open source software industry, the creation of a pool was followed by a modest increase in the number of new open source software products per year for technology fields in which IBM contributed patents to the pool (Ceccagnoli, Forman, and Wen 2012).[5]

Economic history offers opportunities to investigate pools across a broad range of industries and regulatory settings (Gilbert 2004), starting with the first pool in US history, the Sewing Machine Combination (1856–1877). This pool shared key characteristics of pools that are predicted to encourage innovation today: It combined nine complementary patents, which were necessary to build a commercially viable sewing machine, and it resolved the sewing machine patent war between Elias Howe, the Singer Company, and two other manufacturers, which had delayed commercialization. Litigation data confirm that the creation of a pool lowered litigation risks for members (Lampe and Moser 2010, p. 900). The pool also reduced license fees from $25 for Howe's patent to $5 for the bundle of patents for members and $15 for outside firms, confirming theoretical predictions.

Patenting, however, declined after the pool formed and only increased again after the pool dissolved in 1877 (Lampe and Moser 2010, p. 913). A comparison with the British sewing machine industry, which had no patent pool, suggests that

[5] Earlier empirical analyses have focused on the determinants of pool participation (Layne–Farrar and Lerner 2010) and on rules that govern interactions between pool members (Lerner, Strojwas, and Tirole 2007).

Figure 2
Share of Sewing Machine Patents in All Patents: United States versus Britain



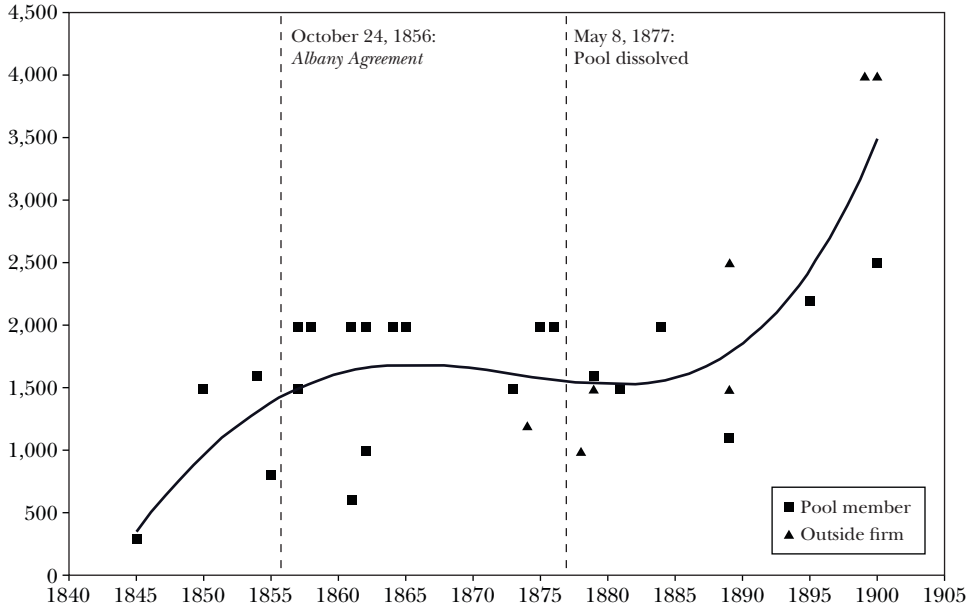
Source: Lampe and Moser (2010).
Notes: US patents granted in USPTO main class 112 ("sewing") and British patents from *A Cradle of Inventions: British Patents from 1617 to 1894*. Series excludes patents for attachments, tables, and stands.

this decline in innovation was a purely American phenomenon, as we can see in Figure 2. In Britain, sewing machine patents continued to increase gradually as a share of all British patents until the early 1874 and experienced no increase after 1877.

To investigate whether this decline in patenting reflected a decline in *innovation*, we collected additional data on objective improvements in the performance of sewing machines. Articles on sewing machines in nineteenth-century magazines, such as the *Scientific American* and the *Ladies' Home Journal* suggest that the key characteristics that consumers valued in a sewing machine were low weight, little noise, and most importantly, a high speed of sewing, measured as the number of stitches per minute that a machine could perform. Data on improvements in sewing speed, which we collected from company records and trade journals in the Smithsonian Institution Library, and shown in Figure 3, indicate that improvements slowed soon after the pool had been established and did not recover until it had dissolved (Lampe and Moser 2010, pp. 916–17).

Whether these results are generalizable to other industries and modern pools is an open question. The unambiguous decline in innovation for sewing machines, however, highlights the need for additional empirical—and theoretical—analyses to guide antitrust policy towards pools. Theoretical models of effects on price are

Figure 3
Stitches per Minute



Sources: Figure from Lampe and Moser (2010). Data from the *Scientific American* (1846–1869), exhibition catalogues, such as the "United States Commissioners Report to the Universal Exposition in Paris," "The Report of the Twenty-seventh Exhibition of American Manufactures, Held in the City of Philadelphia," ads in contemporary trade publications, including "The Textile American;" and historical industry analysis, such as *Uniting the Tailors: Trade Unionism amongst the Tailoring Workers of London and Leeds, 1870–1939.*
Notes: Figure 3 plots improvements in sewing speed based on data collected from company records and trade journals in the Smithsonian Institution Library. The solid line plots a fourth-order polynomial trend.

well developed (Shapiro 2001; Lerner and Tirole 2004), but effects of patent pools on innovation are equally important and less well understood. Existing theoretical models also focus almost exclusively on member firms, but ignore effects on *outside* firms. Patent data, however, indicate that outside firms produced the large majority of patents across industries (Lampe and Moser 2012a), suggesting that their response to the creation of a pool is essential to understanding the welfare effects of pools.

A better understanding of the mechanism by which pools influence the rate and direction of innovation is particularly important as the use of pools expands into innovative research fields with high social value, such as biochemistry, medicines, or energy. The case of the sewing machine industry suggests that the creation of a pool may soften the intensity of competition for member firms, which tend to be larger and more established, at the expense of outside firms, which tend to be smaller and younger than pool members. For example, the sewing machine pool appears to have exacerbated litigation risks for outside firms, even as it reduced

such risks for members (Lampe and Moser 2010, p. 907). The pool also created differential license fees that favored pool members, even though it reduced license fees (as theory predicts). Current antitrust guidelines allow pools to charge differential license fees, unless they have been shown to have direct anticompetitive effects. The experience of the sewing machine pool, however, indicates that differential license fees—which make it harder for outside firms to offer the pool technology at a competitive price—diverted the research investments of outside firms towards technologically inferior substitutes for the pool technologies (Lampe and Moser 2012b). This finding suggests that—in the absence of effective regulation—patent pools may influence not only levels, but also the direction of technical change.

Compulsory Licensing

An alternative mechanism to modify patent systems is compulsory licensing, which weakens the monopoly power of patents by licensing them to competing firms without the consent of patent owners. This policy has moved to the forefront of international trade debates, as international treaties, such as the Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPS) have strengthened foreign-owned patents in developing countries, reducing access to life-saving drugs and other essential innovations (Deardorff 1992; Grossman and Lai 2004; Chaudhuri, Goldberg, and Jia 2006). To address this issue, Article 31 of TRIPS allows national governments to issue compulsory licenses of foreign-owned patents in cases of national emergencies. The World Trade Organization Doha Declaration of 2001 (WT/MIN(01)/DEC/1, Art. 5.b) further specifies that national governments have "the freedom to determine the grounds upon which such licenses are granted." Thailand and Brazil, for example, have used compulsory licensing to procure antiretroviral drugs for millions of patients with HIV/AIDS, and India has used the threat of compulsory licensing to procure vaccines for swine flu (Kremer 2002; Galvão 2002; Gostin 2006; Steinbrook 2007).

Immediate access to foreign-owned inventions may, however, come at the cost of discouraging domestic invention in the licensing country if it displaces domestic research and development. But compulsory licensing may also encourage domestic research and development that is complementary to foreign-owned inventions, and the ability to produce foreign-owned inventions may create opportunities for cumulative innovation (Scotchmer 1991) and learning by doing (Arrow 1962). As a result, the effects of compulsory licensing on domestic invention are theoretically ambiguous. Empirical analyses are complicated by the fact that governments are more likely to use compulsory licensing if demand for foreign-owned inventions is high and if domestic production capacities are advanced enough to produce them; both factors may increase domestic invention irrespective of compulsory licensing.

An episode of compulsory licensing under the US Trading with the Enemy Act (TWEA) as a result of World War I creates a unique opportunity to identify the

effects of compulsory licensing on invention. Passed on November 17, 1917, the TWEA was intended to "dislodge the hostile Hun within our gates" and to place all enemy property "beyond the control or influence of its former owners, where it cannot eventually yield aid or comfort to the enemy" (US Office of Alien Property Custodian 1919, p. 13 and 17). In March 28, 1918, the TWEA was amended to grant the Alien Property Custodian, Mitchell Palmer, the power to sell enemy property, including all enemy-owned patents "as though he were the owner thereof" (US Office of Alien Property Custodian 1919, p. 22). By February 22, 1919, Palmer announced that "practically all known enemy property in the United States has been taken over by me" (US Office of Alien Property Custodian 1919, p. 7). In 1919, the US Chemical Foundation began to issue nonexclusive licenses of enemy-owned patents to US firms.

In Moser and Voena (2012), we exploit this event to examine the effects of compulsory licensing on the patenting activity of US inventors in organic chemistry. Baseline estimates compare changes after 1918 in patent issues per year for 336 technologies with compulsory licensing, with changes for a control group of 7,248 technologies without licensing. Methodologically, the analysis takes advantage of the detailed classification system of the US Patent and Trademark Office to distinguish narrowly defined technologies (measured at the level of subclasses) that were differentially affected by compulsory licensing. Technology fixed effects (at the level of subclasses) and year fixed effects, as well as technology-specific trends make it possible to control for variation in the inventors' use of patents across technologies and over time. The difference-in-differences analyses comparing narrowly defined technologies (at a unit of analysis much below the industry level) make it possible to control for unobservable factors, such as improvements in education, the creation of protectionist tariffs, or the temporary absence of German competitors during the war, which may have encouraged US invention across all types of chemical technologies regardless of compulsory licensing.

Baseline estimates indicate a 20 percent increase in domestic patenting in response to compulsory licensing (Moser and Voena 2012, p. 404). Estimates of time-varying effects indicate that this increase set in with a lag of eight to nine years and remained large and statistically significant throughout the 1930s (Moser and Voena 2012, p. 409).

These results suggest that compulsory licensing may help to *increase* innovation in the licensing countries, even though this increase occurs with some delay if the licensing country lags behind the technology frontier. At the time of the Trading with the Enemy Act, the United States lagged behind Germany in the field of organic chemistry and needed "time to learn" (Arora and Rosenberg 1998, p. 79), even though other branches of US chemical invention were well-developed. For example, the hopes of duplicating German dyes seemed slim for US firms in 1919. Du Pont's initial runs of indigo (which had been developed and patented by the German chemical firm BASF) turned out green (Hounshell and Smith 1988, p. 90). Similarly, countries such as Brazil and India, which are technologically advanced in many fields, seek to license foreign technologies in fields where

domestic invention is weak, and may require some time to catch up to the frontier in these fields.

Learning from patent documents is particularly difficult if information in patent documents is incomplete or obscure. The German BASF, for example, had "effectively bulwarked its discovery [of the Haber–Bosch process of nitrogen fixation] with strong, broad patents which detailed meticulously the apparatus, temperatures and pressures, but cleverly avoided particulars as to the catalysts employed or their preparation" (Haynes 1945, pp. 86–87). "A prolonged learning experience was necessary [for US firms] to understand the two sides of catalysis, the chemical side and the engineering and design side" (Mowery and Rosenberg 1998, p. 75).

In the case of compulsory licensing, these problems are exacerbated because licensees typically cannot access the uncodified knowledge that is embodied in skilled workers and scientists who developed the original improvement. Thus the US Winthrop Chemical Company, which had acquired all of the German company Bayer's production machinery in addition to its patents "could not figure out how to make the sixty-three drugs that were supposed to be [its] stock-in-trade . . . The former German supervisors having been jailed or deported, nobody knew how to run the machines; . . . the patents, which were supposed to specify manufacturing processes, were marvels of obfuscation" (Mann and Plummer 1991, pp. 52–53).

Domestically, regulators have used compulsory licensing as a remedy to restore competition in industries that have become dominated by a small group of firms. For example, Scherer (1977, pp. 47–48) estimates that the US Federal Trade Commission and the US Department of Justice had made thousands of patents available by 1977, in industries ranging from glassware (in the 1946 breakup of the Hartford Empire pool) to copy machines (in the 1975 decision against Xerox). As a mechanism to address anticompetitive patenting behavior in domestic markets, compulsory licensing is expected to increase overall welfare by encouraging competition (Tandon 1982; Gilbert and Shapiro 1990). Survey results and case studies suggest that compulsory licencing may not provoke dramatic changes in rates of patenting and innovation (for example, Scherer 1977, Chien 2003), but more systematic empirical analyses are needed.

Conclusions

Critics of the current patent system argue that a shift towards the strategic use of patents as a "sword" to hold up competitors and extract license fees threatens the effectiveness of patents as a means to encourage innovation (for example, Duhigg and Lohr 2012). The underlying problems with this system, however, may be much broader, and understanding them is critical to the design of patent policies. As early as the 1850s, patentees who did not produce anything were able to hold up entire industries because they had been issued broad patents that had been affirmed in court.

Historical evidence suggests that in countries with patent laws, the majority of innovations occur outside of the patent system. Countries without patent laws have produced as many innovations as countries with patent laws during some time periods, and their innovations have been of comparable quality. Even in countries with relatively modern patent laws, such as the mid-nineteenth-century United States, most inventors avoided patents and relied on alternative mechanisms when these were feasible. Secrecy emerged as a key mechanism to protect intellectual property. The effectiveness of secrecy relative to patents varies with the technological characteristics of innovations across industries and over time. In industries where secrecy was effective, inventors were less likely to use patents. Advances in scientific analysis, which lowered the effectiveness of secrecy, increased inventors' dependency on patents.

Incorporating these basic facts changes the predicted effects of patent laws on innovation. If a substantial share of innovation occurs outside of the patent system, policies that implement even the most drastic shifts towards stronger patents may fail to encourage innovation. If inventors' dependence on patent protection varies across industries, implementing stronger patent rights may alter the direction of technical change. If property rights in ideas encourage inventors to publicize technical information, a shift towards patenting may encourage the diffusion of knowledge.

History also offers a laboratory in which researchers can explore the effectiveness of alternative remedies to problems with the current patent system. For example, patent pools, which allow competing firms to combine their patents, have been proposed as a mechanism to resolve litigation risks as a result of overlapping patent grants, when more than one firm owns patents for the same technology. Historical evidence, however, indicates that pools may discourage and divert research and development by outside firms if the pools create differential litigation risks and licensing schemes that favor their members. Another prominent mechanism is compulsory licensing, which allows competitors to produce patented inventions without the consent of the patent owners. Historical evidence suggests that this policy may encourage innovation by allowing a new set of firms to produce a patented technology, and possibly by increasing competition to improve the technology.

Overall, the weight of the existing historical evidence suggests that patent policies, which grant strong intellectual property rights to early generations of inventors, may *discourage* innovation. On the contrary, policies that encourage the diffusion of ideas and modify patent laws to facilitate entry and encourage competition may be an effective mechanism to encourage innovation. Carefully executed historical analyses can help to shed further light on these pressing issues of patent policy.

■ *I wish to thank David Autor, Eric Hilt, Ryan Lampe, Stephanie Lee, Xing Li, Joel Mokyr, Hoan Nguyen, John List, Paul Rhode, Chang-Tai Hseih, Carlos Serrano, Timothy Taylor, Joel Watson, and especially Gavin Wright for helpful suggestions, and the National Science Foundation for support through NSF Grant SES0921859 and CAREER Grant 1151180.*

References

Abramovitz, Moses. 1989. *Thinking about Growth: And Other Essays on Economic Growth and Welfare.* Cambridge University Press.

Agnew, Ficher W. 1874. *The Law and Practice relating to Letters Patent for Inventions: Together with Notices of the Patent Laws in Force in the Principal Foreign States and in the Colonies.* London: Wildy and Sons.

Allen, Robert C. 1983. "Collective Invention." *Journal of Economic Behavior and Organization* 4(1): 1–24.

Allen, Robert C. 2009. *The British Industrial Revolution in Global Perspective.* University of Chicago Press.

Anton, James J., and Dennis A. Yao. 2004. "Little Patents and Big Secrets: Managing Intellectual Property." *RAND Journal of Economics* 35(1): 1–22.

Aoki, Reiko, and Sadao Nagaoka. 2004. "The Consortium Standard and Patent Pools." *Hi-Stat Discussion Paper 32.*

Arora, A., and N. Rosenberg. 1998. "Chemicals: A US success story." In *Chemicals and Long-Term Economic Growth*, edited by A. Arora, R. Landau, and N. Rosenberg, 71–102. New York: Wiley.

Arrow, Kenneth J. 1962. "The Economic Implications of Learning by Doing." *Review of Economic Studies* 29(3): 155–73.

Belfanti, Carlo. 2004. "Guilds, Patents, and the Circulation of Technical Knowledge: Northern Italy during the Early Modern Age." *Technology and Culture* 45(3): 569–89.

Berichterstattungs-Kommission der Deutschen Zollvereins-Regierungen. 1852–53. *Amtlicher Bericht über die Industrie-Austellung aller Völker zu London im Jahre 1851*, Vols I–III. Berlin, Prussia: Verlag der Deckerschen Geheimen Ober-Hofbuchdruckerei.

Bessen, James, and Michael J. Meurer. 2008. *Patent Failure: How Judges, Bureaucrats, and Lawyers Put Innovators at Risk.* Princeton University Press.

Bilir, L. Kamran, Petra Moser, and Irina Talis. 2011. "Do Patent Treaties Encourage Technology Transfer? Evidence from the Paris Convention" http://ssrn.com/abstract=1893052.

Bissell, Don. 1999. *The First Conglomerate: 145 Years of the Singer Sewing Machine Company.* Brunswick, ME: Audenreed Press.

Bittlingmayer, George. 1988. "Property Rights, Progress, and the Aircraft Patent Agreement." *Journal of Law and Economics* 31(1): 227–48.

Boldrin, Michele, and David K. Levine. 2008. *Against Intellectual Monopoly.* Cambridge University Press.

Boult, Alfred J. 1895. *Digest of British and Foreign Patent Laws*, 2nd edition. London: Boult, Wade & Kilburn.

Ceccagnoli, Marco, Christopher Forman, and Wen Wen. 2012. "Patent Pools, Thickets, and the Open Source Software Entry by Start-up Firms." http://conference.nber.org/confer/2012/IPKE/forman.pdf.

Chaudhuri, Shubham, Pinelopi K. Goldberg, and Panle Gia. 2006. "Estimating the Effects of Global Patent Protection in Pharmaceuticals: A Case Study of Quinolones in India." *American Economic Review* 96(5): 1477–1514.

Chien, Colleen. 2003. "Cheap Drugs at What Price to Innovation: Does the Compulsory Licensing of Pharmaceuticals Hurt Innovation." *Berkeley Techonolgy Law Journal* 18(3):853.

Clark, Gregory. 2006. *A Farewell to Alms: A Brief Economic History of the World.* Princeton University Press.

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

David, Paul A. 1994. "The Evolution of Intellectual Property Institutions." In *Economics in a Changing World*, Vol. 1: *System Transformation: Eastern and Western Assessments*, edited by A. Aganbegyan, O. Bogomolov, and M. Kaser. London: MacMillan.

Deardorff, Alan, V. 1992. "Welfare Benefits of Global Patent Protection." *Economica* 59(233): 35–51.

Department of Justice, and Federal Trade Commission. 1995. "Antitrust Guidelines for the Licensing of Intellectual Property."

Department of Justice, and Federal Trade Commission. 2007. *Antitrust Enforcement and Intellectual Property Rights: Promoting Innovation and Competition.*

Dequiedt, Vianney, and Bruno Versaevel. 2012. "Patent Pools and Dynamic R&D Incentives." Documents De Travail No. 07-03. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=988303.

Duhigg, Charles, and Steve Lohr. 2012. "The Patent, Used as a Sword." *New York Times*, October 7, p. A1. http://www.nytimes.com/2012/10/08/technology/patent-wars-among-tech-giants-can-stifle-competition.html?smid=pl-share.

Dutton, Harold I. 1984. *The Patent System and Inventive Activity during the Industrial Revolution, 1750–1852.* Manchester, UK: Manchester University Press.

Federal Register. 1975. July 23, 40(142): 30848.

Flamm, Kenneth. 2012. "A Tale of Two Standards: Patent Pools and Innovation in the Optical Disk Drive Industry." http://conference.nber.org/confer/2012/IPKE/flamm.pdf.

Galvão, Jane. 2002. "Access to Antiretroviral Drugs in Brazil." *Lancet* 360(9348): 1862–65.

Gilbert, Richard J. 2004. "Antitrust for Patent Pool: A Century of Policy Evolution." *Stanford Technology Law Review*, April 28. http://stlr.stanford.edu/2004/04/antitrust-for-patent-pools/.

Gilbert, Richard J., and Carl Shapiro. 1990. "Optimal Patent Length and Breadth." *RAND Journal of Economics* 21(1): 106–112.

Gostin, Lawrence O. 2006. "Medical Countermeasures for Pandemic Influenza: Ethics and the Law." *JAMA* 295(5): 554–56.

Grossman, Gene M., and Edwin L.-C Lai. 2004. "International Protection of Intellectual Property." *American Economic Review* 94(5): 1635–53.

Habakkuk, H. J. 1962. *American and British Technology in the Nineteenth Century: The Search for Labour-Saving Inventions*. Cambridge University Press.

Haber, Ludwig Fritz. 1958. *The Chemical Industry during the Nineteenth Century: A Study of the Economic Aspect of Applied Chemistry in Europe and North America*. Oxford, UK: Oxford University Press.

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498.

Hall, Bronwyn H., and Rosemarie Ham Ziedonis. 2001. "The Patent Paradox Revisited: An Empirical Study of Patenting in the US Semiconductor Industry, 1979–1995." *RAND Journal of Economics* 32(1): 101–128.

Harhoff, Dietmar, and Karin Hoisl. 2006. "Institutionalized Incentives for Ingenuity—Patent Value and the German Employees' Inventions Act." *Research Policy* 36(8): 1143–62.

Haynes, Williams. 1945. *American Chemical Industry—The World War I Period: 1912–1922*. New York: D. Van Nostrand Company.

Horstmann, Ignatius J., Glenn MacDonald, and Alan Slivinski. 1985. "Patents as Information Transfer Mechanisms: To Patent or (Maybe) Not to Patent." *Journal of Political Economy* 93(5): 837–58.

Hounshell, David. 1985. *From the American System to Mass Production, 1800–1932: The Development of Manufacturing Technology in the United States*. Baltimore, MD: Johns Hopkins University Press.

Hounshell, David A., and John Kenly Smith. 1988. *Science and Corporate Strategy: Du Pont R&D, 1902–1980*. Cambridge University Press.

Jaquet, Eugène, and Chapuis, Alfred. 1945. *La Montre Suisse*. Bale and Olten: Editions Urs Graf.

Khan, B. Zorina. 2005. *The Democratization of Invention: Patents and Copyrights in American Economic Development, 1790–1920*. Cambridge University Press.

Khan, B. Zorina, and Kenneth L. Sokoloff. 1993. "Schemes of Practical Utility: Entrepreneurship and Innovation among 'Great Inventors' in the United States, 1790–1865." *Journal of Economic History* 53(2): 289–307.

Khan, B. Zorina, and Kenneth L. Sokoloff. 1998. "Patent Institutions, Industrial Organization and Early Technological Change: Britain and the United States, 1790–1850." In *Technological Revolutions in Europe*, edited by M. Bergand K. Bruland, Cheltenham, 292–313. UK: Edward Elgar.

Khan, B. Zorina, and Kenneth L. Sokoloff. 2001. "History Lessons: The Early Development of Intellectual Property Institutions in the United States." *Journal of Economic Perspectives* 15(3): 233–46

Kremer, Michael. 2002. "Pharmaceuticals and the Developing World." *Journal of Economic Perspectives* 16(4): 67–90.

Kretschmer, Winfried. 1999. *Geschichte der Weltausstellungen*. Frankfurt: Campus Verlag.

Kroker, Evelyn. 1975. *Die Weltausstellungen im 19. Jahrhundert: Industrieller Leistungsnachweis, Konkurrenzverhalten und Kommunikationsfunktion unter Berücksichtigung der Montanindustrie des Ruhrgebietes zwischen 1851 und 1880*. Göttingen, Germany: Vandhoeck Ruprecht.

Lamoreaux, Naomi R., and Kenneth L. Sokoloff. 1999. "Inventors, Firms, and the Market for Technology in the Late Nineteenth and Early Twentieth Centuries." In *Learning by Doing in Markets, Firms and Countries*, edited by Naomi R. Lamoreaux, Daniel M. G. Raff, and Peter Temin P. University of Chicago Press.

Lampe, Ryan, and Petra Moser. 2010. "Do Patent Pools Encourage Innovation? Evidence from the Nineteenth-Century Sewing Machine Industry." *Journal of Economic History* 70(4): 898–920.

Lampe, Ryan, and Petra Moser. 2012a. "Patent Pools and Innovation? Evidence from 20 U.S. Industries under the New Deal." Stanford Law and Economics Olin Working Paper No. 417. http://ssrn.com/abstract=1967246.

Lampe, Ryan, and Petra Moser. 2012b. "Patent Pools and Innovation in Substitutes—Evidence from the 19th Century Sewing Machine Industry" http://ssrn.com/abstract=1468062.

Landes, David S. 1969. *The Unbound Prometheus. Technological Change and Industrial Development in Western Europe from 1750 to the Present*. Cambridge University Press.

Layne-Farrar, Anne, and Josh Lerner. 2010. "To Join or Not to Join: Examining Patent Pool

Participation and Rent Sharing Rules." *International Journal of Industrial Organization* 29(2): 294–303.

Lerner, Josh. 2000. "150 Years of Patent Protection." NBER Working Paper 7478.

Lerner, Josh, Marcin Strojwas, and Jean Tirole. 2007. "The Design of Patent Pools: The Determinants of Licensing Rules." *RAND Journal of Economics* 38(3): 610–25.

Lerner, Josh, and Jean Tirole. 2004. "Efficient Patent Pools." *American Economic Review* 94(3): 691–711.

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* no. 3 (Special Issue on Microeconomics): 783–832.

Li, Xing, Megan MacGarvie and Petra Moser. 2012. "Dead Poets' Property—The Copyright Act of 1814 and the Price of Books in the Romantic Period." http://ssrn.com/abstract=2170447.

Machlup, Fritz, and Edith Penrose. 1950. "The Patent Controversy in the Nineteenth Century." *Journal of Economic History* 10(1): 1–29.

MacLeod, Christine. 1988. *Inventing the Industrial Revolution: The English Patent System, 1660–1800*. Cambridge University Press.

Mann, Charles C., and Mark L. Plummer. 1991. *The Aspirin Wars: Money, Medicine, and 100 Years of Rampant Competition*. New York: Knopf.

Merges, Robert. 2001. "Institutions for Intellectual Property Exchange: The Case of Patent Pools." In *Intellectual Products: Novel Claims to Protection and their Boundaries*, edited by Rochelle Dreyfuss. Oxford University Press.

Mokyr, Joel. 1990. *The Lever of Riches: Technological Creativity and Economic Progress*. New York: Oxford University Press.

Mokyr, Joel. 2009. *The Enlightened Economy: An Economic History of Britain, 1700–1850*. New Haven, CT: Yale University Press.

Moser, Petra. 2005. "How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs." *American Economic Review* 95(4): 1215–36.

Moser, Petra. 2011. "Do Patents Weaken the Localization of Innovations? Evidence from World's Fairs." *Journal of Economic History* 71(2): 363–82.

Moser, Petra. 2012. "Innovation without Patents: Evidence from World's Fairs." *Journal of Law and Economics* 55(1): 43–74.

Moser, Petra, and Tom Nicholas. 2012. "Publicity is Prize-less. Non-monetary Awards as a Mechanism to Encourage Patenting." http://ssrn.com/abstract=1910220.

Moser, Petra, and Paul W. Rhode. 2012. "Did Plant Patents Create the American Rose?" Chap. 8 in *The Rate and Direction of Inventive Activity Revisited*, edited by Joshua Lerner and Scott Stern. University of Chicago Press.

Moser, Petra, and Alessandra Voena. 2012. "Compulsory Licensing: Evidence from the Trading-with-the-Enemy-Act." *American Economic Review* 102(1): 396–427.

Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2011. "German Jewish Émigrés and U.S. Invention." March 21. Available at SSRN: http://ssrn.com/abstract=1910247 or http://dx.doi.org/10.2139/ssrn.1910247.

Moser, Petra, and Assaf Zimring. 2012. "Patent Laws, Market Size and Innovation." Earlier version available at SSRN: http://ssrn.com/abstract=2010136.

Mowery, David C. and Nathan Rosenberg. 1998. *Paths of Innovation: Technological Change in 20th-Century America*. Cambridge University Press.

Nordhaus, William D. 1969. "An Economic Theory of Technological Change." *American Economic Review* 59(2): 18–28.

North, Douglass C., and Robert P. Thomas. 1973. *The Rise of the Western World: A New Economic History*. Cambridge University Press.

Nuvolari, Alessandro. 2004. "Collective Invention during the British Industrial Revolution: The Case of the Cornish Pumping Engine." *Cambridge Journal of Economics* 28(3): 347–63.

Officer, Lawrence, and Sam Williamson. 2011. "Measures of Worth." MeasuringWorth.com, www.measuringworth.com/worthmeasures.php.

Parton, James. 1867. "History of the Sewing Machine." *Atlantic Monthly*, May.

Penrose, Edith Tilton. 1951. *The Economics of the International Patent System*. Baltimore, MD: Johns Hopkins University Press.

Robb, Harry C., Jr. 1964. "Plant Patents." In *Encyclopaedia of Patent Practice and Invention Management*, edited by Robert Calvert, pp. 641–55. New York: Reinhold.

Rosenberg, Nathan. 1963. "Technological Change in the Machine Tool Industry, 1840–1910." *Journal of Economic History* 23(4): 414–43.

Rosenberg, Nathan. 1969. *The American System of Manufactures: The Report of the Committee on the Machinery of the United States 1855, and the Special Reports of George Wallis and Joseph Whitworth, 1854*. Edinburgh University Press.

Rosenberg, Nathan. 1972. *Technology and American Economic Growth*. New York: Harper.

Rossman, Joesph. 1930. "The Planter Breeder Becomes an Inventor." *Science News-Letter* 18, December 20, pp. 349–95.

Rothbarth, Edwin. 1946. "Causes of the Superior Efficiency of U.S.A. Industry Compared with British Industry." *Economic Journal* 56: 383–90.

Scherer, Frederic. M. 1977. *The Economic Effects of Patent Compulsory Licensing.* New York University, Graduate School of Business Administration, Center for the Study of Financial Institutions.

Schiff, Eric. 1971. *Industrialization without National Patents: The Netherlands 1868–1912, Switzerland 1850–1907.* Princeton University Press.

Schmookler, Jacob. 1962. "Economic Sources of Inventive Activity." *Journal of Economic History* 22(1): 1–20.

Schmookler, Jacob. 1966. *Invention and Economic Growth.* Cambridge: Harvard University Press.

Scotchmer, Suzanne. 1991. "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law." *Journal of Economic Perspectives* 5(1): 29–41.

Shapiro, Carl. 2001. "Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting." Chap. 4 in *Innovation Policy and the Economy*, Vol. 1, edited by A. Jaffe, J. Lerner, and S. Stern. Cambridge, MA: MIT Press.

Sokoloff, Kenneth. 1988. "Inventive Activity in Early Industrial America: Evidence from Patent Records, 1790–1846." *Journal of Economic History* 48(4): 813–50.

Steinbrook, Robert. 2007. "Thailand and the Compulsory Licensing of Efavirenz." *New England Journal of Medicine* 356(6): 544–46.

Stewart, Amy. 2007. *Flower Confidential: The Good, the Bad, and the Beautiful in the Business of Flowers.* Chapel Hill, NC: Algonquin Books of Chapel Hill.

Stubbs, Kevin D. 2002. *Race to the Front: The Material Foundation of Coalition Strategy in the Great War, 1914–1918.* Westport, CT: Greenwood Publishing Group.

Tandon, Pankaj. 1982. "Optimal Patents with Compulsory Licensing." *Journal of Political Economy* 90(3): 470–86.

Thomson, Ross. 2009. *Structures of Change in the Mechanical Age: Technological Innovation in the United States, 1790–1865.* John Hopkins University Press.

US Office of Alien Property Custodian. 1919. "Alien Custodian Report: A Detailed Report by the Alien Property Custodian of All Proceedings Had by Him under the Trading with the Enemy Act during the Calendar Year 1918 and to the Close of Business on February 15, 1919." Washington: Government Printing Office.

Vaughan, Floyd L. 1956. *The United States Patent System: Legal and Economic Conflicts in American Patent History.* Norman, OK: University of Oklahoma Press.

Wallace, Anthony F. C. 1986. *Rockdale: The Growth of an American Village in the Early Industrial Revolution.* New York: Knopf.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 45–66

The New Patent Intermediaries: Platforms, Defensive Aggregators, and Super-Aggregators

Andrei Hagiu and David B. Yoffie

Some assets are traded in liquid markets, at transparent prices, with the help of many thriving intermediaries: houses and apartments, stocks and other financial products, books, DVDs, electronics, and all sorts of collectibles. Intellectual property in general and patents in particular—the focus of this paper—are not among those assets (Gans and Stern 2010). The patent market consists mainly of bilateral transactions, either sales or cross-licenses, between large companies. Such deals are privately negotiated and might involve hundreds or thousands of patents. For example, in June 2011, a consortium of Apple, Microsoft, Sony, and several other large tech companies outbid Google to buy Nortel's 6,000 patents and patent applications for $4.5 billion. Google responded first by buying over 1,000 patents from IBM for an undisclosed price, and then by acquiring Motorola Mobile and its more than 17,000 patents for $12.5 billion. In April 2012, Microsoft bought 925 patents from AOL for $1.1 billion, then sold a portion of that portfolio to Facebook for $550 million. And in September 2012, Samsung lost a $1 billion judgment to Apple and faced a potential injunction from a federal judge in a jury trial over patent infringement. The very real threat of adverse jury rulings or injunctions, which might lead to partial or total shutdown of existing businesses, have led to extremely high willingness-to-pay for some intellectual property.

Outside of these bilateral deals, patent buyers and sellers frequently have a hard time finding each other. There is no eBay, Amazon, New York Stock Exchange, or Kelley's Blue Book equivalent for patents, and when buyers and sellers do find

■ *Andrei Hagiu is Associate Professor and David B. Yoffie is Max and Doris Starr Professor of International Business Administration, both in the Strategy Unit at the Harvard Business School, Boston, Massachusetts. Their email addresses are ahagiu@hbs.edu and dyoffie@hbs.edu.*

http://dx.doi.org/10.1257/jep.27.1.45. doi=10.1257/jep.27.1.45

each other, they usually negotiate under enormous uncertainty: prices of similar patents vary widely from transaction to transaction and the terms of the transactions (including prices) are often secret and confidential.

Inefficient and illiquid markets, such as the one for patents, generally create profit opportunities for intermediaries. In this paper, we begin with an overview of the problems that arise in patent markets, and how traditional institutions like patent brokers, patent pools, and standard-setting organizations have sought to address them. But during the last decade, a variety of novel patent intermediaries has emerged. We will discuss how several online platforms have started services for buying and selling patents but have failed to gain meaningful traction. However, new intermediaries that we call defensive patent aggregators and super-aggregators have become quite influential and controversial in the technology industries they touch. In particular, the rising prominence of a new and powerful patent aggregator called Intellectual Ventures has sparked heated debates about the economic role played by intermediaries in the patent market and their effects on innovation.[1]

One might expect that new intermediaries and competition between them could lead to increased market efficiency. Sometimes, however, intermediaries are able to exploit market inefficiencies without contributing much social value or, worse, they might even exacerbate existing market failures. The goal of this paper is to shed light on the role and efficiency tradeoffs of these new patent intermediaries. In the conclusion, we offer a provisional assessment of how the new patent intermediary institutions affect economic welfare.

Patent Market Failures and Traditional Patent Intermediaries

Why is the market for patents so illiquid and inefficient? While the root causes are well-known to economists and are a subset of market failures that arise in many markets for ideas, it is useful to summarize them briefly here, highlighting the issues most relevant for patent intermediaries. Gans and Stern (2010) offer a review of market failures in the market for ideas, many of which apply to patents.

First, patents are much more difficult to value than most other goods. This problem arises not simply because patents are intangible assets: after all, intangibles such as brand equity are routinely valued. What sets patents apart is that every patent

[1] Because our notion of a patent intermediary is an organization (firm or not-for-profit entity) that *directly* facilitates the sale or licensing of patents from owners-creators to users, we will not discuss here the patent rating, valuation, and search services that aim to create liquidity indirectly by providing useful patent information. An example of such a service is ArticleOne Partners (http://www.articleonepartners.com/). In addition, we focus specifically on patent intermediaries as opposed to other forms of intellectual property and more general notions of markets for technology (Arora, Fosfuri, and Gambardella 2001) and for ideas (Gans and Stern 2010). Thus, our study does not cover firms like InnoCentive and NineSigma, which connect companies with individuals or institutions that can create pre-patent solutions to science or technology problems.

is by definition unique: they lack "comparables," which are used in many markets to estimate a given asset's value. More importantly, patent value in many modern technologies is subject to strong complementarities and portfolio effects (Gans and Stern 2010; Parchomovsky and Wagner 2005). The issue of complementarities arises because in industries like semiconductors and smart phones, products are covered by dozens or even hundreds of interdependent patents. As a result, the value of *individual* patents is heavily discounted. Potential buyers or licensees may not place much value on a given patent sold by itself unless it complements a portfolio that they already own. This greatly reduces the number of buyers and the potential for liquidity. Portfolio effects create asymmetries between large operating firms on one side and individual inventors and small companies on the other side (Jaffe and Lerner 2004). There is a lower probability for smaller inventors to monetize their patents because they lack a large portfolio and because their owners typically have limited financial resources and legal expertise, which severely undermines their ability to bargain effectively. A well-known example (and the subject of the 2008 movie *Flash of Genius*) is that of engineer Robert Kearns, who in 1964 applied for a patent for an intermittent windshield wiper system for automobiles. Manufacturers refused Kearns's requests to sign licensing agreements and began producing cars featuring the wiper system in 1969. Kearns spent decades battling in court for infringement. He eventually earned $30 million in settlements from Ford and Chrysler but, in the process, lost his job, divorced, and suffered multiple nervous breakdowns (Schudel 2005).

Second, both sides of the patent market face high search costs. For patent owners, it is prohibitively costly to find all current users (actual infringers) and all *potential* applications of their patents. For potential patent buyers or users, it is very costly to find all prior art and patents that "read on" (that is, that might cover the technology within) their products, especially when these products are complex and rely on fast-changing technologies. Indeed, although patent offices around the world as well as private databases provide comprehensive and searchable lists of all patents issued, patent applicants typically seek to disclose only the minimum necessary to obtain the patent, and use language that is oftentimes broad and opaque. This makes it very difficult to figure out their relationship with other patents and prior art, particularly with millions of patents in circulation. To illustrate, consider Apple's "bounce-back" utility patent, which was highly publicized during the recent *Apple vs. Samsung* trial settled in August 2012 before a California jury (Gallagher 2012). This patent essentially covers a method for allowing users to scroll beyond the edge of an image, webpage, or list and have it bounce back onto the screen. Despite the highly intuitive nature of this functionality, it is quite hard to identify its scope and the way it is meant to be implemented from the language used in the actual patent.[2]

[2] The patent number is 7469381, and its detailed description is available from the US Patent Office at http://patft.uspto.gov/netacgi/nph-Parser?Sect2=PTO1&Sect2=HITOFF&p=1&u=/netahtml/PTO/search-bool.html&r=1&f=G&l=50&d=PALL&RefSrch=yes&Query=PN/7469381.

Third, patent transactions always happen in the shadow of litigation, which exacerbates valuation problems and creates large transaction costs. Litigation often implies that patents are best viewed as "probabilistic property rights" or "lottery tickets" (Lemley and Shapiro 2005): few patents are litigated, but of those that are, approximately half end up being invalidated. Given this risk, many patent owners and users prefer to settle out of court for amounts that have more to do with their opportunity costs of going to trial and their attitude towards risk than with the "true" economic value of their patents. Is the plaintiff a small company or individual with limited resources who prefers to settle for a small amount rather than face the possibility of years of litigation? What about a competitor who can be countersued and brought to accept a cross-licensing agreement? Or what about a "nonpracticing entity" against which injunctions that they cannot produce the product will not work—because the entity doesn't produce in the first place. Furthermore, some courts have a reputation for bias in favor of small players and against large companies, which makes them attractive patent litigation forums for small players and nonpracticing entities. For example, the Eastern District of Texas received 25 percent of all US patent infringement cases during 2011 and found in favor of patent owners almost 75 percent of the time (Decker 2012). The prospect of choosing a favorable court setting increases the amount of (inefficient) litigation.

The complexity that arises when valuation is intertwined with litigation has been heightened by the emergence of the US International Trade Commission (ITC) as a new forum for patent battles. The ITC is an independent federal agency with powers to do investigations and fact-finding on international trade issues, including import subsidies, dumping cases, and also issues of whether an imported product infringes on existing intellectual property. The ITC typically decides much faster than federal courts: often 12–15 months vs. several years in federal courts (Analysis Group, n.d.; Fisher 2006). It also offers the possibility of relatively quick injunctive relief against defendants: it can require that the offending imports be halted, which can be even more effective in extracting monetary settlements. Qualcomm, for example, was forced to negotiate an $891 million settlement with Broadcom in 2009, after losing a case at the ITC and facing an import ban (Crothers 2009). Thus, the ITC has significantly increased the costs of exposure to potential patent infringement lawsuits for firms in traded goods industries such as semiconductors, smartphones, and computers.

These patent market failures are most problematic for individual inventors or small companies, who represent the majority of patent owners. One study, compiling data from a variety of public sources, found that inventors and small businesses contribute 60 percent of all patents in the United States, but only extract 1 percent of total licensing revenues. The remaining 99 percent of licensing revenue goes to large companies (Hagiu, Yoffie, and Wagonfeld 2011, exhibit 11). Of course, it is not shocking that large companies may tend to focus on higher-value patents, but the disjunction is nonetheless striking.

One possible mechanism for small patent owners to address the problems of getting paid for their ideas would be to incorporate them in start-ups and seek either to compete with incumbent companies or to cooperate with them by licensing or

being acquired (Gans, Hsu, and Stern 2002). In this way, investors, particularly venture capitalists, could mitigate some of these market failures. But many patents are not worth incorporating in a start-up, especially if they are not part of broader portfolios. Furthermore, great inventors are not necessarily great entrepreneurs (Wasserman 2012). In fact, it is arguably more efficient for inventors to specialize in invention rather than to pursue commercialization, a point argued by Lamoreaux and Sokoloff (2003) in the context of late nineteenth century United States and probably equally valid today.

With so many difficulties facing inventors trying to monetize their discoveries, an obvious answer is to create intermediaries that facilitate the sale of patents to users (mainly operating companies), thereby maintaining appropriate incentives for innovation. In the next section, we discuss the three main traditional patent intermediaries studied in the existing economics literature: patent brokers/agents, patent pools, and standard-setting organizations. These traditional patent intermediaries have been around for a long time, but each faces certain limitations which prevent them from solving many of the patent market's problems.

Three Traditional Patent Intermediaries

Patent brokers help patent owners sell or license their technologies in exchange for a fee contingent on successful transfer. Their activity helps reduce search and transaction costs by investing in specific knowledge and connections on both sides of the market. Brokers often facilitate not just the sale or licensing of patents, but broader technology transfers, which include patents and know-how. They also offer consulting services helping patent owners market and sell their assets. There are a large number of patent brokers, which tend to be small companies with fewer than 10 employees. Some examples include Thinkfire (http://www.thinkfire.com/), IPValue (http://www.ipvalue.com/), Pluritas (http://www.pluritas.com/), and Competitive Technologies (http://www.competitivetech.net/).

Such patent brokers have existed since at least the nineteenth century: for example, Lamoreaux and Sokoloff (2003) document the positive effect of brokers on the US market for patented technology between 1870 and 1920. These brokers were typically patent agents or lawyers who matched inventors looking to sell new technologies with investors or buyers eager to commercialize them. At that time, however, there were few products encompassing hundreds of patented technologies like today: thus, the portfolio effects problem was less prevalent, and patents with fuzzy and overlapping boundaries were relatively rare. The job of modern patent brokers is much harder than those of a century ago. Unlike other markets for assets like stocks or real estate, the existence of many brokers in the patent market does not create sufficient liquidity on its own. Indeed, patent brokers are small in scale and tend to focus on facilitating high-end licensing transactions that carry large price tags. Their fees are above 10 percent of the value of the transaction and sometimes reach 20–30 percent (Young 2008), a level high enough to suggest that inefficiencies prevail in the patent market.

Patent pools are formal or informal organizations in which for-profit firms come together to license patents to each other or to third parties (Lerner, Strojwas, and Tirole 2007; Shapiro 2001). Some common examples of patent pools include the historical example of the patent pool for sewing machines (see for example Lampe and Moser 2010), along with more recent technology patent pools such as Bluetooth and MPEG-4. Bluetooth is a technology standard for exchanging data over short distances; the corresponding pool brought together patents from 12 companies including Ericsson, IBM, Intel, Motorola, Nokia, and Toshiba (Layne-Farrar and Lerner 2011). MPEG-4 is a method for compressing audio-visual data; this pool contained 29 companies, including Apple, AT&T, Canon, France Telecom, Fujitsu, Hitachi, Microsoft, RealNetworks, and Sharp (Layne-Farrar and Lerner 2011). Patent pools emerged to solve the "multiple marginalization" problem—also known as "royalty-stacking"—which arises when multiple parties hold market power in a chain of production. If all parties attempt to exercise their market power to the fullest, the resulting prices will typically be above the level that would be set by a single party with market power—and the joint profits and social welfare will be lower than in the case of a single party with market power.

While patent pools can create social value by reducing royalty stacking, it is not clear how well they address the traditional problems of patent markets. First, if patents included in a pool are substitutes rather than complements, the pool may turn out to have anticompetitive effects in the form of higher prices: the pool facilitates price collusion at the expense of price competition (Shapiro 2001; Lerner and Tirole 2004). Second, patent pools can create barriers to entry and innovation, favoring large companies with sizable patent portfolios who are members of the patent pool and discriminating against small companies or individual inventors who find it hard to negotiate their way into the pool. Third, the applicability of patent pools is limited to a small number of markets, where the essential intellectual property to producing a specific product or service is more or less evenly distributed among several large, identifiable players.

Similarly, standard-setting organizations have made it possible for participants in industries where there is an important need for interoperability between many components to come together and voluntarily produce consensus technical standards. Standard-setting organizations create economic value by enabling coordination on (Simcoe 2012) and certification of (Chiao, Lerner, and Tirole 2007; Lerner and Tirole 2006) technical standards. When these organizations endorse a specific technological standard, participants in the relevant industries typically adopt that standard and agree to cross-license or to pay the required royalties to the standard owner(s). The technological standard usually consists of many patents, owned by a patent pool, or on rare occasions by one company or institution. The process of choosing and certifying standards, however, is often subject to conflicts of interest due to interference by large producers (Schmalensee 2009). Furthermore, the scope of standard-setting organizations is limited to a small number of industries and technologies relative to the size of the broad patent market.

Thus, while brokers, patent pools, and standard-setting organizations have a role in bridging some of the gaps in the patent market, their effects are limited, and they have not managed to help small inventors get paid for their ideas. Indeed, small patent owners generally do not participate in pools or standard-setting organizations, and most small patent owners are not worth the time of professional brokers.

Exploiting Market Failures: Nonpracticing Entities

The general lack of depth in patent markets has created a particularly favorable environment for the so-called "nonpracticing entities," which have become the most controversial patent intermediaries. In essence, nonpracticing entities act as arbitrageurs, first acquiring patents, typically from individual inventors or small companies, and then seeking licensing revenues from operating companies through litigation or the threat of litigation. These entities do not innovate themselves, nor do they produce output. In 2001, nonpracticing entities brought 144 lawsuits targeting over 578 operating companies; by 2011, the numbers had increased to 1,211 lawsuits targeting 5,031 operating companies respectively (according to PatentFreedom research at https://www.patentfreedom.com/research-lot.html).

Two main factors account for the explosion in activity of nonpracticing entities. First, the Internet has greatly reduced transaction costs for inventors to find intermediaries to whom they can sell their patents (Spulber 2011). Although nonpracticing entities appeared in the second half of the 1990s, the way they found undervalued intellectual property assets at that time was largely serendipitous—for example, through personal connections to inventors or sales of distressed assets containing obscure patents. Today, with a quick Internet search, any inventor can locate nonpracticing entities directly or contact brokers who can help one do so (Lohr 2009).

Second, the value and prominence of patents have increased along with the revenues and profits associated with intellectual-property-intensive businesses. This growth was fueled in large part by the explosion of the information and communication technology sectors in areas like software, semiconductors, and mobile communications. Not coincidentally, most of the activity of the nonpracticing entities is concentrated in those sectors. These industries produce complex products and services, which involve many interrelated processes and components. For example, manufacturing an integrated circuit requires hundreds of steps, with literally billions of transistors and thousands of complex algorithms. Consequently, the potential for newly issued patents to have "fuzzy boundaries" (in the sense of Besen and Meurer 2008) and to overlap with prior art is very high in these sectors. Furthermore, no firm—even the industry's largest ones—has more than 30 percent of the patents that cover semiconductor design and manufacturing. This fragmented ownership of the relevant intellectual property exacerbates the uncertainty regarding the merits of the many patents involved.

Contrast this situation with the pharmaceutical industry, where patents also play a crucial role, but the boundaries of intellectual property are much more clearly defined. Patent claims on new molecules are easily distinguishable from other patented molecules. Not surprisingly then, nonpracticing entities and other patent merchants have been largely absent from the pharmaceutical sector. In contrast, different patents on smartphone user interfaces oftentimes contain closely related claims. For example, the difference between a horizontal and a vertical swiping mechanism for unlocking a touchscreen smartphone leaves lots of room for interpretation. How a jury might construe these claims create big opportunities for nonpracticing entities.

The arbitrage opportunities available to nonpracticing entities are sizable. As of 2010, the median price paid by nonpracticing entities for a patent was approximately $100,000 and the mean was $400,000 (according to PatentFreedom website, accessed December 2010). On the other side of the market, most patent settlements range between $50,000 and a few million dollars (Sharma and Clark 2008). In a few notable cases, however, nonpracticing entities have managed to extract hundreds of millions of dollars. The best-known example is a 2006 settlement in which Research in Motion (maker of the Blackberry smartphones) agreed to pay $612.5 million to NTP, a Virginia-based nonpracticing entity, which had sued Research in Motion for infringing on eight wireless email patents (Riordan 2004). While precise data on the distributions of prices paid and settlements received by nonpracticing entities is unavailable, it is useful to consider the following back-of-the-envelope calculation based on the numbers above. If 99/100 settlements are uniformly distributed between $50,000 and $1 million and 1/100 settlements are for $100 million, then the net expected payoff for a risk-neutral nonpracticing entity for purchasing a patent is approximately $1.15 million. Even after litigation costs, this offers an attractive arbitrage opportunity.

Nonpracticing entities have attracted financing from investors looking for novel diversification opportunities with high returns. A number of hedge fund, venture capital, and private equity firms either invest in nonpracticing entities or approach small patent-holders directly, offering to finance lawsuits against operating companies in exchange for a cut of any resulting payments (for some examples, see Masnick 2009; Bergelt 2010; McCurdy 2009).

Nonpracticing entities are sometimes pejoratively known as "patent trolls." The originator of the patent troll model is generally agreed to be the company TechSearch and its lawyer Raymond Niro. Beginning in the late 1990s, TechSearch originated the practice of buying up patents and suing companies for infringement to demand payments (Bario 2011). In 2001, Intel's in-house lawyer Peter Detkin referred to Niro as a "patent troll" and popularized the term. (Perhaps ironically, Detkin went on to cofound Intellectual Ventures, the largest nonpracticing entity today, which we discuss below.) The meaning of the term "patent troll" has evolved over time, and there is no commonly agreed-on definition. However, trolls are generally viewed as combining the following characteristics: 1) they acquire intellectual property assets, like patents, solely for the purpose of extracting payments

from alleged infringers; 2) they do not do research or develop any technology or products related to their patents; and 3) they behave opportunistically by waiting until industry participants have made irreversible investments before asserting their claims (Lemley 2008; Schmalensee 2009).

In itself, buying and reselling patents solely for price arbitrage is not necessarily a harmful practice. One could even argue that it increases market efficiency by creating liquidity and a way for small patent owners to get paid, similar to the function performed by dealers and market-makers in financial markets (McDonough 2006; Schmalensee 2009; Spulber 2011). The main reason that nonpracticing entities can instead create economic harm is that they seek to extract disproportionate payments through two practices. First, they typically engage in "nuisance value" litigation: they sue many companies simultaneously for moderate amounts so that targets are more likely to settle instead of risking a costly and uncertain trial. Second, they attempt to hold up (or "ambush") practicing companies by bringing the lawsuits at the most vulnerable times for the targets, like just before the release of a new product, when the target can ill afford a risky trial involving its new product shipments. Memory chip companies, for example, accused Rambus of ambushing the industry with litigation just after a new industry standard had been set (Schmalensee 2009).[3] These two practices exacerbate patent market inefficiencies. The net effect is to create perverse incentives for some small patent owners to seek out nonpracticing entities to acquire and enforce patents of questionable merit. In addition, the expansion of such lawsuits may well produce a defensive backlash by large operating companies against all small patent owners, even the ones that might have a legitimate and valuable claim.

Two-sided Patent Platforms: A Failed Solution

In parallel with the increased activity of nonpracticing entities during the 2000s, a number of companies built two-sided platforms in an attempt to create more efficient ways to bring buyers and sellers of patents together. The goal of two-sided patent platforms was to facilitate patent transactions without taking title or ownership of the patents involved. Two main categories of such platforms have been attempted: online marketplaces and live auctions.

Online patent marketplaces appeared as early as 1998, but replicating in the market for patents what eBay has done for collectibles has proven difficult. Some of the online portals dedicated to facilitating patent search and transactions have been shut down or renamed and redirected towards other services.[4] The online platforms

[3] The law does try to address this problem through the doctrine of "equitable estoppel," which can bar enforcement of patents by someone who has deliberately waited until after an investment decision has been locked-in to assert patents. We are grateful to Douglas Melamed for bringing this to our attention.
[4] For instance, Patent License and Exchange (pl-x) was created in 1998 as an online intellectual property and licensing marketplace. By 2006 it had been renamed PLX Systems and completely dropped the marketplace idea; instead, it provided software solutions for business and financial management of

that are still independent have limited scale, and they function more as brokerage or consulting companies: two representative examples are Yet2 (http://www.yet2.com) and Tynax (http://www.tynax.com). Both websites contain thousands of listings for both sides of the market. Sellers post detailed information about the patents they want to sell, along with any special conditions (for example, perhaps a license must be granted back to the seller) and without revealing their identity. Buyers can find information about patents that are in the market for sale, search by keywords and patent classes, and post descriptions of specific intellectual property assets in which they have an interest, also without revealing their identity. Both Tynax and Yet2 work with Fortune 500 companies, and for both, keeping the identities of buyers and sellers confidential is a key part of their value proposition. Furthermore, they employ various mechanisms like screening through upfront fees and disclosure requirements to mitigate adverse selection in which only weak patents are offered for sale (Dushnitsky and Klueter 2011)—a potentially serious concern for online trading platforms. Indeed, in the absence of fees, the ratio of low-quality to high-quality products is very high on any online marketplace (for example, Craigslist). This clutter significantly raises search costs for buyers, which in turn disincentivizes high-quality product suppliers from participating. The problem is even worse for patents, because search costs are already very high.

Despite the extensive listings on Yet2 and Tynax's online portals, no transactions are completed online. Instead, once a buyer or a seller expresses clear and credible interest in a posting, Tynax or Yet2 manages and facilitates the buyer–seller interaction offline through one of its dealmakers—who is an actual person. The majority of revenues come from commissions on completed transactions: $100,000 to $10 million for Tynax or 15 percent of licensing fees for Yet2. Thus, both Tynax and Yet2 remain essentially patent and technology brokerage firms.

At first glance, auctions might have seemed like a useful mechanism for eliciting market valuations for patents. The fact that Chicago-based Ocean Tomo managed to organize ten live intellectual property auctions between April 2006 and June 2009 generated significant buzz and optimism regarding the potential for bringing liquidity to the patent market via platforms. These auctions functioned like other live auctions—for example, like art at Sotheby's and Christie's—with an auctioneer taking bids for each lot, which could be a single patent, copyright, trademark, or domain name right, or a bundle of such assets. The lots were sold to the highest bidder on condition that the highest bid exceeded the seller's reserve price.

But the auctions struggled to gain traction. The total value of transactions through Ocean Tomo's ten intellectual property auctions was only $114.6 million (Jarosz, Heider, Bazelon, Bieri, and Hess 2010, p. 17). This total is relatively small, especially when compared to, say, the billions of dollars spent on patent portfolios by

intellectual property for the music and entertainment industry. Other online platforms for matching patent sellers or licensors with buyers or licensees that have disappeared include Open-IP.org, TechEx, PricewaterhouseCoopers' IPEX, and Ocean Tomo's "The Dean's List."

Google in 2011–12 alone. The average sales-to-listings ratio over all ten Ocean Tomo auctions was reportedly 38 percent, and the spring 2009 auction only sold six out of 85 lots listed (*Inside IP* 2012). Part of the reason for the lack of activity in spring 2009 was the financial crisis, but all auctions had been characterized by low participation and little bidding (Jarosz, Heider, Bazelon, Bieri, and Hess 2010, p. 20–22). In June 2009, Ocean Tomo sold its transactions line of business (including auctions and the now-closed "The Dean's List" online platform) to ICAP, an interdealer broker, for just $10 million (ICAP 2009). The live intellectual property auctions were subsequently revived in March 2010 under the joint brand ICAP–Ocean Tomo. The spring 2010 auction (the 11th overall) was reported to have generated $14.3 million in transaction value, including buyers' premiums (ICAP 2010).

Thus, while the idea of creating two-sided platforms for matching and facilitating transactions between patent buyers and sellers is appealing in principle, so far none of these platforms has been able to gain significant traction. None is close to creating a sustainable eBay or Sotheby's for intellectual property. One might argue that Tynax and Yet2.com are creating the economic equivalent of Craigslist for patents, but little more. Why is it so hard to establish two-sided platforms for patent transactions?

First, two-sided platforms that attempt to bring together buyers and sellers without ever taking possession of the goods being exchanged face a difficult chicken-and-egg problem. Unlike market-makers who buy and resell, two-sided platforms have to attract a critical mass of both buyers and sellers. Some online platforms managed to attract many listings, but (as pointed out above) they do not facilitate many actual transactions. Ocean Tomo's auctions never achieved sufficient scale to convince buyers and sellers that they would become an important venue for trading patents. Owners of valuable patents did not *expect* these platforms to offer attractive monetization opportunities for their assets compared to other options like licensing directly, selling to nonprofit entities and splitting the proceeds from litigation, or raising venture capital funding and incorporating. In turn, the lack of valuable patents meant that few large operating companies would participate actively, which confirmed the initial negative expectation of sellers-owners. A broad market was never created; instead a handful of nonprofit entities were very active as buyers in Ocean Tomo's auctions (Malek 2009).

Second, while online intellectual property platforms like Tynax and Yet2.com have generated some search cost reductions through their thousands of listings, they have been unable to create significant reductions in transaction costs. The sensitivity of intellectual property information and the need for "close-touch" and often in-person due diligence make potential buyers and sellers reluctant to reveal enough details for completing a patent transactions online. Of course, this is why Tynax and Yet2.com still function as offline brokers for the actual transactions. But if personal dealmakers have to be directly involved in each transaction, their business model cannot easily scale up at low marginal cost. Moreover, the final transaction prices and valuations are private information, which cannot be leveraged to create greater transparency and liquidity in the patent market.

Will two-sided patent platforms remain limited in scope and scale? Even if they overcome the chicken-and-egg hurdle of how to attract the high-quality and high-value patents, patent platforms seem unlikely to solve the liquidity problems that plague the market for patents. Indeed, given the heterogeneity and strategic sensitivity of patent transactions, it is hard to see how one could create the equivalent of an eBay for patents. Furthermore, the strong complementarities and portfolio effects across modern patents imply that two-sided platforms are at an inherent disadvantage relative to other types of patent intermediaries who take ownership of patents and are able to exploit those complementarities directly. By definition, two-sided platforms cannot do so. That diagnosis does not rule out the emergence (or growth) of platforms specializing in reducing search costs—similar to Tynax and Yet2.com. There is value in being able to browse through thousands of patents, bundles of patents, and technologies wanted or for sale in one place and in a unified format. The official Patent Office listings—patents granted or under review and searchable patent abstracts—leave significant scope for quasi-brokers to further reduce search costs with better listings and search functionality. As pointed out above, many official patent abstracts are written in such a way as to discourage workarounds and to make the broadest possible claims, which often makes it hard to identify potential applications. In this context, firms such as Yet2 create their own abstracts written in clear language in order to help potential buyers assess the potential benefits of the patented technology they are investigating.

Defensive Aggregators and Super-aggregators

The rise of nonpracticing entities combined with the failure of patent platforms to bring transparency and liquidity to the patent market (which might have reduced the arbitrage opportunities for nonpracticing entities) have posed a growing threat to operating companies. In response, two new novel patent intermediaries have emerged, which we call defensive aggregators and super-aggregators.

Defensive Aggregators

There are currently two prominent defensive aggregators: RPX (a for-profit firm, publicly traded since May 2011) and Allied Security Trust (a not-for-profit). In essence, defensive aggregators offer an incomplete insurance policy against patent troll risk to large operating companies. Firms such as Barnes & Noble, Best Buy, Cisco, eBay, HTC, IBM, Intel, McAfee, Microsoft, NEC, Nokia, Panasonic, Research In Motion, Samsung, Sony, and Verizon pay RPX annual subscription fees ranging from $65,000 to $6.9 million, depending on operating income (as explained at RPX's website: http://www.rpxcorp.com/index.cfm?pageid=85, accessed May 2012). In exchange, RPX identifies patents that *might* threaten subscribers, acquires those patents (or the right to grant sublicenses) in the open market, and provides all of its subscribers with licenses to those patents. The patents owned by RPX are also made available for use in counterlawsuits against nonmembers who initiate litigation against members.

Unlike a traditional insurance policy, RPX faces no liability if a subscriber is sued or loses a patent case.

Allied Security Trust, known as AST, offers two main variations on the RPX model. First, RPX decides unilaterally (sometimes in consultation with members) which patents to buy and uses its own capital to do so, while AST identifies patents or portfolios of patents and then solicits acquisition bids from its subscribers, who are also its governing members. Within AST, the bids and the identity of the bidders are kept secret from one another, and each member is required to have sufficient funds in an escrow account in order to support every bid it makes (as explained at the Allied Security Trust website at http://www.alliedsecuritytrust.com/Services/AcquisitionModel.aspx). If the sum of the bids for a particular set of patents is sufficient to close the transaction, then only the members who bid for that particular acquisition receive a license to the relevant intellectual property (as explained at http://www.alliedsecuritytrust.com/Services/LicensingModel.aspx). In the case of RPX, all members receive a license to all patents acquired by RPX. AST's licenses are perpetual from the outset, unlike RPX which introduces vesting periods in its licenses. Members who do not bid in the initial acquisition can still subsequently purchase a license to the patents involved, at a price equal to the highest bid.

Second, after AST acquires a set of patents and licenses its bidding members, it looks to sell those patents. It starts by offering each of the original bidders, starting with the highest one, the opportunity to buy out the entire portfolio by reimbursing the other bidders and AST's related expenses. If none of the bidders is interested, AST places the portfolio for sale with a broker (a divestiture process explained at http://www.alliedsecuritytrust.com/Services/DivestitureProcess.aspx). In contrast, RPX only sells patents occasionally, when it deems that they are no longer useful to its subscribers.

For economists, defensive aggregators raise some interesting issues about contracting. First, the value of RPX to its subscribers seems difficult to verify. Unlike traditional insurers who pay customers when "accidents" happen, defensive aggregators get paid to reduce the probability of "accidents"—in this case, lawsuits from nonproducing entities. But how can members know that RPX is effectively reducing litigation risk on their behalf? Presumably, part of the answer lies is the number of relevant patents that RPX buys. But perhaps more importantly, subscribers view RPX as offering a more efficient buying service for patents they have *already* identified as threatening. When patents are critical to their business, operating companies will often buy them on their own. The issue for many firms is what to do about marginally relevant patents: the expected value of the potential damage may not be sufficient to justify the cost of buying the patent unilaterally, but it may be worth the membership fee paid to RPX, who in turn can aggregate payments across multiple subscribers.

Second, defensive aggregators make an intriguing public commitment *never* to litigate in order to extract revenues (for example, see RPX's website http://rpxcorp.com/, accessed May 2012). This commitment helps differentiate them from patent trolls and serves to reassure potential subscribers, but at the same time,

it creates a significant free-rider problem. When RPX buys a patent (say, for Nokia in smartphones), and eliminates the threat from a troll, then nonsubscribers in the same industries (say, Motorola) equally benefit, so they may be less likely to pay RPX's subscription fees. One way in which RPX mitigates this problem is by adopting a "catch-and-release" approach: it acquires a patent, grants its subscribers a license, and then resells the patent on the open market (preferably to a nonpracticing entity), which means nonsubscribers remain exposed to litigation risk (Hansell 2009). Still, reselling the patents acquired reduces the value of subscribing to RPX for *new* members. This approach also complicates the decision for existing members, who have to determine whether to renew their subscriptions.

Third, the defensive aggregator business model faces an inherent limitation by relying exclusively on subscription revenues. RPX has no shot at the huge payoffs that can be achieved by nonpracticing entities (or a super-aggregator like Intellectual Ventures, which we discuss below). In turn, this puts RPX at a disadvantage in acquiring patents. For example, nonpracticing entities can offer payments for patents that are at least partially contingent on what might be received in a later lawsuit—and therefore a much larger potential payoff to owners—whereas RPX can only offer a fixed payment. RPX may also face unreasonable prices from patent owners if the latter interpret an approach by RPX as a sign of interest from its subscribers—who are, after all, large and potentially rich operating companies. This outcome is related to the issue of "awareness-inducing information" in incomplete contract settings studied formally in Tirole (2009). RPX tries to mitigate this problem by forming buying syndicates among its subscribers and then using shell companies to buy patents of interest to the syndicate.

It is still too early to tell whether RPX has managed to address these issues successfully: it was founded in 2008, and most of its members are locked in for a minimum of three years, so there is insufficient data as yet regarding membership renewal rates.

Super-aggregator(s)

A new type of player, which we call a super-aggregator, has emerged as the largest and most controversial type of intellectual property intermediary. Epitomized by Intellectual Ventures, a super-aggregator is a hybrid between a defensive aggregator, a large nonpracticing entity, and a "weapons dealer," who can provide intellectual property to litigants on both sides of a battle. At the time of this writing, Intellectual Ventures seems unique because of its size—the company has raised more than $5 billion from a variety of investors—but other entities are trying to emulate its model by raising similar amounts of capital.

Intellectual Ventures is a nonpracticing entity. Its first investor, Microsoft, has publicly said that Intellectual Ventures delivers a highly valued service for technology firms (Hagiu, Yoffie, and Wagonfeld 2011). However, critics have described Intellectual Ventures as "the world's largest patent troll" because it acquires, creates, and seeks to license patents without directly making any products or services itself (Orey 2006). Founded in 2000 by former Microsoft chief

technology officer Nathan Myrhvold, as of mid-2012 the firm has spent approximately $2 billion building the world's third-largest patent portfolio—roughly 35,000 patents, mostly covering software, semiconductors, communications, and e-commerce. Like a venture capital or private equity firm, Intellectual Ventures is structured as a series of funds. Its two largest funds are dedicated to acquiring existing patents from all possible sources: individual inventors, or small and large companies. Its third fund focuses on developing its own inventions in partnership with scientists; for example, current projects include a new type of nuclear reactor and a laser-based weapon for fighting malaria mosquitoes. A fourth fund is targeted at developing and acquiring pre-filing inventions, mostly from universities in Asia, through a variety of technology transfer deals.

The last two funds distinguish Intellectual Ventures from typical patent trolls, who do not invent. During its first 10 years, Intellectual Ventures also differed from a typical nonpracticing entity in that it had not litigated—at least not directly. The company had instead sought to monetize its patent portfolios through "friendly" licensing deals and, when necessary, by forming shell companies or selling patents to third-party nonpracticing entities who would in turn litigate. This indirect approach changed in December 2010, when Intellectual Ventures started filing direct patent infringement lawsuits against a variety of operating companies. In its first lawsuits, Intellectual Ventures filed three patent infringement suits against nine companies, including McAfee, Symantec, and Hynix Semiconductor. In July 2011, Intellectual Ventures filed its fourth suit against a group of 12 companies, including HP, Dell, Wal-Mart, and Best Buy.

The fundamental feature that sets Intellectual Ventures apart from other nonpracticing entitities is that many of its investors are strategic and include prominent technology companies such as Amazon, American Express, Apple, Cisco, eBay, Google, Intel, Microsoft, Nokia, SAP, Sony, Samsung, and Verizon.[5] For these strategic investors, Intellectual Ventures also functions as a defensive patent aggregator. Indeed, firms that invest in Intellectual Ventures automatically receive licenses for subsets of the patents acquired by the firm (earlier investors receive wider coverage), which serves to shield them against lawsuits from trolls or competitors.

The dual structure of Intellectual Ventures as both a nonpracticing entity and a defensive aggregator means that it has a potentially difficult balance to strike between the economic interests of its two types of investors: its strategic investors, who are operating companies, and its financial investors, who include pension funds and university endowments. This conflict was presumably the reason behind the firm's initial reluctance to litigate directly. The "friendly" licensing approach was aligned with the interests of strategic investors-licensees, while financial investors' interests are conceivably better served by a more aggressive litigation strategy. Suppose, for example, that an operating company is a limited partner in one of Intellectual Ventures' funds, and is only licensed to part of the portfolio. If the

[5] The list of investors in Intellectual Ventures has been revealed in the filings for a lawsuit initiated by Intellectual Ventures against Xilinx (*XILINX, Inc. v. Intellectual Ventures LLC* (N.D. Cal. 2011)).

operating company were infringing on new patents bought by Intellectual Ventures, Intellectual Ventures might be reluctant to bring a lawsuit against this company, thus creating an opportunity cost borne by all of its financial investors.

The fundamental premise of the Intellectual Ventures model is that its unprecedented scale helps reduce search and transaction costs, as well as patent valuation uncertainty, on both sides of the market. Because of its size, Intellectual Ventures can single-handedly create liquidity in the patent market. It has become an attractive outlet for a number of small patent owners, including smaller universities, most of whom do not have the necessary legal and technical expertise, resources, and credibility to monetize their intellectual property on their own. On the other side of the market, Intellectual Ventures provides patent buyers and users with a "one-stop shop" for their licensing needs: similar to RPX, the company is more efficient when it comes to search and negotiating with multiple patent owners. Furthermore, the scale of Intellectual Ventures allows it to capitalize on huge portfolio and learning effects in aggregating patents.

Of course, the super-aggregator model also carries large risks. Even after accounting for complementarities and portfolio effects, the inventory risk remains very high: no matter how effectively Intellectual Ventures filters the patents that it buys, many patents turn out to be of low value or poor quality or both (as many as 19 in 20 or 49 in 50, according to the company's own estimates, as described in Hagiu, Yoffie, and Wagonfeld 2011). Furthermore, sorting through and maintaining tens of thousands of patents may actually create diseconomies of scale. After all, patents are rapidly depreciating assets because their value expires after 20 years, and they require payment of maintenance fees to be kept valid (several hundred to a few thousand dollars to be paid at the end of years 3, 7, and 11 (as explained at the USPTO website, http://www.uspto.gov/patents/process/maintain.jsp).

Finally, the time-horizon for Intellectual Ventures investment funds is relatively long at 15–20 years, and one may question whether the firm will ever be able to generate returns for its investors comparable to other investment vehicles with similar time horizons, like venture capital and private equity. The last concern suggests that Intellectual Ventures is under pressure to engage in more lawsuits. Yet the lawsuits raise their own problems: cost escalation and, even more seriously, the risk of having some patents invalidated by the courts, which might cast doubts on the value of Intellectual Ventures' broader patent portfolio.

Implications and Conclusions

The patent system is inherently a second-best mechanism for trading off the benefits of enhanced future innovation against the costs of temporary distortions of the economic system after innovation has occurred. Furthermore, the practical realities of the patent system have created additional problems: for instance, a substantial number of low-quality, overlapping, and excessively broad patents. Patent intermediaries (including the new ones described in this article) are able

to profit from the patent system's inherent tension by improving the payoffs to innovators and/or by taxing more heavily the fruits of past innovations. Given the organizational complexity of the new patent intermediaries and the multiplicity of channels through which they affect participants in the patent market, it is very difficult to draw clear conclusions about whether they generate net benefits or costs for society. Nevertheless, it is useful to point out that intermediation mechanisms that move the imperfect patent system in the direction of enhancing rewards for innovation are more likely to be a positive, while mechanisms that move the system in the direction of extracting taxes on prior innovation are likely to be a social negative. The new patent intermediaries clearly do both—and in fact, cannot do one without the other. But their organizational structures and business models may be skewed more heavily on one side or the other, which provides some basis for considering their net social value.

While defensive aggregators are completely aligned with the interests of operating companies in reducing the patent troll threat, this orientation does not mean that they improve the overall efficiency of the patent market. To some extent, they facilitate collusion between large operating companies at the expense of small inventors. By definition, their incentives are to acquire relevant intellectual property at the lowest possible cost to defend their subscribers, not to maximize the value of the patents they acquire. Thus, they are likely to exacerbate the bargaining and information asymmetries between small patent owners and large operating companies (a similar effect to that of traditional cross-licensing practices).

Intellectual Ventures (and other future super-aggregators) are significantly more complicated because of their hybrid nature. Let us consider how a super-aggregator affects the incentives of operating companies, financial investors, and small inventors. Operating companies may see their operating costs increase when Intellectual Ventures aggregates and asserts previously "silent" patents against them. But a super-aggregator like Intellectual Ventures may also lower their aggregate search and transaction costs by providing a one-stop group-licensing shop—just like defensive aggregators do for their members. This service is particularly valuable for technology companies in sectors with short innovation cycles. As a consequence, the net effect of Intellectual Ventures on the development and innovation incentives on operating companies is ambiguous. Some operating companies like Microsoft view it as providing a useful patent discovery and licensing service; others view it as a dangerous nonpracticing entity which significantly raises their costs. Small patent owners, individual inventors, and small companies and universities involved in invention unambiguously benefit from the existence of Intellectual Ventures, because it channels more financial rewards to previously undercompensated inventors, which should unambiguously increase their innovation incentives. Similarly, financial (nonstrategic) investors see Intellectual Ventures as a viable vehicle for investing in patents as a new, large, and uncorrelated asset class.

Due to huge economies of scale, it seems most likely that in the long run there will only be a few super-aggregators—or even just one. This concentration raises significant hold-up concerns. A super-aggregator may become nothing more

than a super-troll, able to hold up both sides of the market by extracting excessive payments from operating companies (for example, by strategically disaggregating patent portfolios and enforcing the different parts sequentially) while at the same time paying lower compensation to inventors. Perhaps an even greater source of concern is that super-aggregators' incentives may be skewed towards imposing higher fees on operating companies *current production* activities, rather than facilitating the commercialization of *unproven patents* (a riskier endeavor).

But, perhaps surprisingly, there could also be significant social efficiency gains from super-aggregator market concentration. Scale leads to significant learning effects in assessing the value of patents, which may create a more reliable mechanism for patent valuation (where patent platforms have failed). Furthermore, in the second-best world created by patent market failures, which lead to excessive patent infringement, it may be efficient to have only a few (or one) market-based enforcer(s). A super-aggregator, in theory, can compensate inventors of a given patent (or portfolio) who otherwise would fall through the cracks. When a super-aggregator buys patents in order to assert them against operating companies that attempt to free-ride on the intellectual property, it preserves the incentives for future innovation. Finally, scale and capital structure, and the accompanying large returns promised to financial investors, can act as credible commitments to build valuable patent portfolios and license them broadly to many players in any given industry. In particular, a super-aggregator's ability to sign large numbers of licensees, without the risks of litigation, depends on its reputation. Enforcing even one weak patent for "nuisance value" (like many small nonpracticing entities do) would run the risk of casting doubt over the value of the super-aggregator's broader patent portfolio. This is an instance of the reputation-building mechanism by intermediaries in contexts with goods of uncertain quality, as studied formally by Biglaiser (1993).

The task of empirically measuring the *net* economic impact of any intellectual property intermediary and deciding whether it is harmful to society is inherently difficult. Such an analysis would require measuring the net effect on operating companies, inventors, universities, and financial investors, both in terms of short-run payments made or received and in terms of long-run innovation incentives. These effects seem dauntingly complex to measure. For this reason, most recent empirical studies only estimate the effects on one side of the market—and thus are by definition incomplete. Examples include the Bessen, Meurer, and Ford (2011) estimates of the costs imposed by trolls on operating companies between 1990 and 2010, and Tucker (2012) evaluating the effect of trolls on the adoption of medical imaging technology sold by vendors targeted by trolls.

Part of the problem is the difficulty of measuring net transfers to inventors. In many cases, nonpracticing entities make lump-sum payments to inventors in exchange for control of their patents *before* any litigation occurs; for example, Intellectual Ventures spent over $1 billion dollars acquiring patents from various sources before it began suing publicly in late 2010. These transfers are usually not disclosed publicly, unlike the settlements or damages resulting from lawsuits. In the absence of access to such information, empirical research on intellectual property intermediaries might

tackle some narrower efficiency questions. For example, an important contributing factor to the effect of nonpracticing entities (including super-aggregators) on innovation incentives is whether they seek to enforce *proven* patents on *existing* products or to facilitate the commercialization of *unproven* patents. Thus, perhaps one could categorize and measure the mix of patents monetized by nonpracticing entities (even without transaction prices) to provide a valuable proxy for their likely effect on innovation.

■ *We are grateful to the following people for extremely insightful comments and feedback on earlier drafts of the article: David Autor, Peter Detkin, Chang-Tai Hseih, Josh Lerner, John List, Douglas Melamed, and Timothy Taylor. Britta Kelley provided excellent research assistance. The views expressed in this paper are entirely our own.*

References

Analysis Group. n.d. "Patent Litigation before the International Trade Commission," p. 2. http://www.analysisgroup.com/uploadedFiles/Practice_Areas/Patent_Litigation_Before_the_ITC.pdf (accessed May, 2012).

Arora, Ashish, Andrea Fosfuri, and Alfonso Gambardella. 2001. *Markets for Technology: The Economics of Innovation and Corporate Strategy.* Cambridge, MA: MIT Press.

Bario, David. 2011. "'Original Patent Troll' Is Back after Long Hiatus, with a New Name and New Patents." *Am Law Litigation Daily*, August 17. http://amlawdaily.typepad.com/amlawdaily/2011/08/original-patent-troll-is-back-new-name-new-patents.html.

Bergelt, Keith. 2010. "Patents, Probes and Strength in Unity: Participate in Keeping Open Source Open." Presentation, *LinuxCon* (Boston, MA), August 10. http://events.linuxfoundation.org/linuxcon2010/bergelt.

Bessen, James, and Michael J. Meurer. 2008. *Patent Failure: How Judges, Bureaucrats and Lawyers Put Innovators at Risk.* Princeton, NJ: Princeton University Press.

Bessen, James E., Michael J. Meurer, and Jennifer Laurissa Ford. 2011. "The Private and Social Costs of Patent Trolls." Boston University School of Law, Law and Economics Research Paper No. 11-45. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1930272.

Biglaiser, Gary. 1993. "Middlemen as Experts." *RAND Journal of Economics* 24(2): 212–23.

Chiao, Benjamin, Josh Lerner, and Jean Tirole. 2007. "The Rules of Standard-Setting Organizations: An Empirical Analysis." *RAND Journal of Economics* 38(4): 905–930.

Crothers, Brooke. 2009. "Qualcomm, Broadcom Reach $891 Million Settlement." *CNET News*, April 26, 2009. http://news.cnet.com/8301-13924_3-10227815-64.html (accessed May, 2012).

Decker, Susan. 2012. "A Crackdown on Patently Absurd Lawsuits." *Bloomberg Businessweek*, May 10. http://www.businessweek.com/articles/2012-05-10/a-crackdown-on-patently-absurd-lawsuits.

Dushnitsky, Gary, and Thomas Klueter. 2011. "Is There an eBay for Ideas? Insights from Online Knowledge Marketplaces." *European Management Review* 8(1): 17–32.

Fisher, Jeffrey M. 2006. "The Pros and Cons of Patent Litigation before the International Trade Commission." Farella, Braun, and Martel LLP, November 15. Available at: http://www.fbm.com/The_Pros_and_Cons_of_Patent_Litigation_Before_the_International_Trade_Commission_11-15-2006.

Gallagher, Billy. 2012. "Apple Awarded $1.049

Billion in Damages as Jury Finds Samsung Infringed on Design and Software Patents." Tech Crunch, August 24. http://techcrunch.com/2012/08/24/apple-wins-patent-ruling-as-jury-finds-samsung-infringes/.

Gans, Joshua S., David H. Hsu, and Scott Stern. 2002. "When Does Start-Up Innovation Spur the Gale of Creative Destruction?" *RAND Journal of Economics* 33(4): 571–86.

Gans, Joshua S., and Scott Stern. 2010. "Is There a Market for Ideas?" *Industrial and Corporate Change* 19(3): 805–837.

Hagiu, Andrei, David B. Yoffie, and Alison Wagonfeld. 2011. "Intellectual Ventures." *Harvard Business School Case* No. 710-423, Harvard Business School Publishing.

Hansell, Saul. 2009. "Trolling for Patents to Fight Patent Trolls." *New York Times*, March 30. http://bits.blogs.nytimes.com/2009/03/30/trolling-for-patents-to-fight-patent-trolls/.

ICAP. 2009. "ICAP Reaches Agreement to Acquire Patent Brokerage Business of Ocean Tomo LLC." June 16. http://www.icap.com/news-events/in-the-news/news/2009/icap-reaches-agreement-to-acquite-patent-brokerage-business-of-ocean-tomo-llc.aspx (accessed May, 2012).

ICAP. 2010. "ICAP Ocean Tomo Auction Sees Record Bidding." March 26. http://www.icap.com/news-events/in-the-news/news/2010/icap-ocean-tomo-auction-sees-record-biding.aspx (accessed May 2012).

Inside IP. 2012. "Going . . . Going . . . Gone?: The Rises and Falls of Patent Auctions." June 14. http://www.vennershipley.co.uk/show-news-id-249&p=2.html (accessed May 2012).

Jaffe, Adam B., and Josh Lerner. 2004. *Innovation and Its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to Do About It.* Princeton, NJ: Princeton University Press.

Jarosz, John, Robin Heider, Coleman Bazelon, Christine Bieri, and Peter Hess. 2010. "Patent Auctions: How Far Have We Come?" *les Nouvelles*, March, pp. 17, 20–22.

Lamoreaux, Naomi R., and Kenneth L. Sokoloff. 2003. "Intermediaries in the U.S. Market for Technology, 1870–1920." In *Finance, Intermediaries and Economic Development*, edited by Stanley L. Engerman, Philip T. Hoffman, Jean-Laurent Rosenthal, and Kenneth L. Sokoloff, 209–246. Cambridge: Cambridge University Press.

Lampe, Ryan, and Petra Moser. 2010. "Do Patent Pools Encourage Innovation? Evidence from the Nineteenth-Century Sewing Machine Industry." *Journal of Economic History* 70(4): 898–920.

Layne-Farrar, Anne, and Josh Lerner. 2011. "To Join or Not to Join: Examining Patent Pool Participation and Rent Sharing Rules." *International Journal of Industrial Organization* 29(2): 294–303.

Lemley, Mark A. 2008. "Are Universities Patent Trolls?" *Fordham Intellectual Property, Media and Entertainment Law Journal* 18(3): Article 2.

Lemley, Mark A., and Carl Shapiro. 2005. "Probabilistic Patents." *Journal of Economic Perspectives* 19(2): 75–98.

Lerner, Josh, and Jean Tirole. 2004. "Efficient Patent Pools." *American Economic Review* 94(3): 691–711.

Lerner, Josh, and Jean Tirole. 2006. "A Model of Forum Shopping." *American Economic Review* 96(4): 1091–1113.

Lerner, Josh, Marcin Strojwas, and Jean Tirole. 2007. "The Design of Patent Pools: The Determinants of Licensing Rules." *RAND Journal of Economics* 38(3): 610–625.

Lohr, Steve. 2009. "Patent Auctions Offer Protections to Inventors." *New York Times*, September 20. http://www.nytimes.com/2009/09/21/technology/21patent.html?_r=0.

Malek, Marcus. 2009. "R.I.P. Ocean Tomo: Complete Auction Analysis." Presentation slides, June 23, http://www.slideshare.net/marcusmalek/complete-ocean-tomo-auction-analysis-marcus-malek-intangitopia (accessed May 2012).

Masnick, Mike. 2009. "Patent Holder Sues McAfee, Gets $25 Million . . . But May End up Losing $5 Million due to Everyone It Has to Pay Off." *Techdirt*, November 4. http://www.techdirt.com/articles/20091103/0333046778.shtml.

McCurdy, Daniel P. 2009. "Patent Trolls Erode the Foundation of the U.S. Patent System." *Science Progress*, January 12. http://www.scienceprogress.org/2009/01/patent-trolls-erode-patent-system/.

McDonough, James F. III. 2006. "The Myth of the Patent Troll: An Alternative View of the Function of Patent Dealers in an Idea Economy." *Emory Law Journal* 56(1): 189–228.

Orey, Michael. 2006. "Inside Nathan Myhrvold's Mysterious New Idea Machine." With Moria Herbst. *Bloomberg Businessweek*, July 2, 2006. http://www.businessweek.com/stories/2006-07-02/inside-nathan-myhrvolds-mysterious-new-idea-machine.

Parchomovsky, Gideon, and Richard Polk Wagner. 2005. "Patent Portfolios." *University of Pennsylvania Law Review* 154(1): 1–78.

Patent Freedom. n.d. Website. https://www.patentfreedom.com/research-lot.html.

Riordan, Teresa. 2004. "Contest over BlackBerry Patent." *New York Times*, June 7. http://www.nytimes.com/2004/06/07/technology/07patent.html.

Schmalensee, Richard. 2009. "Standard-Setting,

Innovation Specialists, and Competition Policy." http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1219784.

Schudel, Matt. 2005. "Accomplished, Frustrated Inventor Dies." *Washington Post*, February 26, 2005. http://www.washingtonpost.com/wp-dyn/articles/A54564-2005Feb25.html.

Shapiro, Carl. 2001. "Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting." In *Innovation Policy and the Economy*, Vol. 1, edited by Adam Jaffe, Josh Lerner, and Scott Stern, 119–150. Cambridge, MA: MIT Press.

Sharma, Amol, and Don Clark. 2008. "Tech Guru Riles the Industry by Seeking Huge Patent Fees." *Wall Street Journal*, September 17. http://online.wsj.com/article/SB122161127802345821.html.

Simcoe, Timothy. 2012. "Standard Setting Committees: Consensus Governance for Shared Technology Platforms." *American Economic Review* 102(1): 305–36.

Spulber, Daniel F. 2011. "Should Business Method Inventions Be Patentable?" *Journal of Legal Analysis* 3(1): 265–340.

Tirole, Jean. 2009. "Cognition and Incomplete Contracts." *American Economic Review* 99(1): 265–94.

Tucker, Catherine. 2012. "Patent Trolls and Technology Diffusion." TILEC Discussion Paper No. 2012-030. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2136955.

Wasserman, Noam. 2012. *The Founder's Dilemmas: Anticipating and Avoiding the Pitfalls That Can Sink a Startup*. Princeton, NJ: Princeton University Press.

Young, Eric. 2008. "Patent Marketeers." *San Francisco Business Times*, March 23. http://www.bizjournals.com/sanfrancisco/stories/2008/03/24/focus1.html?page=all.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 67–86

Of Smart Phone Wars and Software Patents

Stuart Graham and Saurabh Vishnubhakat

Among the main criticisms currently confronting the US Patent and Trademark Office are concerns about software patents and what role they play in the web of litigation now proceeding in the smart phone industry. While such criticisms are not new, the realm of smart phones offers an opportunity to examine the evidence on the litigation and the treatment by the Patent Office of patents that include software elements. The term "software patent" is a bit of a misnomer, since computer programming is a general purpose technology. After all, patents that claim software elements can be found in virtually every industry and a broad range of technologies.

More broadly, this article discusses the competing values at work in the patent system and how the system has dealt with disputes that, like the smart phone wars, routinely erupt over time, in fact dating back to the very founding of the United States. We present specific empirical evidence regarding the examination by the Patent Office of software patents, their validity, and their role in the smart phone wars. The article concludes with an outlook for systematic policymaking within the patent system in the wake of major recent legislative and administrative reforms. Principally, the article highlights how the US Patent Office acts responsibly when it engages constructively with principled criticisms and calls for reform, as it has during the passage and now implementation of the landmark Leahy–Smith America Invents Act of 2011.

■ *Stuart J. H. Graham is an Expert Advisor at the US Patent and Trademark Office. He is an Assistant Professor, Scheller College of Business, Georgia Institute of Technology, Atlanta, Georgia, and is an attorney licensed in the State of New York. Saurabh Vishnubhakat is an Expert Advisor at the US Patent and Trademark Office. He is an Adjunct Professor at the Northern Virginia Community College, Alexandria, Virginia, and is an attorney licensed in the state of Illinois.*

http://dx.doi.org/10.1257/jep.27.1.67. doi=10.1257/jep.27.1.67

Constitutional Background

For more than 235 years, a national investment in the future has been part of the formal social contract of the United States, providing patent rights limited in time and scope in exchange for a full and timely disclosure of new and useful innovations. The US Constitution in Article I, Section 8, Clause 8 expressly empowers Congress to manage this investment by establishing a patent system. The constitutional design of the US patent system recognizes that the marketplace tends to undersupply knowledge, particularly when up-front costs of discovery or development are high and marginal costs of copying are low. A period of limited exclusivity is meant to provide an incentive to make such investments.

Accordingly, the revealed national preference of the United States has been to forgo some immediate economic benefit in favor of creating incentives for new generations of advances in scientific knowledge and technological application. There is a natural tendency for innovation—which requires investment for long-term benefit—to interfere with access in the form of consumption in the short term. But even in the here and now, this constitutionally protected ability of innovators in the United States to leverage their patent rights for attracting investment capital, creating jobs, and expanding into new markets reflects a preference for ensuring that individuals and small start-up firms have an opportunity to thrive and grow. To be sure, patent rights are only part of a suite of legal and nonlegal appropriability options available to innovators. Still, patents have been a useful and oft-used means of protecting innovations since the country's inception.

The Patent Office Role

The US Patent and Trademark Office's primary responsibility is to support the innovation system by examining patent and trademark applications. In the US system, patents on mechanical, electronic, and chemical technologies are generally "utility patents." Utility patent applications submitted to the Patent Office by inventors may cover processes, machines, articles of manufacture, or compositions of matter. Upon being accepted as complete, applications are given a technology classification and assigned to an examiner group. Patent examiners are specialized technology employees with training and experience in various science and engineering backgrounds related to different kinds of inventions. Examiners are also public officers who have a legal duty to grant a patent so long as the inventor has met the requirements for patentability set by Congress and the federal courts. In fact, Congress demands that "[a] person shall be entitled to a patent unless" the examiner is able to find a basis to refuse the application. To find possible bases for rejection, the examiner compares the claimed invention to the existing state of knowledge as reflected in the prior art, consisting of patent documents and the scientific and commercial literature.

Because a patent is a series of claims that define the scope of the invention, an examiner uses the results of that search to determine whether these claims

delineate what the law demands: that the invention be new, useful, and adequately described and enabled. A legally sufficient claim must also be nonobvious to persons "having ordinary skill in the art" of the invention, so as to encourage inventions that will serve the cumulative advancement of technology at the frontiers of knowledge.

Through such search and examination of the application, examiners serve the public by clearing out patent claims that do not meet these legal requirements. Their labor is highly complex and knowledge-intensive, since it requires both scientific and legal understanding. In 2011, the Patent Office received over 500,000 applications.

The most fundamental and important contribution the Patent Office makes to improving the patent system involves focusing on, and investing in, higher-quality and more timely processing of patent applications (US Patent and Trademark Office 2010). In this context, "quality" refers to patent claims being clearly defined and consistently interpreted under the law, and "timeliness" encompasses a reduction in delays and pendency during examination. Both goals reduce uncertainty, and allow for more efficient investment and transactions in the market for innovation. Scholars have consistently supported these goals to improve the operation of the innovation system (National Research Council 2004).

Economic Research

This focus on reducing uncertainty—an economic concept—raises an important question about what role the patent system plays in economic growth. It is clear that a substantial share of national growth in the United States has been driven by innovation and the deployment of new technologies, which have, in turn, produced higher standards of living along with better, longer lives for people. Economists have struggled over the years to discover what role intellectual property rights play in the supply of innovation and the productivity improvements and economic growth that new technologies have ushered in. The task has been made difficult by endogeneity problems, in that patenting is correlated with other important drivers of performance. Good instruments to help us untangle this complexity are rare to nonexistent, and apart from some fine historical examples (Moser 2005), increasing international harmonization of patent laws minimizes the opportunity to observe the results of natural experiments in the real world.

That said, a body of economic research has demonstrated a positive role for patenting in economic performance. Gould and Gruben (1996), for instance, utilize cross-country data on patent protection to find that intellectual property protection is a significant determinant of economic growth. Branstetter and Saggi (2011) contribute to this general finding, showing that increased intellectual property protection in developing countries leads to more inbound foreign direct investment, a greater local production share of the global basket of goods, and higher real wages for local labor.

The mechanisms through which patent rights work to drive growth have also been a subject of research. In addition to the classical view that intellectual property rights provide an incentive to create knowledge (Arrow 1962), scholars have found that the issuing of patents is a significant determinant of commercializing inventions through licensing (Gans, Hsu, and Stern 2008). This latter view is consistent with work on the role of intellectual property rights in providing a transactional platform that facilitates a more efficient transfer of knowledge assets and gains from trade in the markets for technology (Arora, Fosfuri, and Gambardella 2004).

At the level of the firm, too, patents have been found to have an economically and statistically significant impact on firm-level productivity and market value (Bloom and Van Reenen 2002). Evidence provided by several surveys of managers at technology firms supports the notion that patents are valuable and serve a range of purposes, in preventing copying, earning profits, and engaging in effective technology competition. In a survey of research and development managers at firms across the US economy, researchers discovered that patents are widely used by firms in industry and are routinely cited as being important for profiting from innovation, although not ubiquitously so (Cohen, Nelson, and Walsh 2000). The respondents suggested that when patents were used, they served heterogeneous purposes, including protecting inventions from copying, earning licensing revenues, supporting negotiations, and enhancing reputation. A more recent survey of young technology startups basically confirmed these findings, although the respondents commonly cited the importance of building patent portfolios to facilitate inward capital investment and increasing the likelihood of successful exit events such as initial public offerings and acquisitions by other firms (Graham, Merges, Samuelson, and Sichelman 2009).

While a growing body of evidence finds that patent protection supports innovation and growth, some critics contend that the patent system should be dismantled wholesale. However, large systematic changes of the kind advocated by these critics are best interpreted in light of Oliver Williamson's (2009) "remediableness" criterion, to wit: an existing practice for which no superior feasible alternative can be described and implemented with expected net gain is presumed to be efficient. Without strong evidence of the superiority of such a large change in the institutional environment in which innovation and economic activity occurs, a "do away with patents" alternative cannot be fairly categorized as a hypothetical ideal. And even if, *arguendo*, such an alternative were hypothetically ideal, the large transaction costs associated with moving an innovation system and an economy to this new equilibrium would have to be considered in netting out the possible gains. Advocates for this view have made little progress in carrying either of these very heavy burdens.

The Patent System Has Faced and Still Faces Problems Arising from New Technologies and Uncertainty in Legal Treatment

The view that society would do better by rejecting patent incentives as both unnecessary and obstructing to knowledge consumption in the short term is closely

related to criticisms that have been made regarding the patent infringement litigation of recent years among firms in the smart phone industry. Such arguments suggest that these so-called "smart phone wars" arise from overbroad and improperly issued software patents, and thus reflect flaws in patent eligibility doctrine, a too-permissive treatment at the Patent Office of software patents, and economic waste in litigation. Such criticisms are not new. They commonly reflect the recurring difficulty the patent system has experienced when facing the legal and market uncertainty associated with the eruption of discontinuous technological change.

History is a guide to us in this regard, since over time the United States patent system has been met with new challenges in technology and industrial organization, but has adapted. At times, the resolution has come legislatively, as in the 1836 patent act. Under the 1793 patent statute, patent examination was not just permissive, it was nonexistent: the Patent Office granted any patent properly applied for, leaving to society and the courts the costs of clarifying patent rights through piecemeal litigation. To mitigate the social costs, the 1836 patent act reintroduced substantive examination of patent applications for novelty and utility.

In this century, important changes in the patent law intended to deal with the demands of a changing innovation environment have occurred in 1930,[1] in 1952,[2] in 1970,[3] in 1982,[4] in 1994,[5] in 1999,[6] and again recently in the sweeping changes required in the 2011 America Invents Act (which we discuss below). Often the change has come about because of compelling factual circumstances, such as regional or national economic concerns or even the exigencies of war. Such changes have been messy and with contradiction, and often against the backdrop of patent litigation around the valuable technologies at stake. But a well-developed economic history suggests that this is what we should expect when institutional systems supporting economic activity respond to new circumstances (North 1981).

In the history of the United States, society has repeatedly celebrated seminal inventions while bemoaning the patent disputes that emerged around them. For example, Eli Whitney patented the mechanical cotton gin in 1794, ushering in huge productivity gains, but was unable to prevent wholesale infringement for many years since local juries, who resented Whitney for taking large royalties from farmers,

[1] Plant Patent Act of 1930, Pub. L. No. 71-245, 46 Stat. 376 (making patent-eligible certain new varieties of plants).

[2] Patent Act of 1952, Pub. L. No. 82-593, 66 Stat. 792 (broadly codifying and clarifying existing patent case law).

[3] Plant Variety Protection Act of 1970, Pub. L. No. 91-577, 84 Stat. 1542 (extending exclusive plant breeder rights to sexually reproduced and tuber-propagated varieties).

[4] Federal Courts Improvement Act of 1982, Pub. L. No. 97-164, Apr. 2, 1982, 96 Stat. 25. (creating the US Court of Appeals for the Federal Circuit and giving it exclusive jurisdiction over appeals both from patent litigations in the district courts and from administrative patent appeals in the US Patent and Trademark Office).

[5] Uruguay Round Agreements Act, Pub. L. No. 103-465, 108 Stat. 4809 (among other things, setting the patent term as 20 years from filing date rather than 17 years from issue date).

[6] American Inventors Protection Act, Pub. L. No. 106-113, 113 Stat. 1501 (among other things, adjusting patent term to partly compensate for certain examination delays, and requiring publication of most applications at 18 months from filing date).

would rule against him. It was not until the patent law was amended in 1800 that Whitney's legal rights were vindicated, and even then with limited economic reward.

A half-century later, when Elias Howe in 1846 patented his eye-pointed needle sewing machine, contributing to productivity gains and new economic freedoms for women, it began a period of extensive litigation among industry rivals. In response to moves like those of Isaac Singer, who tried unsuccessfully to invalidate the Howe patent, the legal landscape changed again, with the emergence in 1856 of one of the first US patent pools, in which major producers cross-licensed their mutually blocking patents. Notably, Howe was not himself a manufacturer of sewing machines, but rather a patent-holder interested in licensing his invention—by modern standards, a nonpracticing entity.

Another half-century later, Orville and Wilbur Wright patented the wing and steering designs of their flying machine, in 1906, and showed their work to the Aerial Experiment Association, founded the following year by another celebrated inventor and well-known patent litigant, Alexander Graham Bell. Having refused a license from Glenn Curtiss for his engine, the Wrights were soon mired in litigation when airplanes built by Curtiss and other industry players that infringed on the Wright brothers' steering patents met with commercial and reputational success. While the infringement dispute ended with a verdict for the Wrights, the broader business dispute was resolved only when Assistant Navy Secretary Franklin Roosevelt in 1917 pressured the rivals to allow unrestricted production of airplanes for the war effort. The scale of the dispute was larger than ever, but the lessons of the sewing machine wars had not been lost, and the airplane patents were cross-licensed through a patent pool administered by Manufacturer's Aircraft Association.

Still another half-century later, in 1957, Columbia University student Gordon Gould made some rough calculations and a sketch in his notebook of the first LASER (that is, Light Amplification by Stimulated Emission of Radiation). Gould soon left Columbia for a private research firm, and other scientists independently developed the same technology about three months after Gould, igniting a 30-year series of disputes. When Gould ultimately prevailed, the controversy over invention priority gave way to industry resistance during the 1970s and 1980s to the enforcement of Gould's patents that had been pending as applications for long periods of time. Such so-called "submarine" patents can be problematic in instances like lasers, where the patented technology becomes widely adopted across industries without firms knowing that fundamental patents exist. As before, the system righted itself, in this instance by reducing the ability of inventors to "submarine" their inventions, in 1995 by changing the measurement and length of the patent term, and in 1999 by publishing patent applications 18 months after filing.

Now, again about 50 years later after the struggles over the laser, we are embroiled in the smart phone wars. When we take the long view, this controversy does not look like a dispute for the ages, but instead a kind of controversy that has arisen periodically throughout the history of the US innovation system. The resolution of each crisis has been a refinement and reform of the patent law to meet modern needs, particularly as innovation has over time commanded increasing priority to national

economic health. That same recalibration appears to be at work in how the system is dealing with smart phone patents. In fact, fair examination of the available evidence shows that the smart phone patent wars are not about low-quality software patents, nor about errors in software patent examination or issuance.

Smart Phone Wars and "Software Patents:" Some Empirical Evidence

The smart phone patent wars have produced a large number of US lawsuits involving major industry players like Samsung, Google's Motorola Mobility division, and Apple, with many counterpart suits filed overseas. Yet across these many lawsuits involving smart phones, some important questions have gone unanswered. How credible are the lawsuits? How far have these suits progressed, and how many patents are actually involved? And, given that many critics have suggested the culprit is low-quality software patents, what technologies were actually covered by the patents involved, and how did the patents fare?

In attempting to answer these questions, we examined the US patents involved in some of the high-profile litigation among four major firms in the smart phone industry: Motorola, Microsoft, Apple, and Samsung. While 133 patents were initially asserted across 13 lawsuits, a substantial share was dismissed from the cases and, as of November 2012, only 73 patents remained in controversy. A technology expert at the Patent Office reviewed these 73 patents, determining whether any of the claims could be fairly characterized as involving "software" inventions. We found that 65 of the patents included at least one software-related claim. Thus, while many of the patents asserted in the Motorola, Microsoft, Apple, and Samsung suits involved software claims, not all of the claims were to software elements, and in fact some of the patents asserted had no software elements at all. This finding is not surprising, as smart phones contain much innovation beyond software—for example, display technology, microprocessor technology, signal processing technology, signal transmission technology, and compression technology.

Of the 65 software patents still involved in this litigation, thus far only 21 of them—less than one-third—have received court decisions of the type that provide some indication of their validity or likely validity. Of those, only four patents have had decisions indicating they are invalid or likely invalid. The remaining 17 software patents evaluated so far in these cases have been declared by a court to be valid or likely valid. This 80 percent favorability ratio is not consistent with the pronouncements that the smart phone wars are being driven by low-quality software patents. In fact, this rate of validity determinations compares favorably with other technology areas. In summary, the US federal district courts, which are the principal reviewers of Patent Office decision-making, are finding in a large share of these cases that prior Patent Office examinations of the software patents involved in the smart phone litigation have been completed properly.

While that finding is positive, we were interested in digging deeper and asking other relevant questions. The recent US Supreme Court case of *Bilski v. Kappos*

(130 S. Ct. 3218 [2010]), which overturned a lower-court ruling that patents needed to involve a machine (or apparatus) and/or a transformation of an article to a different state or thing, has implications for the patent-eligibility of software and is but the latest in a series of reminders that this area of law continues to evolve. The Patent Office will continue to reassess its granting and legal treatments of patents that include software elements.

The patenting of software has created much controversy, and the underlying arguments go back long before courts in the 1980s and 1990s affirmed patent-eligibility for software and, relatedly, for patents for "business methods" (Graham and Mowery 2003). Since as early as the Morse telegraph patent disputes in the 1840s, the US patent system has grappled with abstract ideas such as mathematical principles and laws of nature on the one hand, and implementations of these ideas on the other—particularly in nascent technologies where both scientific and legal uncertainty is high. In rejecting one part of Samuel Morse's patent claim—the part concerning the use of electromagnetic power for marking characters at any distance—the US Supreme Court in *O'Reilly v. Morse* (56 US 62 [1853]) noted that Morse had described and enabled only the use of galvanic repeater circuits to preserve a signal over long distances. Without disclosing and teaching more, the Court found, his patent could not cover future applications of electromagnetic force: he could claim his way of transmitting signals, but not *signal transmission itself.*

This emphasis on knowledge diffusion and the patent *quid pro quo*, particularly in emerging and legally uncertain technological environments, has remained integral in US patent law to rewarding particular innovative solutions to problems without foreclosing the problems themselves. Similarly, the patent law leaves available to the public the intellectual tools that follow-on innovators can use to solve such problems.

In the context of software inventions, these principles have been applied by the US Supreme Court. It has denied patent-eligibility to bare algorithms for converting binary-coded decimal numbers into pure binary numerals (*Gottschalk v. Benson*, 409 US 63 [1972]) and for smoothing fluctuations in process variable trends (*Parker v. Flook*, 437 US 584 [1978]), but affirmed patent-eligibility for the physical implementation by a rubber-molding press of the Arrhenius equation (*Diamond v. Diehr*, 450 US 175 [1981]). At the Patent Office, the examining of patent applications for software-related inventions has emphasized, through exacting review of the written description and enabling disclosure of the application, that the invention as claimed must be commensurate with the invention as taught to the public.

Before examining data on patenting of software inventions, first comes a definitional question: What is a software patent, and can we identify it? As any patent examiner can confirm, applications across virtually all major technology areas can include software elements, and among economic researchers, no common definition has emerged for conducting empirical analysis (Layne-Farrar 2005). Part of the difficulty stems from software having some of the characteristics of a general purpose technology. As outlined by Bresnahan and Trajtenberg (1995), these technologies are "pervasive," being widely adopted across many technologies and heterogeneous

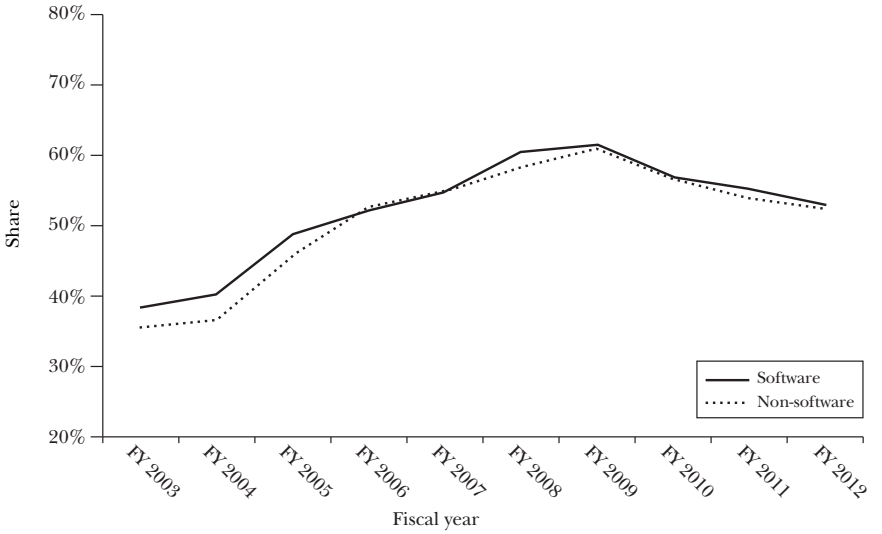
sectors in the economy. In fact, one way in which researchers measure "generality" in empirical analyses of patenting is to examine how widely adopted patents are in later, heterogeneous patented technologies (Hall and Trajtenberg 2004). Related to such pervasiveness, an accurate "software patent" definition is elusive because many patents have software elements mixed with non-software elements.

While the relatively small number of patents involved in the smart phone wars allowed us, above, to employ an expert to read the claims, that method is neither reproducible nor feasible for large-scale empirical analyses. We therefore relied on methods commonly used in the prior literature to identify "software patents," by employing patent classifications (Graham and Mowery 2003; Hall and MacGarvie 2010). Still, identifying patents with software elements can be a tricky business.

To conduct the following analyses in this paper, Patent Office experts examined all US patent classes and subclasses and determined which were likely to contain patents applications or issued patents containing some element of either general purpose software or software that is specific to some form of hardware.[7] While this definition will certainly be both over-inclusive and under-inclusive, the method is calibrated to help us identify classes in which patents with software claims are most likely to be found. As shorthand, we refer to those applications or patents which fall into these classes and subclasses as "software" applications or patents, and to those

[7] The class-subclass pairs are as follows. Class 29: Subclasses 026000-065000, 560000-566400, 650000-650000; Class 73: Subclasses 455000-487000, 570000-669000; Class 84: Subclasses 600000-746000; Class 235; Class 236; Class 244: Subclasses 003100-003300, 014000; Class 250; Class 257; Class 307; Class 315; Class 318: Subclasses 700000-832000; Class 320; Class 323; Class 324; Class 326; Class 327; Class 330; Class 331; Class 340: Subclasses 850000-870440; Class 340: Subclasses 002100-010600, 825000-825980; Class 340: Subclasses 286010-693900, 901000-999000; Class 340: Subclasses 815400-815730, 815740-815920; Class 341: Subclasses 020000-035000, 173000-192000; Class 341: Subclasses 001000-017000, 050000-172000, 200000-899000; Class 342: Subclasses 001000-465000; Class 343; Class 345: Subclasses 001100-215000, 418000-428000, 440000-472300, 473000-475000, 501000-517000, 518000-689000, 690000-698000, 699000; Class 348; Class 353; Class 355; Class 356: Subclasses 002000-003000, 004090-004100, 006000-027000, 030000-139000, 140000, 142000-151000, 153000-900000; Class 358: Subclasses 001100-003320, 260000-517000, 518000-540000; Class 359: Subclasses 326000-332000; Class 361: Subclasses 001000-270000, 437000; Class 363; Class 365; Class 367: Subclasses 001000-008000, 009000, 010000-013000, 014000-080000, 081000-085000, 086000, 087000-092000, 093000-094000, 095000-191000, 197000-199000, 900000-910000, 911000-912000; Class 368; Class 369: Subclasses 001000-032000, 043000-054000, 058000-062000, 064000, 069000-070000, 083000-095000, 097000, 100000-126000, 128000-152000, 174000-175000, 275100-276000, 300000; Class 370; Class 374; Class 375; Class 378: Subclasses 004000-020000, 210000-901000; Class 379: Subclasses 067100-088280, 188000-337000; Class 380; Class 381; Class 382; Class 385; Class 386; Class 396: Subclasses 028000, 048000-304000, 310000-321000, 373000-386000, 406000-410000, 421000, 449000-501000, 505000-510000, 529000-533000, 563000; Class 398; Class 438: Subclasses 009000, 689000-698000, 704000-757000; Class 455; Class 463: Subclasses 001000-047000, 048000-069000; Class 473: Subclasses 065000, 070000, 136000, 140000-141000, 151000-156000, 407000; Class 482: Subclasses 001000-009000, 051000-053000, 057000-065000, 069000-070000, 112000-113000; Class 600: Subclasses 001000-015000, 019000-041000, 300000-406000, 407000-480000, 481000-507000, 529000-595000, 920000-921000; Class 606: Subclasses 001000-052000, 163000-164000; Class 623: Subclasses 024000-026000; Class 700; Class 701; Class 702; Class 703: Subclasses 001000-010000, 011000-012000, 013000-999000; Class 704; Class 705; Class 706; Class 707; Class 708; Class 709; Class 710; Class 711; Class 712; Class 713; Class 714: Subclasses 001000-100000, 699000-824000; Class 715; Class 716; Class 717; Class 718; Class 719; Class 725; Class 726; Class 901; Class 902.

Figure 1
Share of US Patent Office First Final Actions that Were Rejections, FY 2003–FY 2012



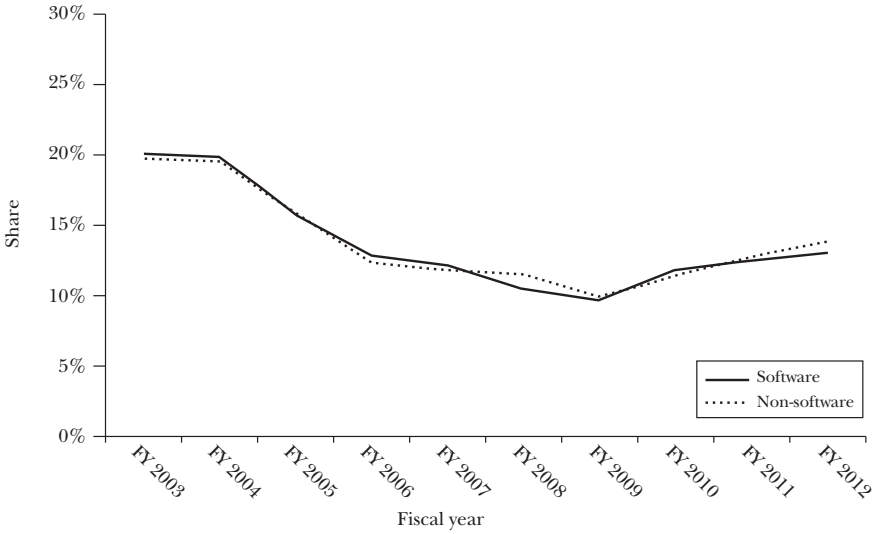
Source: Authors.
Notes: After an examiner initially rejects claims as unpatentable and the applicant responds with arguments or amendments, the examiner issues a "final action:" either an allowance or rejection. This is termed a "first final action" because the applicant may seek continued examination, leading to further iterations of nonfinal and final actions subsequent to the "first" one.

which fall outside as "non-software," with the understanding that this nomenclature is one of convenience, and will not be accurate in all cases.

Having a definition of convenience in hand, we can then proceed to some questions. How does our rejection rate for software applications compare with that of applications in the other technologies? Conversely, how does our rate of allowance on a first-action by the examiner compare with that of applications in the other technologies? How often are our examiners' rejection decisions upheld by our board of patent appeals (the principal reviewer within the Patent Office of examiners' denying patent protection)? How has our reviewing court, the US Court of Appeals for the Federal Circuit, treated our rejection decisions compared with our own board of patent appeals? In other words, when the board of patent appeals upholds examiner rejections, how does the US Court of Appeals treat those determinations?

First, as regards final rejections, the Patent Office used to reject software applications at a higher rate than non-software applications, as shown in Figure 1. Ten years ago, the rate of final rejection for software applications was 38.4 percent, 2.8 percentage points higher than for non-software applications. Over time, the final rejection rates for both software and non-software applications had risen, exceeding 60 percent by 2009. Thereafter, these rates declined to below 55 percent.

Figure 2
Share of US Patent Office First Actions that Were Allowances, FY 2003–FY 2012



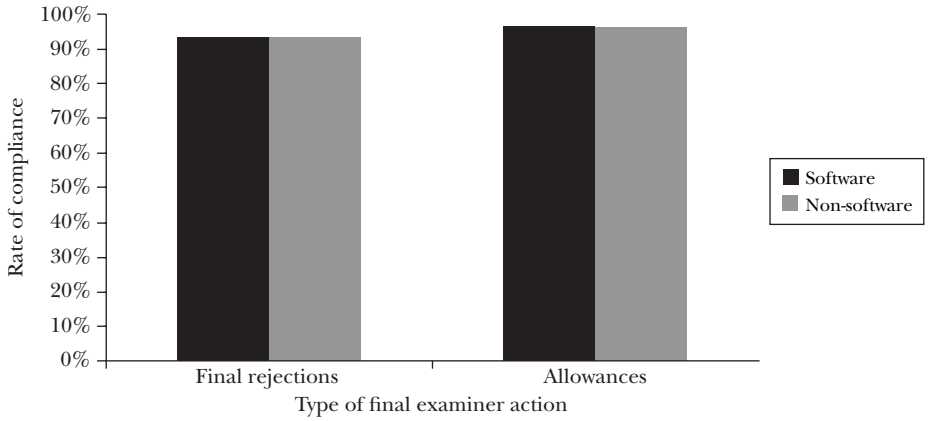
Source: Authors.

Among first final actions[8] per year from fiscal years 2003 through 2012, rejections were more likely for software applications than for non-software applications by an average difference of just 1.4 percentage points. These annual differences were significant at the 95 percent confidence interval for every year observed except 2006, 2007, and 2010. Cumulatively, rejections were more likely for software applications than for non-software applications by a difference of 2.4 percentage points, significant at the 95 percent confidence interval. The final rejection rate for software applications in 2012 is 53.1 percent, only 0.7 percentage points higher than for non-software applications, again a difference significant at the 95 percent confidence interval. Over the last decade, it appears that there has been relatively little difference in the treatment of software and non-software patent application rejections in the Patent Office.

As a comparison, we also examine the likelihood that a patent application will be allowed during the first action on the merits by the examiner during fiscal years 2003 through 2012. As illustrated in Figure 2, while first-action allowances were sometimes more, and sometimes less, likely for software patents than for non-software patents, the annual differences were small, and significant at the 95 percent

[8] After an examiner initially rejects claims as unpatentable and the applicant responds with arguments or amendments, the examiner issues a "final action:" either an allowance or rejection. This is termed a "first final action" because the applicant may seek continued examination, leading to further iterations of nonfinal and final actions subsequent to the "first" one.

Figure 3
Findings from USPTO Quality Assurance Review: Final Actions on Software and Non-software Applications, Rate of Compliance with Applicable Laws and Regulations Governing Patent Examination, FY 2007–FY2012



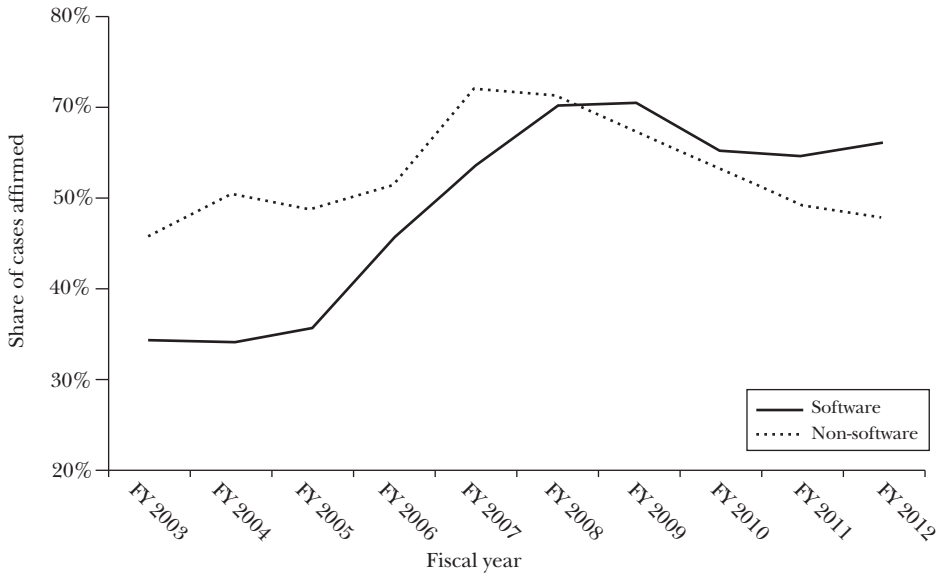
Source: Authors.

confidence interval for 2006–2008 and 2010–2012. Cumulatively during the entire period, these first-action allowances were less likely for software applications than for non-software applications by a difference of 0.5 percentage points, significant at the 95 percent confidence interval. Over the last decade, it again appears that there has been relatively little difference in the treatment of software and non-software patent application allowances in the Patent Office.

Several different explanations could account for these trends, particularly the recent decline in final rejection rates. One is that the Patent Office's focus on more compact and effective interaction between the applicant and the examiner has resulted in dispositions without the need for final rejections. Another is that guidelines, best practices, and outreach to the applicant community regarding obviousness, written description, and other examination issues have resulted in higher-quality applications being filed in the first place, a selection effect that would result in fewer final rejections.

But one explanation that the evidence does not support is that fewer final rejections reflect low-quality examination by the Patent Office. In fact, data from Patent Office internal quality assurance reviews on nearly 29,000 random examination audits over six years show that, for both software and non-software applications, the overwhelming majority of allowances and final rejections correctly apply the patent laws and examination standards. Allowances across both software and non-software applications were correctly issued over 95 percent of the time each of the last six years. Final rejections across both software and non-software applications in 2012 were correctly issued about 96 percent of the time, increasing meaningfully from 90 percent six years prior. Figure 3 shows that cumulatively, over the same

Figure 4
Affirmance of Administrative Appeals from USPTO Examiner Rejections in Software and Non-Software Applications, FY 2003–FY 2012



Source: Authors.

six years, allowances for software applications were correctly issued in 96.8 percent of cases and for non-software applications in 96.5 percent of the cases, Similarly, final rejections for software applications were correctly issued in 93.6 percent of the cases and for non-software applications in 93.5 percent of the cases. These differences in allowance and final rejection are not statistically significant, showing that software applications are being examined in the same manner as applications in all other technologies, and upon independent review, examiners are found to have correctly followed all laws and regulations in a very high percentage of the cases.

We next consider the review by the Patent Office board of patent appeals of examiner rejections during fiscal years 2003–2012, as shown in Figure 4. Data from the later years, 2008–2012, show that the board of patent appeals affirms (in whole or in part) our examiners' rejections of software applications in 57.0 percent of cases, about 2.2 percentage points higher than the rate of affirmance for denial decisions across other technologies. This share shows a narrowing of the difference between software and non-software compared with 2003–2007, when the board of patent appeals affirmed non-software rejections notably more often than software rejections, with statistically significant annual differences at the 95 percent confidence interval. In 2008, the affirmance rate for software application appeals was essentially the same as with the overall rate, and has since declined more slowly than the affirmance rate among non-software

application appeals. The affirmance rate among software appeals is currently 8.2 percentage points higher than that of other technologies, significant at the 95 percent confidence interval.

As yet another institutional check on the work being completed at the Patent Office, Congress has mandated that the decisions of the board of patent appeals can be submitted for review by the US Court of Appeals for the Federal Circuit. We also examined these federal-court appeals, and while there are relatively few instances in which the US Court of Appeals has substantively evaluated the rejection of software applications, that court has upheld such rejections in over 95 percent of the cases during 2003–2012. Cumulatively, the decisions of the US Court of Appeals on cases appealed from the Patent Office board of patent appeals have not meaningfully differed as regards the review of software and non-software applications.

These data demonstrate that it is not fair to conclude the Patent Office is "soft" on software patent applications. In fact, our investigation of rejection rates shows that Patent Office software application rejections are proper, as judged by comparison to other technology areas as well as when reviewed by our board of patent appeals. Moreover, the work of Patent Office examiners is being upheld by a wide margin in the US federal courts that review their decision making.

Our analysis thus does not provide support for the statements many have made concerning the origin of the smart phone patent wars and the work of the Patent Office. But that is not to say that the US Patent and Trademark Office believes all is perfect in the world of software patents. There are things the Patent Office should address, and is addressing, principally through the implementation of the America Invents Act of 2011, the most sweeping patent reform legislation in at least a generation.

The America Invents Act as an Intervention

The America Invents Act of 2011 was the outcome of major compromise, and thus a source of both satisfaction and disappointment to all parties. Taken as a whole, the act in both substance and implementation addresses a host of issues raised by software patent critics. Among the provisions especially applicable to software inventions are the new laws enabling individuals and firms to challenge the validity of issued patents. These "post-grant" challenge options include: post-grant review, *inter partes* (or third party) review, and "covered business method" patents review. These challenge procedures are handled by a panel of administrative judges, each of whom is highly skilled in both technology and patent-law issues. Moreover, all three options are statutorily mandated to be completed in one year, thereby offering substantial cost savings over litigation and ensuring resolution of validity disputes far faster than possible in the federal courts. This speedy resolution of controversies is particularly relevant to the software industry where product life cycles are often measured in months, not years. Furthermore, the Patent Office

regulations implementing all three options are built on a common streamlined platform to promote simplicity, speed, and cost-effectiveness, all critical to software innovators of any size who may want to contest patents.

In the new process of post-grant opposition, patents can be challenged on all grounds, including eligibility and clarity. The new "covered business method" review procedure will also be useful in the software area, since it allows a party actually sued, or threatened with suit on any existing business method patent (no matter how recently issued), to challenge its validity. Also, in interpreting the meaning of "business methods" under the new statute, the Patent Office has adopted an inclusive interpretation of that term to ensure that business methods implemented in software are eligible for review.

The *inter partes*, or third-party submission, allows any member of the public to participate by submitting documents and commentary for use by patent examiners. Because deep knowledge is commonly housed in the electronic records of software experts outside the Patent Office, this provision can help ensure patent examiners have access to the most relevant documents when examining software patent applications. Again, the Patent Office has implemented the third-party submission provision in a simple, streamlined, and open fashion, providing an Internet-enabled path for third parties to make submissions at no cost for the first three or fewer documents.

In these and other ways, the America Invents Act seeks to address many of the principal concerns surrounding software patent quality, approaching them in new and powerful ways. While the law continues to take effect, the Patent Office has been using the flexibility it has within its operational and regulatory scope to grant only valid software-related patents.

The Patent Office: Responsibilities and Responses

Among the core drivers of software patent quality, there are perhaps two overarching considerations: 1) the correspondence between the scope of the patent disclosure—the explanation of what was invented and how it works; and 2) the scope of the patent claims—the boundaries of the legal protection provided to the patentee. For the patent bargain to work, to incentivize rather than to inhibit innovation, legal protection must be commensurate with scope of disclosure. Otherwise, an inventor who describes only one way to solve a problem may obtain patent coverage for many ways, or all ways, to solve the problem. Worse yet, a patent that describes no clear problem and solution does society no good at all. Those who work at the Patent Office struggle every day to get this correspondence right, and see it as a primary responsibility.

While the disclosure-claim balance must be struck across all inventions in all fields, it has proven particularly difficult in the software area, where terminology has tended to shift and can be imprecise, and where functional language is frequently used to describe ideas that themselves are inherently functional in nature (leading

to a kind of "generalization on generalization" phenomenon). Moreover, during the 1990s while software patent filings were escalating, the courts as well as the Patent Office were primarily focused on other parts of the patentability equation, and less on the tight correspondence between disclosure and claims.

More recently, participants in the patent system have paid renewed attention to disclosure–claim correspondence. Courts have issued a series of decisions strengthening requirements, and the Patent Office has increased the time allotted to examiners for each patent application review while providing them with the training and tools to place more focus on disclosure requirements. In a further move, the Patent Office issued specific internal guidelines focusing examiners on disclosure clarity and claim–disclosure correspondence. Patent Office review of examiner actions shows an increase in the incidence of examiners raising clarity and claim–disclosure correspondence issues. More training, measurement, and refinement is underway to ensure continued improvement.

Along with the disclosure–claim correspondence, another vital component in ensuring that only appropriate software patents are issued is the strong application by examiners of the legal doctrine of "obviousness." Obviousness governs the circumstances under which a patent applicant's claim, judged against the body of relevant prior art documents predating a patent application, is merely obvious or an advance that merits patent protection. Here again, key court decisions during the last several years have significantly changed the law in a direction enabling tighter examination practices by the Patent Office. The seminal case was the Supreme Court's decision in *KSR Int'l Co. v. Teleflex, Inc.* (550 US 398 [2007]) in which the Court rejected a narrow, rigid conception of obviousness, and instead set forth a broader set of inquiries to find whether patent claims should be treated as obvious.

The *KSR* decision, along with subsequent cases in the courts, have enabled patent examiners to consider software-related claims more carefully, taking advantage of the analogous nature of so much software and the ability of skilled programmers to draw from separate algorithms in creating new solutions. And the Patent Office has taken advantage of the heightened standard by developing appropriate examination guidelines, educating examiners to use them, and ensuring usage. The goal is to produce more technical prior art available for examiners to apply, more appropriate ways to apply it, and ultimately the granting of software patents that more accurately reflect substantial innovation.

A Systemic Approach to Patent System Health

With these changes duly noted, there remains concern about an overhang of patents that were issued in the past. While some of the provisions of the America Invents Act—such as expanded post-grant review—may help, policy advocates have made other legislative and judicial proposals. Some have called upon Congress to expand the new "covered business method review" to include software, thereby giving competitors the opportunity to use evidence that has come to light in recent

years to challenge existing patents in a quick and cheap administrative hearing. Others are proposing the SHIELD Act that would adopt an English rule of cost-shifting in litigation, thus putting the cost burden of defending a suit on the loser and creating disincentives to enforce low-quality patents. Similarly, courts continue to be asked to act on issues such as enhanced scrutiny of patent claims and experimentation safe harbors, among others.

While the Patent Office has not taken an official position on these recommendations, these ongoing disputes do reflect a reality that the patent system is just that—a system. Different institutions work together to produce it. The Patent Office, constrained by available resources and laws, cannot solve all possible problems. Importantly, the Patent Office is often forced by circumstances to operate in areas of legal and technological uncertainty, like making decisions on the patentability of embryonic technologies at a point when prior art is not well developed. It routinely takes many years before the courts begin to settle legal questions, and before scientific progress resolves uncertainty about technological relationships. As history has shown, the Patent Office is routinely called upon to act before all possible bases of uncertainty are resolved.

To those who speculate on the costs of moving quickly in the face of uncertainty, economics teaches us to consider the counterfactual—we cannot know what growth and innovation would have looked like in the face of a "wait and see" approach. In this context, the biotechnology industry offers a notable example. The US moved quickly to make artificial life forms patent-eligible in the *Diamond v. Chakrabarty* decision (404 US 303 [1980]), signaling that research in emerging fields ranging from recombinant genetics to bioinformatics would be a sound investment. Other industrialized nations have spent decades trying to catch up to the growth and value that the United States created in this sector.

Conclusion

Such results reflect the ongoing balance sought by the US patent system, a balance most recently struck with the innovative reforms of the America Invents Act and the operational improvements of the Patent Office to provide more robust and transparent examination. And as the data in this article show, the recent track record at the Patent Office of examining patents containing software-related claims is an important counterweight to suggestions that the balance being struck is not appropriate. Accordingly, the smart phone patent wars, like other large-scale patent disputes in the past, may not reflect a patent system that is broken, but rather a patent system that has helped to cultivate a groundbreaking body of advances in communications technology, advances that have invited market entry by competitors. Still, just as patents are a meaningful incentive to innovate, so also is the enforcement of patents a reasonable exercise in appropriating value from innovation. That reality is at the heart of how the constitutional and legislative system of patent rights is intended to operate.

The history of the US patent system reflects a cycle of disruption—occasioned by discontinuous technological change and market adaptation in its wake—and the ensuing search for a new institutional balance. The new balance has sometimes arisen from market solutions such as cross-licensure in patent pools, or legislative solutions such as patent term reform and pre-grant publication, or judicial solutions such as revised doctrines of nonobviousness and adequate disclosure. The store of knowledge has grown, whether in textiles with the sewing machine, or in high technology with the laser, or in biotechnology with engineered bacteria. Consumers have received not merely the now-inexpensive innovations of the past, but also a reliable promise of innovation for the future. To be sure, such a commitment to the long-term benefits of innovation is a struggle against demands for access in the short term, but it is one that eventually pays for itself.

References

Arora, Ashisa, Andrea Fosfuri, and Alfonso Gambardella. 2004. *Markets for Technology: The Economics of Innovation and Corporate Strategy*. Cambridge, MA: MIT Press.

Arrow, Kenneth J. 1962. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, 609–625. Princeton, NJ: Princeton University Press.

Bloom, Nicholas, and John Van Reenen. 2002. "Patents, Real Options and Firm Performance." *Economic Journal* 112(478): C97–C116.

Branstetter, Lee, and Kamal Saggi. 2011. "Intellectual Property Rights, Foreign Direct Investment and Industrial Development." *Economic Journal* 121(555): 1161–91.

Bresnahan, Timothy F., and Manuel Trajtenberg. 1995. "General Purpose Technologies: 'Engines of Growth'?" *Journal of Econometrics* 65(1): 83–108.

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

Gans, Joshua S., David H. Hsu, and Scott Stern. 2008. "The Impact of Uncertain Intellectual Property Rights on the Market for Ideas." *Management Science* 54(5): 982–97.

Gould, David M., and William C. Gruben. 1996. "The Role of Intellectual Property Rights in Economic Growth." *Journal of Development Economics* 48(2): 323–50.

Graham, Stuart J. H., Robert P. Merges, Pamela Samuelson, and Ted M. Sichelman. 2009. "High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey." *Berkeley Technology Law Journal* 24(4): 255–327.

Graham, Stuart J. H., and David C. Mowery. 2003. "Intellectual Property Protection in the U.S. Software Industry." In *Patents in the Knowledge-Based Economy*, National Research Council, edited by W. M. Cohen and S. A. Merrill, 219–258. Washington: National Academies Press.

Hall, Bronwyn H., and Megan MacGarvie. 2010. "The Private Value of Software Patents." *Research Policy* 39(7): 994–1009.

Hall, Bronwyn H., and Manuel Trajtenberg. 2004. "Uncovering GPTS with Patent Data." NBER Working Paper 10901.

Layne-Farrar, Anne. 2005. "Defining Software Patents: A Research Field Guide." AEI-Brookings Joint Center Working Paper 05-14.

Moser, Petra. 2005. "How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs." *American Economic Review* 95(4): 1214–36.

National Research Council. 2004. *A Patent System for the 21st Century*. Washington: National Academies Press.

North, Douglass C. 1981. *Structure and Change in Economic History*. New York: Norton & Company.

US Patent and Trademark Office. 2010. *United States Patent and Trademark Office 2010-2015 Strategic Plan*. "Goal I: Optimize Patent Quality and Timeliness," p. 8–20. http://www.uspto.gov/about/stratplan/USPTO_2010-2015_Strategic_Plan.pdf.

Williamson, Oliver E. 2009. "Transaction Cost Economics: The Natural Progression." Nobel Prize Lecture, December 8. http://www.nobelprize.org/nobel_prizes/economics/laureates/2009/williamson_lecture.pdf.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 87–102

Markets for Pollution Allowances: What Are the (New) Lessons?

Lawrence H. Goulder

About 45 years ago a few economists offered the novel idea of trading pollution rights as a way of meeting environmental goals. Such trading was touted as a more cost-effective alternative to traditional forms of regulation, such as specific technology requirements or performance standards. The principal form of trading in pollution rights is a cap-and-trade system, a system whose essential elements are few and simple. First, the regulatory authority specifies the cap—the total pollution allowed by all of the facilities covered by the regulatory program. Second, the regulatory authority needs to distribute the allowances, either by auction or through free provision. Third, the system provides for trading of allowances.

The idea of cap and trade was implicit in the classic work of Ronald Coase (1960) on how well-defined property rights can assure efficient outcomes despite the presence of externalities. It then took on shape in journal contributions by Crocker (1966), Dales (1968), and Montgomery (1972). The concept materialized into policy starting in 1974, when the US Environmental Protection Agency allowed companies to trade emissions reductions among sources within the firm so long as total, combined emissions did not exceed an aggregate limit (Tietenberg 1985; Hahn and Hester 1989; Foster and Hahn 1995). The EPA's "offset" program, introduced in 1997, went further in allowing for trading across firms. These systems applied to various local pollutants, including volatile organic compounds, carbon monoxide, sulfur dioxide, and nitrogen oxides.

■ *Lawrence H. Goulder is Shuzo Nishihara Professor of Environmental and Resource Economics, Stanford University, Stanford, California. He is also a University Fellow, Resources for the Future, Washington, D.C. and a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is goulder@stanford.edu.*

http://dx.doi.org/10.1257/jep.27.1.87. doi=10.1257/jep.27.1.87

Since the 1980s the use of cap and trade has grown substantially. The three other papers in this symposium reveal and assess some of the most important applications. Schmalensee and Stavins indicate that cap and trade has been a principal part of the US Environmental Protection Agency's efforts to reduce US emissions of sulfur dioxide (SO_2) under the Clean Air Act. Newell, Pizer, and Raimi show how cap and trade applied to emissions of greenhouse gases has become an important instrument for climate change policy at the regional (state), national, and international levels.[1] And Fisher-Vanden and Olmstead describe how emissions trading is being used to control water pollution. Cap and trade was also applied to accomplish the phasedown of leaded gasoline in the United States during the 1980s. It has been employed at the municipal level as well, to control a range of pollutants including carbon monoxide, volatile organic compounds, SO_2, and nitrogen oxides (NO_x). An example is the Regional Clean Air Incentives Market (RECLAIM) in the Los Angeles airshed, a program introduced in 1994.[2]

In addition, principles similar to cap and trade have promoted cost-effective environmental protection in programs involving trading of commodities other than pollution. At least 10 nations have implemented programs of individual transferrable fishing rights, in which a limited supply of permits to catch fish is allocated among fishing operators. And some US states have instituted programs involving tradable land-development rights as a way of conserving natural habitats and protecting biodiversity.[3]

The provision for trading of allowances is the key to achieving desired emission reductions at a lower cost than with other, less-flexible, approaches. The separate sources of pollution will tend to have a range of different marginal costs for abating pollution. Facilities with the highest costs of reducing emissions will find it advantageous to reduce their costs by buying additional allowances from other facilities rather than trying to meet the pollution limits given by their original holdings of allowances. Likewise, the facilities for which it is relatively inexpensive to reduce emissions will find it profitable to sell some of their allowances. Even though this obliges them to reduce emissions even more, the returns from the sale of allowances will exceed the additional abatement (pollution-reduction) costs.

[1] Regional programs include the carbon dioxide emissions trading ("cap-and-trade") program in the US Northeast under the nine-state Regional Greenhouse Gas Initiative, which went into effect in 2008. A cap-and-trade program is slotted to go into effect in California in January 2013. National programs include carbon emissions cap-and-trade systems in Australia and New Zealand, and the European Union's 27-country cap-and-trade program. International trading in greenhouse gas emissions is allowed for under the Kyoto Protocol, the international treaty to reduce greenhouse gases.

[2] Cap and trade is not the only form of pollution trading, although it is the one that has gained most attention and been implemented the most. Another trading approach allows firms to receive credits for reducing emissions below some stipulated level, even though they are not penalized if their emissions exceed that level. Here the regulator offers a one-sided option, and there is no cap on aggregate pollution from the covered facilities. This approach has been considered for bringing about greater participation by developing countries in efforts to reduce greenhouse gases (Millard-Ball forthcoming).

[3] For an analysis of a range of issues associated with individual transferable fishing rights and tradable habitats, see Arnason (2012) and Crocker (2005), respectively.

Thus, trading leads to more abatement by those facilities that can reduce emissions most cheaply. It tends to bring marginal abatement costs toward equality, a condition for cost minimization. Regulators do not need to know the marginal abatement costs of individual facilities: they can let the market promote equality in marginal abatement costs. This is a potential advantage over technology requirements or performance standards because regulators generally will not have sufficient information to set the requirements or standards at levels that assure equal marginal abatement costs across the covered entities.

In this overview article, I consider some key lessons about when cap-and-trade programs work well, when they perform less effectively, how they work compared with other policy options, and how they might need to be modified to address issues that had not been anticipated.

I distinguish two types of lessons. The first are, essentially, confirmations of prior theoretical predictions. The second are insights that emerge in response to previously unanticipated circumstances or problems, or as a result of recent analytical contributions. I consider each type of lesson in turn.

Some (Mostly) Reassuring Outcomes

1) In national and subnational cap-and-trade programs applied to local air pollutants, effective monitoring and compliance have enabled cap-and-trade programs to succeed in limiting emissions to specified targets. Difficulties of monitoring have limited the use of cap-and-trade programs aimed at water pollution, and problems of compliance have hampered the effectiveness of cap-and-trade programs under the international Kyoto Protocol.

For the early proponents of cap and trade, one of the touted attractions was that this regulatory approach would establish and maintain clear limits on total emissions of pollution by the covered sectors, with the limit in each period given by the specified cap (or total number of allowances in circulation). The ability to specify an aggregate limit on emissions distinguishes cap and trade from other regulatory approaches: neither limits on the emissions at the firm- or plant-level, nor mandates for the use of certain technologies for pollution abatement, nor sector- or economy-wide pollution taxes specify a total quantity of emissions.

Imposing a limit on total emissions and letting the market determine the price is not necessarily more efficient than imposing a price on emissions and letting the market determine the quantity—as under a pollution tax. Weitzman's (1974) seminal article indicates that the relative advantage of setting the quantity or setting the price depends on the nature of uncertainty about marginal benefits and costs from pollution reductions. But allowing the regulator to choose the quantity of pollution explicitly has considerable practical political appeal.

The promise of keeping aggregate pollution within the stipulated overall cap has been fulfilled in most of the cap-and-trade systems introduced for air pollution control. For example, the US Environmental Protection Agency's programs to reduce sulfur dioxide and nitrogen oxides under the Clean Air Act and the

RECLAIM program for curbing these same pollutants in the Los Angeles region can claim success in reducing emissions to the targeted levels. In addition, the European Union's Emissions Trading Scheme has largely managed to keep greenhouse gas emissions from covered sectors within the levels targeted (although this program is likely to have stimulated a partially offsetting increase in emissions outside of the European Union, a "leakage" phenomenon I discuss below).

Two factors have contributed to these successes. First, emissions of the air pollutants involved have proved relatively easy to monitor, or at least to estimate with some accuracy. In addition, the programs have included strong incentives for compliance. For example, under Europe's Emissions Trading System, the noncompliance penalty is 100 euros per ton, considerably higher than the market price of allowances, which has seldom exceeded 15 euros, and compliance in fact appears to have been very good in all of these programs.[4]

In contrast, under the Kyoto Protocol of 1997, serious problems of compliance have arisen and remain. This largely reflects the lack of significant enforcement capabilities under the Protocol. This is a problem common to many international agreements, rather than any inherent weakness of cap and trade. Under the Protocol, 37 nations committed themselves to maximum levels of emissions of greenhouse gases in the first commitment period, 2008–2012. Parties that did not meet their targets in the first commitment period were required to make up the difference plus 30 percent more in the anticipated second commitment period. However, several parties that are expected to miss their initial targets—including Japan, Canada, and Russia—have simply announced they will not continue to abide by the Protocol in the second commitment period.

In the context of water pollution, the accomplishments are somewhat limited. Cap and trade has enjoyed success in restricting the effluent pollution from regulated point sources. Currently, there are about 13 trading programs, with most of them arising since the turn of the century. As pointed out by Fisher-Vanden and Olmstead, trading of water pollution permits generally has embraced only those sources that are easy to monitor—namely large industrial establishments and municipal sewage treatment plants. The agriculture sector is an important contributor to water pollution, but in general this sector is not covered by enforceable effluent regulations under the Clean Water Act. This reflects the difficulty of monitoring the effluent from these so-called nonpoint sources. It is worth noting that any sort of pollution control, whether via market-based approaches or by way of more conventional approaches, is challenging with nonpoint sources. The absence of cap and trade applied to water pollution from agriculture also reflects the considerable political opposition by the agriculture industry to limits on pollution.[5]

[4] The qualifier "appears" is used because the successful cheaters, by definition, are not observed.

[5] Fisher-Vanden and Olmstead (this issue) point out another important challenge to the application of cap and trade to water pollution: water pollutants often are not uniformly mixed. As discussed by these authors, a simple cap-and-trade system, where given releases of effluent are all traded at the same price, can produce undesirable environmental outcomes.

2) Cap-and-trade programs have brought significant cost reductions relative to conventional regulatory approaches.

The evidence for cost savings from a cap-and-trade policy must always be indirect since researchers never observe the counterfactual world in which an alternative program is introduced under otherwise identical economic and environmental conditions. Moreover, there are not enough instances of cap and trade and other regulatory approaches in roughly similar settings to allow the impact of cap and trade to be identified econometrically.

Still, economists have managed to arrive at plausible estimates of cost savings by estimating the marginal abatement cost curves of the covered facilities, assessing the extent to which marginal abatement costs would differ across facilities under conventional regulation (often the previously prevailing form of regulation), and then calculating the extent to which these differences are eliminated (and total abatement costs reduced) by a cap-and-trade program. The analyses generally rely on the assumption that the market for trading allowances is effective in bringing marginal abatement costs to equality across facilities. Behind this assumption is the implicit assumption that transactions costs are low.

A review by Chan, Stavins, Stowe, and Sweeney (2012) of various analyses using this approach indicates that sulfur dioxide allowance trading under the Clean Air Act yielded cost savings in the range of 15 to 90 percent relative to the costs under conventional forms of regulation. There is some evidence that transactions costs are fairly low (Stavins 1995) and the trading market is fairly fluid, which would support these findings.

Using a similar approach, an analysis of sulfur dioxide and nitrogen oxides trading in the Los Angeles area RECLAIM market claimed cost savings of 46 percent relative to the costs of achieving the same aggregate reductions under the prior air quality management program, which involved fixed emissions caps and no trades. The estimates for recent savings may be overestimated, however, as various restrictions on trades have been introduced since the analysis was performed. In addition, some analyses suggest that the efficiency of the trading equilibrium was compromised as a result of interactions between cap-and-trade systems and rate-of-return regulation faced by utilities, an issue to which I return below. Ellerman, Convery, and de Perthius (2010) estimate that Europe's Emissions Trading System achieved cost reductions in the range of 2–5 percent. For other pollution trading markets, the quantitative evidence for cost savings is limited. However, even in these other markets the qualitative conclusion that cap and trade has lowered costs is tacitly supported by the mere existence of trading, as trading shifts responsibility for pollution reduction to facilities that can do so relatively cheaply.

Overall, these considerations suggest some success for many of the cap-and-trade systems that have been introduced. But some important qualifications are in order. To a large extent, these empirical studies show the cost savings compared to a relatively inflexible form of conventional regulation—fixed emissions caps. They show the savings from trading relative to the same regulation without trading. They do not assess cost-savings relative to other, more flexible, nonmarket instruments (such

as performance standards) or relative to an alternative market-based instrument: namely, a pollution tax. In addition, the initial assessments of cost savings ignore factors whose importance has only recently come to light. I address these issues below.

Surprises, Challenges, and New Lessons

3) The environmental effectiveness and cost-effectiveness of cap and trade can be significantly compromised by interactions with other regulations.

Virtually all analyses of environmental policies have ignored interactions with other policies. This is particularly important in the case of cap and trade. Economic theory as well as recent experience shows that these interactions can significantly reduce both environmental effectiveness and cost-effectiveness.

One difficulty arises when regulations in one jurisdiction are "nested" within a cap-and-trade system introduced in a higher-level jurisdiction. Suppose, for example, a cap-and-trade system was introduced at the national level in the United States with a national emissions cap. Now suppose that a given state desires further emissions reductions by firms within its boundaries, beyond those that would result from the federal program: through cap and trade or some other instrument, the state prompts further reductions by facilities within its borders. As a result of this state's action, firms within this state will now have excess federal allowances, which they will sell to firms in other states that do not have tougher standards. Since nationwide emissions continue to be determined by the unchanged national cap, the one state's imposition of tougher environmental rules leads to no overall reduction for the nation: it just causes "emissions leakage"—offsetting increases in emissions elsewhere. By affecting the distribution of emissions, these adjustments can raise or lower aggregate environmental damage, depending on how they alter the geographical pattern of pollution concentrations. The national cap effectively prevents lower-level jurisdictions from eliciting further emissions reductions.[6]

The issue came to life when the United Kingdom recently decided to impose a tax on carbon dioxide emissions by electric power generators in the country. For each unit of emissions, these generators will need to pay this tax in addition to the price that they pay for emissions allowances from the EU Emissions Trading System (ETS). Although the tax will likely cause greater abatement by generators within the United Kingdom, it will not cause greater overall abatement in Europe, since overall European abatement is determined by the Europe-wide cap under the ETS. The UK initiative will reduce the UK's demands for emissions allowances from the ETS,

[6] For further discussion of these issues, see Fankhauser, Hepburn, and Park (2010), Burtraw and Shobe (forthcoming), and Goulder and Stavins (2012). The same issue can arise within a single jurisdiction. For example, California introduced a cap-and-trade system as part of its Global Warming Solutions Act. To the extent that other regulations such as a standard for low-carbon fuel aim to achieve further reductions, the affected firms will have excess allowances, and these allowances will be sold to other covered entities. Statewide emissions from the covered sectors will not be reduced further, as they are determined by the state's cap. For discussion of other interactions within a single jurisdiction, see Levinsohn (2012).

putting downward pressure on allowance prices and prompting increased emissions in the rest of Europe. CDC Climate Research (2011) offers a quantitative assessment of the impacts.

The issue also arose when 14 US states attempted to impose tighter limits on greenhouse gases per mile from automobiles below the level implied by existing federal Corporate Average Fuel Economy standards. The 14-state initiative would have caused automobile manufacturers in those states to more than meet the federal corporate auto fuel economy (CAFE) standards, allowing them to sell less-fuel-efficient cars in other states and still remain within the national standard. In Goulder, Jacobsen, and von Benthem (2012), my coauthors and I estimate that about 75 percent of reduction in greenhouse gases achieved in the 14 states would have been offset by increased emissions in other states. As it turned out, the 14-state initiative helped put pressure on automobile manufacturers to accept tighter requirements at the federal level in exchange for elimination of the tougher action by these states.

These difficulties are relevant to recent US initiatives to institute a federal-level tradable clean electricity standard, since some states may wish to impose standards tougher than the federal one.

A second problem arises when firms within the cap-and-trade system are subject to other *non*environmental regulations that affect demands for allowances and the distribution of emissions-abatement effort across firms. This issue arose in the South Coast Air Quality Management District's RECLAIM program to reduce emissions of sulfur dioxide and nitrogen oxides in the Los Angeles area. Electric power generators were important contributors to these emissions: however, these generators were also subject to rate-of-return regulation under the local public utilities commission. As shown by Kolstad and Wolak (2003), these vertically integrated firms *benefited* from higher allowance prices, because the higher prices could be incorporated in the rate base determining the prices that could be charged to consumers. The higher rate base implied higher prices for electricity, which yielded increments to profits despite the higher prices of allowances. These interactions implied a shift in the distribution of wealth from ratepayers to owners of utilities. They also implied a shift in ownership of allowances and abatement effort toward utilities and away from other emitters. This shift compromised cost-effectiveness, as some low-cost abatement by entities other than utilities was crowded out.

The Clean Air Act's sulfur dioxide allowance trading market offers yet another case where the cap-and-trade system was vulnerable to other regulations, as detailed in the accompanying article by Schmalensee and Stavins. In this case, the other regulation was the Clean Air Interstate Rule, which was promulgated in 2005, well after the cap-and-trade program's implementation in 1990. This rule imposed stringent emissions-reduction requirements that eventually led to significant reductions in the demand for sulfur dioxide allowances in the trading market. As a result, the cap in the sulfur dioxide trading program became no longer binding, and allowances prices subsequently have collapsed. Although the Clean Air Interstate Rule accomplished significant reductions (which many might

applaud), the neutering of the cap-and-trade program suggests that the reductions were not accomplished as cost-effectively as would have been the case if instead the reductions had been achieved by a tightening of the cap (which would have required Congressional action).

Schmalensee and Stavins (this issue), along with Burtraw (forthcoming), claim that a key lesson from this episode is the importance of building flexibility into cap-and-trade systems. The absence of institutional rules permitting adjustments of the cap in the face of new information contributed to the need to invoke different, potentially less-efficient, regulations. Making it easier to adjust the cap might have some drawbacks, however. Greater flexibility could adversely affect the credibility of the government's commitment to a given time profile for the emissions cap and introduce new uncertainties into the system.

In sum, interactions with other regulations can compromise cap-and-trade's environmental effectiveness, distort the demands for allowances, or make a cap-and-trade program irrelevant. Ignoring regulatory interactions can be imprudent, just as a doctor in prescribing a medication without knowing what other medications the patient is taking would be reckless.

4) Volatility of allowance prices has been a significant concern.

Under cap and trade, the supply of allowances is highly inelastic in the short term, changing only as a result of government policy decisions (that one hopes are predictable). With highly inelastic supply, shifts in demand can cause significant price changes, and irregular shifts in demand can produce price volatility.

Some existing cap-and-trade systems have displayed considerable allowance price volatility. The energy supply crisis in California in the summer of 2000 gave power companies incentives to bring online some older power generators in the Los Angeles region. This led to a significant increase in the demand for emissions allowances for nitrogen oxides under the RECLAIM program, since allowances were needed to validate the emissions produced by these generators. As a consequence, NO_x allowance prices rose from about $400 per ton to an average in the year 2000 of over $40,000 per ton—with the average allowance price reaching $70,000 in the peak month of 2000 (Ellerman, Joskow, and Harrison 2003).

There was also significant price volatility in the first (that is, the pilot) phase of cap and trade under the European Union's Emissions Trading Scheme. About a year after its implementation, emissions allowance prices dropped dramatically with the release of information that indicated that the Phase I permit allocations were generous in the sense that they barely constrained the covered sources. The December 2008 futures prices fell from 32.25 euros to 17.80 euros between April 19 and May 12, 2006. There was even greater volatility for the Phase I permit prices contained in December 2007 contracts. These prices dropped from 31.65 euros on April 19, 2006, to 11.95 euros on May 3, 2006. When Phase II of the program began in 2008, allowance prices rose to more than 20 euros in the first half of 2008 and averaged 22 euros in the second half of 2008. In the first half of 2009, they fell to 13 euros. Since then, allowance prices have remained below 13 euros.

Is price volatility a problem? Critics of cap and trade point out that it is hard for producers to make sound investment decisions when the prices of allowances (and associated costs of production) fluctuate and are subject to uncertainty. Others claim that unstable allowance prices can produce macroeconomic disruptions. On the other hand, the ups and downs of allowance prices can play a beneficial countercyclical role. During economic downturns, the demand for allowances will fall, putting downward pressure on allowance prices. Lower allowance prices soften the impact of the pollution regulation on firms during the difficult economic times.

Reflecting the idea that significant swings in allowance prices should be avoided, policymakers have come up with ways to limit price volatility. One is to incorporate within the trading system an allowance price floor, price ceiling, or both. To impose a ceiling, the regulator can make available for sale additional allowances once the price reaches a given level. This prevents allowance prices from rising further. To enforce a price floor, the regulator buys allowances (and removes them from circulation) whenever the floor price is reached, thereby preventing prices from falling further.

The presence of a price ceiling implies that once the ceiling is reached, overall emissions no longer are constrained to the level of the original cap, because new allowances are being introduced to maintain the ceiling price. Thus, certainty about the total level of emissions is sacrificed for the sake of reduced uncertainty about allowance prices. Some interested parties have questioned whether this swap is worthwhile.

Another way to reduce potential price volatility is to allow for intertemporal banking and borrowing of allowances. With intertemporal borrowing, firms can credit toward present emissions the allowances allocated to them for future time periods. With intertemporal banking, firms can apply to future periods the allowances they do not use in the current period. Such intertemporal flexibility makes the current supply of allowances more elastic in any given period, which helps dampen price volatility. Of the major tradable allowance systems tried in the United States, RECLAIM offered the fewest opportunities for banking allowances. Stavins (2007) and Ellerman and Joskow (2008) suggest that much of the allowance price volatility experienced by RECLAIM was due to the absence of provisions for banking. Similarly, volatility in allowance prices for Phase I of Europe's Emissions Trading system has been attributed in part to the fact that the program prevented banking of allowances from the first phase to the second (Market Advisory Committee 2007; Schmalensee and Stavins, this issue).

In contrast, unlimited banking in the US Sulfur Dioxide Allowance Trading Program is generally viewed to have been a successful design feature of that program, as it mitigated issues of price volatility and led firms to reduce emissions faster than they would have without banking (Ellerman, Joskow, and Harrison 2003). Banking is also considered responsible for a large share of the gains from trade under the program.

That said, allowing intertemporal banking is not a panacea. Nordhaus (2007) finds that sulfur dioxide allowance prices between 1995 and 2006 were about as volatile as oil prices, and that they were much more volatile than prices of stocks, other

assets such as houses, and most consumer goods. Sulfur dioxide allowance prices were particularly volatile in the late 2000s, as a series of court and regulatory decisions changed expectations about the future stringency of the cap (Schmalensee and Stavins, this issue; Palmer and Evans 2009; Bravender 2009).

5) Because of interactions with the fiscal system, certain decisions about the design of a cap-and-trade system—namely, the choice between auctioning and freely allocating allowances, and the way that any auction revenues are returned to the economy—significantly affect policy costs. Indeed, these decisions can determine whether a cap-and-trade program is more cost effective than some more conventional pollution control approaches.

The early assessments of cap and trade tended to be partial equilibrium in nature. Since the early 1990s, however, several studies have examined cap and trade (and other environmental policies) in a general equilibrium framework. These studies reveal that general equilibrium connections between cap and trade and the fiscal system have a first-order impact on the costs of cap and trade.

One of the key findings concerns the method of introducing emissions allowances into circulation. The regulating authority can give out all allowances free, auction them all out, or use a combination of free allocation and auctioning. A time-honored notion in economics is that while this choice affects the distribution of wealth, it does not affect cost-effectiveness because no matter how the allowances are initially distributed, the process of trading will assure that reductions in emissions happen in a cost-effective manner.

In a general equilibrium framework that accounts for interactions with the fiscal system, this logic no longer holds. By yielding government revenue, auctioning has the potential to reduce the government's reliance on distortionary taxes—such as income, sales, and payroll taxes—to finance its expenditures. The implied reductions (or avoided increases) in distortionary taxes can confer a benefit in terms of economic efficiency. In contrast, when allowances are given out free, the government does not receive these revenues, and society does not enjoy this potential benefit. The word "potential" is important here: if the revenues are recycled in ways that do not reduce marginal rates of prior taxes or that do not avoid increases in marginal rates of these taxes, this benefit is not realized.[7]

The potential benefits are substantial. Parry and Williams (2010) provide general formulas suggesting that auctioning can reduce the costs of meeting a given target for emissions reductions by almost half compared to a program with free permits. In a model focusing on the US economy in Goulder, Hafstead, and Dworsky (2010), we find that the costs of achieving a 42 percent reduction in carbon dioxide emissions under cap and trade are about 33 percent lower under 100 percent auctioning with

[7] While the choice between auctioning and free allocation has implications for cost-effectiveness, the choice about how to distribute the allowances *within a program involving free allocation* does not influence the cost-effectiveness of that program. This property was implicit in Coase (1960) and was first emphasized by Montgomery (1972).

recycling of revenues in the form of cuts in distortionary taxes as compared with 100 percent free allocation.

Historically, cap-and-trade policy has relied principally on free allocation. This is changing, however, especially for cap-and-trade programs aiming to cap greenhouse gas emissions. The European Union's Emissions Trading Scheme, the Regional Greenhouse Gas Initiative in the northeastern United States, and the State of California's new climate change policy all are moving toward auctioning more than half of their allowances. This change offers the potential for very large benefits in terms of economic efficiency, although the political motivation for these changes appears to have been a concern about distributional implications—the view that continued reliance on free allocation would generate windfalls to the recipient firms—as well as interest in obtaining funds to support various environmental programs. Economic analysis indicates that the concern about potential windfalls has merit. Studies of nitrogen oxide allowance trading under the US Clean Air Act (Bovenberg, Goulder, and Gurney 2005) and of potential carbon dioxide allowance trading in the United States (Bovenberg and Goulder 2001; Smith, Ross, and Montgomery 2002), suggest that the rents from 100 percent free allocation would substantially overcompensate firms for the costs they would otherwise face under these programs. In fact, these studies show that a fairly small share of the allowances—generally less than 30 percent—needs to be freely allocated to provide sufficient rents to prevent an overall decline in firm equity values.

In fact, the decision about whether to auction or freely allocate emissions allowances can determine whether a cap-and-trade program is more cost effective than certain more conventional regulatory alternatives. As we show in Parry, Williams, and Goulder (1999), to the extent that the cost of environmental policies are shifted forward to consumers (in the form of higher prices paid for pollution-intensive goods and services), the consumer price level will rise, implying a reduction in real factor returns. This depresses factor supply, and the resulting efficiency loss in factor markets (termed the "tax-interaction effect") raises the costs of environmental policies. In Goulder, Parry, Williams, and Burtraw (1999), we show that the tax-interaction effect is larger under emissions-pricing policies like cap and trade than for performance standards or technology mandates, which do not raise consumer prices as much. This potential disadvantage of cap and trade is overcome when cap and trade involves an auction and auction revenues are used to finance cuts in pre-existing distortionary taxes. In that case, cap and trade is more cost effective than these alternatives. But cap and trade can be more costly than the alternatives when allowances are given out free or when auction revenues are not used to finance cuts in prior tax rates.

Thus, the method of introducing allowances and the way that any revenues from the system are recycled importantly influence the cost-effectiveness of a cap-and-trade system. It can determine whether cap and trade is more or less cost effective than more conventional policy instruments. For cost-effectiveness, the design of a cap-and-trade system is of first-order importance.

These considerations do not contradict the idea that cap and trade generally has lowered the costs of pollution control. This is because cap and trade often

has substituted for some of the more costly methods of control, such as fixed facility-level caps on emissions. But these broader concerns show that cap and trade needs to be carefully designed to assure lower costs than other regulatory alternatives. Auctioning and judicious revenue-recycling are needed to assure greater cost-effectiveness than some of the relatively flexible alternatives such as performance standards.

6) Should cap and trade displace other approaches?

Cap and trade cannot achieve all the efficiency-related goals of environmental policy. If the concern is economic efficiency, then in many settings it should complement, rather than substitute for, other instruments for environmental protection. The reason is that cap and trade cannot address all of the market failures responsible for pollution that is excessive from an efficiency point of view. And the same point applies to a pollution tax. As a form of emissions pricing, cap and trade addresses the market failure stemming from the emissions-related externality: it establishes a price for the otherwise external costs associated with pollution. But several other important market failures are not confronted by cap and trade (or by a pollution tax).

For example, an "innovation market failure" is associated with the spillover knowledge and the associated external benefits resulting from knowledge-generating activities. Additional measures—for example, a subsidy to research and development—are called for to confront this market failure directly. In its early history, some analysts touted cap and trade as the preferred instrument not only for encouraging conservation by consumers and substitution to cleaner known production processes by firms, but also for stimulating technological change—in particular, the invention of cleaner technologies. By raising the relative price of pollution-intensive production methods, cap and trade can provide incentives for innovation.[8] But efficiency calls for supplementing cap and trade with another instrument that directly addresses the innovation market failure. It is a common principle of policy analysis that multiple market failures generally call for multiple policy instruments.[9] Cap and trade is an excellent instrument for dealing with the externality associated with emissions, yet it should not displace other approaches that address other market failures.

But is cap and trade the best instrument for confronting the emissions externality? The main alternative is a pollution tax. A number of authors have analyzed the relative strengths and limitations of the cap-and-trade and pollution-tax options (for example, Metcalf 2007; Stavins 2007; Metcalf and Weisbach 2009; Goulder and Schein 2012). Although numerous issues are involved, perhaps the first point to emphasize is that both approaches offer similar advantages relative to conventional approaches for curbing emissions. Both approaches effectively impose, at the margin, a price

[8] However, as pointed out by Gans (2012), in general equilibrium, cap and trade (or, more generally, emissions pricing) can sometimes reduce incentives to innovate.

[9] For quantitative assessments of the significance of this principle in the context of environmental regulation, see Goulder and Schneider (1999), Fischer and Newell (2008), and Acemoglu, Aghion, Bursztyn, and Hemous (2011).

for each unit of emissions. This is the case for cap and trade even when allowances are initially given out free to the covered entities. After all, even when allowances are received for free, each additional unit of emissions carries an opportunity cost: one more unit of pollution either reduces the number of allowances the firm can sell, or it raises the number of allowances the firm will need to buy to remain in compliance. By establishing one price for pollution that facilities must face, both approaches encourage equality of abatement costs at the margin across facilities, which works toward cost-effectiveness.

Moreover, there is no inherent difference between the two approaches in terms of the distributional impacts on facilities. Under a cap-and-trade system, free allocation of allowances can cushion the impact of the regulation on covered firms, shifting the burden onto the general public (since more free allocation implies less revenue collected by the auction). Under a pollution tax, offering inframarginal exemptions to the tax yields the same opportunities for altering the distribution of impacts.

The two approaches do differ in some important ways, however. A pollution tax avoids the problem of emissions price volatility. On the other hand, the pollution tax does not impose a predetermined cap on aggregate emissions; some would regard this as a disadvantage.

It has often been suggested that a cap-and-trade system would be more costly to administer than a pollution tax. One claim is that administrative costs are higher because a cap-and-trade program would involve more entities whose emissions must be tracked. This claim is incorrect. The number of covered entities depends on where the cap-and-trade system or pollution tax is imposed—upstream, midstream, or downstream—and both approaches can be introduced at any of these levels. Still, recent experience suggests that a cap-and-trade system might involve somewhat greater administrative challenges for two reasons: 1) there are costs of setting up a market for auctioning and trading allowances (which may be higher than the costs of incorporating a pollution tax within the existing tax-collection institutions), and 2) under a cap-and-trade system, the regulator must not only keep track of the emissions of covered facilities, but also establish a registry to record changes in ownership of allowances as a result of allowance purchases or sales.

At the same time, current policy conditions and political economy considerations might favor cap and trade, at least in the climate policy context. Given the existence of other cap-and-trade systems overseas, it might be easier to achieve international harmonization through a US cap-and-trade program than with a US carbon tax (Jaffe, Ranson, and Stavins 2010; Metcalf and Weisbach 2009). Cap and trade has been an easier political sell than a pollution tax, partly because cap and trade is less costly to the covered firms than a pollution tax would be.[10] It is also partly because the public, often averse to any new tax, has tended to view a cap-and-trade program as something very different from a tax measure. However, this political advantage

[10] This statement assumes that the pollution tax policy does not include inframarginal exemptions. Such exemptions would function much like free allowances under a cap-and-trade system, lowering the costs to the covered firms.

seems to be waning, at least in the United States, where opponents of cap-and-trade policies for limiting carbon emissions have started to refer to them as "cap and tax" policies (for example, "The Cap and Tax Fiction," in the *Wall Street Journal* 2009).

The bottom line is that neither a pollution tax nor a cap-and-trade approach clearly dominates. The degree of efficiency in reducing emissions seems to depend more on the extent of emissions pricing (under either form) and on the particular design of the emissions-pricing instrument (for example, the degree to which a cap-and-trade program relies on auctioning of allowances).

Conclusions

Trading rights to pollute—which was just an idea in the minds of a few economists 45 years ago—has now taken form in many locales and for many types of pollution. This novel approach has largely lived up to its basic promises: that is, in most places where it has been tried, it has succeeded in bringing down pollution to the targeted levels and has achieved those emissions reductions at lower cost than would have been possible under many of the more conventional forms of regulation. At national and subnational levels, the environmental targets have largely been met under cap-and-trade systems for local pollutants including sulfur dioxide and nitrogen oxide compounds, as well as for carbon dioxide, the principal greenhouse gas.

Important challenges remain, however. The application of cap and trade for control of water pollution has been limited by difficulties of tracking the nonpoint sources, particularly the water pollution generated by the agricultural sector. The international-level use of cap and trade to limit greenhouse gas emissions has been limited by difficulties in enforcement.

We have reached a much deeper understanding of the potential environmental and economic impacts of cap and trade. Research reveals how the simple textbook version of cap-and-trade system can be modified to address potential difficulties such as the problem of price volatility. It also makes clear how the impacts of cap and trade depend on interactions with other regulations and with the existing tax system. These interactions are of first-order importance: they influence whether cap and trade manages to reduce pollution, and they indicate that the particular design of a cap-and-trade system makes a substantial difference to its cost. Indeed, the design can determine whether the program yields efficiency gains.

Cap and trade has some advantages and some drawbacks relative to the chief alternative form of emissions pricing—a pollution tax. Neither approach dominates the other. When well designed, either form of emissions pricing will offer several advantages over conventional forms of regulation. Yet neither cap and trade nor a pollution tax is a cure-all for environmental problems: emissions pricing does not eliminate the need to engage other environmental policy instruments to address environment-related market failures other than the one stemming from the emissions externality.

■ *I am very grateful to Dallas Burtraw, Steven Cliff, Sheila Olmstead, Christian de Perthuis, Robert Stavins, Roberton Williams III, and the editors of this journal for helpful comments and suggestions, and to Santiago Saavedra Pineda for excellent research assistance.*

References

Acemoglu, Daron, Philippe Aghion, Leonardo Bursztyn, and David Hemous. 2011. "The Environment and Directed Technical Change." *American Economic Review* 102(1): 131–66.

Arnason, Ragnar. 2012. "Property Rights in Fisheries: How Much Can Individual Transferable Quotas Accomplish?" *Review of Environmental Economics and Policy* 6(2): 217–36.

Bovenberg, A. Lans, and Lawrence H. Goulder. 2001. "Neutralizing the Adverse Industry Impacts of CO_2 Abatement Policies: What Does It Cost?" In *Behavioral and Distributional Effects of Environmental Policy*, edited by C. Carraro and G. Metcalf, 45–85. University of Chicago Press.

Bovenberg, A. Lans, Lawrence H. Goulder, and Derek J. Gurney. 2005. "Efficiency Costs of Meeting Industry-Distributional Constraints under Environmental Permits and Taxes." *RAND Journal of Economics* 36(4): 950–70.

Bravender, Robin. 2009. "Acid Rain Credits Nosedive on CAIR Concerns." *Greenwire*, March 27. http://www.eenews.net/public/Greenwire/2009/03/27/4.

Burtraw, Dallas. Forthcoming. "The Institutional Blind Spot in Environmental Economics." *Daedalus*.

Burtraw, Dallas, and William Shobe. Forthcoming. "Rethinking Environmental Federalism in a Warming World." *Climate Change Economics*.

CDC Climate Research. 2011. "Carbon Price Flaw? The Impact of the UK's CO_2 Price Support on the EU ETS." Climate Brief No. 6, June.

Chan, Gabriel, Robert N. Stavins, Robert Stowe, and Richard Sweeney. 2012. "The SO_2 Allowance Trading System and the Clean Air Act Amendments of 1990: Reflections on 20 Years of Policy Innovation." *National Tax Journal* 65(2): 419–52.

Coase, Ronald H. 1960. "The Problem of Social Cost." *Journal of Law and Economics* 3(October): 31–44.

Crocker, Thomas D. 1966. "The Structure of Atmospheric Pollution Control Systems." In *The Economics of Air Pollution*, edited by H. Wolozin, 61–86. New York: W. W. Norton.

Crocker, Thomas D. 2005. "Markets for Conserving Biodiversity Habitat: Principles and Practice." Chap. 7 in *Species at Risk: Using Economic Incentives to Shelter Endangered Species on Private Lands*, edited by Jason F. Shogren. Austin: University of Texas Press.

Dales, John H. 1968. *Pollution, Property and Prices: An Essay in Policy-making and Economics*. University of Toronto Press.

Ellerman, A. Denny., Frank J. Convery, and Christian de Perthuis. 2010. P*ricing Carbon: The European Union Emissions Trading Scheme*, especially chap. 6, "Emissions Abatement." Cambridge University Press.

Ellerman, A. Denny, and Paul L. Joskow. 2008. "The European Union's Emissions Trading System in Perspective." Pew Center on Global Climate Change.

Ellerman, A. Denny, Paul L. Joskow, and David L. Harrison, Jr. 2003. *Emissions Trading: Experience, Lessons, and Considerations for Greenhouse Gases*. Washington, DC: Pew Center on Global Climate Change.

Fankhauser, Samuel, Cameron Hepburn, and Jisung Park. 2010. "Combining Multiple Climate Policy Instruments: How Not to Do It." *Climate Change Economics* 1(3): 1–17.

Fischer, Carolyn, and Richard G. Newell. 2008. "Environmental and Technology Policies for Climate Mitigation." *Journal of Environmental Economics and Management* 55(2): 142–62.

Foster, Vivien, and Robert W. Hahn. 1995. "Designing More Efficient Markets: Lessons from Los Angeles Smog Control." *Journal of Law and Economics* 38(1): 19–48.

Gans, Joshua. 2012. "Innovation and Climate Change Policy." Unpublished Paper.

Goulder, Lawrence H., Marc A. C. Hafstead, and Michael Dworsky. 2010. "Impacts of Alternative Emissions Allowance Allocation Methods under a Federal Cap-and-Trade Program." *Journal of Environmental Economics and Management* 60(3): 161–81.

Goulder, Lawrence H., Marc R. Jacobsen,

and Arthur van Benthem. 2012. "Unintended Consequences from Nested State and Federal Environmental Regulation: The Case of the Pavley Greenhouse-Gas-Per-Mile Limits." *Journal of Environmental Economics and Management* 63(2): 187–207.

Goulder, Lawrence H., Ian W. H. Parry, Roberton C. Williams III, and Dallas Burtraw. 1999. "The Cost-Effectiveness of Alternative Instruments for Environmental Protection in a Second-Best Setting." *Journal of Public Economics* 72(3): 329–60.

Goulder, Lawrence H., and Andrew Schein. 2012. "Carbon Taxes vs. Cap and Trade." Unpublished paper, Stanford University.

Goulder, Lawrence H., and Stephen H. Schneider. 1999. "Induced Technological Change and the Attractiveness of CO_2 Emissions Abatement Policies." *Resource and Energy Economics* 21(3–4): 211–53.

Goulder, Lawrence H., and Robert N. Stavins. 2012. "Interactions between State and Federal Climate Change Policies." Chap. 7 in *The Design and Implementation of U.S. Climate Policy*, edited by Don Fullerton and Catherine Wolfram. University of Chicago Press.

Hahn, Robert W., and Gordon Hester. 1989. "Where Did All the Markets Go? An Analysis of EPA's Emissions Trading Program." *Yale Journal on Regulation* 6(1): 109–153.

Jaffe, Judson, Matthew Ranson, and Robert N. Stavins. 2010. "Linking Tradable Permit Systems: A Key Element of Emerging International Climate Policy Architecture." *Ecology Law Quarterly* 36(4): 789–808.

Kolstad, Jonathan, and Frank Wolak. 2003. "Using Enivronmental Emissions Permit Prices to Raise Electricity Prices: Evidence from the California Electricity Market." CSEM Working Paper 113, Center for the Study of Energy Markets, University of California, Berkeley.

Levinsohn, Arik. 2012. "Belts and Suspenders: Interactions among Climate Policy Regulations." Chap 8 in *The Design & Implementation of U.S. Climate Policy*, edited by D. Fullerton and C. Wolfram. University of Chicago Press.

Market Advisory Committee to the California Air Resources Board. 2007. *Recommendations for Designing a Greenhouse Gas Cap-and-Trade System for California.* Report to the California Air Resources Board.

Metcalf, Gilbert E. 2007. "A Proposal for a U.S. Carbon Tax Swap: An Equitable Tax Reform to Address Global Climate Change." Discussion Paper 2007-12, The Hamilton Project.

Metcalf, Gilbert E., and David Weisbach. 2009. "The Design of a Carbon Tax." *Harvard Environmental Law Review* 33(2): 499–556.

Millard-Ball, Adam. Forthcoming. "The Trouble with Voluntary Emissions Trading." *Journal of Environmental Economics and Management.*

Montgomery, W. David. 1972. "Markets in Licenses and Efficient Pollution Control Programs." *Journal of Economic Theory* 5(3): 395–418.

Nordhaus, William D. 2007. "To Tax or Not to Tax: Alternative Approaches to Slowing Global Warming." *Review of Environmental Economics and Policy* 1(1): 26–44.

Palmer, Karen, and David A. Evans. 2009. "The Evolving SO_2 Allowance Market: Title IV, Cair, and Beyond." Resources for the Future Weekly Policy Commentary, July 13. www.rff.org/Publications/WPC/Pages/090713-Evolving-SO2-Allowance-Market.aspx.

Parry, Ian W. H., and Roberton C. Williams III. 2010. "What Are the Costs of Meeting Distributional Objectives for Climate Policy?" *B. E. Journal of Economic Analysis & Policy* 10(2).

Parry, Ian W. H., Roberton C. Williams III, and Lawrence H. Goulder. 1999. "When Can Carbon Abatement Policies Increase Welfare? The Fundamental Role of Distorted Factor Markets." *Journal of Environmental Economics and Management* 37(1): 52–84.

Smith, Anne E., Martin E. Ross, and W. David Montgomery. 2002. "Implications of Trading Implementation Design for Equity–Efficiency Tradeoffs in Carbon Permit Allocations." Unpublished paper, Charles River Associates.

Stavins, Robert N. 1995. "Transactions Costs and Tradeable Permits." *Journal of Environmental Economics and Management* 29(2): 133–47.

Stavins, Robert N. 2007. "A US Cap-and-Trade System to Address Global Climate Change." Discussion Paper 2007-13, The Hamilton Project.

Tietenberg, Thomas H. 1985. *Emissions Trading: An Exercise in Reforming Pollution Policy.* Washington, DC: Resources for the Future.

Wall Street Journal. 2009. "The Cap and Tax Fiction." June 26, A12.

Weitzman, Martin L. 1974. "Prices *vs.* Quantities." *Review of Economic Studies* 41(4): 477–91.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 103–122

The SO_2 Allowance Trading System: The Ironic History of a Grand Policy Experiment

Richard Schmalensee and Robert N. Stavins

In the late 1980s, there was growing concern in the United States and other countries that acid precipitation—the result of emissions of sulfur dioxide (SO_2) and, to a lesser extent, nitrogen oxides (NO_x) reacting in the atmosphere to form sulfuric and nitric acids—was damaging forests and aquatic ecosystems, particularly in the US Northeast and southern Canada. In the United States, flue gas emissions from coal-fired, electric generating plants were the primary source of SO_2 emissions and a major source of NO_x emissions. In response to this and other concerns, the US Congress passed and President George H. W. Bush signed into law the Clean Air Act Amendments of 1990. Title IV of this law (which took up only 16 percent of its total pages) launched a grand experiment in market-based environmental policy: the path-breaking SO_2 allowance trading program.

The concept of allocating permits to emit a certain quantity of pollution that would phase down over time, while allowing permit-holders to trade their permits, is now broadly familiar. But two decades ago, this cap-and-trade approach to environmental protection was quite novel. Many in the environmental community—with the prominent exception of the Environmental Defense Fund—were hostile to the notion of trading "rights to pollute"; others doubted the workability of such a scheme. Nearly all pollution regulations took a much more prescriptive

■ *Richard Schmalensee is the Howard W. Johnson Professor of Economics and Management, Emeritus, Massachusetts Institute of Technology, Cambridge, Massachusetts. Robert N. Stavins is the Albert Pratt Professor of Business and Government at the Harvard Kennedy School, Cambridge, Massachusetts, and a University Fellow of Resources for the Future, Washington, D.C. Both authors are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are rschmal@mit.edu and robert_stavins@harvard.edu.*

http://dx.doi.org/10.1257/jep.27.1.103. doi=10.1257/jep.27.1.103

"command-and-control" approach, either by setting uniform emission rate limits on classes of emitters or by specifying the type of pollution-control equipment to be installed. Of course, such inflexible regulations impose the same abatement path upon a range of heterogeneous facilities and ignore the fact that the costs of compliance might vary widely across individual facilities depending on their age, technology characteristics, operating conditions, and characteristics of fuel used.

By the close of the twentieth century, the SO_2 allowance trading system had come to be seen as both innovative and successful (for discussion in this journal, see Schmalensee, Joskow, Ellerman, Montero, and Bailey 1998; Stavins 1998). It has become exceptionally influential, leading to a series of policy innovations in the United States and abroad to address a range of environmental challenges, including the threat of global climate change (Stavins 2003). Most prominent among these innovations has been the European Union Emission Trading System, a carbon dioxide (CO_2) cap-and-trade system adopted in 2003 that is by far the world's largest environmental pricing regime (European Commission 2012).

However, the design and implementation of the landmark SO_2 cap-and-trade system have led to a number of striking ironies, which are the focus of this essay. First, subsequent research indicates that in enacting an ambitious—and successful—policy to reduce SO_2 emissions in order to curb acid rain, the government essentially did the right thing for the wrong reason. Second, although the program appears to have been successful, a substantial source of its cost-effectiveness was an unanticipated consequence of the deregulation of railroad rates in the late 1970s and early 1980s. Third, market-based, cost-effective policy innovation in environmental regulation—in particular, cap-and-trade—was championed and implemented by Republican administrations from that of President Ronald Reagan to that of President George W. Bush, but in recent years Republicans have led the way in demonizing cap-and-trade (as an approach to limiting carbon emissions). Fourth and finally, court decisions and subsequent regulatory responses have led to the virtual collapse of the SO_2 market, demonstrating that what the government gives, the government can take away. In order to explore these four ironies, we first briefly review highlights of the system's design and performance.

A fifth, long-recognized irony deserves brief mention. Acid rain itself was largely a consequence of compliance with national ambient air quality standards set in the 1970s for SO_2 and other localized pollutants. In order to reduce local concentrations of these pollutants, electric utilities built more than 400 tall smokestacks, many greater than 500 feet in height (Regens and Rycroft 1988), which successfully dispersed the stack gases, but did so by injecting them high enough into the atmosphere that they precipitated out tens or hundreds of miles downwind as acidified rain, snow, or particles.

Design

Any cap-and-trade policy must face two basic decisions, the level of pollution to be permitted over time and how the initial allocation of permits will be set. The

objective of the SO_2 trading program was to reduce total annual US SO_2 emissions by 10 million tons relative to 1980. Phase I (1995–1999) of the trading program required significant emissions reductions from the 263 most-polluting coal-fired electric generating units, almost all located east of the Mississippi River. Phase II, which began in 2000, placed an aggregate national emissions cap on approximately 3,200 electric generating units—nearly the entire fleet of fossil-fueled plants in the continental United States (Ellerman, Joskow, Schmalensee, Montero, and Bailey 2000). This cap—affecting almost exclusively the power sector—represented a 50 percent reduction from 1980 levels. The permits were demarcated by vintage, with the total number decreasing for successive vintages, thereby achieving a declining cap. (The discussion in this section draws on Chan, Stavins, Stowe, and Sweeney 2012; also see Ellerman et al. 2000.)

How was this target selected? When the policy was enacted, no credible estimates of economic benefits of alternative target levels were available. (Actually, this is true of most environmental policies.) Instead, the target was selected largely based on what was believed to be the "elbow" of the abatement cost curve—that is, a level of abatement that was possible at relatively low costs, and above which the marginal costs of reducing emissions would climb dramatically. This process was consistent with the Baumol and Oates (1971) model of policy making, whereby a politically acceptable target is chosen with an eye toward avoiding regions of steep change in the policy's impact on social welfare. Also, there was a political desire to choose a target level of reductions that was big enough to gain the support of the environmental community and to be seen as satisfying a campaign pledge of newly elected President George H. W. Bush.

The government *gave* permits to emit called "allowances"—denominated in tons of SO_2 emissions—to power plants covered by the law. (The term "permit," which is standard in the economics literature, had another long-established meaning in US environmental law, so the new term "allowance" was coined and used instead.) If annual emissions at a regulated facility exceeded the allowances allocated to that facility, the owner could buy allowances or reduce emissions, whether by installing pollution controls, changing the mix of fuels used to operate the facility, or scaling back operations. If emissions at a regulated facility were reduced below its allowance allocation, the facility owner could sell the extra allowances or, since damages were understood to reflect cumulative emissions over time rather than annual emissions, bank them for future use. EPA's role was essentially to keep score by monitoring emissions on a continuous basis, tracking the ownership of all outstanding allowances (that is, recording initial allocations and subsequent trades), and withdrawing allowances corresponding to each facility's emissions from its account annually. As opposed to a command-and-control regulatory scheme that would have specified an across-the-board timeline for reductions in emissions or dictated specific technologies for pollution control, a cap-and-trade system created incentives to find ways to reduce SO_2 emissions at the lowest cost and to take advantage of low-cost abatement options as soon as they became available.

The free allocation of allowances posed some tradeoffs. After all, government auctioning of allowances would have generated revenue that could, in principle, have been used to reduce distortionary taxes, thereby reducing the program's social cost (Goulder 1995). But this efficiency argument was not advanced at the time; and the affected utilities and their customers' representatives would have strongly opposed auctioning.

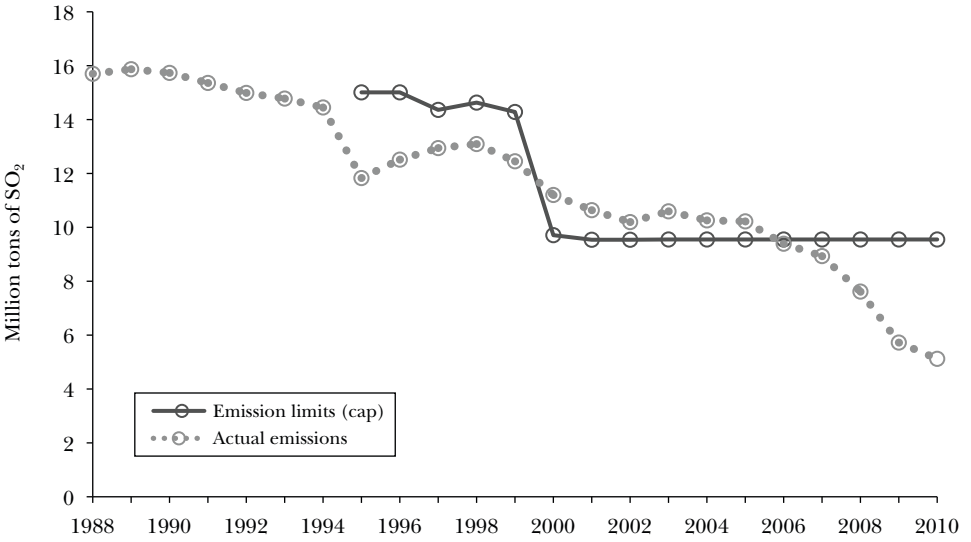
The case for free allocation rested on several arguments. Because cost-of-service regulation characterized the entire investor-owned electric utility industry in 1990, it was assumed that the value of free allowances would be passed on to consumers and would not generate windfall profits for providers. (The use of any allowance involves an opportunity cost because the allowance could be sold instead of used. Absent regulation, output prices would be expected to increase to reflect these opportunity costs, and because the allowances were in fact freely allocated, windfall profits would result.) As important, the political value of being able to allocate free allowances to address differential economic impacts across regions, states, and Congressional districts as well as other concerns was substantial (Joskow and Schmalensee 1998). This was possible because the equilibrium allocation of pollution permits, after trading has occurred, is independent of the initial allocation (Coase 1960; Montgomery 1972)—at least barring particularly problematic types of transaction costs (Stavins 1995; Hahn and Stavins 2011). This means that the initial allocation of allowances could be designed to ensure the greatest political support without fear that this would jeopardize the system's environmental performance or economic cost.

Performance

Beginning in 1995 and over the subsequent decade, the SO_2 allowance trading program performed exceptionally well along all relevant dimensions. SO_2 emissions from electric power plants decreased 36 percent—from 15.9 million to 10.2 million tons—between 1990 and 2004 (US Environmental Protection Agency 2011b), even though electricity generation from coal-fired power plants *increased* 25 percent over the same period (US Energy Information Administration 2012). The program's long-term annual emissions goal was achieved in 2006, and by 2010, SO_2 emissions had declined further, to 5.1 million tons, as shown in Figure 1.

Overall, the program delivered emissions reductions more quickly than expected, as utilities, particularly Phase I units, took advantage of the freedom to bank allowances for future use. (Phase I units were expected, in aggregate, to have lower costs of emissions reduction than Phase II units). Hence, emissions from Phase I units fell well below their cap from 1995 to 1999 and then total emissions temporarily exceeded their cap as banked allowances were used for compliance. After 2006, total emissions (from all units combined) dropped to well below the aggregate cap because of other regulations that imposed tighter restrictions, as we discuss later. With the program's $2,000/ton statutory fine for any emissions exceeding allowance holdings and continuous emissions monitoring, compliance was nearly 100 percent.

Figure 1
SO_2 Caps and Emissions, 1988–2010



Source: Ellerman (2003); US Environmental Protection Agency (2012).
Notes: The emission limits shown for the period 1995–1999 are equal to the Phase 1 units' cap plus Phase 2 units' emissions. Actual emissions shown for all years are the sum of emissions from Phase 1 and Phase 2 units.

The costs of achieving these environmental objectives with cap-and-trade were significantly less than they would have been with a command-and-control regulatory approach. Cost savings were at least 15 percent, and perhaps as much as 90 percent, compared with counterfactual policies that specified the means of regulation in various ways and for various portions of the program's regulatory period (Carlson, Burtraw, Cropper, and Palmer 2000; Ellerman et al. 2000; Keohane 2003). In addition to static cost effectiveness, there is evidence that the program brought down abatement costs over time by providing incentives for innovation and diffusion that were generally much stronger than those provided by traditional command-and-control regulation. Utilities learned how to burn cost-effective mixtures of different types of coal,[1] how to take allowance prices into account in operating decisions, and how to build more cost-effective flue gas desulfurization devices, called "scrubbers" (Ellerman et al. 2000, pp. 235–48; Popp 2003; Bellas and Lange 2011; Frey 2013).

[1] Coal is often divided into three categories: anthracite, bituminous, and lignite. Anthracite is the highest-quality coal, burning with the most energy. Much eastern coast coal is bituminous, and is of intermediate quality. Much of the Powder River Basin coal is "sub-bituminous." Lignite is the lowest quality.

While the SO_2 trading program was less costly than a conventional approach, the costs may or may not have been as low as they could have been. There was significant trading—about 20.3 million tons of allowances were bought and sold by March 1998 (Ellerman et al. 2000, p. 176)—but the implications of this large amount of trade are not obvious. The efficient volume of trade depends on the difference between the initial allocation of allowances and the efficient distribution of emissions among regulated entities, thus very low volumes of trading could also be consistent with overall cost minimization. That said, marginal abatement costs did vary significantly across facilities, at least in the program's first two years (Carlson, Burtraw, Cropper, and Palmer 2000).

There is evidence that the intertemporal allocation of abatement cost (via allowance banking) was at least approximately efficient (Ellerman and Montero 2007), with greater-than-required emissions reductions in Phase I used (via banking) to delay more expensive reductions by Phase II units. In addition, the pattern of voluntary compliance was consistent with cost-effective compliance strategies (Montero 1999). Finally, it is worth noting that the volume of trading grew substantially during the program's early years as utilities gained experience, from 1.5 million tons in the April 1994 to March 1995 period, to 8.4 million tons in the April 1997 to March 1998 period (Ellerman et al. 2000, p. 176).

The following factors could have kept costs above the theoretical minimum, though their influence has been debated: 1) certain provisions in the 1990 legislation that encouraged early use of scrubbers instead of switching to low-sulfur coal, provisions included in an attempt to limit effects of the legislation on high-sulfur coal producers (Ellerman et al. 2000, pp. 301–302); 2) lack of information about marginal abatement costs on the part of market participants, particularly in the early years; 3) state regulations intended to protect domestic high-sulfur coal interests that, particularly in the early years of the program, had the effect of distorting or constraining utilities' responses to federal environmental regulation (Arimura 2002; Bohi and Burtraw 1992; Ellerman et al. 2000, pp. 190–95); 4) interactions between the SO_2 program and other federal regulations, such as New Source Review and New Source Performance Standards, which constrained the program's operation (Gruenspecht and Stavins 2002); and 5) policy uncertainty when regulators and policy makers subsequently considered further reductions in the national SO_2 cap, as we discuss later.

The program can also be evaluated based on the geographic distribution of impacts. Recall that the program came into being mainly in response to concerns about acid rain in the US Northeast. Although it was clear at the time the program was enacted that emissions from different plants had different impacts, the Title IV emissions trading scheme ignored this fact. Most coal-fired power plants were located east of the Mississippi, and model-based analyses predicted that the largest share of cost-effective emissions reductions would come from plants having the greatest impact on lakes and forests in the Northeast. Nonetheless, some worried that emissions would end up disproportionately concentrated and would produce "hot spots" of unacceptably high SO_2 concentrations. Despite these concerns, the

geographic pattern of emissions reductions was broadly consistent with model predictions, and the program did not generate significant hot spots (Ellerman et al. 2000, pp. 130–31; Swift 2004).[2]

In sum, the SO_2 allowance trading system's actual costs, even if they exceeded the cost-effective ideal for a cap-and-trade system, were much lower than would have been incurred with a comparable traditional regulatory approach. The program's goals were achieved with less litigation (and thus less uncertainty) than is typical for traditional environmental programs, both because firms that found it particularly costly to reduce emissions had the option of buying allowances and because firms could not complain about the exercise of administrative discretion by the US Environmental Protection Agency, as the law gave it essentially no discretion. Overall, there is broad agreement that the SO_2 allowance trading system provided a compelling demonstration of the advantages of a market-based approach to environmental protection. With this background on design and performance, we turn to four significant ironies.

Doing the Right Thing for the Wrong Reason

The central purpose of the SO_2 allowance trading program was to reduce the acidification of forest and aquatic ecosystems by cutting precursor SO_2 emissions, primarily in the northeastern United States (National Acid Precipitation Assessment Program 1998). The goal of reducing SO_2 emissions was met and exceeded. However, it turns out that the ecological benefits of the program have been relatively small, largely because it takes much longer than thought to reverse the acidification of ecosystems (National Acid Precipitation Assessment Program 2005). On the other hand, other completely unanticipated benefits of the program have been massive.

Whereas some studies at the time of the program's enactment predicted that its benefits would be approximately equal to its costs (Portney 1990), more recent estimates suggest annual benefits of between $59 and $116 billion, compared with annual costs of $0.5 to $2 billion, as shown in Table 1. However, more than 95 percent of these benefits are associated not with ecological impacts—including acidification of aquatic ecosystems—but instead with human health impacts of reduced levels of airborne fine sulfate particles less than 2.5 micrometers in diameter ($PM_{2.5}$), particles which derive from SO_2 emissions. Epidemiological evidence of the harmful human health effects of these fine particulates mounted rapidly in the decade *after* the CAAA was enacted (Chestnut and Mills 2005).

Estimates of these health benefits vary widely, but they appear to be on the order of $50 billion to more than $100 billion per year (Burtraw, Krupnick, Mansur,

[2] Muller and Mendelsohn (2009) suggest that the use of damage-based trading ratios, where allowances might be adjusted for the marginal environmental damage each source of emissions would do, rather than using a single allowance price, could have been welfare-improving. Of course, the practical challenges of setting such ratios—particularly in a political environment—would be serious.

Table 1
Estimated Annual US Benefits and Costs of the SO_2 Allowance Trading Program; Title IV, Clean Air Amendments of 1990
(billions of US 2000 Dollars)

Benefits	
Mortality	50–100
Morbidity	3–7
Recreational visibility	2–3
Residential visibility	2–3
Ecosystem effects	0.5
Total	**59–116**
Costs	**0.5–2.0**
Net benefits	**58–114**

Source: Burtraw, Krupnick, Mansur, Austin, and Farrell (1998); Burtraw (1999); Chestnut and Mills (2005); Banzhaf, Burtraw, Evans, and Krupnick (2006).

Austin, and Farrell 1998; Burtraw 1999; Chestnut and Mills 2005; National Acid Precipitation Assessment Program 2005; Shadbegian, Gray, and Morgan 2005; US Environmental Protection Agency 2011a).[3] As Table 1 shows, strict ecosystem benefits are probably considerably less than program costs, though at least one study (Banzhaf, Burtraw, Evans, and Krupnick 2006) suggests that ecosystem benefits alone have exceeded costs. But estimated human health benefits of the program may have exceeded annual costs by a factor of more than fifty! With its mandated 50 percent cut in SO_2 emissions, the government did what turned out to be the right thing for the wrong reason.

An Unanticipated Consequence of Deregulation

The realized costs of the SO_2 allowance trading program were substantially less than forecasts made prior to implementation (National Acid Precipitation Assessment Program 2005).[4] Part of this discrepancy was due to technological innovation and the speed with which the allowance market matured. But another major factor in low realized compliance costs was the emergence of input substitution,

[3] The lower end of this range of benefit estimates is linked with the possibly nonlinear relationship between cuts in SO_2 emissions and reductions in $PM_{2.5}$ deposition (West, Ansari, and Pandis 1999).
[4] A revolutionary aspect of the cap-and-trade approach was that for the first time regulators had instantaneous information in a summary statistic (the allowance price) of the marginal cost of compliance, but the program's design did not allow for any response to that information, such as changing the cap.

from high- to low-sulfur coal, as a cost-effective strategy for reducing SO_2 emissions. Indeed, the attractiveness of switching to low-sulfur coal was increasing *before* the program even went into effect due to a public policy change *unrelated to the environment* and initiated long before 1990.

The three major coal deposits in the United States are located in the Powder River Basin of Wyoming and Montana, the Illinois Basin, and Central Appalachia. Of these, Powder River Basin coal is cheapest to mine and has the lowest sulfur content (though considerable low-sulfur coal was also produced in the East, particularly after the acid rain program took effect). However, the majority of coal-fired power plants in the United States are located along or east of the Mississippi River, making Powder River Basin the most distant option for major sources of demand.

Prior to 1976, the Interstate Commerce Commission set rates for freight rail, which is the main way in which coal is transported. The Railroad Revitalization and Regulatory Reform Act of 1976 and the Staggers Rail Act of 1980 gave rail carriers the ability to set their own rates and legalized private railroad-shipper contracts. As a result, shipping rates for coal (and other products) declined significantly (Winston 2005; US Government Accountability Office 2007). The cost of bringing coal from the Powder River Basin to centers of high demand east of the Mississippi River fell dramatically (Ellerman et al. 2000)—even though the existence of only two major rail lines kept shipping costs above competitive levels (Busse and Keohane 2007).

Deregulation gave the freight carriers flexibility and incentive to contract with eastern utilities, and, as noted above, these same utilities developed cost-effective ways to burn sub-bituminous coal (which had lower energy content as well as lower sulfur content) (Ellerman et al. 2000, pp. 243–45). The average sulfur content of coal burned at electric generating units began to fall. In fact, SO_2 emissions at units covered by the allowance trading program were actually falling from 1985 to 1993, *before* the acid rain program took effect (Ellerman and Montero 1998). The main source of this decline was the increased use of Powder River Basin coal, with average rail rates of shipping that coal from Wyoming and Montana to Midwest generators falling by over 50 percent from 1979 to 1993 (Gerking and Hamilton 2008).

For some power plants, fuel-switching from high- to low-sulfur coal was cost-effective even without acid rain legislation; and for many other eastern power plants, rail deregulation made fuel-switching less expensive than installing scrubbers in response to the legislation. Of the 263 units regulated in Phase I of the allowance trading program, 52 percent primarily pursued fuel-switching or blending low-sulfur coal with higher-sulfur coal, accounting for 59 percent of emissions reductions; and scrubbers were installed at about 10 percent of the units, accounting for 28 percent of emissions reduction (US Energy Information Administration 1997).[5] About one-third of SO_2 emissions reductions in the early years of the program were due to

[5] In addition, 32 percent of the units complied by obtaining additional allowances as well as switching to lower-sulfur coal, accounting for 9 percent of emissions reductions; 3 percent of the units were retired, accounting for 2 percent of emissions reductions; and 3 percent of the units used other compliance methods, accounting for 2 percent of emissions reductions (US Energy Information Administration 1997).

prior railroad deregulation and two-thirds to the SO_2 allowance trading program (Ellerman et al. 2000, p. 122).

It could be argued that because these reductions in delivered fuel costs would have occurred in the absence of the SO_2 allowance trading program and would have reduced the costs of a command-and-control SO_2 program as well, the cost savings attributed to the SO_2 allowance trading program (relative to a command-and-control system) should be adjusted downward (Carlson et al. 2000). This point has some validity, but it is also true that a prescriptive regulatory approach—say, a policy that required installing scrubbers at all power plants—might have prevented electricity companies from taking advantage of some of these alternative compliance options. In any event, it is clear that significant shares of the emissions reduction—about one-third in the early years—and of the cost savings associated with the SO_2 allowance trading system were actually an unanticipated consequence of an earlier, unrelated public policy change.

Conservatives Demonize Their Own Innovation

For a long time, market-based approaches to environmental protection, such as cap-and-trade, bore a Republican label. In the 1980s, President Ronald Reagan's Environmental Protection Agency put in place a trading program to phase out leaded gasoline. It produced a more rapid elimination of leaded gasoline from the marketplace than had been anticipated, and at a savings of some $250 million per year compared with a conventional no-trade, command-and-control approach (Stavins 2003). Not only did President George H. W. Bush successfully propose the use of cap-and-trade to cut US SO_2 emissions, his administration advocated in international forums the use of emissions trading to cut global CO_2 emissions, a proposal initially resisted but ultimately adopted by the European Union. In 2005, President George W. Bush's EPA issued the Clean Air Interstate Rule, aimed at reducing SO_2 emissions by a further 70 percent from their 2003 levels. Cap-and-trade was again the policy instrument of choice. (More about this rule below.)

When the Clean Air Act Amendments were being considered in the US Congress in 1989–1990, political support was not divided on partisan lines. Indeed, environmental and energy debates from the 1970s through much of the 1990s typically broke along geographic, rather than partisan, lines,[6] with key parameters being degree of urbanization and reliance on specific fuel types, such as coal versus natural gas. Thus, the Clean Air Act Amendments of 1990 passed the US Senate by a vote of 89–11 with 87 percent of Republican members and 91 percent of Democrats voting yea, and the legislation passed the House of Representatives by a vote of 401–21 with 87 percent of Republicans and 96 percent of Democrats voting in support.

[6] The same was true of trade policy debates until the early 1990s, that is, they were driven by economic impacts on various sectors and populations, which resulted in geographic, not partisan, divisions.

However, 20 years later when climate change legislation was receiving serious consideration in Washington, environmental politics had changed dramatically, with Congressional support for environmental legislation coming mainly to reflect partisan divisions.[7] In 2009, the US House of Representatives passed the American Clean Energy and Security Act of 2009 (H.R. 2454)—often known as the Waxman–Markey bill—that included an economy-wide cap-and-trade system to cut carbon dioxide (CO_2) emissions. The Waxman–Markey bill passed the House by a narrow margin of 219–212, with support from 83 percent of Democrats, but only 4 percent of Republicans. In July 2010, the US Senate abandoned its attempt to pass companion legislation. In the process of debating this legislation, conservatives (largely Republicans and some coal-state Democrats) attacked the cap-and-trade system as "cap-and-tax,"[8] much as an earlier generation of liberals had denigrated cap-and-trade as "selling licenses to pollute."

Many conservatives in the Congress undoubtedly opposed climate policies because of disagreement about the threat of climate change or the costs of the policies, but instead of debating those risks and costs, they chose to launch an ultimately successful campaign to demonize and thereby tarnish cap-and-trade as an instrument of public policy, rendering it "collateral damage" in the wider climate policy battle. This scorched-earth approach could come back to haunt conservatives if future environmental initiatives with widespread support are enacted without making use of the power of the marketplace to reduce compliance costs. It is ironic that conservatives chose to demonize their own market-based creation. It is perhaps even more ironic that this tactic seems to have been effective despite their creation's excellent performance.

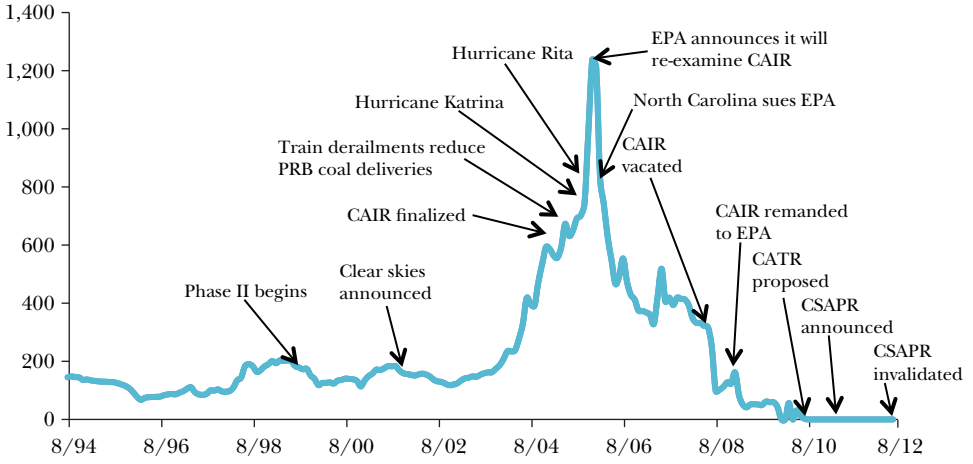
What the Government Gives, It Can Take Away

A major source of uncertainty about any government-created market is that the government can undo what it created—possibly unintentionally. In essence, this happened in the SO_2 allowance market. Through a series of new Clean Air Act regulations, court rulings, and regulatory responses, the courts affirmed that EPA could not set up a new interstate trading system or modify the Title IV system in the absence of new legislation from Congress. In response, state-level and source-level constraints were put in place that ultimately rendered the SO_2 cap-and-trade system itself nonbinding and effectively closed down the allowance market.

[7] This polarization between the two political parties on environmental issues (Shipan and Lowry 2001) was and is part of a gradually widening gulf between the parties on virtually all issues (Fleisher and Bond 2004; Poole and Rosenthal 1997, 2007). Moderates have been gradually disappearing for decades (Lowry and Shipan 2002; Theriault 2008).

[8] They may have been helped by President Obama's February 2009 budget message to Congress, which provided for revenues from an auction of 100 percent of the allowances under such a scheme (Chan, Stavins, Stowe, and Sweeney 2012).

Figure 2
SO_2 Allowance Prices and the Regulatory Environment, 1994–2012
(1995 dollars per ton)



Source: Data on spot prices compiled by Power & Energy Analytic Resources (PEAR) Inc. from Cantor Fitzgerald until September 11, 2001, and from ICAP United thereafter.
Notes: CAIR is "Clean Air Interstate Rule." CATR is "Clean Air Transport Rule." CSAPR is "Cross-State Air Pollution Rule."

Prices for SO_2 allowances were remarkably stable throughout the program's first decade, as shown in Figure 2, and then we see a steep spike. What happened? It was widely recognized by the late 1990s that SO_2 reductions in excess of those resulting from the trading program of Title IV would be required by other provisions in the Clean Air Act dealing with air quality standards because of the significant adverse health effects of fine particulates associated with SO_2 emissions. But the law did not give the EPA authority to adjust the Title IV program, such as by tightening the overall cap, in response to new information about the benefits (or costs) of emissions reductions. This crucial fact drove the chain of events leading to the ultimate collapse of the SO_2 allowance trading program.

In early 2002, President George W. Bush proposed the Clear Skies Act, which would have greatly tightened the SO_2 cap. Prices in the allowance market did not immediately budge, however, which suggests it was no surprise to market participants when this proposal died in March 2005, having failed to move out of committee. The Bush administration then promulgated its Clean Air Interstate Rule in May 2005, with the same purpose of lowering the cap on SO_2 emissions (to 70 percent below the 2003 emissions level). This rule sought to apply more stringent emission requirements on states that were contributing to violations of EPA's primary ambient air quality standards for fine particulates in the eastern United States (Palmer and Evans 2009). It required sources within those states to surrender two additional allowances for every ton of SO_2 emissions—effectively reducing the cap by two-thirds. Because

the Clean Air Interstate Rule provided that firms could bank their existing SO_2 allowances for use in the new program, prices rose further in anticipation of this more stringent cap, with spot prices increasing from $273 per ton in EPA's 2004 auction to $703 in the 2005 auction.[9]

After peaking in 2005 at more than $1,200 per ton (see Figure 2), SO_2 allowance prices dropped just as fast as they had risen, aided by an announcement from the US Environmental Protection Agency that it would reexamine the Clean Air Interstate Rule (Samuelsohn 2005) and speculation about impending legal challenges (Samuelsohn 2006a; Kruse 2009).[10] On June 26, 2006, North Carolina and other states and a number of utilities sued the Environmental Protection Agency over the Clean Air Interstate Rule (Samuelsohn 2006b). The states argued that the interstate trading allowed under the rule was inconsistent with Section 110(a) of the Clean Air Act, which obliges each state to prevent emissions that interfere with any other state's attainment or maintenance of air quality standards. This meant that the EPA could not set up a new trading program built on the SO_2 allowance trading system by regulatory means and would therefore have to focus on source-level or other types of regulation in its efforts to reduce emissions below the limits established in Title IV in order to meet new local-air-quality standards. Because the new, required regulation, rather than Title IV, would become the binding constraint on emissions, trading under the original SO_2 allowance trading system would be rendered unimportant.

Two years later, on July 11, 2008, the Circuit Court of Appeals for the District of Columbia (*State of North Carolina v. Environmental Protection Agency*, 531 F. 3d 896 [D.C. Cir. 2008]) vacated the Clean Air Interstate Rule in its entirety on the grounds that, under the Clean Air Act, the Environmental Protection Agency could not ignore the relationship between sources and receptors in matters involving air quality standards (US Environmental Protection Agency 2011a). Thus, without new legislation, the Title IV program, with interstate trading at its core, could not be

[9] An array of other factors contributed to the run-up and eventual spike in SO_2 allowance prices, including Hurricanes Katrina (August 2005) and Rita (September 2005), which impaired petroleum refining and natural gas capacity. In addition, delivery of low-sulfur coal from the Powder River Basin to Midwestern power plants was disrupted by track failures (May 2005) on both the Union Pacific and Burlington Northern Santa Fe railroads, which caused low-sulfur coal prices in the Midwest to peak in December 2005, at a level three times greater than a year earlier. As a result, some power companies switched to higher-sulfur coal from the east, increasing demand for SO_2 allowances. A final factor was features of the allowance trading program's design that interacted with the tax system and utility regulation to restrict the number of allowances actually available for trading at any time (the "float"), thus compounding the price impacts of the other factors (Parsons, Ellerman, and Feilhauer 2009).

[10] Also contributing to the fall in allowances prices from their peak was a drop in natural gas prices, the restoration of refining and gas capacity in the Gulf of Mexico following Hurricanes Katrina and Rita, and the realization of a likely adequate supply of allowances and installed scrubber capacity to comply with the Clean Air Interstate Rule (Burtraw and Szambelan 2009). In addition, many expected an economy-wide CO_2 cap-and-trade system, which all three major Presidential candidates—John McCain, Hillary Clinton, and Barack Obama—in 2007 supported and which would have led to an exogenous, long-run decline in coal usage and thus in SO_2 emissions, and hence to a decline in the value of banked allowances.

modified to drive further reductions in SO_2 emissions to meet air quality standards. On that single day, the SO_2 allowance price fell from $315 to $115 (Burtraw and Szambelan 2009). The Bush administration, followed by the subsequent Obama administration, chose not to appeal that ruling. The court allowed the Clean Air Interstate Rule to remain in effect while the EPA devised a replacement that addressed its concerns, but it remained clear that unlimited interstate trading was doomed. Prices continued to fall, returning to the range of their pre-2004 levels. At the 2009 auction, spot allowances (which could be used in 2009 or later) sold for $70 per ton, compared with $390 a year earlier (Burtraw and Szambelan 2009).

In July 2010, the Obama administration proposed an alternative rule to limit annual SO_2 (and NO_x) emissions in 28 states, as a replacement for the Clean Air Interstate Rule. The proposed rule established state-specific emissions caps for power plant SO_2 emissions, thereby limiting interstate trading. The rule was finalized in July 2011 as the Cross-State Air Pollution Rule, allowing only intrastate trading and limited trading between two groups of states. Predictably, this rule too was challenged in court, by 27 states and 18 other parties; in August 2012, the US Court of Appeals for the D.C. Circuit invalidated the rule (*EME Homer City Generation, L.P. vs. Environmental Protection Agency, et al.*, No. 11-1302).

While the SO_2 allowance market functioned well, the broader regulatory environment served to end its effective life. The allowance market remains nominally in place, but the imposition of state-level and source-specific prescriptive regulation has virtually eliminated the demand for federal SO_2 allowances. By the time of the Environmental Protection Agency's 2012 auction, market-clearing prices had fallen to $0.56 in the spot auction and $0.12 in the seven-year advance auction.[11] Those states with binding caps for SO_2 under the Cross-State Air Pollution Rule must still reduce their emissions, whether by mandating the use of scrubbers, retiring coal-fired power plants, or setting up intrastate trading of emission allowances.

In essence, the series of regulations, court rulings, and regulatory responses that followed Congress's rejection of the George W. Bush administration's Clear Skies Act affirmed that: 1) EPA cannot set up an interstate trading system under the Clean Air Act in the absence of specific legislation from Congress (which, of course, it had for the SO_2 allowance trading system under Title IV of the Clean Air Act amendments of 1990); and 2) consequent state-level and source-level constraints following the Clean Air Interstate Rule rendered the SO_2 cap-and-trade system itself nonbinding.

One more irony: the SO_2 program's success may have weakened the case for continuing the allowance market by reducing the heterogeneity of abatements costs across sources, thus reducing potential gains from trade (Newell and Stavins 2003).

[11] When new Mercury and Air Toxics Standards affecting coal-fired power plants take effect—likely in 2015–2016—they will likely be so stringent that SO_2 constraints under the Cross-State Air Pollution Rule will be rendered nonbinding in one of the two SO_2 trading zones. Further, the Mercury and Air Toxics Standards explicitly do not allow trading, and so assuming these rules are finalized and implemented as expected, there will be only a minimal market for SO_2 (Burtraw, Palmer, Paul, Beasley, and Woerman 2012).

When the government creates a market, it can also destroy it, possibly fostering a legacy of increased regulatory uncertainty and reduced investor confidence in future cap-and-trade regimes, and hence reduced credibility of pollution markets more broadly.

Conclusions

More than 20 years ago, the Clean Air Act amendments of 1990 launched the path-breaking SO_2 allowance trading system, the world's first large-scale market-based environmental initiative. That grand experiment in public policy continues to enjoy its reputation around the world as a great success. Although it is true that the system performed at least as well as its advocates had anticipated through its first decade of operation—reducing emissions cost-effectively—it is also true that reflections from our current perspective yield a considerably more nuanced assessment of performance. The actual costs of compliance turned out to be lower than expected, but this was in substantial part an unintended consequence of other, nonenvironmental policy innovations: specifically, the earlier deregulation of US railroads that allowed less-expensive delivery of low-sulfur coal from the Powder River Basin to the Midwest. The actual benefits turned out to be substantially greater than originally expected but not because of ecological benefits. Rather, reductions in SO_2 emissions resulted in substantial decreases in downwind concentrations of small particulates, thereby producing great benefits to human health.

What appeared in 1990 to be a quintessential moderate Republican approach to environmental protection—cap-and-trade—generated great hostility from conservatives 20 years later. In the process of opposing Congressional climate policy initiatives in 2009–2010, conservatives demonized cap-and-trade proposals as "cap-and-tax" and may have thereby tarnished this market-based approach to environmental protection for years to come. Ironically, an attempt by a Republican administration to use the cap-and-trade approach to reduce the SO_2 emissions cap eventually led, through a series of court cases and regulatory responses, to the virtual closure of the SO_2 allowance market.

What are some lessons of this history of the SO_2 allowance trading program for future market-based and other public policies? First, much is often learned over time regarding any policy's benefits and costs. What may appear to be wise initially may not turn out to be wise in the long term, and what appears to be unwise initially may turn out to be very attractive in the long term. Thus it can be important for policies to be flexible and responsive to changes in knowledge and technology. On the other hand, policy stability encourages efficient investment, so unnecessary changes can be destructive. It can be argued that the SO_2 cap-and-trade system provided valuable stability, but the legislation also made it impossible to make what would have been responsive, effective, and efficient changes in the policy.

Second, unintended consequences of policies are almost inevitable. They can sometimes be beneficial, as in the case of the effects of rail deregulation on the

performance of the SO_2 allowance trading system. They can sometimes be negative, as when regulatory responses to invalidation of the Clean Air Interstate Rule led to the virtual collapse of the SO_2 market. But by definition, such changes are almost impossible to predict. The implication is to be very careful and modest with forecasts and assessments. This can be demonstrated by a retrospective review of initial (under)estimates of the consequences of the Staggers Rail Act of 1980 and (overly hopeful) assessments of the promulgation of the Clean Air Interstate Rule in 2005.

Third, in most cases, politics trumps science and economics. The target of Title IV to reduce SO_2 emissions by 50 percent was set neither on the basis of the science, drawing on the findings of the National Acid Precipitation Assessment Program, nor economics, drawing on a comparison of anticipated benefits and costs. The implication is not to ignore politics but, rather to design policies that are likely to succeed in real-world political settings. Cap-and-trade systems can facilitate sound performance in political settings because of their ability to build constituencies of political support through free allocation of allowances without this negatively affecting the system's aggregate performance, either environmentally or economically.

Fourth, market-based policies have great cost and feasibility advantages, but like any public policy, the government can change or repeal these initiatives, or render them irrelevant. Market-based and other public policies can be constrained by other policies. Economists and other analysts tend to examine policies one at a time, but this misses potential interactions, which can be exceptionally important (Goulder and Stavins 2011).

Finally, what are the implications for future climate change policy? The bad news seems to be that "cap-and-tax" rhetoric may make it hard to use this approach in the United States to deal with climate change. Emissions of CO_2 from coal-fired power plants will no doubt be reduced by EPA rules on SO_2, NO_x, mercury, coal fly-ash, and cooling-water withdrawals that are working their way through the regulatory process and that will drive up the cost of generating electricity with coal. But these rules, and those likely to be adopted by the Environmental Protection Agency in response to the US Supreme Court decision in *Massachusetts et al. v. Environmental Protection Agency et al.* (549 US 497 [2007] (at http://www.supremecourt.gov/opinions/06pdf/05-1120.pdf) that it regulate CO_2 under the Clean Air Act, are unlikely to be cost-effective policies for reducing greenhouse gas emissions in the long run. At a time when environmental protection in general and climate policy in particular have become highly partisan in the US Congress, the outlook for an efficient and effective national climate policy is not very promising.

The good news, however, is that cap-and-trade is no longer just a subject for academic seminars and journal articles; it is a proven, viable option for tackling large-scale environmental problems. It is now being used around the world, including for addressing CO_2 emissions linked with global climate change. Even if the SO_2 allowance trading program's performance was enhanced by unanticipated benefits and declines in coal prices, and even if it has been essentially wiped out by later policy changes, the fact is that the allowance trading program achieved its

target emissions reductions rapidly and cost-effectively. Few other environmental programs of any sort have performed as well.

■ *We thank Richard Sweeney for research assistance, and Dallas Burtraw, Juan-Pablo Montero, Sean Theriault, Tom Tietenberg, and especially Denny Ellerman, as well as the Editor, Co-editors, and Managing Editor of this journal for valuable comments on a previous version of this paper. Any errors and all opinions are our own.*

References

Arimura, Toshi H. 2002. "An Empirical Study of the SO_2 Allowance Market: Effects of PUC Regulations." *Journal of Environmental Economics and Management* 44(2): 271–89.

Banzhaf, H. Spencer, Dallas Burtraw, David Evans, and Alan Krupnick. 2006. "Valuation of Natural Resource Improvements in the Adirondacks." *Land Economics* 82(3): 445–64.

Baumol, William J., and Wallace E. Oates. 1971. "The Use of Standards and Prices for Protection of the Environment." *Swedish Journal of Economics* 73(1): 42–54.

Bellas, Allen S., and Ian Lange. 2011. "Evidence of Innovation and Diffusion under Tradable Permit Programs." *International Review of Environmental and Resource Economics* 5(1): 1–22.

Bohi, Douglas R., and Dallas Burtraw. 1992. "Utility Investment Behavior and the Emission Trading Market." *Resources and Energy* 14(1–2): 129–53.

Burtraw, Dallas. 1999. "Cost Savings, Market Performance and Economic Benefits of the U.S. Acid Rain Program. *Pollution for Sale: Emissions Trading and Joint Implementation*, edited by Steve Sorrell and Jim Skea. Cheltenham, UK: Edward Elgar.

Burtraw, Dallas, Alan Krupnick, Erin Mansur, David Austin, and Deirdre Farrell. 1998. "Cost and Benefits of Reducing Air Pollutants Related to Acid Rain." *Contemporary Economic Policy* 16(4): 379–400.

Burtraw, Dallas, Karen L. Palmer, Anthony Paul, Blair Beasley, and Matthew Woerman. 2012. "Reliability in the Electricity Industry under New Environmental Regulations." Discussion Paper 12-18, Resources for the Future.

Burtraw, Dallas, and Sarah Jo Szambelan. 2009. "U.S. Emissions Trading Markets for SO_2 and NO_x." Discussion Paper 09-40, Resources for the Future.

Busse, Meghan R., and Nathaniel O. Keohane. 2007. "Market Effects of Environmental Regulation: Coal, Railroads, and the 1990 Clean Air Act." *RAND Journal of Economics* 38(4): 1159–79.

Carlson, Curtis, Dallas Burtraw, Maureen L. Cropper, and Karen Palmer. 2000. "Sulfur Dioxide Control by Electric Utilities: What are the Gains from Trade?" *Journal of Political Economy* 108(6): 1292–1326.

Chan, Gabriel, Robert Stavins, Robert Stowe, and Richard Sweeney. 2012. *The SO_2 Allowance Trading System and the Clean Air Act Amendments of 1990: Reflections on Twenty Years of Policy Innovation*. Cambridge, MA: Harvard Environmental Economics Program.

Chestnut, Lauraine G., and David M. Mills. 2005. "A Fresh Look at the Benefits and Costs of the U.S. Acid Rain Program." *Journal of Environmental Management* 77(3): 252–66.

Coase, Ronald H. 1960. "The Problem of Social Cost." *Journal of Law and Economics* 3(October): 1–44.

Ellerman, A. Denny. 2003. "Ex Post Evaluation of Tradable Permits: The U.S. SO_2 Cap-and-Trade Program." Working Paper CEEPR 03-003, Massachusetts Institute of Technology.

Ellerman, A. Denny, Paul L. Joskow, Richard Schmalensee, Juan-Pablo Montero, and Elizabeth M. Bailey. 2000. *Markets for Clean Air: The U.S. Acid Rain Program*. Cambridge, UK: Cambridge University Press.

Ellerman, A. Denny, and Juan-Pablo Montero.

1998. "The Declining Trend in Sulfur Dioxide Emissions: Implications for Allowance Prices." *Journal of Environmental Economics and Management* 36(1): 26–45.

Ellerman, A. Denny, and Juan-Pablo Montero. 2007. "The Efficiency and Robustness of Allowance Banking in the U.S. Acid Rain Program." *Energy Journal* 28(4): 47–71.

European Commission. 2012. *Emissions Trading System (EU ETS).* Available at http://ec.europa. eu /clima/policies/ets/index_en.htm.

Fleisher, Richard, and John R. Bond. 2004. "The Shrinking Middle in the U.S. Congress." *British Journal of Political Science* 34(3): 429–51.

Frey, Elaine. 2013. "Technology Diffusion and Environmental Regulation: The Adoption of Scrubbers by Coal-Fired Plants." *Energy Journal* 34(1): 177–205.

Gerking, Shelby, and Stephen F. Hamilton. 2008. "What Explains the Increased Utilization of Powder River Basin Coal in Electric Power Generation?" *American Journal of Agricultural Economics* 90(4): 933–50.

Goulder, Lawrence H. 1995. "Environmental Taxation and the 'Double Dividend': A Reader's Guide." *International Tax and Public Finance* 2(2): 157–83.

Goulder, Lawrence H., and Robert N. Stavins. 2011. "Challenges from State–Federal Interactions in US Climate Change Policy." *American Economic Review* 101(3): 253–57.

Gruenspecht, Howard K., and Robert N. Stavins. 2002. "New Source Review under the Clean Air Act: Ripe for Reform." *Resources*, Issue 147, pp. 19–23.

Hahn, Robert W., and Robert N. Stavins. 2011. "The Effect of Allowance Allocations on Cap-and-Trade System Performance." *Journal of Law and Economics* 54(4): S267–S294.

Joskow, Paul L., and Richard Schmalensee. 1998. "The Political Economy of Market-based Environmental Policy: The U.S. Acid Rain Program." *Journal of Law and Economics* 41(1): 37–83.

Keohane, Nathaniel O. 2003. "What Did the Market Buy? Cost Savings under the U.S. Tradeable Permits Program for Sulfur Dioxide." Working Paper YCELP-01-11-2003, Yale Center for Environmental Law and Policy.

Kruse, Elizabeth. 2009. "North Carolina v. Environmental Protection Agency." *Harvard Environmental Law Review* 33(1): 283–96.

Lowry, William R., and Charles R. Shipan. 2002. "Party Differentiation in Congress." *Legislative Studies Quarterly* 27(1): 33–60.

Montero, Juan-Pablo. 1999. "Voluntary Compliance with Market-Based Environmental Policy: Evidence from the U.S. Acid Rain Program." *Journal of Political Economy* 107(5): 998–1033.

Montgomery, W. David. 1972. "Markets in Licenses and Efficient Pollution Control Programs." *Journal of Economic Theory* 5(3): 395–418.

Muller, Nicholas Z., and Robert Mendelsohn. 2009. "Efficient Pollution Regulation: Getting the Prices Right." *American Economic Review* 99(5): 1714–39.

National Acid Precipitation Assessment Program. 1998. *Biennial Report to Congress: An Integrated Assessment.* Washington, DC: National Science and Technology Council, Committee on Environment and Natural Resources.

National Acid Precipitation Assessment Program. 2005. *National Acid Precipitation Assessment Program Report to Congress: An Integrated Assessment.* Washington, DC: National Science and Technology Council, Committee on Environment and Natural Resources.

Newell, Richard G., and Robert N. Stavins. 2003. "Cost Heterogeneity and the Potential Savings from Market-Based Policies." *Journal of Regulatory Economics* 23(1):43–59.

Palmer, Karen, and David A. Evans. 2009. "The Evolving SO_2 Allowance Market: Title IV, CAIR, and Beyond." *Resources for the Future Weekly Policy Commentary*, July 13.

Parsons, John E., A. Denney Ellerman, and Stephen Feilhauer. 2009. "Designing a U.S. Market for CO_2." *Journal of Applied Corporate Finance* 21(1): 79–86.

Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting.* New York: Oxford University Press.

Poole, Keith T., and Howard Rosenthal. 2007. *Ideology and Congress.* New Brunswick: Transaction Publishers.

Popp, David. 2003. "Pollution Control Innovations and the Clean Air Act of 1990." *Journal of Policy Analysis and Management* 22(4): 641–60.

Portney, Paul R. 1990. "Policy Watch: Economics of the Clean Air Act." *Journal of Economic Perspectives* 4(4): 173–81.

Regens, James L, and Robert W. Rycroft. 1988. *The Acid Rain Controversy.* University of Pittsburgh Press.

Samuelsohn, Darren. 2005. "EPA Agrees to Re-examine CAIR Power Plant Standard." *Greenwire*, November 23.

Samuelsohn, Darren. 2006a. "Bush Regs Facing a Year of Legal Challenges." *Greenwire*, January 6.

Samuelsohn, Darren. 2006b. "North Carolina Sues EPA over Upwind Emissions." *Greenwire*, June 27.

Schmalensee, Richard, Paul L. Joskow, A. Denny

Ellerman, Juan-Pablo Montero, and Elizabeth M. Bailey. 1998. "An Interim Evaluation of Sulfur Dioxide Emissions Trading." *Journal of Economic Perspectives* 12(3): 53–68.

Shadbegian, Ronald J., Wayne B. Gray, and Cynthia L. Morgan. 2005. "Benefits and Costs from Sulfur Dioxide Trading: A Distributional Analysis." Working Paper 05-09. U.S. Environmental Protection Agency, National Center for Environmental Economics.

Shipan, Charles R., and William R. Lowry. 2001. "Environmental Policy and Party Divergence in Congress." *Political Research Quarterly* 54(2): 245–63.

Stavins, Robert N. 1995. "Transaction Costs and Tradeable Permits." *Journal of Environmental Economics and Management* 29(2): 133–48.

Stavins, Robert N. 1998. "What Can We Learn from the Grand Policy Experiment? Lessons from SO_2 Allowance Trading." *Journal of Economic Perspectives* 12(3): 69–88.

Stavins, Robert N. 2003. "Experience with Market-Based Environmental Policy Instruments." Chap. 9 in *Handbook of Environmental Economics*, Vol. I, edited by Karl-Göran Mäler and Jeffrey Vincent, 355–435. Amsterdam: Elsevier Science.

Swift, Byron. 2004. "Emissions Trading and Hot Spots: A Review of the Major Programs." *Environment Reporter* 35(19): 1–16.

Theriault, Sean M. 2008. *Party Polarization in the Congress*. New York: Cambridge University Press.

US Energy Information Administration. 1997. *The Effects of Title IV of the Clean Air Act Amendments of 1990 on Electric Utilities: An Update*. http://www.eia.gov/cneaf/electricity/clean_air_upd97/exec_sum.html.

US Energy Information Administration. 2012. "Electricity Net Generation: Total (All Sectors), 1949–2011." Table 8.2a in *Annual Energy Review 2011*. http://www.eia.gov/totalenergy/data/annual/pdf/aer.pdf.

US Environmental Protection Agency. 2011a. *Clean Air Interstate Rule, Acid Rain Program, and Former NO_x Budget Trading Program 2010 Progress Report*. Washington, D.C.: U.S. Environmental Protection Agency. www.epa.gov/airmarkt/progress/ARPCAIR10_01.html.

US Environmental Protection Agency. 2011b. "National Emissions Inventory (NEI) Air Pollutant Emissions Trends Data: 1970–2011." Available at: www.epa.gov/ttn/chief/trends/index.html.

US Environmental Protection Agency. 2012. "Air Markets Program Data." http://ampd.epa.gov/ampd/ (accessed December 8, 2012.)

US Government Accountability Office. 2007. *Freight Railroads: Updated Information on Rates and Other Industry Trends*. Publication no. GAO-07-291R.

West, J. Jason, Asif S. Ansari, and Spyros N. Pandis. 1999. "Marginal $PM_{2.5}$: Nonlinear Aerosol Mass Response to Sulfate Reductions in the Eastern United States." *Journal of the Air & Waste Management Association* 49(12): 1415–24.

Winston, Clifford. 2005. "The Success of the Staggers Rail Act of 1980." Related Publication 05-24, AEI-Brookings Joint Center for Regulatory Studies.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 123–146

Carbon Markets 15 Years after Kyoto: Lessons Learned, New Challenges

Richard G. Newell, William A. Pizer, and Daniel Raimi

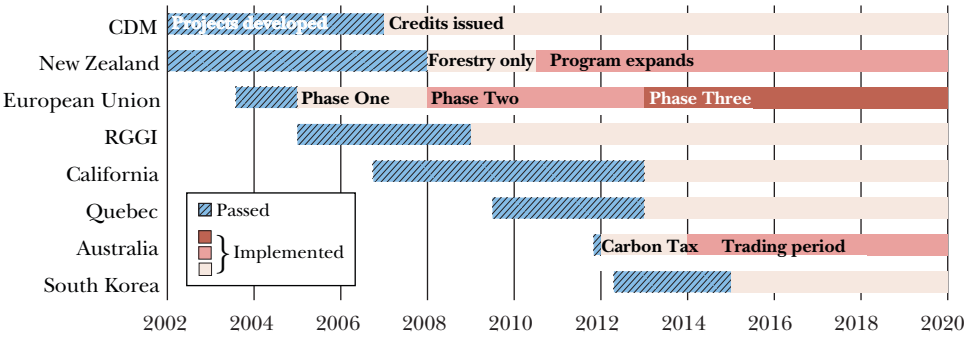
Back in 1997, when 37 industrialized countries and the European Community committed themselves to reducing emissions of carbon dioxide and other greenhouse gases as part of the Kyoto Protocol, the public debate focused largely on how to design a single global market for trading carbon permits as "the" vehicle to address global climate change. Because one ton of a greenhouse gas emitted anywhere in the world has the same climate change consequences for everyone, a single global market would be an economically desirable outcome, equalizing incentives to reduce emissions everywhere. However, this late-1990s dream of a top-down global design now seems far away, if not impossible.

Instead, we see a multiplicity of regional, national, and even subnational markets emerging, most notably the Emissions Trading System set up by the European Union in 2005, but also including the Regional Greenhouse Gas Initiative in the northeastern United States, the New Zealand Emissions Trading Scheme, and (on the horizon) California, Quebec, Australia, and South Korea, as shown in Figure 1.

■ *Richard G. Newell is the Gendell Associate Professor of Energy and Environmental Economics, Nicholas School of the Environment, and Director of the Duke University Energy Initiative, both at Duke University, Durham, NC. He is also a Research Associate at the National Bureau of Economic Research and a Member of the Board of Directors of Resources for the Future. William A. Pizer is Associate Professor, Sanford School of Public Policy, and Faculty Fellow, Nicholas Institute for Environmental Policy Solutions, both at Duke University, Durham, NC. He is also a University Fellow at Resources for the Future, a Research Associate at the National Bureau of Economic Research, and a Nonresident Fellow at the Center for Global Development. Daniel Raimi is a Research Analyst, Duke University Energy Initiative, Durham, NC. Their email addresses are richard.newell @duke.edu, william.pizer@duke.edu, and daniel.raimi@duke.edu.*

http://dx.doi.org/10.1257/jep.27.1.123. doi=10.1257/jep.27.1.123

Figure 1
Timeline for Selected Greenhouse Gas Emissions Trading Programs



Source: Authors.
Notes: "CDM" stands for the Clean Development Mechanism, which was set up as part of the Kyoto Protocol. "RGGI" stands for the Regional Greenhouse Gas Initiative, which operates in the northeastern United States.

The Clean Development Mechanism, set up as part of the Kyoto Protocol, has its own significant market for certified emission reductions undertaken by developing countries that can be used for compliance in other programs.

Thus, unlike back in 1997, we now have experience with actual carbon markets. Carbon markets are now the largest class of environmental or emissions trading markets in the world, in terms of both volume and market value, by a very wide margin. (Although other greenhouse gases may be included, we use the term "carbon market" because carbon dioxide is the dominant gas in terms of its overall contribution to global warming and because the units of trade are always denominated in terms of "carbon dioxide equivalent.")

This turn of events raises interesting questions of why the Kyoto model has not panned out, and why a growing number of jurisdictions nonetheless continue to pursue emission reductions in the absence of an agreement among all major emitters to reduce emissions. We will not try to answer those questions here, but we direct interested readers to work by Aldy and Stavins (2007) on international climate architectures, and Victor (2008) and Nordhaus and Boyer (1998) on problems with the Kyoto approach. Instead, we want to focus on what we have learned about the design and operation of carbon markets, and what new challenges we face.

In the next section, we begin with an overview of the major existing carbon markets (along with several incipient markets) and some of their key design features. With this background in place, we then spell out a number of lessons gleaned from the functioning of these markets—lessons about the reductions in carbon emissions; effects on end-users of energy; the risk of "leakage" of carbon emissions to jurisdictions not included in the carbon market; reducing the risks of overly high or volatile prices for carbon allowances; the role for banking of emissions credits; the

role for "offsets," which reduce emissions among unregulated sources; and the role for government regulatory oversight of these new financial markets.

The growth of a multiplicity of carbon trading programs has also raised questions that were not fully anticipated or understood during the design stages of existing carbon market systems. Now that these separate carbon markets exist, how might they be linked? How should carbon markets address the inevitable need for occasional changes in the underlying government-set rules? As countries approach carbon abatement with a mixture of different policy tools—an emission trading program, an emission tax, a performance standard, or traditional regulation—how can the overall intensity of different countries' abatement efforts be compared? In the decentralized, bottom-up carbon market and climate policy landscape that is emerging, how can international negotiations best contribute to further progress?

The importance of understanding the strengths and weaknesses of carbon markets as they emerge is enormous, in both environmental and economic terms. Carbon dioxide is a fundamental product of the combustion of fossil fuels, and fossil fuels are the source for over 80 percent of US and global energy consumption. More than 30 billion metric tons of carbon dioxide per year are emitted globally from fossil fuel combustion (Boden, Marland, and Andres 2011). The market value of one year of allowances for these emissions at $10 per ton of carbon, for example, would be $300 billion; at $25 per ton it would be $750 billion.[1] For higher allowance prices, or aggregating across several vintages of allowances, the potential value of these hypothetical allowances is easily in the trillions of dollars. Whether these numbers are taken to represent the value of the environmental impact of carbon dioxide emissions or the potential shifts in wealth as those emissions are constrained and property rights conveyed, the numbers are large. Moreover, the lessons from carbon markets could be relevant elsewhere as market mechanisms are applied to tackle other environmental and nonenvironmental problems.

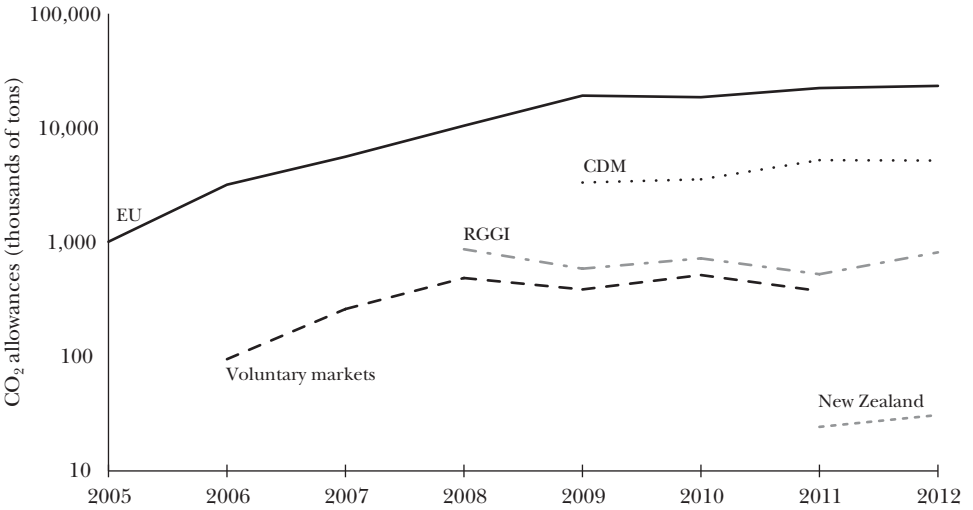
The Current Status of Carbon Markets and Some Key Design Choices

As of the end of 2012, the vast majority of carbon markets around the world took place in five arenas (each of which will be discussed below): the European Union's Emissions Trading System (ETS); the Clean Development Mechanism (developed under the Kyoto Protocol); the Regional Greenhouse Gas Initiative (northeastern United States); New Zealand's Emissions Trading Scheme; and voluntary markets.

The volume of trades in these markets is shown in Figure 2. The vertical axis showing the volume of trades is a logarithmic scale, and the figure demonstrates that the European Union's Emissions Trading System (ETS) has to this point

[1] It would take us far afield to discuss climate change impacts and the many challenges of measuring mitigation benefits, but recent estimates by the US government suggesting a net present value of expected global benefits of roughly $25 per ton of carbon dioxide reduced provide a useful reference point (Interagency Working Group on Social Cost of Carbon 2009).

Figure 2
Volume of CO_2 Allowance Trades
(daily average)



Source: Authors.
Notes: As of the end of 2012, the vast majority of carbon markets around the world took place in five arenas: the European Union's Emissions Trading System; the Clean Development Mechanism (developed under the Kyoto Protocol); the Regional Greenhouse Gas Initiative (northeastern United States); New Zealand's Emissions Trading Scheme; and voluntary markets. The volume of trades in these markets is shown in the figure. Exchange-traded volumes are through June 30, 2012 as reported by Point Carbon (http://www.pointcarbon.com/), RGGI CO_2 Auction Tracking System (https://rggi-coats.org/eats/rggi/), Ecosystem Marketplace/Bloomberg New Energy Finance (http://www.ecosystemmarketplace.com/pages/dynamic/our_publications.landing_page.php). Our voluntary market data is based on year-end reports, and thus we have no data for 2012.

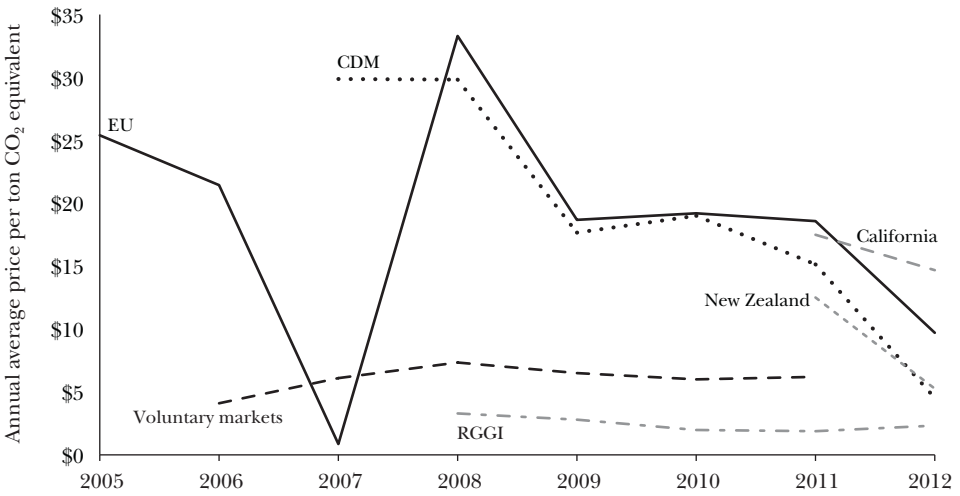
dominated the marketplace, with far greater volumes and liquidity than any other market. Volumes have been increasing, both in terms of activity within markets as well as the creation of new markets.

The average annual price per ton of carbon dioxide is shown in Figure 3. Carbon prices in all markets have been falling since 2008 in response to the global recession. Figure 3 also includes information on futures contracts in California (whose first compliance period begins in 2013). The following discussion provides an overview of major carbon markets and mentions other carbon trading programs that are scheduled to begin operating in the next year or so; for more details on other proposed and existing programs, Hood (2010) is a useful starting point.

European Union Emissions Trading System

The European Union has created by far the world's largest market-based system to reduce greenhouse gas emissions: the Emissions Trading System, which began operating in 2005 (total emissions under the cap were roughly 2.1 billion

Figure 3
CO_2 Allowance Prices
(nominal)



Source: Authors.
Notes: "CDM" stands for the Clean Development Mechanism, which was set up as part of the Kyoto Protocol. "RGGI" stands for the Regional Greenhouse Gas Initiative, which operates in the northeastern United States. Exchange-traded prices are through June 30, 2012 as reported by Point Carbon, RGGI COATS, Ecosystem Marketplace/Bloomberg New Energy Finance. Our voluntary market data is based on year-end reports, and thus we have no data for 2012.

metric tons in 2011). The program has operated in phases, with a pilot phase from 2005–2007 covering the power sector and certain heavy industry, a second phase from 2008–2012 expanding coverage slightly, and a third phase set for 2013–2020 that will add a significant range of industrial activity. Under the first two phases, each of 27 EU nations (later expanded to include Norway, Iceland, and Liechtenstein) submitted National Allocation Plans for total emissions of greenhouse gases to the European Commission. Once the plans were finalized, nations had significant discretion over how to distribute emissions credits to different sectors of their economies (Ellerman, Convery, and de Perthius 2010; European Commission 2012a).

The pilot phase from 2005–2007 was something of a test. Modest emissions reduction goals were enacted, but the primary goal of the pilot phase was to prepare for 2008, when the program would help the EU comply with its obligations under the Kyoto Protocol. The vast majority of allowances were allocated free of charge in the pilot and second phases, and each nation determined the level and distribution of free allocation to different sectors of the economy. These national-level plans also specified the number of offset credits emitters in each nation could purchase from carbon abatement projects in developing economies through the Clean Development Mechanism (discussed in the next section), with limits ranging from 0–20 percent of each firm's eventual compliance obligation.

Additionally, the European system is the only one where a significant secondary market for carbon has developed, with market participants buying and selling standardized contracts up to five years in advance on a variety of exchanges. While trading in the European system began mostly with nonstandardized over-the-counter transactions, exchange-based trading likely surpassed over-the-counter volumes sometime in 2008, indicating increased levels of standardization and liquidity (Ellerman, Convery, and de Perthius 2010).

The European Union system has evolved in a number of important ways as it enters its third phase in 2013. First, rules for distributing allowances have become more harmonized across the EU, with national-level plans now being largely a thing of the past. Second, the program has expanded to cover additional sectors of the economy, such as aviation and petrochemicals, along with additional greenhouse gases, such as nitrous oxide from certain industrial activities.

Probably the biggest hiccup for the Emissions Trading System to this point is visible in the price data in Figure 3: namely, in 2007 the price of carbon emissions collapsed to essentially zero. This situation was created by a confluence of several factors. First, the goals for emission reduction in the pilot program were constructed under time pressure with a shortage of reliable data and were supposed to be relatively modest (Ellerman, Convery, and de Perthius 2010). Second, aggregate emission data was unavailable until almost halfway through the pilot program, and when the first tranche of actual emissions data was released in 2006 by the EU Commission, market participants realized aggregate emission levels were low vis-à-vis allowance supply. Third, emissions allowances in this pilot first phase of the program could only be used between 2005 and 2007 and could not be further banked. The too-late realization of oversupply coupled with an inability to use excess allowances sparked a dramatic fall in prices. The rationale for not allowing banking was the desire to separate Phase II (which coincided with the first Kyoto compliance period starting in 2008) from the pilot program period—but the consequences of this decision were clear: by the final quarter of 2007, spot prices were essentially zero, at €0.06/ton, even while contract futures prices for Phase II allowances hovered above €20/ton (Point Carbon 2012). Banking is now allowed in all current and future phases.

The Clean Development Mechanism

The Clean Development Mechanism (CDM) is not a cap-and-trade program, per se, but a vehicle for translating emissions reduction efforts in developing countries into credits that can be used to offset capped emissions elsewhere. In 2011, roughly 300 million tons of offsets were issued under the CDM. The CDM was created as part of the 1997 Kyoto Protocol in order to provide additional flexibility for industrialized countries to meet their specified targets (United Nations Framework Convention on Climate Change 2012b). Credits generated through the CDM, called Certified Emission Reductions, now represent the second-largest market of carbon-denominated assets and are being used as offsets in a variety of jurisdictions. (A related but smaller program called Joint Implementation was created

for emission reduction projects in the former Soviet Union and Eastern Europe. Discussions of CDM sometimes include this program as well.)

The number of proposed and implemented Clean Development Mechanism projects has grown substantially over the past five years. Over 6,200 CDM projects have been approved and more than 1 billion offset credits have been issued (authors' analysis of data from CDM/JI Pipeline, http://cdmpipeline.org). The distribution of offset credits was slanted heavily in early years towards a small number of projects that reduce industrial gases with massive global warming effects. (This focus turned out to be problematic, an issue we discuss in our "lessons" section.) The distribution of projects overall has been led by renewable energy such as wind, solar, or biomass, a trend that has only increased in recent years. As the industrial gas projects with large numbers of credits per project have become more limited, most credits are now issued for renewable energy, energy efficiency, and projects that capture fugitive methane emissions from landfills and other locations. However, the winding down of these high-volume industrial gas projects has also led to a decline in overall issuance of CDM credits since 2007 even as the volume of projects continues to rise.

The European Union Emissions Trading System has been the main purchaser of Clean Development Mechanism credits. Their use, however, is limited by regulations to a fraction of each member state's cap (and, in turn, the compliance obligations of each facility). In aggregate, use of Certified Emissions Reduction credits for compliance across all EU nations from 2008 through mid 2012 was roughly 6 percent of their total compliance obligation (based on the authors' analysis of data from European Environment Agency, at http://www.eea.europa.eu/data-and-maps/data/data-viewers/emissions-trading-viewer). In New Zealand, emitters purchase CDM credits for up to 100 percent of their compliance obligation, and in Australia, the relevant figure is 12.5 percent (with up to another 50 percent coming from domestic offset programs). In Japan, the government has purchased over 100 million CDM credits to reach its target under the first round of the Kyoto Protocol; collectively, governments are expected to purchase roughly one-third of total CDM credits through 2012 (World Bank 2012). The CDM's projects and rules continue to evolve. Currently, a variety of project types face review from the UN body overseeing the CDM as well as the European Union, which does not accept CDM credits generated from certain project types.

Regional Greenhouse Gas Initiative

In 2005, seven northeastern US states became the first collection of jurisdictions in the United States to agree to an emissions trading program: Connecticut, Delaware, Maine, New Hampshire, New Jersey, New York, and Vermont. Maryland joined in 2006, Massachusetts and Rhode Island joined in 2007, and New Jersey withdrew at the end of 2011. Total capped emissions were roughly 150 million metric tons in 2011. Known as the Regional Greenhouse Gas Initiative (RGGI, pronounced as "Reggie"), this program only covers large electricity generators, and seeks to reduce emissions from this sector by 10 percent below 2009 levels by 2018. Revenues from allowances—almost all of which are auctioned—go to state governments, which have invested

most revenues in local renewable energy or energy efficiency projects, while roughly 18 percent of revenue goes to state deficit reduction. Offsets for emitters are limited to just 3.3 percent and must come from projects within RGGI states, although no offsets have been used for compliance to date (Regional Greenhouse Gas Initiative 2012b).

After several years of operation, the program has exceeded its initial reduction targets, largely due to fuel switching from coal-fired power to low-priced natural gas. In 2011, overall emissions were 33 percent below the program cap (Regional Greenhouse Gas Initiative 2012a). Allowance prices have not collapsed, thanks to an established floor price in the allowance auctions. However, allowances have gone unsold in the auctions and have generally traded at roughly the floor price during this time.

Voluntary Markets

Voluntary carbon markets refer to a variety of organizations that allow individuals or businesses to purchase offsets from emissions reduction projects located around the world. Since 2002, voluntary markets have grown from $43 million in revenues to a peak of $705 million in 2008, and stood at $572 million as of 2011 (Ecosystem Marketplace and Bloomberg New Energy Finance 2008–2012). Estimated reductions in 2011 were 95 million metric tons. Dozens of organizations offer voluntary carbon offsets, with a variety of procedures and standards for monitoring and verification of carbon reductions.

One important issue for these markets is that their standards for evaluating and monitoring greenhouse gas reduction projects are typically less stringent than, say, the Clean Development Mechanism. On one side, less stringent standards reduce bureaucracy and lead to lower project costs. On the other side, weaker standards could also lead to certification of projects that do not provide their stated benefits (Benessaiah 2012).

New Zealand Emissions Trading System

New Zealand launched an emissions trading program in 2008 that by 2011 covered roughly 32 million metric tons. The program will eventually cover almost all New Zealand emissions, with caps based on New Zealand's 2008–2012 commitment under the Kyoto Protocol. Since New Zealand is a small economy, the program was built around the idea of linking to other markets; this initially includes the Clean Development Mechanism but could be expanded to other national or regional carbon markets (such as the European Union or Australia). This feature has made the program vulnerable to international policy uncertainty and to issues surrounding the CDM. The program covers a relatively small number of large emitters who must reduce emissions, purchase domestic or international offsets, or pay $25 (New Zealand) per ton of emissions. The program has no price floor and prices have steadily declined through 2012, generally following the movements of CDM prices. Industries facing international competition, horticulture, and fishing receive up to 90 percent free allocation, but the power sector, transportation, and forestry do not receive free allocations (New Zealand Government 2012).

California, Australia, and Others

Two new carbon markets are just in the process of emerging and gaining experience. First, in California, a new cap-and-trade program held its first auction in December 2012 in advance of its 2013 start date (over-the-counter contracts have been traded since at least December 2011). The trading program will initially cover the power sector and some heavy industry, with two-thirds of allowances auctioned. By 2015, it will expand to cover transportation fuels and auction 80 percent of allowances. Emitters may meet up to 8 percent of their obligations through approved domestic offsets and in the future possibly through international forestry offsets. Given California's stature as the world's sixth-largest economy, this is a significant new market with annual capped emissions of roughly 160 million metric tons in 2013 and roughly 400 million metric tons by 2015.

In Australia, after a long and contentious political process, a 2011 law passed that requires an emissions trading program to begin in 2015. In the meantime, major carbon emitters will pay a steadily increasing carbon tax set by the legislation, though many large emitters will receive government support in the form of a large share of free allowances (Australian Government 2012). Much of the government revenues from the tax and subsequent auctions will go towards new spending on energy efficiency, renewables, and technology programs, and at least half of the revenues will go towards increased pension payments, increased tax credits, and decreased income taxes for households. However, the opposition Liberal party has made repealing the carbon price "the top priority" on its agenda, calling into question the policy's viability moving forward (Australia Liberal Party 2012).

Carbon markets also exist at smaller scales, and some large ones are brewing in other jurisdictions. The Canadian province of Quebec has developed a market which will link with California. Recent legislation passed in South Korea and Mexico has laid the groundwork for new national-level programs beginning in 2015. China has established a series of regional pilot programs, while other programs under discussion in India, Japan, Vietnam, and Thailand indicate an interest in cap and trade across much of Asia. Other emissions-trading proposals are currently under discussion or development in Brazil and Chile, among other places (Hood 2010).

Lessons from the Early Carbon Markets

Emissions Fall, But How Much is Unclear

The presence of a consistent and significantly positive price on carbon suggests that these trading programs should be having at least some effect on behavior that reduces emissions levels, but research on the extent of these reductions remains limited.

One way to approach the abatement question is to estimate emissions reductions based on elasticities derived from related policy simulations. A rough analysis of projections from the emissions trading program in the proposed US Waxman–Markey 2009 legislation suggests that for each $10/ton increase in the price of

US carbon dioxide allowances, emissions from 2012–2015 would fall between 1.5 to 6 percent compared with a scenario with no price on carbon dioxide emissions.[2] If similar economic dynamics are at play in Europe's Emissions Trading System, an allowance price of $16/ton (the Phase I average) would suggest that the program resulted in reductions of 2–9 percent compared with business as usual. Indeed, empirical research on Phase I of Europe's ETS suggests that during 2005–2007, emissions fell by 2–5 percent compared with business as usual (Ellerman, Convery, and de Perthius 2010).

A key question for—and sometimes criticism of—current market-based policies concerns the degree to which they encourage long-term investment in new technologies rather than solely short-term fuel-switching and energy conservation. Early research into Europe's Emissions Trading System suggests that such long-term investments may be limited (Leiter, Paolini, and Winner 2011). However, carbon markets may be still too new to inspire the long-term confidence to make those investments.

Allowance Allocation in the Power Sector Can Involve Important Distributional Effects

Emission allowances can be auctioned, allocated for free, or some combination of the two. There are both distributional as well as efficiency consequences to allowance allocation, and these can be substantial given the sizable economic rents at stake. The power sector is a particularly important area of concern because of its large share of emissions, its universal inclusion in all existing programs, and the complexity of both power markets and the power market regulation influencing the distribution of costs.

In deregulated power markets where fossil-fueled generation tends to be the marginal producer and to set the market price, economists would expect competitive pressure to lead power prices to reflect the price that is placed on carbon content, regardless of any free allocation. Consequently, end users of electricity would ultimately end up paying for compliance costs. In Germany, for example, power generators received carbon allowances for free, and then passed along the opportunity costs of these free allowances to their customers, allowing generators to extract rents roughly comparable to their proportion of freely allocated allowances (Ellerman and Joskow 2008; Sijm, Hers, Lise, and Wetzelaer 2008; Ellerman, Convery, and de Perthius 2010). This market outcome was completely predictable, though not warmly received by the public (Gow 2006; Harrison 2009).

There are several possible responses to concern over the costs of free allowances being passed along to consumers in the form of higher prices. After providing substantial free allocations early on, the European Commission has more recently limited free allocations of carbon allowances to electricity generators, and it will sharply increase the proportion of allowances sold at auction in its third phase (European Commission 2012c). Other programs have varied in their approach,

[2] Authors' analysis of data from Energy Information Administration, "Energy Market and Economic Impacts of H.R. 2454, the American Clean Energy and Security Act of 2009." Similar ranges can be estimated from other analyses.

with New Zealand giving no free allocations to the power sector and the Regional Greenhouse Gas Initiative giving very few, while Australia will give substantial but temporary free allocations to its coal-heavy power fleet.

When free allocations in the power sector are eliminated, governments take the impact of the emissions trading program on consumer power prices as given and redistribute the rents from auctioning in a more acceptable manner. The opposite approach is to try to limit the higher prices to consumers. For example, cost-of-service regulation could prevent generators from passing through the opportunity costs of carbon permits to consumers even with free allocations to generators. In California's program, free allowances will be provided to the power sector on the condition that they use those allowances to reduce costs for ratepayers. Other proposals to direct free allocations to local power distribution companies or to pursue tradable performance standards, instead of cap and trade, reflect similar efforts to alter the distributional impacts on electricity consumers (for example, Aldy 2011). By limiting the effect on consumer power prices, however, all of these approaches reduce the incentive to conserve electricity.

Part of the motivation for depressing consumer power prices is that carbon pricing, or anything else that raises power prices, disproportionately harms low-income households (Hassett, Mathur, and Metcalf 2009). Rather than limiting the increase in power prices, a number of mechanisms have been proposed to ameliorate the regressiveness of carbon pricing, including lump-sum rebates to households (so-called "cap-and-dividend") and parallel offsetting changes to income or social security taxes (Burtraw, Walls, and Blonz 2010). From an efficiency point of view, these are better compensation mechanisms vis-à-vis depressing power prices and, in the case of tax reform, take advantage of opportunities to lower distortionary taxes.

Significant Competitiveness Effects and Emissions Leakage Have Not Yet Emerged

Another motivation for depressing the impact of carbon pricing on energy prices has been the concern that emission-related activities, particularly energy-intensive industries facing outside competition, will relocate to an unregulated jurisdictions when faced with an emissions trading program that raises production costs. This concern involves an environmental angle—that emission reductions are simply being shifted outside the boundaries of the trading program—referred to as emissions "leakage." It also involves an economic angle—that local industries are being harmed to the advantage of industries abroad, who can be viewed as skirting their environmental responsibilities (Jaffe, Peterson, Portnoy, and Stavins 1995). Facing the practical constraint of a less-than-global response to a global externality, efforts to limit price changes and leakage through various allocation incentives may even be cost-effective (Fischer and Fox 2009). Rather than depress local price increases, programs could also attempt to adjust foreign prices at the border, although this approach raises controversial legal and practical issues (van Asselt and Brewer 2010).

Even without explicit efforts among existing programs to depress carbon-related energy price increases, significant competitiveness impacts and leakage have yet to

emerge. For the early phases of Europe's Emissions Trading System, a (limited) empirical literature indicates that competitive losses appear to have been small. Ellerman, Convery, and de Perthius (2010) found "no observed impact" on competitiveness in the oil refining, cement, aluminum, or steel sectors during Phase I. Demailly and Quirion (2008) found that Phase I created only a small loss of competitiveness in the iron and steel sectors. Lacombe (2008) found a similar limited impact on the refining sector during Phase I. An analysis of Europe's aluminum sector by Reinaud (2008) found no statistical evidence of negative competitiveness impacts from the program. The only countervailing evidence comes from a survey of firm managers (215 respondents across all affected industries in the European Union) where 55 percent of metals manufacturers and 44 percent of pulp/paper and cement/lime/glass manufacturers stated they have either moved or are considering moving out of Europe's carbon market compliance zone; 14 percent of the remaining firms stated they have moved or are considering such a move (Point Carbon 2011).

These observed competitiveness impacts generally fall below the levels predicted by some earlier analyses (Aldy and Pizer 2008; Ho, Morgenstern, and Shi 2008; Interagency Competitiveness Analysis Team 2009). This may reflect the modest targets for greenhouse gas reduction implemented in the first phase of the European System. It may also reflect the consequences of free allocation to many energy-intensive industrial sectors. Despite the above-noted trend towards auctioning allowances in the power sector, these industrial sectors continue to receive significant free allocations. For a detailed discussion of this issue, see Ellerman, Convery, and de Perthius (2010, chap. 4).

Limited evidence also suggests that leakage in the Regional Greenhouse Gas Initiative has been small, despite some early concerns and analysis to the contrary. Some research had suggested leakage rates could range from 28 percent with $3/ton prices to 90 percent with $7/ton prices (Chen 2009; Wing and Kolodziej 2009). However, low carbon prices resulting from a weak economy and historically low natural gas prices appear to have prevented extensive leakage in RGGI (Kindle, Shawhan, and Swider 2011).

A Variety of Tools Can Be Used to Manage Concerns about Costs and Volatility

Newly started carbon markets face substantial uncertainty over costs and, even though many markets have seen low prices in 2012, program designers still seek to prevent the risk that allowance prices might exceed economically and politically tolerable levels. Research on climate policy instrument choice under uncertainty also suggests that policies exhibiting stable prices and less-certain emissions, as typically associated with a carbon tax, have higher expected net benefits than policies where emissions are fixed and prices fluctuate—as in a rigid cap-and-trade system (Pizer 2002; Newell and Pizer 2003).

Carbon trading programs have typically turned to one or more of the following three types of cost management. First, regulators can impose a price ceiling, allowing emitters to purchase unlimited (or a relatively large volume of) allowances directly from the government at the ceiling price. For example, participants

in the California and Quebec programs will be able to purchase credits from the government for $40–$50/ton, essentially capping trading prices (Western Climate Initiative 2012). Australia has established a carbon tax for the first two years of their program, allowing unlimited emissions at $23 (in Australian dollars) and placing a binding cap on emissions only in 2015 (Australian Government 2012).

Second, regulators can employ price floors to prevent market prices from falling below a certain level. Auction price floors—where allowances are kept out of circulation unless purchasers are willing to pay a minimum price—have been used in the Regional Greenhouse Gas Initiative and in California, and are part of anticipated programs in Australia and Quebec. In California's November 2012 allowance auction, for example, only 14 percent of the 2015 allowances sold at the minimum price of $10, leaving 86 percent unsold (California Air Resources Board 2012). Price floors clearly reduce cost uncertainty by limiting low-cost outcomes. But in limiting the possibility of very low prices, these mechanisms can unlock opportunities for negotiation on other features—such as the cap, offset provisions, and/or price ceilings—to reduce the possibility of high costs. As we have seen in the Regional Greenhouse Gas Initiative, price floors can continue to provide an incentive for emissions reductions even if the imposed cap is not binding. Supporting these efforts, theoretical work has showed that price-like modifications within a cap-and-trade program—ceilings and floors on the allowance price or otherwise adjusting the cap to accommodate cost shocks—can help to achieve the same outcomes as a carbon tax, where the cost is certain (Newell, Pizer, and Zhang 2005; Murray, Newell, and Pizer 2009).

A third approach is to allow high carbon market prices to trigger provisions that relax constraints of the program other than the cap itself. If carbon prices reach $7/ton, for example, the Regional Greenhouse Gas Initiative allows emitters to purchase more carbon offsets to meet their compliance needs than is otherwise allowed. If prices reach $10/ton, emitters may purchase still more offsets to reach their targets. Unlike explicit expansion or contraction of the emission cap through allowance sales at a fixed price, the exact impact of these mechanisms is less transparent. On the one hand, the capacity of offset markets to expand in response to newly triggered RGGI demand may not be sufficient to ward off higher prices. On the other hand, if offset markets do respond quickly, prices could spike then fall, creating additional volatility.

The Flexibility to Trade Allowances over Time—Banking and Borrowing—Can Smooth Uncertain Cost Shocks with Minimal Environmental Consequence

Emissions of carbon dioxide and most of the other greenhouse gases remain in the atmosphere for decades if not centuries, and the accumulated stock of such emissions is what leads to environmental problems. In other words, the timing of emissions in terms of day, month, or year is not consequential for climate impact. This intuition lies behind the aforementioned preferences for stable prices. Allowing flexibility through banking or borrowing of allowances across time, even without turning to price floors and ceilings, can smooth out prices and costs, increasing

cost-effectiveness without additional harm to the climate (Fell, MacKenzie, and Pizer 2012).

In this way, banking carbon allowances can be a partial response to concerns about uncertain costs (although the problem remains if costs are headed permanently higher or lower). Without trading between periods, cost shocks have to be absorbed immediately. Europe's experience during the first phase of its Emissions Trading System, which did not allow banking, provides a prime example. Facing unexpectedly low compliance costs, prices for carbon allowances collapsed. Unlimited banking is now allowed in all carbon trading programs, though few allow borrowing. An emerging question is how much banking an emissions trading system can (and perhaps should) support. For example, recent estimates suggest market participants in the European Union system are banking nearly 2.5 billion allowances, roughly 119 percent of Phase II's annual cap, for carryover into Phase III (Neuhoff, Schopp, Boyd, Stelmakh, and Vasa 2012). For reference, the US sulfur dioxide cap-and-trade program saw banking levels of over 6 million tons in 1998, or roughly the volume of the annual cap (Stavins 1998).

Policies that allow banking of carbon allowances do face some challenges. Banking links expectations over time, so prices today depend on expected prices tomorrow. Depending on the government's level of commitment to the policy and the public's perception of that commitment, this can be a good or bad thing. (We return to the issue of future policy adjustments below.) Recent low prices in Europe, for example, have been linked to questions about whether an aggressive renewables policy will depress carbon prices in the future (Grubb 2012).

Another issue raised by the potential movement of allowances across time is the trading ratio that should be applied to banked or borrowed allowances, and how this rate should be applied. Theory suggests that the optimal trading ratio between periods is equal to one plus the discount rate, minus the desired rate of change in permit prices (Leiby and Rubin 2001). In addition to this formula, a discount rate is required, which raises a distinct set of analytical challenges, for both the estimation of damages and the rate at which the carbon price should rise (Interagency Working Group on Social Cost of Carbon 2009; Aldy et al. 2010; National Research Council 2010). In practice, banking has faced a trading ratio of unity, sometimes coupled with very limited amounts of borrowing. Where allowed, large-scale borrowing has typically faced a trading ratio equivalent to one plus a discount rate (in Kyoto, this was a trading ratio of 1.3 over five years; under the Waxman-Markey legislation, the discount rate would have been 8 percent per year).

Offsets Can Provide Low-Cost Mitigation Options, but Raise Complex Issues

Offsets allow mitigation activities outside a cap-and-trade system to count against the cap, expanding the scope of potential responses and thereby lowering costs. Developing-country emissions offsets, in particular, offer a very large potential pool of inexpensive compliance opportunities for industrialized nations, relative to reducing greenhouse gas emissions within their own borders (Weyant and Hill 1999). Domestic or local offsets can also offer cost savings while keeping investments and cash flow

at home, but represent a smaller universe of activities compared to international offsets. Although specific provisions and restrictions vary, all programs to date employ offsets in some capacity. However, difficulties arise in assuring that offsets provide actual reduction in emissions and that the subsidy effect from offset crediting is not creating perverse outcomes. In addition, as financial flows to offset projects grow, attention can shift to questions of distribution as well as efficiency.

For offsets to reduce emissions, credits can only be given to projects (and for measurable reductions) that would not have occurred without the offset credit program. At the same time, rigorous screening creates transaction costs that eat into potential cost savings. In practice, offset programs must strike a balance, and a variety of approaches have emerged (Hall 2007). As the world's largest offset program, the Clean Development Mechanism has pioneered many of these approaches, and considerable research indicates that it has resulted in real emission reductions taken as a whole. However, it is easy to find subcategories of projects where researchers question whether the reductions were real (Lambert 2011; Zhang and Wang 2011).

The most problematic example from the Clean Development Mechanism involves HFC-23, a compound produced primarily as a by-product in the production of refrigerants in developing countries. As a by-product, HFC-23 is typically vented to the atmosphere where it has roughly 10,000 times the global warming potential (ton-for-ton) compared to carbon dioxide (United Nations Framework Convention on Climate Change 2012a). Because of its high global warming potential, projects that destroy the HFC-23 by-product receive large amounts of credits—enough to make it profitable to increase operations that emit HFC-23 in the present, just to destroy more HFC-23 in the future. Lambert (2011) found evidence of such behavior in the first few years of the program, leading the European Commission (and later Australia and New Zealand) to disallow such credits and encouraging the CDM itself to revise its guidelines concerning HFC-23 and similar gases.

Forestry offset projects have also been a particularly thorny issue, as carbon stored in stands of trees is—by its nature—difficult to guarantee and deforestation avoided in one area can easily crop up in another. Yet the allure of preserving forests while sequestering carbon has attracted tremendous interest. Currently, California allows certain forestry offsets, as do many voluntary programs. The European system does not allow forestry or land-use change projects (Kim et al. 2008). New Zealand has included domestic forestry under its national cap. The complexity in dealing with forestry has led to a variety of proposals that continue to be debated (Murray, Galik, Mitchell, and Cottle 2012).

As the use of international offsets grows, a variety of distributional questions can arise due to the transfer of resources from higher-income nations to developing nations. Through 2011, by far the largest share of projects and credits were going to China and India. In fact, between 2006 and 2011, over half of each year's Clean Development Mechanism credits went to projects in China—topping out at 75 percent in 2007 (data from CDM/JI Pipeline http://www.cdmpipeline.org). Since China is the world's largest carbon emitter, it is not surprising that a large share of carbon reduction projects would occur there. However, those nations or political stakeholders that

believe China should commit to a more stringent emissions reduction plan, or see China as a competitor, may object to the transfers enabled by the CDM (International Energy Agency 2012). More broadly, whether the CDM projects are meeting broader development objectives, such as economic growth or technology transfer, remains uncertain (Dechezlepretre, Glachant, and Menier 2008; Popp 2011).

One solution to the distributional issue is to focus international offsets on poorer countries; another is to emphasize regional or local offset programs. The Regional Greenhouse Gas Initiative, and California, Quebec, and Australia each encourage and accept for compliance regional or local offsets, and in the case of Australia, a domestic offset program that encourages emissions reductions on farms is a central tenet of the overall program (Australian Government 2012). Even if local offsets are more costly, they are sometimes favored by local authorities. For example, regional programs in North America have thus far given preference to offsets from regional or domestic emissions reduction projects—although these programs also allow offsets to make up just a small share of compliance.

It is impossible to conclude a discussion of offsets without at least noting the collapse of the Clean Development Mechanism market at the end of 2012. After remaining around €10–15 per metric ton for most of 2009, 2010, and the first half of 2011, CDM prices fell steadily to less than €1 in November and December 2012. This has been ascribed to increased limitations on the use of CDM credits in the European Union, uncertainty about future demand, and increasingly robust supply—issues that will need to be sorted out for investors to continue to have confidence in offset markets.

Market Monitoring and Oversight Must Be an Integral Feature of Cap-and-Trade Programs

After the 2008 financial crisis, virtually all financial markets came under new scrutiny. Carbon markets were no exception, and new proposals for trading programs in the United States came with calls for strong oversight. In fact, the 2010 Dodd–Frank financial reform and consumer protection bill created an interagency working group to conduct a study on maintaining and increasing transparency for carbon markets (Interagency Working Group for the Study on Oversight of Carbon Markets 2011). Similarly, an EU directive adopted in 2011 will significantly expand oversight of carbon markets (European Commission 2012b). Primary goals for market oversight include facilitating price discovery, ensuring transparency and access to information, and preventing manipulation or abuse in the marketplace. However, European governments have come under criticism for not releasing timely and detailed data on individual allowance trades and holdings (de Perthius 2011).

The European Emissions Trading System has faced three high-profile market controversies, two of which were not specific to emissions trading markets. One of these two cases involved traders manipulating value-added tax laws in different countries to defraud governments of over €1 billion from 2008–2009, while the second involved cyber-attacks which likely stole over €50 million worth of allowances on spot exchanges in 2011 (de Perthius 2011; Frunza, Guegan, and Lassoudiere 2011). The

one major controversy unique to emission markets occurred when Clean Development Mechanism credits previously collected by the Hungarian government for compliance re-entered the market. It appears that the Hungarian government simply swapped the CDM credits for another type of carbon asset under the Kyoto Protocol that it needed to sell. While the swap was legal under the Kyoto Protocol, it was surprising to many participants in the European Emissions Trading System and created the appearance of possible credit "recycling" that would have negated relevant carbon reductions and diminished the integrity of the trading system (de Perthius 2011). The European Commission has since revised its rules to address each of these concerns.

As for the Regional Greenhouse Gas Initiative, the program's independent market monitor has found no major irregularities since trading began in 2008 (Potomac Economics 2009, 2010, 2011). Market and auction data is released by RGGI regularly, and allowance holdings are traceable online through the program's CO_2 Allowance Tracking System (see http://www.rggi-coats.org for details).

The Future of Carbon Markets: New Issues

A more general lesson from the past decade is that climate policy and carbon markets are not static concepts, but are instead constantly evolving. The vision of a single, top-down global trading system has morphed into the reality of various national and subnational trading programs. These programs are themselves evolving over time as are views about the relative role of carbon markets vis-à-vis other policy responses. Against this backdrop, carbon markets face a variety of emerging issues.

Linking Carbon Markets

Front and center in the discussion of current carbon markets is how, whether, and when different markets can be "linked" so that regulated entities in one jurisdiction can use allowances or credits from another jurisdiction for compliance (Jaffe, Ranson, and Stavins 2009). It might seem as if linking two carbon markets must always be a universally positive step by adding additional flexibility for trading; but when carbon markets have certain characteristics, this conclusion is incorrect. For example, Fischer (2003) shows that linking a system that is indexed to output with an ordinary capped system almost always increases emissions. Researchers have begun to think about exactly which features have to be aligned to avoid such issues, and which do not (Mace et al. 2008).

In practice, linkages may be one-way or two-way (Mehling and Haites 2011). In a one-way linkage, credits in one system can be used for compliance in another, but not vice-versa. In a two-way linkage, both systems mutually allow the other's credits to be used for compliance.[3] It is also useful to think about even one-way linkages in

[3] Linkages can also be indirect: If system A links to B and B links to C, A will have an indirect linkage with C. For example, A's credits can be used for compliance in B, freeing up B's credits to move into C. The net result would be credits leaving A and entering C.

terms of buy-linkages and sell-linkages: that is, a buy-linkage represents the decision by one trading system to accept for compliance allowances or credits created and offered for sale by another system, while a sell-linkage represents a decision by one jurisdiction to allow or encourage other jurisdictions to use its allowances or credits for compliance. A complete two-way linkage really involves two distinct decisions about buying and selling by each of two jurisdictions.

Linkages among trading systems have proceeded relatively slowly so far, for three main reasons. First, buyers tend to be concerned about environmental integrity and so will be careful in accepting that purchased allowances are valid for compliance in their system (Mace et al. 2008). Second, the necessary harmonization of certain design features also means that one or the other system is giving up some sovereign control. The result often depends on who has more power in the linking negotiation, which is frequently a function of the relative market size. Currently, for example, the European Union set the terms for Norway, Iceland, and Lichtenstein to enter the Emissions Trading System. In a different model, emitters under Australia's program will be able to purchase European allowances overseen by the European Commission in 2015, while European emitters will be allowed to purchase allowances from Australia in 2018 (Reklev 2012). In North America, Quebec's program embraces many aspects of the California design and will likely soon link to that much larger market (Carroll 2012).

Third, distributional concerns tend to arise, particularly in the selling system. For the buying system, linking lowers allowance prices with the same environmental outcome—something many programs desire. The main downside is faced by investors holding allowances without any corresponding obligations, or the government in the case of auctions. But for the selling system, linking raises allowance prices for carbon allowances, increasing costs for those with compliance obligations as well as their downstream consumers. For this reason, Australia initially planned to restrict international sales of its allowances, despite the net gains from trade (Jotzo and Betz 2011).

New Information and Program Revision

One of the defining characteristics of climate change is uncertainty about both mitigation costs and benefits as economic conditions, technologies for carbon abatement, and scientific knowledge advance. Occasional revisions to carbon market policies are essential to long-term efficiency (Murray, Newell, and Pizer 2009). While markets and affected stakeholders may crave certainty, governments cannot guarantee certainty where it does not fundamentally exist. Carbon market policies are certain to be revised and even overhauled as time passes.

Carbon market policy revisions have the potential to create financial gains and losses in the carbon market. At any point in time, carbon market participants have both carbon assets in the form of allowances, and liabilities in the form of expected emissions. As expected carbon prices change, so do the balance sheets of these economic actors. While the same would be true for changes in carbon taxes, the existence of banking provisions—which link carbon prices over time through

the potential for arbitrage—imply that any change in future price expectations should also affect current prices. Like expected shifts in conventional regulation, expected changes in carbon market policies also affect incentives for investing in new, emission-related physical capital and technology, as well as the value of the current capital stock.

For example, as Europe's Emissions Trading System enters its third phase in 2013, it is reportedly considering a delay in auctioning a large share of allowances (roughly 900 million from 2013–2015); a delay would likely drive up prices until the auction date becomes certain (Allan 2012; Szabo 2012). Most European governments expected to gain from the plan, as higher prices offset lower auction volumes; Poland's government has opposed the delay, however, because its auction volume is small and it expects to lose revenue. The Regional Greenhouse Gas Initiative saw a decrease in the size of its market when New Jersey announced in May 2011 that it would withdraw at the end of the year (Christie 2011). However, prices were unaffected, perhaps in part because they were already trading at the established price floor. In New Zealand, rules were revised to allow only one allowance to be used for two tons of emissions (rather than one ton) during a transition phase (Fallow 2009). This did not substantially impact New Zealand allowances prices, which are closely tied to Clean Development Mechanism prices determined internationally, but halves the emission reduction incentive for New Zealand firms.

If the holders of allowances are largely the same agents who face compliance obligations, the net effect of price changes on firms' balance sheets could be relatively small, as the market value of allowances will fluctuate along with the cost of their future compliance obligation. However, the specifics of how allowances are valued on balance sheets can create problems even for these businesses. For allowance holders that have no compliance obligations, and for those with obligations but no allowances, the financial consequences of large price changes could be substantial.

Policy revisions cannot be avoided, but governments should strive to make them transparent and orderly. Regulatory agencies, courts, legislatures, and central banks all face the need to pursue market-sensitive decisions in a way that allows all market participants equal access to information as well as advance notice of the sequence and timing of the decision process. For example, one legislative proposal for a US carbon market (Low Carbon Economy Act of 2007, S. 1766, 110th Congress) would have implemented a specific schedule for periodic five-year reviews and revisions, with presidential submission of recommendations shortly after the compliance year ends and then expedited Congressional action within six months.

Another option might be to put these decisions into the hands of an oversight entity, similar to a central bank (Pizer and Tatsutani 2008); Newell, Pizer, and Zhang 2005). Such an entity would be responsible for periodic reviews and changes to the emission limit or other rules, and would have the flexibility to do so outside the explicitly political sphere. However, climate change is an issue with a continuing divergence of views about the appropriate level of response, even among experts, and the independence of an oversight entity cannot solve that problem.

Alternative Policies and Comparability

In addition to finding ourselves in a world of multiple emission trading regimes with varying rules, many jurisdictions are pursuing alternative policy approaches such as a carbon tax or more traditional regulation. For example, policy-related emission reductions in the United States over the past few years have arisen from tighter regulations on automobile fuel economy and tailpipe greenhouse gas emissions, renewable electricity capacity additions associated with federal and state subsidies and mandates, and new power plant emission regulations from the US Environmental Protection Agency. The European system only covers roughly half of European emissions, with traditional regulation used elsewhere (for example, with automobiles). Several European nations, such as Great Britain, Ireland, the Netherlands, and Norway, also apply carbon taxes to certain fuel types. Australia is temporarily using a carbon tax in advance of emissions trading.

This diversity of policy approaches was not altogether unexpected. Under the Kyoto Protocol, there is no requirement to use a carbon market as the sole tool to implement a domestic emissions reduction program. When the United States seemed closest to establishing its own cap-and-trade program in 2009 and sought to assuage domestic concerns about competitiveness, the proposed legislation asked other countries to have a "nationally enforceable and economy-wide greenhouse gas emissions reduction commitment for that country that is at least as stringent as that of the United States" without specifying emissions trading (see H.R. 2454, §767(c)(1), 111th Congress).

This diversity of approaches raises the need to measure the "comparability" of policies. Among jurisdictions with carbon markets, comparability is necessary for jurisdictions to consider linking. More generally, comparability among jurisdictions with and without carbon markets is necessary for countries to justify continued domestic action on a global problem and, more specifically, to avoid escalating concerns over competitiveness and emission leakage that could threaten the sustainability of policy actions. Most discussions look at emission reduction efforts in one of six ways: 1) emission reductions versus some year in the past; 2) reductions versus what would happen with a business-as-usual baseline; 3) reductions in emissions per unit of output (gross domestic product, energy use, power generation); 4) reductions in emissions per capita; 5) the realized carbon price; or 6) energy prices or price effects. There is no agreement on which metric is best, many raise practical issues like conversion of carbon prices among currencies or calculation of business-as-usual forecasts, and different metrics yield dramatically different messages. This question of comparability is compounded when evaluating actual implementation of policies and their outcomes, as opposed to pledges.

International Negotiations

Earlier, Kyoto-style negotiations focused on a sequence of top-down, larger-to-smaller emission trading issues—national emission caps, trading rules, and then further details, such as the Clean Development Mechanism. However, the new

negotiations in the aftermath of the Durban conference in late 2011 will necessarily focus on the tools for a bottom-up approach. On the one hand, a new agreement will need to support concerns over comparability and transparency of effort. Those countries already engaged in or pursuing carbon markets will want assurances that other jurisdictions will do their fair share.

A new international agreement also needs to focus on ways to provide institutional support for markets themselves. For example, some developing countries might benefit from "model rules" for establishing a domestic trading program that would presumptively link to developed country programs already utilizing the Clean Development Mechanism. While rules for carbon markets and other abatement programs can and may emerge organically without an anchor in international agreements, creating model rules could be valuable, particularly for the many countries that will be too small to pursue an entirely customized approach. There are also questions about the future of the CDM itself. Decisions in December 2012 will limit future access to the CDM to countries participating in the next phase (2013–2020) of the Kyoto Protocol. This approach steers the CDM away from a role in a decentralized global carbon market by limiting its relevance to the subset of Kyoto participants. To achieve efficiency, future negotiations should be creating opportunities for linkages, not blocking them.

Conclusion

Fifteen years after the signing of the Kyoto Protocol and the creation of the first major platform for carbon markets, the prospect for a unified global carbon trading system in the foreseeable future is essentially finished. However, carbon markets are a reality and the design of carbon markets is benefiting from actual experience. Experience with windfall profits from free allowance allocation has led to an increased use of auctions. Jurisdictions are learning to handle market-sensitive information in a more transparent and orderly manner, although progress remains to be made. Efforts to moderate both high and low prices are providing lessons on what works. Perhaps most importantly, we are seeing that carbon allowance trading can support emission reductions and send market signals for future investment. The challenge now is to figure out how carbon markets can work in a much more complex—but clearly more realistic—world.

■ *David Autor, Dallas Burtraw, Denny Ellerman, Chang-Tai Hseih, Suzi Kerr, John List, Timothy Taylor, Robert Stavins, and Jonathan Wiener provided invaluable comments on an earlier draft. A longer version of this paper is available online from the National Bureau of Economic Research (W18504) and Resources for the Future (DP-12-51).*

References

Aldy, Joseph E. 2011. "Promoting Clean Energy in the American Power Sector." The Hamilton Project Discussion Paper 2011-04.

Aldy, Joseph E., Alan J. Krupnick, Richard G. Newell, Ian W. H. Parry, and William A. Pizer. 2010. "Designing Climate Mitigation Policy." *Journal of Economic Literature* 48(4): 903–34.

Aldy, Joseph E., and William A. Pizer. 2008. "Issues in Designing U.S. Climate Change Policy." Resources for the Future Discussion Paper 08-20.

Aldy, Joseph E., and Robert N. Stavins. 2007. *Architecture for Agreement.* Cambridge University Press.

Allan, Andrew. 2012. "Market Split on Impact of EU Price Support Plan." Point Carbon. http://www.pointcarbon.com/news/1.1937474.

Australia Liberal Party. 2012. "Our Plan to Abolish the Carbon Tax." http://www.liberal.org.au/Pages/Our-Plan-to-Abolish-the-Carbon-Tax.aspx.

Australian Government. 2012. "Securing a Clean Energy Future: The Australian Government's Climate Change Plan." Canberra. http://www.cleanenergyfuture.gov.au/wp-content/uploads/2011/07/Consolidated-Final.pdf.

Benessaiah, Karina. 2012. "Carbon and Livelihoods in Post-Kyoto: Assessing Voluntary Carbon Markets." *Ecological Economics* 77 (May): 1–6.

Boden, Tom, Gregg Marland, and Bob Andres. 2011. "Global CO_2 Emissions from Fossil-Fuel Burning, Cement Manufacture, and Gas Flaring: 1751–2008." A table. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory. http://cdiac.ornl.gov/ftp/ndp030/global.1751_2008.ems.

Burtraw, Dallas, Margaret Walls, and Joshua Blonz. 2010. "Distributional Impacts of Carbon Pricing Policies in the Electricity Sector." In *U.S. Energy Tax Policy*, edited by G. E. Metcalf, 10–40. Cambridge University Press.

California Air Resources Board. 2012. "California Air Resources Board Quarterly Auction 1." http://www.arb.ca.gov/cc/capandtrade/auction/november_2012/auction1_results_2012q4nov.pdf.

Carroll, Rory. 2012. "Quebec Govt Approves Link to California CO_2 Market." Point Carbon, December 14. Available at: http://www.pointcarbon.com/news/1.2102367?date=20121214&sdtc=1.

Chen, Yihsu. 2009. "Does a Regional Greenhouse Gas Policy Make Sense? A Case Study of Carbon Leakage and Emissions Spillover." *Energy Economics* 31(5): 667–75.

Christie, Chris. 2011. "Governor Christie: New Jersey's Future Is Green." Video and transcript of a speech at State of New Jersey website: http://www.nj.gov/governor/news/news/552011/approved/20110526a.html.

de Perthius, Christian. 2011. "Carbon Markets Regulation: The Case for a CO_2 Central Bank." Paris Dauphine University Climate Economics Chair Information and Debates Series, vol. 10.

Dechezlepretre, Antoine, Matthieu Glachant, and Yann Meniere. 2008. "The Clean Development Mechanism and the International Diffusion of Technologies: An Empirical Study." *Energy Policy* 36(4): 1273–83.

Demially, Damien, and Philippe Quirion. 2008. "European Emission Trading Scheme and Competitiveness: A Case Study on the Iron and Steel Industry." *Energy Economics* 30(4): 2009–27.

Ecosystem Marketplace and Bloomberg New Energy Finance. 2008–2012. "State of the Voluntary Carbon Markets" reports. http://www.forest-trends.org/publications.php.

Ellerman, A. Denny, Frank J. Convery, and Christian de Perthius. 2010. *Pricing Carbon.* Cambridge University Press.

Ellerman, A. Denny, and Paul L. Joskow. 2008. "The European Union's Emissions Trading System: In Perspective." Pew Center on Global Climate Change. http://www.c2es.org/docUploads/EU-ETS-In-Perspective-Report.pdf.

Energy Information Administration. 2009. "Energy Market and Economic Impacts of H.R. 2454, the American Clean Energy and Security Act of 2009." http://www.eia.gov/oiaf/servicerpt/hr2454/pdf/sroiaf%282009%2905.pdf.

European Commission. 2012a. "Climate Action." http://ec.europa.eu/dgs/clima/mission/index_en.htm.

European Commission. 2012b. "Ensuring the Integrity of the European Carbon Market." http://ec.europa.eu/clima/policies/ets/oversight/index_en.htm.

European Commission. 2012c. "Questions and Answers on the Revised EU Emissions Trading System." http://ec.europa.eu/clima/policies/ets/faq_en.htm.

Fallow, Brian. 2009. "Business Backs Emission Plan Changes." *New Zealand Herald*, September 15.

Fell, Harrison, Ian MacKenzie, and William Pizer. 2012. "Prices versus Quantities versus Bankable Quantities." *Resource and Energy Economics* 34(4): 607–23.

Fischer, Carolyn. 2003. "Combining Rate-Based

and Cap-and-Trade Emission Policies." Resources for the Future Discussion Paper 03-32.

Fischer, Carolyn, and Alan K. Fox. 2009. "Combining Rebates with Carbon Taxes." Resources for the Future Working Paper 09-12.

Frunza, Marius-Cristian, Dominique Guegan, and Antonin Lassoudiere. 2011. "Missing Trader Fraud on the Emissions Market." *Journal of Financial Crime* 18(2): 183–94.

Gow, David. 2006. "Power Tool: European Energy Groups Involved in Carbon Capture Are Manipulating the Scheme for Profit." *The Gaurdian,* UK, May 17.

Grubb, Michael. 2012. "Strengthening the EU ETS: Creating a Stable Platform for EU Energy Sector Investment." London: Climate Strategies.

Hall, Daniel S. 2007. "Greenhouse Gas Emissions and the Fossil Fuel Supply Chain in the United States." In *Assessing U.S. Climate Policy Options,* ed. R. J. Kopp and W. A. Pizer, 23–38. Washington, DC: Resources for the Future.

Harrison, Pete. 2009. "Carbon Windfall Profits Seen for EU Industry." http://www.reuters.com/article/2009/05/19/us-eu-climate-industry-analysis-idUSTRE54I2LO20090519.

Hassett, Kevin A., Aparna Mathur, and Gilbert E. Metcalf. 2009. "The Incidence of a U.S. Carbon Tax: A Lifetime and Regional Analysis." *Energy Journal* 30(2): 155–77.

Ho, Mun S., Richard Morgenstern, and Jhih-Shyang Shih. 2008. "Impact of Carbon Price Policies on U.S. Industry." Resources for the Future Discussion Paper DP-08-37.

Hood, Christina. 2010. "Reviewing Existing and Proposed Emissions Trading Systems." IEA Energy Paper no. 2010/13, International Energy Agency.

Interagency Competitiveness Analysis Team. 2009. "The Effects of H.R. 2454 on International Competitiveness and Emission Leakage in Energy-Intensive Trade-Exposed Industries." Washington, DC: US EPA.

Interagency Working Group for the Study on Oversight of Carbon Markets. 2011. "Report on the Oversight of Existing and Prospective Carbon Markets." http://www.cftc.gov/ucm/groups/public/@swaps/documents/file/dfstudy_carbon_011811.pdf.

Interagency Working Group on Social Cost of Carbon. 2009. "Social Cost of Carbon for Regulatory Impact Analysis under Executive Order 12866." US EPA Technical Support Document. http://www.epa.gov/oms/climate/regulations/scc-tsd.pdf.

International Energy Agency. 2012. "Global Carbon-Dioxide Emissions Increase by 1.0 Gt in 2011 to Record High." http://www.iea.org/newsroomandevents/news/2012/may/name,27216,en.html.

Jaffe, Adam B., Steven R. Peterson, Paul R. Portnoy, and Robert N. Stavins. 1995. "Environmental Regulation and the Competitiveness of U.S. Manufacturing: What Does the Evidence Tell Us?" *Journal of Economic Literature* 33(1): 132–63.

Jaffe, Judson, Matthew Ranson, and Robert N. Stavins. 2009. "Linking Tradable Permit Systems: A Key Element of Emerging International Climate Policy Architecture." *Ecological Law Quarterly* 36(4): 789.

Jotzo, Frank, and Regina Betz. 2011. "Australia's Emissions Trading Scheme: Opportunities and Obstacles for Linking." *Climate Policy* 9(4): 402–14.

Kim, Man-Keun, Bruce A. McCarl, and Brian C. Murray. 2008. "Permanence Discounting for Land-Based Carbon Sequestration." *Ecological Economics* 64(4): 763–69.

Kindle, Andrew G, Daniel L. Shawhan, and Michael J. Swider. 2011. "An Empirical Test for Inter-State Carbon-Dioxide Emissions Leakage Resulting from the Regional Greenhouse Gas Initiative." New York Independent Systems Operator. http://www.nyiso.com/public/webdocs/media_room/publications_presentations/Other_Reports/Other_Reports/ARCHIVE/Report_on_Empirical_Test_for_Interstate_CO2_Emissions_Leakage_04202011_FINAL.pdf.

Lacombe, Romain H. 2008. "Economic Impact of the European Union Emission Trading Scheme: Evidence from the Refining Sector." Dissertation, MIT.

Lambert, Richard Schneider. 2011. "Perverse Incentives under the CDM: An Evaluation of HFC-23 Destruction Projects." *Climate Policy* 11(2): 851–64.

Leiby, Paul, and Jonathan Rubin. 2001. "Intertemporal Permit Trading for the Control of Greenhouse Gas Emissions." *Environmental and Resource Economics* 19(3): 229–56.

Leiter, Andrea M., Arno Paolini, and Hannes Winner. 2011. "Environmental Regulation and Investment: Evidence from European Industry Data." *Ecological Economics* 70(4): 759–70.

Mace, M. J., Ilona Millar, Christoph Schwarte, Jason Anderson, Derik Broekhoff, Robert Bradley, Catherine Bowyer, and Robert Heilmayr. 2008. *Analysis of the Legal and Organizational Issues Arising in Linking the EU Emission Trading Scheme to Other Existing and Emerging Emission Trading Schemes.* Washington, DC: World Resources Institute.

Mehling, Michael, and Erik Haites. 2011. "Mechanisms for Linking Emissions Trading Schemes." *Climate Policy* 9(2): 169–84.

Murray, Brian C., Christopher S. Galik, Stephen Mitchell, and Phil Cottle. 2012. "Alternative Approaches to Addressing the Risk of Non-Permanence in Afforestation and Reforestation Projects

under the Clean Development Mechanism." Nicholas Institute for Environmental Policy Solutions: Duke University.

Murray, Brian C., Richard G. Newell, and William A. Pizer. 2009. "Balancing Cost and Emissions Certainty: An Allowance Reserve for Cap-and-Trade." *Review of Environmental Economics and Policy* 3(1): 84–103.

National Research Council. 2010. "Climate Change." Chap. 5 in *Hidden Costs of Energy: Unpriced Consequences of Energy Production and Use.* Washington, DC: National Academies Press.

Neuhoff, Karsten, Anne Schopp, Rodney Boyd, Kateryna Stelmakh, and Alexander Vasa. 2012. "Banking of Surplus Emisssions Allowances: Does the Volume Matter?" DIW Berlin Working Paper No. 1196, German Institution for Economic Research.

New Zealand Government. 2012. "The New Zealand Emissions Trading Scheme." http://www.climatechange.govt.nz/emissions-trading-scheme/.

Newell, Richard G., and William A. Pizer. 2003. "Regulating Stock Externalities under Uncertainty." *Journal of Environmental Economics and Management* 45(2): 416–32.

Newell, Richard G., William A. Pizer, and Jiangfeng Zhang. 2005. "Managing Permit Markets to Stabilize Prices." *Environmental and Resource Economics* 31(2): 133–57.

Nordhaus, William D., and Joseph G. Boyer. 1998. "Requiem for Kyoto: An Economic Analysis of the Kyoto Protocol." Cowles Foundation Discussion Paper no. 1201.

Pizer, William A. 2002. "Combining Price and Quantity Controls to Mitigate Global Climate Change." *Journal of Public Economics* 85(3): 409–34.

Pizer, William A., and Marika Tatsutani. 2008. "Managing Costs in a U.S. Greenhouse Gas Trading Program." Resources for the Future Discussion Paper 08-23.

Point Carbon. 2011. "Carbon 2011." Annual report.

Point Carbon. 2012. "EU ETS Market Data." http://www.pointcarbon.com/news/marketdata/euets/forward/eua/.

Popp, David. 2011. "International Technology Transfer, Climate Change, and the Clean Development Mechanism." *Review of Environmental Economics and Policy* 5(1): 131–52.

Potomac Economics. 2009, 2010, 2011. "Annual Report on the Market for RGGI CO_2 Allowances." Annual reports available at: http://www.rggi.org/market/market_monitor.

Regional Greenhouse Gas Initiative Inc. 2012a. "97% of RGGI Units Meet First Compliance Period Obligations." Press Release. http://www.rggi.org/docs/PR060412_Compliance.pdf.

Regional Greenhouse Gas Initiative Inc. 2012b. "Program Design." http://www.rggi.org/design.

Reinaud, Julia. 2008. "Climate Policy and Carbon Leakage: Impacts of the European Emissions Trading Scheme on Aluminum." International Energy Agency.

Reklev, Stian. 2012. "Australia to Link with EU CO_2 Market, Drops Price Floor." Point Carbon. http://www.pointcarbon.com/news/1.1967394.

Sijm, Jos, Sebastian Hers, Wietze Lise, and Bas Wetzelaer. 2008. "The Impact of the EU ETS on Electricity Prices." European Commission Publication no. ECN-E-08-007.

Stavins, Robert N. 1998. "What Can We Learn from the Grand Policy Experiment? Lessons from SO_2 Allowance Trading." *Journal of Economic Perspectives* 12(3): 69–88.

Szabo, Michael. 2012. "EU CO_2 Permit Auction Delay Illegal: German Lawyer." Point Carbon. http://www.pointcarbon.com/news/1.1929012.

United Nations Framework Convention on Climate Change. 2012a. "Global Warming Potential." A chart: http://unfccc.int/ghg_data/items/3825.php.

United Nations Framework Convention on Climate Change. 2012b. "Kyoto Protocol." http://unfccc.int/kyoto_protocol/items/2830.php.

van Asselt, Harro, and Thomas Brewer. 2010. "Addressing Competitiveness and Leakage Concerns in Climate Policy: An Analysis of Border Adjustment Measures in the US and the EU." *Energy Policy* 38(1): 42–51.

Victor, David G. 2008. *The Collapse of the Kyoto Protocol and the Struggle to Slow Global Warming.* Princeton University Press.

Western Climate Initiative. 2012. "History." http://www.westernclimateinitiative.org/history.

Weyant, John P. and Jennifer N. Hill. 1999. "The Costs of the Kyoto Protocol: A Multi-Model Evaluation: Introduction and Overview." *The Energy Journal*, Special Issue on the Costs of the Kyoto Protocol: a Multi-Model Evaluation, pp. vii–xliv.

Wing, Ian Sue, and Marek Kolodziej. 2009. "The Regional Greenhouse Gas Initiative: Emission Leakage and the Effectiveness of Interstate Border Adjustments." http://people.bu.edu/isw/papers/rggi_leakage.pdf.

World Bank. 2012. *State and Trends of the Carbon Market 2012.* Washington, D.C: Carbon Finance at the World Bank.

Zhang, Junjie, and Can Wang. 2011. "Co-benefits and Additionality of the Clean Development Mechanism: An Empirical Analysis." *Journal of Environmental Economics and Management* 62(2): 140–54.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 147–172

Moving Pollution Trading from Air to Water: Potential, Problems, and Prognosis

Karen Fisher-Vanden and Sheila Olmstead

The primary means for achieving water quality goals in the United States is the requirement under the Clean Water Act that point sources of pollution—mostly industrial facilities and municipal wastewater treatment plants—obtain permits delineating maximum discharge quantities. Several studies suggest that this law produced significant net benefits between 1972 and the late 1980s, while important local water pollution problems from point sources were remedied, but that around 1990, marginal costs began to exceed marginal benefits (Carson and Mitchell 1993; Lyon and Farrow 1995; Freeman 2000). More than ten years ago, the US Environmental Protection Agency (2001) estimated that expanded use of water quality trading could significantly reduce Clean Water Act compliance costs.

While nearly three dozen water pollution trading programs have been established in the United States, many have seen no trading at all, and few are operating on a scale that could be considered economically significant (Breetz, Fisher-Vanden, Garzon, Jacobs, Kroetz, and Terry 2004; Morgan and Wolverton 2005). The global experience with water quality trading is not much more extensive, though there are active programs in Australia and Canada (Selman, Greenhalgh, Branosky, Jones, and Guiling 2009). While water quality trading holds substantial promise, many challenges remain to be worked out by economists and by environmental managers. These challenges involve both physical aspects of water pollution problems that require modifications to the typical structure of pollution trading as practiced for

■ *Karen Fisher-Vanden is Associate Professor of Agricultural Economics and Rural Sociology, Pennsylvania State University, University Park, Pennsylvania. Sheila Olmstead is a Fellow at Resources for the Future, Washington, DC. Their email addresses are fishervanden@psu.edu and olmstead@rff.org.*

http://dx.doi.org/10.1257/jep.27.1.147. doi=10.1257/jep.27.1.147

air quality, as well as constraints imposed by current regulatory approaches to water pollution control that limit market function, including the implied assignment of rights to pollute.

This paper seeks to assess the current status of water quality trading and to identify possible problems and solutions. We begin with some background on US water pollution regulation, and then present an informal assessment of the current status of water quality trading. We describe six criteria for successful pollution trading programs and consider how these apply to standard water quality problems, as compared to air quality. We then highlight some important issues to be resolved if current water quality trading programs are to function as the "leading edge" of a new frontier in cost-effective pollution permit trading in the United States.

Background on US Water Quality Regulation and the Role of Trading

Water quality concerns were a major impetus for the establishment of the Environmental Protection Agency in 1970; for example, the infamous fire on the Cuyahoga River near Cleveland, Ohio, occurred in 1969—though in truth it was the tenth such fire that had occurred since the mid-1800s, and not the worst. The Federal Water Pollution Control Act, commonly known as the Clean Water Act, became law in 1972. The stated goals of the Clean Water Act were: 1) the attainment of fishable and swimmable waters by July 1, 1983; and 2) the elimination of *all* discharges of pollutants into navigable waters by 1985 (Freeman 2000). Obviously, those deadlines have been postponed through amendments, and distinctions have since been made between different types of pollutants. However, one should not underestimate the degree to which these original goals have influenced regulation under the law.[1]

The Clean Water Act's main tool is a set of effluent standards, implemented through point-source permitting. The National Pollutant Discharge Elimination System (NPDES) specifies quantitative effluent limits by pollutant, for each point source, based on available control technologies. For the most part, industrial point source compliance with these permits has been high (Freeman 2000). Municipal sewage treatment has also expanded dramatically, resulting in impressive improvements in urban water quality—for examples, see Boston Harbor and the Hudson River near New York City. But the gains from point source controls are reaching their limits. Even if all point sources were to achieve zero discharge, only 10 percent of US river and stream miles would rise one step or more on EPA's water quality ladder (Bingham et al. 2000).

Nonpoint source pollution such as agricultural and urban runoff, atmospheric deposition, and runoff from forests and mines has become the major concern of

[1] Statistical analyses attempting to link water quality improvements with the Clean Water Act itself suggest that the Act's impact on water quality has been somewhat small, though this remains an empirically ambiguous question (Bingham et al. 2000; Gianessi, Peskin, and Young 1981).

water pollution abatement efforts. In fact, nonpoint source pollution from agricultural activities is now the *primary* source of impairment in US rivers and streams (US Environmental Protection Agency 2009). Nonpoint source pollution involving nutrients like nitrogen and phosphorus causes excessive aquatic vegetation and algae growth and eventual decomposition, which deprives deeper waters of oxygen, creating hypoxic or "dead" zones, fish kills, and other damages. This problem is geographically widespread; seasonal dead zones in US coastal waters affect Puget Sound, the Gulf of Mexico, the Chesapeake Bay, and Long Island Sound.

However, agricultural nonpoint source pollution is essentially unregulated by the Clean Water Act, creating a de facto property rights distortion that strongly affects the ability to attain water quality goals. Although the Act does not address this issue directly, an important provision is Section 303(d), which requires states to establish a Total Maximum Daily Load (TMDL)—basically a "pollution budget"—for each water body that does not meet ambient water quality standards for its designated use, despite point source controls. Designated uses include recreational use, public water supply, and industrial water supply, and each designated use has an applicable water quality standard. State courts began ordering the development of TMDLs in the 1980s and 1990s in response to lawsuits by environmental groups.[2] Since 1996, the states in cooperation with the Environmental Protection Agency have completed thousands of TMDLs. Establishing a TMDL is a "holistic accounting exercise" in which all permitted sources and land uses within a watershed drainage area, including agriculture and urban runoff, are inventoried and allocated responsibility for portions of the pollution budget (Boyd 2000).

While regulators cannot implement enforceable caps on agricultural pollution through this process, they have recognized the importance of incorporating agricultural abatement into clean-up processes, and water quality trading is one tool they have employed for this purpose. Not surprisingly, marginal abatement costs for point sources which have faced stringent regulation over the past 40 years tend to be high relative to those for nonpoint sources, which have been unregulated. Thus, allowing point sources of water pollution to offset their effluent, or to trade credits for abatement by farms and other entities responsible for nonpoint source pollution, could be cost-effective (Stephenson and Shabman 2011). In almost all water quality trading programs established in the United States, the regulatory driver has been the establishment (or anticipated establishment) of a Total Maximum Daily Load. The US Environmental Protection Agency (2001) estimated that expanded use of water quality trading between point and nonpoint sources could reduce compliance costs associated with TMDL regulations by $1 billion or more annually between 2000 and 2015.

The Environmental Protection Agency established a "draft framework for watershed-based trading" in 1996, and many water quality trading programs were

[2] The Environmental Protection Agency offers a survey of TDML lawsuits and outstanding obligations under these lawsuits at http://water.epa.gov/lawsregs/lawsguidance/cwa/tmdl/lawsuit.cfm and http://ofmpub.epa.gov/tmdl_waters10/attains_nation_cy.control?p_report_type=T#APRTMDLS.

established during the 1990s. EPA formalized its overarching policy toward water quality trading in January 2003 (US Environmental Protection Agency 2003). At that time, the agency also funded 11 pilot trading projects across the United States. A few important specifics in the 2003 policy continue to shape US water quality trading programs. First, once a Total Maximum Daily Load has been established, all trading must take place "within a watershed or a defined area for which a TMDL has been approved." Second, the policy supports trading of nutrients (nitrogen and phosphorus) and sediment, but notes that trading of other pollutants will trigger increased scrutiny and can only be implemented with prior approval. Third, point sources cannot typically use trading to fulfill their National Pollutant Discharge Elimination System permit requirements; instead trading or offsets can only be applied to a source's effort to comply with the additional TMDL-related restrictions. This rule has been relaxed in some cases, allowing some industrial and municipal point sources—on a case-by-case basis—to purchase water quality abatement from other sources (usually farms), to reduce their cost of compliance with permits issued under the National Pollutant Discharge Elimination System.

Current Status of Water Quality Trading

Water pollution permit trading programs tend to be small, diffuse, and low-profile, and have rarely been comprehensively described and analyzed in the peer-reviewed literature. Including active programs and completed or otherwise inactive programs, we identify approximately three dozen initiatives. We assessed the status of current programs using existing sources (Breetz et al. 2004; Industrial Economics 2008; Selman, Greenhalgh, Branosky, Jones, and Guiling 2009), as well as extensive use of phone calls and Internet searches for program websites.

Table 1 describes several characteristics of the 21 current active and pilot programs.[3] We divide these active programs into two categories: those that involve actual trades, and pure offset programs. As we define them, trading programs must involve multiple recipients and multiple sources. The offset programs, by contrast, with one exception, all involve a single recipient of water quality credits from one source or multiple sources. Typically, the offset credit recipient invests directly in credit-generating projects rather than purchasing credits outright. Within each category, programs in Table 1 are ordered by their year of establishment.

[3] The list of active programs in Table 1 differs considerably from others in the literature. In some cases, programs described as active by earlier researchers are clearly inactive as of 2012; in others, we were not able to gather enough information on program characteristics to justify their inclusion. We were fairly conservative in our definition of what counts as an active water quality trading or offset program; a more liberal definition would have resulted in closer to 60 such programs (Selman et al. 2009). A more comprehensive list of programs including maps showing all trading programs at the state and watershed levels can be found at: http://www.envtn.org/State_Programs___Rules.html.

Table 1
Active Water Quality Trading and Offset Programs

Program name	*Year est.*	*Location*	*Types of trades/ offsets*	*Pollutants*	*Trading or offset structure*
Trading programs					
Tar-Pamlico Nutrient Trading	1990	NC, US	PS-PS/NPS	N/P	Bilateral/ Clearinghouse
South Creek Bubble Licensing	1996	NSW, Austr.	PS-PS	N/P	Bilateral
Cherry Creek Reservoir Watershed Phosphorus Trading	1997	CO, US	PS-PS/NPS	P	Clearinghouse
Chatfield Reservoir Trading	1999	CO, US	PS-PS/NPS	P	Bilateral/ Clearinghouse
South Nation River Watershed Trading	2000	ONT, Can.	PS-NPS	P	Clearinghouse
Long Island Sound Nitrogen Credit Exchange	2002	CT, US	PS-PS	N	Clearinghouse
Neuse River Basin Total Nitrogen Trading	2002	NC, US	PS-PS/NPS	N	Bilateral/ Clearinghouse
Hunter River Salinity Trading	2004	NSW, Austr.	PS-PS	Salinity	Exchange market
Great Miami River Watershed Trading Pilot	2006	OH, US	PS-NPS	N/P	Clearinghouse
Minnesota River Basin Trading	2006	MN, US	PS-PS	P	Bilateral
Maryland Water Quality Trading	2008	MD, US	PS-PS/NPS	N/P/sediment	Exchange Market/ Bilateral
Pennsylvania Nutrient Credit Trading	2010	PA, US	PS-PS/NPS	N/P/sediment	Exchange Market/ Bilateral
Chesapeake Bay Watershed Nutrient Credit Exchange	2011	VA, US	PS-PS/NPS	N/P	Clearinghouse/ Bilateral
Offset programs					
Rahr Malting	1997	MN, US	PS-NPS	CBOD5	Bilateral
Pinnacle Foods	1998	DE, US	PS-NPS	N, P	Bilateral
Southern Minnesota Beet Sugar Cooperative	1999	MN, US	PS-NPS	P	Clearinghouse
Bear Creek	2001	CO, US	PS-PS	P	Bilateral
Piasa Creek Watershed Project	2001	IL, US	PS-NPS	Sediment	Bilateral
Clean Water Services/Tualatin River	2005	OR, US	PS-PS/NPS	BOD/NH4/temp.	Bilateral
Red Cedar River Nutrient Trading Pilot	2007	WI, US	PS-NPS	P	Bilateral
Alpine Cheese Company/Sugar Creek	2008	OH, US	PS-NPS	P	Bilateral

Notes: Abbreviations in column 4 refer to point sources (PS) and nonpoint sources (NPS). In column 5, abbreviations refer to nitrogen (N), phosphorus (P), biochemical oxygen demand (BOD), 5-day carbonaceous biochemical oxygen demand (CBOD5), ammonia (NH4), and temperature (temp.).

Active Trading Programs

Nine of the 13 active trading programs in Table 1 have been established since 2000, with the remainder established during the 1990s. With the exception of Australia's Hunter River Basin salinity trading program, the pollutants traded in all active programs in Table 1 are nutrients, or a combination of nutrients and sediment.

We distinguish between three market structures in Table 1 (adapted from Woodward, Kaiser, and Wicks 2002; Selman, Greenhalgh, Branosky, Jones, and Guiling 2009): bilateral, clearinghouse, and exchange markets. In bilateral programs, participants engage in individual negotiations to arrange trades or offsets. The required one-to-one negotiations lead to higher transaction costs than

other structures, though these cost differences vary across trading programs. In the Pennsylvania Nutrient Credit Trading program, per-pound transaction costs for bilateral nitrogen trades are estimated at about two times those for auction trades (Ribaudo and McCann 2012).

In clearinghouse programs, a single broker or intermediary may generate credits; for example, in the Neuse River program in North Carolina, point-source participants may engage in bilateral trades with other point sources, or they may pay into a state wetland restoration fund, which funds nonpoint source abatement projects. The intermediary in a clearinghouse program may also convert the abatement activities of diffuse nonpoint sources into a uniform "credit currency" that can be purchased by point sources. For example, in the Great Miami River program in Ohio, farmers submit applications for "best management practices" projects to generate credits, and a public clearinghouse holds a reverse auction to fund the most cost-effective projects from these applications. Credits are then allocated to participating point sources in proportion to their investments in the aggregate credit bank. Nguyen, Shortle, Reed, and Nguyen (forthcoming) find that a clearinghouse market structure is more efficient at facilitating trades between point and nonpoint sources than bilateral trading. Finally, two active programs, the Hunter River Salinity Trading program in Australia and the Pennsylvania Nutrient Credit Trading program, have established true exchange markets, where buyers and sellers trade uniform credits at transparent prices. (A third program, Maryland Water Quality Trading, is set up as an exchange market, but there hasn't been activity on this market.)

Table 2 offers further detail on market participants. In all but one of the active trading programs, point source participants include municipal wastewater treatment plants. Several programs also involve industrial point sources: Tar-Pamlico, South Nation, Minnesota River Basin, Pennsylvania Nutrient Credit Trading, and Chesapeake Bay Watershed Nutrient Credit Exchange. With one exception, the Neuse River Basin Total Nitrogen Trading program, nonpoint source participants in the active trading programs are agricultural sources.

Trading activity is very limited in most of these programs, however. In this section, we describe six of the most active markets (two of them, in Pennsylvania and Virginia, developed under a single Total Maximum Daily Load for the Chesapeake Bay Watershed).

1) Hunter River Salinity Trading. Since 2004, the Australian state of New South Wales, in Southeast Australia, has operated a trading program to control salinity in the Hunter River Basin. Sources of salinity include agricultural irrigation, disposal of brine from coal mining, and water diversions for cooling in electricity generation which concentrates salts in the water remaining instream. The river is divided into numbered blocks, measured by units of water that will flow past Singleton, New South Wales (the downstream endpoint of the trading scheme) on a particular day. Daily caps are established through continuous monitoring of ambient salinity concentrations and flow levels, with the goal of meeting a maximum allowable salinity concentration at Singleton. The Hunter River Salinity Trading

Table 2
Participants and Trading Ratios in Active Trading and Offset Programs

Program name	*Participants*	*NPS:PS trading ratio, if any*
Trading programs		
Tar-Pamlico Nutrient Trading	POTWs, ind. PS, ag. NPS	2.1:1
South Creek Bubble Licensing	POTWs	N/A
Cherry Creek Reservoir Watershed Phosphorus Trading	POTWs, ag. NPS	≥2:1
Chatfield Reservoir Trading	POTWs, ag. NPS	2:1
South Nation River Watershed Trading	POTWs, ind. PS, ag. NPS	4:1
Long Island Sound Nitrogen Credit Exchange	POTWs	N/A[a]
Neuse River Basin Total Nitrogen Trading	POTWs, wetland restoration fund (NPS)	None[b]
Hunter River Salinity Trading	Ind. PS	N/A
Great Miami River Watershed Trading Pilot	POTWs, ag. NPS	1:1–3:1
Maryland Water Quality Trading	POTWs, ind. PS, ag. NPS	TBD
Minnesota River Basin Trading	POTWs, ind. PS, ag. NPS	1.1:1–1.2:1[c]
Pennsylvania Nutrient Credit Trading	POTWs, counties, ind. PS, ag. NPS	1.1:1[c]
Chesapeake Bay Watershed Nutrient Credit Exchange	POTWs, ind. PS, ag. NPS	2:1
Offset programs		
Rahr Malting	Single ind. PS, multiple ag. NPS	2:1
Pinnacle Foods	Single ind. PS, multiple ag. NPS	2.3:1 for N, 7.9:1 for P[d]
Southern Minnesota Beet Sugar Cooperative	Single ind. PS, multiple ag. NPS	2.6:1
Bear Creek	Two POTWs	N/A
Piasa Creek Watershed Project	Single drinking water system, ag. NPS	2:1
Clean Water Services/Tualatin River	Single POTW with two facilities, ag. NPS	2:1[e]
Red Cedar River Nutrient Trading Pilot	Single POTW, ag. NPS	2:1
Alpine Cheese Company/Sugar Creek	Single ind. PS, ag. NPS	3:1

Notes: Abbreviations in column 2 indicate publically owned treatment works (POTWs), industrial (ind.), agricultural (ag.), point sources (PS), and nonpoint sources (NPS). In column 3, N/A indicates that the program does not involve PS-NPS trades or offsets.
[a] PS-PS trading ratios are based on distance of each facility to hypoxic zones in Long Island Sound.
[b] Clearinghouse sets NPS nitrogen abatement price/lb. greater than average marginal cost of PS abatement, but no formal trading ratio.
[c] PS-PS trading ratios are unique to trading pairs, using a formal trading ratio system.
[d] No formal ratios; reported ratios are averages for transacted offsets.
[e] Refers to ratio for NPS:PS temperature offsets; ratios for other contaminants unknown.

Scheme restricts saline discharges by coal mines and power plants to periods when river flows are high, and to amounts less than or equal to a facility's salinity credit allocation. If discharges exceed credits, participants may purchase credits from other facilities.

2) Long Island Sound Nitrogen Credit Exchange. In response to a 2001 Total Maximum Daily Load for dissolved oxygen for Long Island Sound, the state of Connecticut established the Long Island Sound Nitrogen Credit Exchange in

2002, with 79 municipal sewage treatment plants participating.[4] The Long Island Sound program is structured as a clearinghouse, where the annual price is set by regulators based on the estimated average cost of nitrogen removal among participating plants. Because source location affects the environmental impact of a unit of nitrogen discharged, the program uses a system of trading ratios based on geographic trading zones. Abatement cost differentials are generally driven by plant size, as there are significant economies of scale in municipal sewage treatment. The Connecticut Department of Environmental Protection (2010) estimates cost savings from trading through 2009 of $300–$400 million. Though no study has definitively linked the trading program with improved water quality in the Sound (given the significant annual variation in water conditions), the general trend of summer hypoxia incidents is decreasing, despite several years of record-setting warmth since the program began (Connecticut Department of Environmental Protection 2011).

3) South Nation River Watershed Trading. A local watershed organization in Ontario, Canada, South Nation Conservation, developed a phosphorus trading program for the South Nation River watershed in 2000. Participants include 16 municipal and industrial dairy wastewater lagoon operators, who are allowed to expand their effluent discharge to waterways only if they invest in offsetting reductions in nonpoint source agricultural runoff. The watershed organization acts as a clearinghouse in this program, collecting payments from dischargers, investing the proceeds in abatement projects that it identifies on specific farms, and distributing phosphorus credits in exchange. South Nation Conservation estimates that the trading program reduces abatement costs per kilogram of phosphorus for participating dischargers by about 40 percent, compared with the traditional wastewater treatment methods that would otherwise be required (O'Grady 2010).

4) Minnesota River Basin Trading. To address low dissolved oxygen levels caused by algae blooms related to high phosphorus concentrations, the Minnesota Pollution Control Agency issued in 2005 a single National Pollutant Discharge Elimination System permit (updated in 2009) for phosphorus discharges to the Minnesota River, applicable to 47 permitted sources—mostly municipal sewage treatment plants and some industrial point sources (Minnesota Pollution Control Agency 2009). While the general permit does not relieve individual facilities of obligations under individual permits before the Total Maximum Daily Load was implemented, it allows facilities to trade phosphorus abatement allocations required by the new limit. A system of facility-specific trading ratios is used. In 2011, 17 facilities participated in seasonal trades.[5] Unlike a market or clearinghouse approach, trades in the Minnesota River program are made through bilateral negotiations between point sources.

[4] Although the Total Maximum Daily Load requires both Connecticut and New York to reduce nitrogen loads to Long Island Sound by 58.5 percent between 2001 and 2014, New York opted to not create a trading program. The Connecticut program does allow participating municipal sewage treatment plants to sell excess credits to facilities in New York or industrial point sources in Connecticut if total nitrogen loading falls below the aggregate cap, but this option has not yet been exercised.

[5] Current numbers of sources and facilities trading from Lisa McCormick, Minnesota Pollution Control Agency, personal communication, August 23, 2012.

5) State-level trading under the Chesapeake Bay Total Maximum Daily Load. Table 1 includes three active water quality trading programs related to the 2010 Chesapeake Bay Total Maximum Daily Load, which limits allowable discharges of nitrogen, phosphorus, and sediment to rivers and streams in the watershed by six states (Delaware, Maryland, New York, Pennsylvania, Virginia, West Virginia) and the District of Columbia. Pennsylvania, Virginia, and Maryland have chosen to implement water quality trading programs to reduce compliance costs for the required effluent abatement, with an additional program under development in West Virginia. A small amount of nutrient trading had been taking place in some of these states, but the establishment of these new markets could be a gateway to large-scale trading to lower compliance costs throughout the region.

A prospective study estimated the potential cost savings from water quality trading to achieve compliance with the Chesapeake Bay Total Maximum Daily Load at $78 million per year if point sources are allowed to trade only with other point sources within a river basin and within a state—a 20 percent decrease in costs relative to no trading (Van Houtven, Loomis, Baker, Beach, and Casey 2012). However, if trading was allowed watershed-wide across state and basin boundaries and between all sources, compliance costs could be reduced by almost half relative to no trading. The major gains come from allowing trades between point sources and agricultural nonpoint sources.

Both the Pennsylvania and the Virginia programs allow trading of nitrogen and phosphorus. Pennsylvania's program, which started in 2010, thus far includes municipal sewage treatment plants, counties, industrial point sources, and several brokers or credit aggregators of nonpoint source abatement. Trades are facilitated through online auctions or through bilateral negotiation between point and nonpoint sources (Ribaudo and McCann 2012). In Virginia's program, which started in 2011, participating point sources meet their allocations through their own abatement, purchase of credits from other point sources, or payments made to a state water quality improvement fund used for agricultural abatement projects. Trades can either be negotiated bilaterally between point sources, or can be made through a clearinghouse organization of municipal and industrial point sources.

West Virginia and Maryland are also setting up water quality trading programs to lower compliance costs. Maryland's program, listed in Table 1, is quite well-developed, though to our knowledge, no trades have yet taken place.[6] There is little information available on West Virginia's program (and it is likely to be much smaller in scale, given that state's relatively small portion of the Chesapeake Bay watershed), thus it is excluded from Table 1. Current plans suggest that both of these will be exchange markets, which would double the current number of water

[6] The possibility of Maryland's program developing significant trading activity may be hampered by the fact that its program is more restrictive than Pennsylvania's and Virginia's. For example, municipal sewage treatment plants will not be able to purchase credits to meet their allocations, but must implement specific nutrient removal technologies, instead (Van Houtven et al. 2012; Branosky, Jones, and Selman 2011).

quality trading exchange markets in existence and place three out of four in the Chesapeake Bay watershed.

Trading Programs with Minimal Activity

The remaining seven trading programs in Table 1 are much less active. Some of the programs are quite new, like the Maryland Water Quality Trading Program, briefly mentioned above, and the Great Miami River Watershed Trading Pilot in Ohio (Newburn and Woodward 2012), and activity may increase in the future. Others are small-scale because they target local pollution problems involving only a few sources (the South Creek Bubble Licensing program in New South Wales, Australia is an example).

In other cases, problems with program design limit participation. For example, Colorado's Cherry Creek and Chatfield Reservoir phosphorus trading programs are among the oldest such programs in the United States. However, in their 13–15 years of operation, the Cherry Creek program has produced only four trades, and Chatfield Reservoir, only seven trades (Selman, Greenhalgh, Branosky, Jones, and Guiling 2009). Later in this paper, we consider program design issues that may limit trading.

Active Offset Programs

Most of the active US water quality offset programs in Table 1 were created through a modification to a single National Pollutant Discharge Elimination System permit, giving a regulated point source the flexibility to more cheaply reduce discharge and achieve compliance through direct investments in abatement projects outside its own facility. The active US offset programs were established between 1997 and 2008. Of the US programs, all but Bear Creek, a very small annual phosphorus trading arrangement between two Colorado municipal sewage treatment facilities, involve a single point source offsetting nutrient-related permit requirements through investments in off-site abatement. All but one are bilateral exchanges, where the regulated point source negotiates directly with farms regarding investment in projects sufficient to meet its permit requirements.

Since the programs are generally quite similar, a few examples suffice for description of program design. In the Alpine Cheese Company program, a cheese manufacturer was required, as part of its plan to expand production, to reduce phosphorus discharge from its wastewater treatment plant from 225 parts-per-million to 1 part-per-million. Through a treatment plant upgrade, the company reduced phosphorus discharge to 3 parts-per-million. The cost of reducing further to 1 part-per-million was sufficiently high that the firm sought to achieve these remaining reductions through investments in nonpoint source agricultural abatement (Wood 2011). Thus, the company paid 25 local dairy farmers to reduce phosphorus discharge. In a much larger program, the Southern Minnesota Beet Sugar Cooperative, a beet processor, pays its 256 grower-members to invest in phosphorus-reducing land management changes so that the processor can meet its permit requirements for expanded production (Werblow 2007; Fang, Easter, and Brezonik 2005). In this

case, the beet growers and the processing facility are treated under the processor's permit as a single source to meet an overarching phosphorus effluent cap. The structure of Wisconsin's Red Cedar River program is similar, except that the regulated point source investor is a municipal wastewater treatment plant, which pays local farmers to reduce discharge, thereby avoiding a costly treatment plant upgrade.

Inactive or Completed Trading and Offset Programs

Table 3 offers descriptive information on 12 additional trading and offset programs that are currently inactive. In some cases, a very small amount of trading or offset activity occurred before the programs became inactive (for example, Grassland Area Farmers Tradable Loads Program, Lake Dillon Trading Program). In other cases, early studies suggested that trading was unlikely to be successful, so programs were never formally implemented (for example, Vermillion River, Non-Tidal Passaic River Trading Program, Charles River Flow Trading Program). In the remaining cases inactivity has been due to a number of factors including delays in the development of the Total Maximum Daily Load due to litigation, unresolved scientific modeling issues, or lack of demand for pollution credits. In the next section of this paper, where we discuss criteria for successful trading systems, we consider the limitations of some of these programs in more detail.

Applying Criteria for Successful Pollution Permit Trading Systems to Water

The primary objective of any trading program is to meet or exceed the environmental goal at least cost. The environmental goal is best achieved if two conditions are met: a) the pollutant is uniformly mixed to avoid the potential for hot spots; and b) the pollutant can be easily measured and monitored, allowing enforcement to be effective at deterring noncompliance. The cost-effectiveness goal is best achieved if three additional conditions are met: c) sources have significant cost differentials so that the potential gains from trade are large; d) the number of polluting sources is large enough and the regulatory driver stringent enough to generate sufficient trading volume; and e) there is flexibility in when, where, and how reductions and trades are made.[7] In this section, we discuss challenges related to these five criteria in the case of water quality trading, as well as possible solutions.

Uniformly Mixed Pollutants and Non-uniform Mixing

In the case of climate change, the location of greenhouse gas reductions is not important, since these gases are uniformly mixed—that is, the environmental impact of a ton emitted in one location is equal to the impact of a ton emitted elsewhere. In contrast, marginal damages from water pollution may vary dramatically

[7] For alternative lists of criteria for successful pollution trading, see Stavins (1998), Benkovic and Kruger (2001), and Schary and Fisher-Vanden (2004).

Table 3

Selected Inactive or Completed Water Quality Trading and Offset Programs and Pilot Programs

Program name	*Year est.*	*Location*	*Types of trades/offsets*	*Pollutants*	*Reason(s) for inactivity*
Lake Dillon (Dillon Reservoir) Trading Program	1984	Colorado	PS-NPS	P	Low/no credit demand[a]
Boulder Creek Trading Program	1990	Colorado	PS-NPS	NH4	NPS uncertainty[b]
Kalamazoo: Gun Lake Tribe Trading Initiative	1996	Michigan	PS-NPS	P	No regulatory driver
Fox-Wolf Basin	1997	Wisconsin	PS-NPS	P	No regulatory driver; low credit demand/supply[c]
Rock River	1997	Wisconsin	PS-NPS	P	Low credit demand[d]
Grassland Area Farmers Tradable Loads Program	1998	California	NPS-NPS	Selenium	Low/no credit demand[e]
Lower Boise River Effluent Trading	1998	Idaho	PS-NPS	P	No regulatory driver[f]
Upper Middle Snake Rock Subbasin	2001	Idaho	PS-PS	P	Litigation
Charles River Flow Trading Program	2003	Massachusetts	PS-PS	P, flow	Unsuccessful pilot[g]
Non-Tidal Passaic River Trading Program	2005	New Jersey	PS-PS	P	Trading not cost-effective[h]
Vermillion River	2006	Minnesota	NPS-NPS	Temp.	Trading not cost-effective[h]
Willamette Partnership: Counting on the Environment	2006	Oregon	PS-NPS	Temp.	Litigation

Notes: Due to lack of available information, we exclude many programs that were studied but never established. Abbreviations in column 4 refer to point sources (PS) and nonpoint sources (NPS). In column 5, abbreviations refer to nitrogen (N), phosphorus (P), ammonia (NH4), and temperature (temp).

[a] New removal technology reduced the cost of point source abatement, driving credit demand to zero.

[b] Inconclusive evidence for agricultural best management practice effectiveness.

[c] Full set of Total Maximum Daily Loads (TMDLs) has not been developed for the basin, point-source controls are cheaper than expected, and agricultural sources have not participated.

[d] Point sources are able to cost-effectively reduce effluent below required levels without trading. May change if final TMDL results in more stringent limits.

[e] Nine trades took place in 1998–1999, then regional irrigation water re-use project, subsidized by federal and state governments, has kept selenium below cap, with no need for trading since then (Wallace 2007).

[f] TMDL development held up for many years.

[g] Difficulty obtaining cooperation from local municipalities; point sources felt that lack of regulations for nonpoint source pollution was unfair.

[h] For each of these programs, research determined that trading would not result in cost-effective pollution reductions, so program did not develop. For the Passaic River program, cost savings over uniform standard was 2–3%, excluding administrative costs, due to abatement cost homogeneity among participating sewage treatment plants (Obropta, Goldfarb, Strom, Uchrin, Kardos, Boisvert, Poe, and Potent, n.d., http://www.water.rutgers.edu/Projects/trading/FINAL_Water%20quality%20trading%20report_Mar-201003242010.pdf). For the Vermillion River program, research showed agricultural nonpoint source controls did not affect temperature, reducing pool of potential participants to developers and small private property owners, who had homogeneous abatement costs. Transaction costs among many small participants would also have been high (http://www.vermillionriverwatershed.org/index.php?option=com_content&view=article&id=52&Itemid=61).

with the location of discharge, depending on the characteristics of receiving waters and other factors that influence the effectiveness of reductions made in different locations in a watershed. In this case, establishing location-based trading ratios for each pair of polluters, in the manner of exchange rates, has been shown to be an efficient approach (Oates, Krupnick, and Van de Verg 1983; Tietenberg 1985; Rodríguez 2000; Hung and Shaw 2005; Farrow, Schultz, Celikkol, and Van Houtven 2005; Konishi, Coggins, and Wang forthcoming). In fact, Montgomery's (1972) original article introducing the theory of markets for pollution control considered the case of non-uniform mixing and developed a trading-ratio-based system.

Several current water trading programs use systems of trading ratios to ensure that credits traded have equivalent impacts on the water quality problem of concern at a particular location or set of locations. Examples include the three largest active programs discussed in the previous section: nitrogen trading in Long Island Sound, salinity trading in Australia's Hunter River Basin, and phosphorus trading in the Minnesota River.

The cost savings from trading, relative to a prescriptive uniform standard, are reduced when trading ratios are introduced (they are, after all, constraints on trade). However, getting trading ratios right can also increase the benefits of water quality regulation if high-damage sources also have high abatement costs and, without trading ratios, would engage in little abatement when the option to purchase (cheaper) permits is available. Thus, this is an important concern not just for cost-effectiveness, but for efficiency as well. In sum, while non-uniform mixing poses a system design problem that may be generally more significant for water pollution than for air pollution problems, the problem can be addressed in plausible ways.

Measurement, Monitoring, and Enforcement

Effluent from point sources of water pollution, like wastewater treatment plants, is easily measured and monitored. However, the large potential gains from trade in water quality will likely be realized in systems where point sources are net buyers of abatement by nonpoint sources, since these sources are the "low-hanging fruit" of water pollution abatement options. The measurement and monitoring of water pollution from nonpoint sources is challenging, however. In at least one case, the inactive Boulder Creek Trading Program, inconclusive evidence for the effectiveness of agricultural pollution controls was a direct reason for the program's failure.

There are three main sources of uncertainty in measuring and monitoring nonpoint source pollution. First, nonpoint source pollution is inherently more stochastic than point source pollution, because it depends more heavily on weather-related factors such as rainfall and temperature (Shortle and Dunn 1986). Second, there may be scientific or technical uncertainty regarding the effectiveness of abatement projects affecting nonpoint sources, which can lead actual reductions to be less than expected (Harrington et al. 1985). Third, the technical estimation of expected abatement may be correct, but flaws in a project's

implementation and/or operation (institutional uncertainty) may produce less abatement than expected.

Regulators have typically dealt with these uncertainties by requiring more than one unit of abatement in nonpoint source pollution in exchange for each credit toward abatement of point source pollution. These requirements are typically referred to as "trading ratios," a term that economists use to describe the systems of exchange rates used to account for locational differences in damages from pollution. However, economic theory does not necessarily support the use of high trading ratios between point and nonpoint sources to address uncertainty. Indeed, the optimal trading ratio from point to nonpoint sources could be either greater than or less than 1:1 in the presence of stochastic pollutant loading (Shortle 1990; Malik, Letson, and Crutchfield 1993). If regulators are risk-averse, then investments to reduce nonpoint source pollution are doubly beneficial from the perspective of social welfare, because nonpoint source pollution imposes both direct damages from pollution and additional risk, given its inherently stochastic nature (Horan 2001). Thus, regulators seeking to address this aspect of uncertainty ought to skew trades in favor of reductions in nonpoint source pollution, rather than against them (Shortle 1990; Malik, Letson, and Crutchfield 1993). Alternatively, one might try to define water pollution abatement credits in terms of *expected* units of abatement, which consider both the mean and the variance of nonpoint source abatement (Horan and Shortle 2011).

Scientific or technical uncertainty, in contrast, may be more efficiently handled through improved liability rules. Segerson and Wu (2006) suggest a hybrid instrument that includes an ambient tax imposed if nonpoint source abatement projects do not result in real reductions. Regulatory agencies might also develop a preapproved list of "best management practices" for agricultural nonpoint source reductions, which gives point sources certainty over the amount of credit for a specific investment (Schary and Fisher-Vanden 2004). Regulatory agencies could also fund implementation of pilot projects and new scientific research to resolve some uncertainty.

The current approach, in contrast, stacks the deck against nonpoint source reductions. Consider the point/nonpoint source trading ratios (reported in Table 2) for the many programs that involve this component, ranging from 1:1 to 4:1 for active trading programs, and from 2:1 to almost 8:1 for active offset programs.[8] In addition, it may discourage trades that could have reduced pollution and lowered compliance costs. For example, Ontario's South Nation trading program uses a 4:1 trading ratio. As a result, while the cost per kilogram of phosphorus removal through nonpoint sources is 85 percent lower than traditional wastewater treatment

[8] Given that high trading ratios between point and nonpoint sources cannot be justified by theory, what explains their prevalence? Regulators may seek to maximize abatement, rather than minimize pollution damages (Horan 2001). It is also possible that high trading ratios could be optimal in a second-best setting, where trading ratios are set independent of allowable discharge; in an efficient setting, these would be jointly determined (Horan and Shortle 2005).

by point sources, with the 4:1 trading ratio, the cost saving per pound of phosphorus abatement is only about 40 percent (see Environment Canada, n.d.).

The difficulty of measuring nonpoint source pollution abatement clearly makes enforcement difficult, too. Imposing liability rules if promised abatement does not occur provides one potential mechanism for dealing with this problem. However, unlike the air case, where the performance and effectiveness risk of projects to generate emission credits is borne by credit sellers, in many water quality trading programs involving trades between point and nonpoint sources, liability for nonperformance or ineffectiveness lies with point-source credit buyers. Some attribute low trading volumes in current water quality trading programs to this problematic assignment of liability (Stephenson and Shabman 2011; Jarvie and Solomon 1998). However, the current liability structure of trading is a reflection of the fact that agricultural sources are unregulated, and the Environmental Protection Agency can only impose penalties on permitted point sources through the National Pollutant Discharge Elimination System.

Private bilateral contracts between point and nonpoint sources could include provisions that subject nonpoint sources to penalties for nonperformance, though the transaction costs associated with this approach would be high. As an alternative, in some "clearinghouse" water quality trading programs—like Ontario's South Nation River program, and the Great Miami River program in Ohio—the same public or nonprofit third parties that facilitate trades may also assume liability in the case of nonperformance by nonpoint sources. Finally, in other programs, such as the Pennsylvania Nutrient Credit Trading program, private for-profit brokers or credit aggregators work directly with agricultural nonpoint sources on abatement, and then sell credits to point sources, assuming liability for nonperformance.

Abatement Cost Differentials

The larger the differences in marginal abatement costs, the greater the potential gains from trading pollution permits (Newell and Stavins 2003). Such cost differentials among point sources may stem from differences in industrial sector, process, technology, or other source characteristics. Abatement cost differentials can be large between point sources of water pollution: three of the most significant water quality trading programs discussed earlier (Hunter River, Long Island Sound, and Minnesota River) all involve exclusively trades between point sources. However, at least two of the inactive programs mentioned in Table 3, the Vermillion River and Non-Tidal Passaic River programs, failed to mature due to insufficient abatement cost heterogeneity among potential participants.

More significant gains from trade are likely to come from trades between point and nonpoint sources. Faeth (2000) summarizes the list of control options for both point and nonpoint sources of phosphorous, and generates a least-cost curve for reducing phosphorous in three watersheds. In each case, reductions of up to 50 percent can be achieved through low-cost changes in agricultural practices. After this point, the least-cost curve rises sharply, as low-cost agricultural options are exhausted and higher-cost point source controls (phosphorus removal and

filtration) are implemented. Faeth (2000) estimates that savings of 40–80 percent in the per-pound cost of phosphorus removal are achievable through trading between point and nonpoint sources. If US water quality trading programs are to expand significantly, particularly in smaller watersheds, they will need to increase possibilities for trading or offsets between point and nonpoint sources.

Sufficient Trading Volume

Adoption of permit trading for air pollutants has, in most cases, been prompted by a significant increase in regulatory stringency, creating demand for permits among regulated entities. For example, the Clean Air Act Amendments of 1990 used tradable pollution permits to achieve a required 50 percent reduction in sulfur dioxide emissions from coal-fired power plants, and the European Union Emissions Trading Scheme was born of a need to meet European nations' carbon dioxide emissions reduction targets under the Kyoto Protocol. In contrast, during the era in which pollution trading has been widely accepted and practiced, the goals of the Clean Water Act have been relatively unchanged, with the important exception of the Total Maximum Daily Load provisions. These "pollution budgets" for impaired water bodies have been the most important regulatory impetus for water quality trading in the United States. In fact, the regulatory driver for all but two of the US trading programs in Table 1 was the development or anticipated development of a TMDL. The two exceptions are the Colorado programs—Cherry Creek and Chatfield Reservoir—both of which were prompted by the state environmental regulatory agency's development of a total maximum annual load for nutrients, a very similar framework.

However, disputes over Total Maximum Daily Loads have affected the ability of this regulatory driver to prompt trading activity. Litigation over the scientific models that underlie the development of these pollution budgets is common. In the Upper Middle Snake Rock Subbasin program in Idaho, and the Willamette Partnership program ("Counting on the Environment") in Oregon, litigation over TMDLs prevented the initiation of trading programs. In other cases, backlogs in TMDL development for impaired water bodies and delays that occur for other reasons can prevent trading programs from operating. For example, Idaho's Lower Boise River Effluent Trading program began development in 1998. Trading was initially expected to commence by 2001, but a downstream TMDL (for the Snake River/Hell's Canyon) was not approved until 2004. An "implementation plan" was subsequently prepared for the Lower Boise in lieu of a TMDL, but the Environmental Protection Agency discourages water quality trading without an official TMDL.[9] Given that the Snake River TMDL calls for a 79 percent reduction in phosphorus loading from the Lower Boise, the inability to use water quality trading to bring down the costs of such a large pollution reduction will likely have significant economic

[9] Marti Bridges, Idaho Department of Environmental Quality, personal communication, June 16, 2012.

implications. But without a strong regulatory driver, the demand for permits simply does not materialize.

Even where regulation is sufficient to prompt market creation, existing water quality trading markets do not generally have a large number of buyers and sellers in comparison to their counterparts in air pollution regulation. While it operated, the sulfur dioxide allowance trading program comprised approximately 2,500 sources. In contrast, of the six significant programs described in detail earlier, the numbers of trading participants (thus far) are as follows: Hunter River, 23 point sources; Long Island Sound, 79 point sources; South Nation, 16 point sources, plus a single clearinghouse; Minnesota River, 45 point sources; Pennsylvania Nutrient Credit Trading, nine point sources and three brokers for nonpoint sources of effluent; and Virginia's Chesapeake Bay Watershed Nutrient Credit Exchange, 34 point sources, plus the clearinghouse and a few nonpoint source participants.

Thin participation in these programs could be attributed to the restriction of trading within watershed boundaries or simply to the fact that some of the programs are quite new. Nonetheless, even after accounting for the small number of participants, trading volume has been strikingly low. A number of other explanations have been offered, some already discussed: some farmers do not trust the programs, even if participation may be to their financial benefit (Breetz, Fisher-Vanden, Jacobs, Schary 2005); liability rules may discourage participation of point sources (Stephenson and Shabman 2011; Jarvie and Solomon 1998); lack of regulatory stringency may limit demand (King and Kuch 2003); and the existence of agricultural subsidies for nutrient reductions, which may substitute for participation in trading programs, could hamper supply of credits (King and Kuch 2003). In the future, if comprehensive Total Maximum Daily Loads are developed for very large watersheds—the 2010 Chesapeake Bay TMDL is a good example—these markets may grow to encompass many more participants than has been the norm.

Trading Flexibility

An optimal approach to pollution trading is first to allocate allowances to sources and then to grant full discretion to these sources to decide how reductions will be made. Shabman and Stephenson (2007) distinguish between two types of flexibility: waste control flexibility, which allows the source to decide how reductions will be made; and exchange flexibility, which allows sources to trade across time and location. Although water quality trading, in most cases, allows for waste control flexibility, exchange flexibility is limited. Sources are not typically allowed to trade across watershed boundaries or time, and trades are subject to discounting through trading ratios. There are logical reasons for some of these restrictions, often involving non-uniform mixing over space and time. For example, hypoxia from nutrient pollution tends to be a warm-weather phenomenon, thus regulators may not want to allow ambient reductions in winter months to be banked to allow higher discharge in summer months. That said, exchange restrictions do reduce trading volumes and potential cost savings.

One particular issue of inflexibility is that in trades involving point and nonpoint sources, nonpoint sources may be required to prove that reductions have been made before credit is awarded to point sources. In most water quality offset programs, regulators must approve each credit purchase by point sources through modification of their existing permits, raising transaction costs significantly and stifling the cost-effectiveness potential of this approach (Schary and Fisher-Vanden 2004). The preapproval requirement adds to transaction costs and inhibits trading.

Policy Recommendations and Future Research

Since 1990, the Environmental Protection Agency, prodded by the courts, has pushed forward on the development and enforcement of ambient Total Maximum Daily Load "pollution budgets" for impaired water bodies. At the same time, the EPA has shown increasing support for experimentation with market-based approaches applied to water pollution control, including water quality trading and offset programs.

Some of these water quality trading programs face natural limits. For example, the offset programs summarized at the bottom of Table 1 are small-scale examples of what is achievable through more flexible regulatory approaches under the Clean Water Act, but they generally involve case-by-case negotiation between point sources and regulators, resulting in unique modifications to a single discharge permit, or a very small set of permits. The cost savings from this approach will always be disappointing in comparison to the potential savings from larger-scale trading programs envisioned by regulators in their promotion of water quality trading (US Environmental Protection Agency 2001, 2003). Given the very high transaction costs associated with such an approach, it is unlikely that these offset programs will expand to make a significant dent in the total cost of compliance with US water quality regulations.

There is greater cause for optimism when we consider the trading programs summarized at the top of Table 1, particularly those active programs described in detail earlier in the paper. However, the active programs developed thus far are significantly thinner than what might be optimal from an economic perspective. The reasons for this can be described along the two primary dimensions raised at the beginning of the paper: 1) challenges due to the physical characteristics of water pollution problems; and 2) challenges posed by the implied rights to pollute created by the current regulatory environment. We summarize these primary challenges below and note that some are easier to surmount than others.

Challenges Inherent to Water Quality Problems

Damages from water pollution can vary significantly with the location of the discharge, in contrast to the more straightforward cases described by economists in the original theory of environmental policy instrument choice, and in contrast to

many of the applications of cap-and-trade programs thus far (to greenhouse gases and, to some extent, sulfur dioxide and nitrogen oxides). This spatial heterogeneity inherent to water quality problems requires water quality trading programs to institute spatial trading ratios, zones, or other mechanisms to ensure that environmental goals are met. This is a surmountable problem and has been addressed both in theory and in practice.

A related problem has to do with the spatial scope of trading programs. Due in part to the problem of location-specific damages, water quality trading takes place within specific watersheds, limiting the potential number of participating sources, abatement cost heterogeneity, and other market dimensions (US Environmental Protection Agency 2003). An obvious way to increase trading volume is to combine multiple Total Maximum Daily Loads across watersheds (Faeth 2000) to the extent allowed by the particular water quality problem under concern. A recent example is the broad Chesapeake Bay TMDL, which encompasses the Potomac, Susquehanna, James, and Rappahannock river basins (and five smaller basins), to address hypoxia related to nutrient runoff in the Bay. This single TMDL has given rise to nascent trading programs in at least three states and may be a model for a broader vision of how trading can work for particular regional water quality problems. The challenge within such systems is to ensure that local upstream water quality standards are met (for example, in the Chesapeake Bay's individual river basins) while reducing the cost of achieving regional water quality goals (in the Bay, itself) as much as possible through trading.

Another related issue is that nonpoint source pollution—the most significant remaining source of water pollution in the United States and many industrialized countries, and the source of the lowest-cost abatement opportunities for many contaminants—is characterized by significant uncertainty in measurement and monitoring. Unfortunately, to address this uncertainty, regulators often require two, three, four, or more times as much in abatement from nonpoint sources in exchange for a reduction of one unit in point source discharge. This is not the economically optimal way to address such uncertainty, and in recent interviews conducted as part of a water quality trading evaluation, respondents suggested that this approach was a major barrier to increased point–nonpoint source trading (Industrial Economics 2008). Several potential solutions have been offered: regulators can develop a preapproved list of "best management practices" for reducing agricultural nonpoint source pollution, with accepted performance metrics (Schary and Fisher-Vanden 2004); trading programs could denominate abatement units that incorporate both the mean and variance of expected abatement in nonpoint sources (Horan and Shortle 2011); and point–nonpoint source trading could be accompanied by a tax that kicks in if nonpoint source pollution is not reduced as projected (Segerson and Wu 2006). Only the first of these three solutions has been applied in practice, but even where regulators develop a preapproved list of best management practices, they still require more than one unit of abatement from these efforts in exchange for one unit of credit to the purchasing point source.

Challenges Due to Implicit Rights to Pollute and Other Regulatory Barriers

Perhaps the most significant barrier to expanding water quality trading in the United States (and, indeed, to improving water quality at all) is the effective exclusion of agricultural nonpoint sources from direct water quality regulation. As a result, these large collective contributors to water quality problems are not obligated to abate, and must instead be offered incentives to engage in abatement. This exclusion seems at present to be politically nonnegotiable. The water quality implications of agricultural runoff are dealt with primarily by the US Department of Agriculture through federal subsidies for best management practices designed to entice farmers to produce environmental quality along with their other outputs. The use of Total Maximum Daily Loads in combination with water quality trading offers a mechanism for improving water quality in a more cost-effective manner, by allowing regulated point sources with much higher abatement costs to purchase credits from nonpoint sources.

However, this approach has its limits. First, recall that little remains to be achieved through point source abatement in many US rivers and streams (Bingham et al. 2000). Thus, it is not just the low-hanging fruit that is in short supply among point sources of water pollution; the fruit, altogether, is becoming scarce. In many watersheds, the remaining point source pollution problem is not a significant fraction of overall pollution, the vast majority of which is contributed by nonpoint sources. Thus, even if point sources are required to purchase many units of nonpoint source reductions for every unit of credit they receive (as most programs are structured), the net result for water quality may not be a significant improvement.[10]

Second, though participation in water quality trading programs would appear to be in many farmers' financial interest, it has often proven difficult to encourage them to participate. Many farmers have a historic mistrust of regulators, or they may worry that the monitoring required for participation in water quality trading is a step toward full incorporation in the regulatory structure; thus, it may be necessary to work through trusted third parties or existing relationships such as cooperatives or irrigation districts to deal with these issues (Breetz, Fisher-Vanden, Jacobs, and Schary 2005). In addition, the federal US Department of Agriculture subsidies for water quality measures on farms are often an appealing alternative to participation in water quality trading, limiting farmers' interest in participation. Program objectives differ significantly from those of water pollution regulation under the Clean Water Act, so combining them would be difficult (Breetz and Fisher-Vanden 2007). For example, one concern is whether and how farmers should "double dip," receiving Department of Agriculture subsidy payments as well as payments from credit buyers in a water quality trading program (Woodward 2011; Horan, Shortle, and Abler 2004). Economists have proposed reasonably low-information ways to deal with the problem of uncertain nonpoint source pollution flows in the context of a water pollution tax (Segerson 1988; Xepapadeas 1991, 1992; Herriges, Govindasamy, and

[10] We owe this point to Leonard Shabman at Resources for the Future.

Shogren 1994; Horan, Shortle, and Abler 1998; Hansen 1998). Working out how a water pollution abatement subsidy (like those received by farmers from US Department of Agriculture programs) might be integrated with a trading system in which credit buyers face an enforced cap would be an important step forward.

Third, aside from the distortions introduced by the exclusion of major agricultural pollution sources from the "caps" represented by ambient water quality standards, there are other challenges to expanding water pollution trading related to the structure of regulations. Interviewees in a recent water quality trading program evaluation suggest that the National Pollutant Discharge Elimination System permitting process should be modified and made more flexible to better support trading (Industrial Economics 2008). Currently, when a point source wishes to use purchased credits to offset a portion of the discharge limit specified in its permit, all effluent covered in the entire permit must be reopened for discussion. However, there are now some examples in which a single permit has been issued for a particular contaminant, across many point sources. These "aggregate permits" are similar to "emissions bubble" approaches for air quality in that permitted sources are jointly responsible for meeting a standard and may engage in cost-reducing trades in order to do this. Aggregate permits are critical to the functioning of promising water quality trading programs in the Tar-Pamlico, Neuse, and Minnesota Rivers.

Beyond these specific suggestions for addressing structural challenges related to water quality problems themselves and to the institutions that manage them, some of the larger water quality trading programs could be analyzed empirically. Such measurement after the programs have taken place occurs only rarely in regulatory settings and was only incorporated into routine US regulatory functions in January 2011 (Executive Order No. 13563, 2011). Little is known, however, about how well any of these trading programs has actually worked in terms of both environmental impacts and abatement cost savings, though there is a good before-the-program analysis of the potential cost savings from trading under the Chesapeake Bay program as a whole (Van Houtven, Loomis, Baker, Beach, and Casey 2012). The Hunter River, Long Island Sound, and Minnesota River programs may be particularly good candidates for such analysis, since they are reasonably large and have been operating for several years. If more were known quantitatively about environmental outcomes and cost-effectiveness, regulators might demonstrate more flexibility in the future design and implementation of water quality trading programs.

Similarly, the development of efficient and effective trading programs could be helped by new field research targeted at developing a better understanding of factors such as: the effectiveness of nonpoint source controls; the impact of behavioral incentives for farmers to engage in trading with point sources of water pollution; or the potential for alternative approaches such as trading in the mean and variance of pollution or the use of nonperformance taxes as an alternative to high point–nonpoint source trading ratios that hamper trading. Field experimentation using randomization is particularly useful in sorting out policy complications like these that may be hard to understand or control (List 2011).

The scope for water quality trading will always be significantly smaller than for permit trading related to air quality for reasons inherent to water pollution problems, such as the need to limit some trading programs to within a watershed, significant non-uniform mixing of pollutants, and difficulties in measuring and monitoring nonpoint sources. However, the economic performance of market-based instruments in practice—regardless of the environmental objective—may always be disappointing relative to the theoretical ideal (Tietenberg 1990). Many current barriers to expanding trading regionally have more to do with program design than the physical characteristics of water pollution, and can potentially be overcome. Today's trading programs may serve as important laboratories for researchers in economics, supporting analysis that leads to better future program design and eventual expansion in the use of these cost-effective policy instruments.

■ *We are grateful for excellent research assistance from Anne Riddle. Jim Shortle and Claire Schary provided helpful comments on an earlier draft.*

References

Benkovic, Stephanie, and Joseph Kruger. 2001. "To Trade or Not to Trade? Criteria for Applying Cap and Trade." *The Scientific World* (2001)1.

Bingham, Tayler H., Timothy R. Bondelid, Brooks M. Depro, Ruth C. Figueroa, A. Brett Hauber, Suzanne J. Unger, and George L. Van Houtven. 2000. *A Benefits Assessment of Water Pollution Control Programs since 1972: Part 1, The Benefits of Point Source Controls for Conventional Pollutants in Rivers and Streams.* EPA Contract Number 68-C6-0021. Research Triangle Park, NC: Research Triangle Institute.

Boyd, James. 2000. "The New Face of the Clean Water Act: A Critical Review of the EPA's New TMDL Rules." *Duke Environmental Law and Policy Forum* 11(1): 39–87.

Branosky, Evan, Cy Jones, and Mindy Selman. 2011. "Comparison Tables of State Nutrient Trading Programs in the Chesapeake Bay Watershed." WRI Fact Sheet. Washington, DC: World Resources Institute, May.

Breetz, Hanna L., and Karen Fisher-Vanden. 2007. "Does Cost-Share Replicate Water Quality Trading Projects? Implications for a Partnership." *Review of Agricultural Economics* 29(2): 201–215.

Breetz, Hanna L., Karen Fisher-Vanden, Laura Garzon, Hannah Jacobs, Kailin Kroetz, and Rebecca Terry. 2004. *Water Quality Trading and Offset Initiatives in the U.S.: A Comprehensive Survey.* Database developed at Dartmouth College, Hanover, New Hampshire. Available at: http://agsci.psu.edu/enri/centers/multi-college/wqt-database-8_5_04.

Breetz, Hanna L., Karen Fisher-Vanden, Hannah Jacobs, and Claire Schary. 2005. "Trust and Communication: Mechanisms for Increasing Farmers' Participation in Water Quality Trading. *Land Economics* 81(2): 170–90.

Carson, Richard T., and Robert Cameron Mitchell. 1993. The Value of Clean Water: The Public's Willingness to Pay for Boatable, Fishable, and Swimmable Quality Water. *Water Resources Research* 29(7): 2445–54.

Connecticut Department of Environmental Protection. 2010. "Connecticut's Nitrogen Credit Exchange—An Incentive-based Water Quality Trading Program." Connecticut Department of Environmental Protection, Bureau of Water Protection and Land Reuse, Hartford, CT.

Connecticut Department of Environmental

Protection. 2011. "Report of the Nitrogen Credit Advisory Board for Calendar Year 2010 to the Joint Standing Environment Committee of the General Assembly." September 30. Connecticut Department of Environmental Protection, Hartford, CT.

Environment Canada. n.d. "Agents of Change: South Nation Conservation." http://www.ec.gc.ca/p2/default.asp?lang=En&n=21E379B9-1.

Executive Order No. 13563. *Federal Register* 76(14): 3821–23. Available at: http://www.gpo.gov/fdsys/pkg/FR-2011-01-21/pdf/2011-1385.pdf.

Faeth, Paul. 2000. "Fertile Ground: Nutrient Trading's Potential to Cost-Effectively Improve Water Quality." World Resources Institute, Washington, DC.

Fang, Feng, K. William Easter, and Patrick L. Brezonik. 2005. "Point–Nonpoint Source Water Quality Trading: A Case Study in the Minnesota River Basin." *JAWRA: Journal of the American Water Resources Association* 41(3): 645–58.

Farrow, R. Scott, Martin T. Schultz, Pinar Celikkol, and George L. Van Houtven. 2005. "Pollution Trading in Water Quality Limited Areas: Use of Benefits Assessment and Cost-Effective Trading Ratios." *Land Economics* 81(2): 191–205.

Freeman, A. Myrick III. 2000. "Water Pollution Policy." In *Public Policies for Environmental Protection*, 2nd edition, edited by Paul R. Portney and Robert N. Stavins, 169–213. Washington, DC: Resources for the Future.

Gianessi, Leonard P., Henry M. Peskin, and G. K. Young. 1981. "Analysis of Water Pollution Control Policies: 1. A National Network Model." *Water Resources Research* 17(4): 796–802.

Hansen, Lars Gårn. 1998. "A Damage Based Tax Mechanism for Regulation of Non-Point Emissions." *Environmental and Resource Economics* 12(1): 99–112.

Harrington, Winston, Alan J. Krupnick and Henry M. Peskin. 1985. "Policies for Nonpoint-source Water Pollution Control." *Journal of Soil and Water Conservation* 40(1): 27–32.

Herriges, Joseph A., Ramu Govindasamy, and Jason F. Shogren. 1994. "Budget-Balancing Incentive Mechanisms." *Journal of Environmental Economics and Management* 27(3): 275–85.

Horan, Richard D. 2001. "Differences in Social and Public Risk Perceptions and Conflicting Impacts on Point/Nonpoint Trading Ratios." *American Journal of Agricultural Economics* 83(4): 934–41.

Horan, Richard D., and James S. Shortle. 2005. "When Two Wrongs Make a Right: Second-Best Point–Nonpoint Trading Ratios." *American Journal of Agricultural Economics* 87(2): 340–52.

Horan, Richard D., and James S. Shortle. 2011. "Economic and Ecological Rules for Water Quality Trading." *JAWRA: Journal of the American Water Resources Association* 47(1): 59–69. DOI: 10.1111/j.1752-1688.2010.00463.x.

Horan, Richard D., James S. Shortle, and David G. Abler. 1998. "Ambient Taxes when Polluters have Multiple Choices." *Journal of Environmental Economics and Management* 36(2): 186–99.

Horan, Richard D., James S. Shortle, and David G. Abler. 2004. "The Coordination and Design of Point–Nonpoint Trading Programs and Agri-Environmental Policies." *Agricultural and Resource Economics Review* 33(1): 61–78.

Hung, Ming-Feng, and Daigee Shaw. 2005. "A Trading-Ratio System for Trading Water Pollution Discharge Permits." *Journal of Environmental Economics and Management* 49(1): 83–102.

Industrial Economics, Incorporated. 2008. *EPA Water Quality Trading Evaluation: Final Report.* Evaluation performed for the EPA's Office of Policy, Economics and Innovation (OPEI) under Contract EP-W-04-023. Washington, DC: US Environmental Protection Agency, Office of Policy, Economics and Innovation.

Jarvie, Michelle, and Barry Solomon. 1998. "Point–Nonpoint Effluent Trading in Watersheds: A Review and Critique." *Environmental Impact Assessment Review* 18(2): 135–57.

King, Dennis M., and Peter J. Kuch. 2003. "Will Nutrient Credit Trading Ever Work? An Assessment of Supply and Demand Problems and Institutional Obstacles." *Environmental Law Reporter* 33(5): 10352–68.

Konishi, Yoshifumi, Jay Coggins, and Bin Wang. Forthcoming. "Water Quality Trading: Can We Get the Prices of Pollution Right?" *Journal of Environmental Economics and Management.*

List, John A. 2011. "Why Economists Should Conduct Field Experiments and Fourteen Tips for Pulling One Off. *Journal of Economic Perspectives* 25(3): 3–16.

Lyon, Randolph M., and Scott Farrow. 1995. "An Economic Analysis of Clean Water Act Issues." *Water Resources Research* 31(1): 213–23.

Malik, Arun S., David Letson, and Stephen R. Crutchfield. 1993. "Point/Nonpoint Source Trading of Pollution Abatement: Choosing the Right Trading Ratio." *American Journal of Agricultural Economics* 75(4): 959–67.

Minnesota Pollution Control Agency. 2009. National Pollutant Discharge Elimination System (NPDES) and State Disposal System (SDS) Permit MNG420000 (Minnesota River Basin General Phosphorus Permit Phase I), modified December 1, 2009. St. Paul, Minnesota. (A permit.) http://www.pca.state.mn.us/index.php/view-document.html?gid=5997.

Montgomery, David W. 1972. "Markets in Licenses and Efficient Pollution Control Programs." *Journal of Economic Theory* 5(3): 395–418.

Morgan, Cynthia, and Ann Wolverton. 2005. "Water Quality Trading in the United States." National Center for Environmental Economics Working Paper no. 05-07, US Environmental Protection Agency. June.

Newburn, David A., and Richard T. Woodward. 2012. "An Ex Post Evaluation of Ohio's Great Miami Water Quality Trading Program. *JAWRA: Journal of the American Water Resources Association* 48(1): 156–69.

Newell, Richard N., and Robert N. Stavins. 2003. Cost Heterogeneity and the Potential Savings from Market-based Policies. *Journal of Regulatory Economics* 23(1): 43–59.

Nguyen, Nga, James Shortle, Patrick M. Reed, and Trung T. Nguyen. Forthcoming. "Water Quality Trading with Asymmetric Information, Uncertainty and Transaction Costs: A Stochastic Agent-based Modeling Framework." *Resource and Energy Economics.*

Obropta, Christopher, William Goldfarb, Peter Strom, Christopher Uchrin, Richard Boisvert, Gregory Poe, and Jeffrey Potent. 2010. *Development and Water Quality Model Validation of a Phosphorus Trading Program for the Non-Tidal Passaic River Basin.* http://www.water.rutgers.edu/Projects/trading/FINAL_Water%20quality%20trading%20report_Mar-201003242010.pdf.

O'Grady, Dennis. 2010. "Sociopolitical Conditions for Successful Water Quality Trading in the South Nation River Watershed, Ontario, Canada." *JAWRA: Journal of the American Water Resources Association* 47(1): 39–51.

Oates, Wallace E., Alan J. Krupnick, and Eric Van de Verg. 1983. "On Marketable Air-Pollution Permits: The Case for a System of Pollution Offsets." *Journal of Environmental Economics and Management* 10(3): 233–47.

Ribaudo, Marc, and Laura McCann. 2012. "Accounting for Transaction Costs in Point/Nonpoint Water Quality Trading Programs in the Chesapeake Bay Watershed." Poster prepared for the Agricultural and Applied Economics Association Annual Meeting, Seattle, WA, 12–14 August.

Rodríguez, Fernando. 2000. "On the Use of Exchange Rates as Trading Rules in a Bilateral System of Transferable Discharge Permits." *Environmental and Resource Economics* 15(4): 379–95.

Schary, Claire, and Karen Fisher-Vanden. 2004. "A New Approach to Water Quality Trading: Applying Lessons from the Acid Rain Program to the Lower Boise River Watershed." *Environmental Practice* 6(4): 281–95.

Segerson, Kathleen. 1988. "Uncertainty and Incentives for Nonpoint Pollution Control." *Journal of Environmental Economics and Management* 15(1): 87–98.

Segerson, Kathleen, and JunJie Wu. 2006. "Nonpoint Pollution Control: Inducing First-Best Outcomes through the Use of Threats." *Journal of Environmental Economics and Management* 51(2): 165–84.

Selman, Mindy, Suzie Greenhalgh, Evan Branosky, Cy Jones, and Jenny Guiling. 2009. "Water Quality Trading Programs: An International Overview." WRI Issue Brief, no. 1, World Resources Institute, Washington, DC.

Shabman, Leonard, and Kurt Stephenson. 2007. "Achieving Nutrient Water Quality Goals: Bringing Market-like Principles to Water Quality Management." *JAWRA: Journal of the American Water Resources Association* 43(4): 1076–89.

Shortle, James S. 1990. "The Allocative Efficiency Implications of Water Pollution Abatement Cost Comparisons." *Water Resources Research* 26(5): 793–97.

Shortle, James S., and James W. Dunn. 1986. "The Relative Efficiency of Agricultural Source Water Pollution Control Policies." *American Journal of Agricultural Economics* 68(3): 668–77.

Stavins, Robert N. 1998. "What Can We Learn from the Grand Policy Experiment? Lessons from SO_2 Allowance Trading." *Journal of Economic Perspectives* 12(3): 69–88.

Stephenson, Kurt, and Leonard Shabman. 2011. "Rhetoric and Reality of Water Quality Trading and the Potential for Market-like Reform." *JAWRA: Journal of the American Water Resources Association* 47(1): 15–28.

Tietenberg, Tom H. 1985. *Emission Trading: An Exercise in Reforming Pollution Policy.* Washington, DC: Resources for the Future.

Tietenberg, Tom H. 1990. "Economic Instruments for Environmental Regulation." *Oxford Review of Economic Policy* 6(1): 17–33.

US Environmental Protection Agency. 2001. *The National Costs of the Total Maximum Daily Load Program.* EPA-841-D-01-003. Washington, DC: Environmental Protection Agency.

US Environmental Protection Agency. 2003. Final Water Quality Trading Policy. (Office of Water, Water Quality Trading Policy, January 13, 2003.) http://water.epa.gov/type/watersheds/trading/finalpolicy2003.cfm.

US Environmental Protection Agency. 2009. *National Water Quality Inventory: Report to Congress,* 2004 Reporting Cycle. Washington, DC: Environmental Protection Agency.

Van Houtven, George, Ross Loomis, Justin Baker, Robert Beach, and Sara Casey. 2012. "Nutrient Credit Trading for the Chesapeake

Bay: An Economic Study." Report prepared for the Chesapeake Bay Commission, May 2012, RTI International, Research Triangle Park, NC.

Wallace, Katherine Hay. 2007. "Trading Pollution for Water Quality: Assessing the Effects of Market-based Instruments in Three Basins." Unpublished master's thesis, Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA.

Werblow, Steve. 2007. "Water Quality Trading Update." *Partners: A Quarterly Publication of the Conservation Technology Information Center* October, 25(3). http://partnersarchive.ctic.org/partners/090107/emerging.asp.

Wood, Michelle. 2011. "Alpine Cheese Water Quality Trading Project Proves Successful." *Farm and Dairy*, January 13. http://www.farmanddairy.com/columns/holmes-co-alpine-cheese-water-quality-trading-project-proves-successful/19895.html.

Woodward, Richard T. 2011. "Double-dipping in Environmental Markets." *Journal of Environmental Economics and Management* 61(2): 153–69.

Woodward, Richard T., Ronald A. Kaiser, and Aaron-Marie Wicks. 2002. "The Structure and Practice of Water Quality Trading Markets." *JAWRA: Journal of the American Water Resources Association* 38(4): 967–80.

Xepapadeas, A. P. 1991. "Environmental Policy under Imperfect Information: Incentives and Moral Hazard." *Journal of Environmental Economics and Management* 20(2): 113–26.

Xepapadeas, A. P. 1992. "Environmental Policy Design and Dynamic Nonpoint-Source Pollution." *Journal of Environmental Economics and Management* 23(1): 22–39.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 173–196

Thirty Years of Prospect Theory in Economics: A Review and Assessment

Nicholas C. Barberis

In 1979, two Israeli psychologists, Daniel Kahneman and Amos Tversky, already famous for their work on judgment heuristics, published a paper in the journal *Econometrica* titled "Prospect Theory: An Analysis of Decision under Risk." The paper accomplished two things. It collected in one place a series of simple but compelling demonstrations that, in laboratory settings, people systematically violate the predictions of expected utility theory, economists' workhorse model of decision making under risk. It also presented a new model of risk attitudes called "prospect theory," which elegantly captured the experimental evidence on risk taking, including the documented violations of expected utility.

More than 30 years later, prospect theory is still widely viewed as the best available description of how people evaluate risk in experimental settings. Kahneman and Tversky's papers on prospect theory have been cited tens of thousands of times and were decisive in awarding Kahneman the Nobel Prize in economic sciences in 2002. (Tversky would surely have shared the prize had he not passed away in 1996 at the age of 59.)

It is curious, then, that so many years after the publication of the 1979 paper, there are relatively few well-known and broadly accepted applications of prospect theory in economics. One might be tempted to conclude that, even if prospect theory is an excellent description of behavior in experimental settings, it is less relevant outside the laboratory. In my view, this lesson would be incorrect. Rather, the main reason that it has taken so long to apply prospect theory in economics is that, in a sense that I make precise in the next section, it is hard to know exactly *how* to apply

■ *Nicholas C. Barberis is the Stephen and Camille Schramm Professor of Finance, Yale School of Management, New Haven, Connecticut.*

http://dx.doi.org/10.1257/jep.27.1.173. doi=10.1257/jep.27.1.173

it. While prospect theory contains many remarkable insights, it is not ready-made for economic applications.

Over the past decade, researchers in the field of behavioral economics have put a lot of thought into how prospect theory should be applied in economic settings. This effort is bearing fruit. A significant body of theoretical work now incorporates the ideas in prospect theory into more traditional models of economic behavior, and a growing body of empirical work tests the predictions of these new theories. In this essay, after first reviewing prospect theory and the difficulties inherent in applying it, I discuss some of this recent work. It is too early to declare this research effort an unqualified success, but the rapid progress of the last decade makes me optimistic that at least some of the insights of prospect theory will eventually find a permanent and significant place in mainstream economic analysis.

The Prospect Theory Model

The original version of prospect theory is described in Kahneman and Tversky (1979). While this paper contains all of the theory's essential insights, the specific model it proposed has some limitations: it can be applied to gambles with at most two nonzero outcomes, and it predicts that people will sometimes choose dominated gambles. In 1992, Kahneman and Tversky published a modified version of their theory known as "cumulative prospect theory" which resolves these problems. This version is the one typically used in economic analysis, and it is the version I briefly review here.

Consider a gamble

$$(x_{-m}, p_{-m}; x_{-m+1}, p_{-m+1}; \ldots; x_0, p_0; \ldots; x_{n-1}, p_{n-1}; x_n, p_n),$$

where the notation should be read as "gain x_{-m} with probability p_{-m}, x_{-m+1} with probability p_{-m+1}, and so on," where the outcomes are arranged in increasing order, so that $x_i < x_j$ for $i < j$, and where $x_0 = 0$. For example, a 50:50 bet to lose \$100 or gain \$200 would be expressed as $(-\$100, \frac{1}{2}; \$200, \frac{1}{2})$. Under expected utility theory, an individual evaluates the above gamble as

$$\sum_{i=-m}^{n} p_i U(W + x_i),$$

where W is current wealth and $U(\cdot)$ is an increasing and concave utility function. Under cumulative prospect theory, by contrast, the gamble is evaluated as

$$\sum_{i=-m}^{n} \pi_i v(x_i),$$

where $v(\cdot)$, the "value function," is an increasing function with $v(0) = 0$, and where π_i are "decision weights."[1]

This formulation illustrates the four elements of prospect theory: 1) reference dependence, 2) loss aversion, 3) diminishing sensitivity, and 4) probability weighting. First, in prospect theory, people derive utility from *gains and losses,* measured relative to some reference point, rather than from absolute levels of wealth: the argument of $v(\cdot)$ is x_i, not $W + x_i$. Kahneman and Tversky motivate this assumption, known as "reference dependence," with explicit experimental evidence (see, for example, Problems 11 and 12 in their 1979 paper), but also by noting that our perceptual system works in a similar way: we are more attuned to *changes* in attributes such as brightness, loudness, and temperature than we are to their absolute magnitudes.

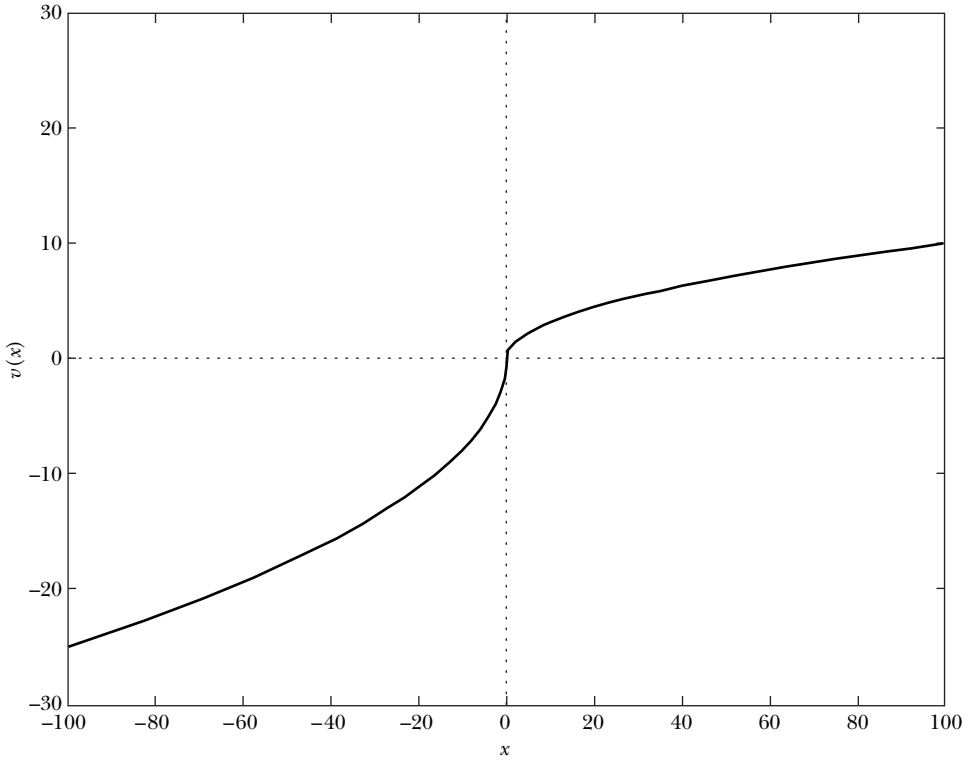
Second, the value function $v(\cdot)$ captures "loss aversion," the idea that people are much more sensitive to losses—even small losses—than to gains of the same magnitude. Informally, loss aversion is generated by making the value function steeper in the region of losses than in the region of gains. This can be seen in Figure 1, which plots a typical value function; the horizontal axis represents the dollar gain or loss x, and the vertical axis, the value $v(x)$ assigned to that gain or loss. Notice that the value placed on a \$100 gain, $v(100)$, is smaller in absolute magnitude than $v(-100)$, the value placed on a \$100 loss. Kahneman and Tversky infer loss aversion from the fact that most people turn down the gamble $\left(-\$100, \frac{1}{2}; \$110, \frac{1}{2}\right)$. As Rabin (2000) shows, it is very hard to understand this fact in the expected utility framework: the dollar amounts are so small relative to typical wealth levels that under expected utility the gamble is evaluated in an essentially risk-neutral way; given its positive expected value, it is therefore attractive. For a loss-averse individual, however, the gamble is unappealing: the pain of losing \$100 far outweighs the pleasure of winning \$110.

Third, as shown in Figure 1, the value function is concave in the region of gains but convex in the region of losses. This element of prospect theory is known as diminishing sensitivity because it implies that, while replacing a \$100 gain (or loss) with a \$200 gain (or loss) has a significant utility impact, replacing a \$1,000 gain (or loss) with a \$1,100 gain (or loss) has a smaller impact. The concavity over gains captures the finding that people tend to be risk averse over moderate probability gains: they typically prefer a certain gain of \$500 to a 50 percent chance of \$1,000. However, people also tend to be risk *seeking* over losses: they prefer a 50 percent chance of losing \$1,000 to losing \$500 for sure. This motivates the convexity over losses.[2]

[1] In taking $U(\cdot)$ to be increasing and concave and its argument to be the level of wealth, I am following the standard convention in applications of expected utility. The assumptions about the form of $U(\cdot)$ capture a simple intuition: that people prefer more wealth to less, and that an additional dollar has a smaller utility impact at higher wealth levels. The concavity assumption generates risk aversion: it predicts that people will prefer a gamble's expected value to the gamble itself.

[2] While the convexity of the value function over losses captures one important psychological intuition, it ignores another. An individual facing a loss that represents a large fraction of wealth will be *very* sensitive, not insensitive, to any additional losses. For some applications, it is important to take this into account.

Figure 1
The Prospect Theory Value Function

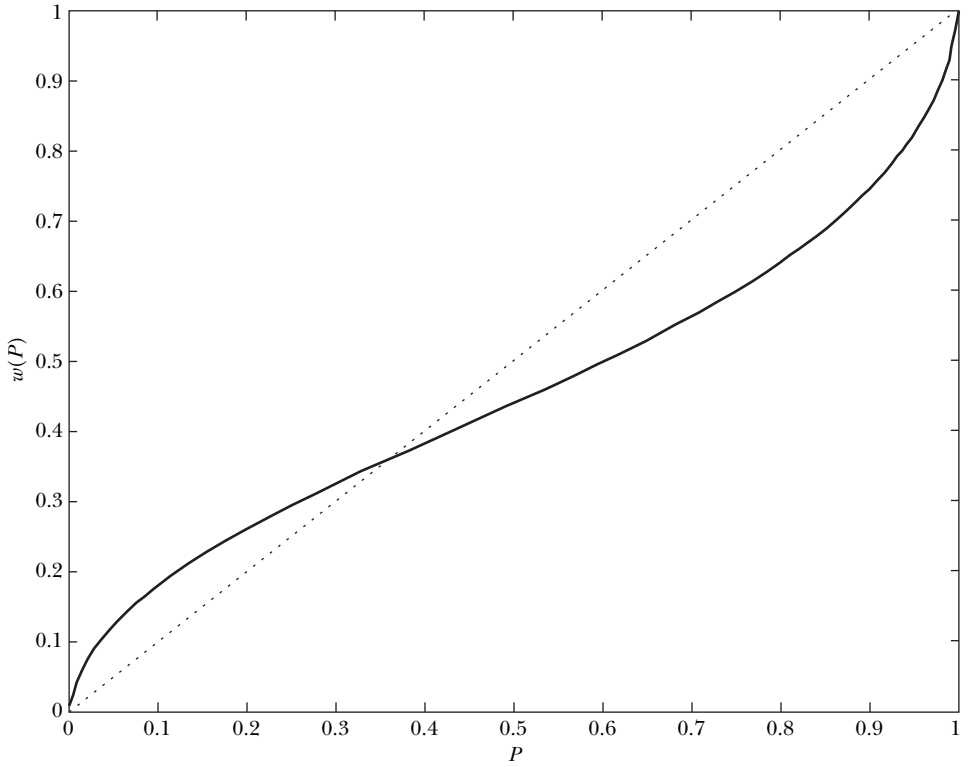


Notes: The graph plots the value function proposed by Tversky and Kahneman (1992) as part of cumulative prospect theory, namely $v(x) = x^{\alpha}$ for $x \geq 0$ and $v(x) = -\lambda(-x)^{\alpha}$ for $x < 0$, where x is a dollar gain or loss. The authors estimate $\alpha = 0.88$ and $\lambda = 2.25$ from experimental data. The plot uses $\alpha = 0.5$ and $\lambda = 2.5$ so as to make loss aversion and diminishing sensitivity easier to see.

The fourth and final component of prospect theory is probability weighting. In prospect theory, people do not weight outcomes by their objective probabilities p_i but rather by transformed probabilities or decision weights π_i. The decision weights are computed with the help of a weighting function $w(\cdot)$ whose argument is an objective probability. The solid line in Figure 2 shows the weighting function proposed by Tversky and Kahneman (1992). As is visible in comparison with the dotted line—a 45 degree line, which corresponds to the expected utility benchmark—the weighting function overweights low probabilities and underweights high probabilities.

In cumulative prospect theory, the weighting function is applied to cumulative probabilities—for example, to the probability of gaining *at least* \$100, or of losing \$50 *or more*. For the purposes of understanding the applications I describe later, the main thing the reader needs to know about probability weighting is that it leads the individual to overweight the *tails* of any distribution—in other words, to overweight unlikely extreme outcomes. Kahneman and Tversky infer this, in

Figure 2
The Probability Weighting Function



Notes: The graph plots the probability weighting function proposed by Tversky and Kahneman (1992) as part of cumulative prospect theory, namely $w(P) = P^{\delta}/(P^{\delta} + (1 - P)^{\delta})^{1/\delta}$, where P is an objective probability, for two values of δ. The solid line corresponds to $\delta = 0.65$, the value estimated by the authors from experimental data. The dotted line corresponds to $\delta = 1$, in other words, to linear probability weighting.

part, from the fact that people like both lotteries and insurance—they prefer a 0.001 chance of $5,000 to a certain gain of $5, but also prefer a certain loss of $5 to a 0.001 chance of losing $5,000—a combination of behaviors that is difficult to explain with expected utility. Under cumulative prospect theory, the unlikely state of the world in which the individual gains or loses $5,000 is overweighted in his mind, thereby explaining these choices. More broadly, the weighting function reflects the certainty equivalents people state for gambles that offer $100, say, with probability p. For example, in an experimental study by Gonzalez and Wu (1999), subjects state an average certainty equivalent of $10 for a 0.05 chance of $100, and $63 for a 0.9 chance of $100. These findings motivate the overweighting of low tail probabilities and the underweighting of high tail probabilities, respectively.

Kahneman and Tversky emphasize that the transformed probabilities π_i do not represent erroneous beliefs; rather, they are decision weights. In the

framework of prospect theory, someone who is offered a 0.001 chance of winning $5,000 knows exactly what it means for something to have a 0.001 probability of occurring; however, when evaluating the gamble, this person weights the $5,000 by more than 0.001.[3]

Subsequent to Tversky and Kahneman's (1992) paper on cumulative prospect theory, several studies have used more sophisticated techniques, in conjunction with new experimental data, to estimate the value function $v(\cdot)$ and the weighting function $w(\cdot)$ more accurately (Gonzalez and Wu 1999; Abdellaoui 2000; Bruhin, Fehr-Duda, and Epper 2010). These studies confirm the properties of these functions identified by Kahneman and Tversky: the loss aversion and diminishing sensitivity features of the value function, and the inverse S-shape of the weighting function. They provide especially strong support for probability weighting.

Challenges in Applying Prospect Theory

I noted earlier that the reason that developing applications of prospect theory in economics is taking a long time is because it is not always obvious how, exactly, to apply it. The central idea in prospect theory is that people derive utility from "gains" and "losses" measured relative to a reference point. But in any given context, it is often unclear how to define precisely what a gain or loss is, not least because Kahneman and Tversky offered relatively little guidance on how the reference point is determined.

An example from finance may help to make this difficulty more concrete. Suppose that we want to predict what kind of portfolio an investor with prospect theory preferences will hold. Right away, we need to specify the "gains" and "losses" the investor is thinking about. Are they gains and losses in overall wealth, in the value of total stock market holdings, or in the value of specific stocks? If the investor's focus is on gains and losses in the value of his stock market holdings, does a "gain" in the stock market simply mean that the return on the stock market was positive? Or does it mean that the stock market return exceeded the risk-free rate, or the return the investor *expected* to earn? And is the investor thinking about annual gains and losses or about monthly or even weekly fluctuations?

Some researchers have been scared off by the lack of a clear answer to these questions. Other researchers, however, have grasped the challenge of trying to understand how people conceptualize gains and losses in different contexts. The best way to tackle this question—and the main approach researchers are taking—is to derive the predictions of prospect theory under a variety of plausible definitions

[3] For more information about the mechanics of probability weighting, see Tversky and Kahneman (1992), Wakker (2010), or Barberis (2012). It is interesting to think about the psychological foundations of probability weighting. Tversky and Kahneman (1992) and Gonzalez and Wu (1999) offer an interpretation based on the principle of diminishing sensitivity, while Rottenstreich and Hsee (2001) give an affect-based interpretation. More recently, Bordalo, Gennaioli, and Shleifer (2012) argue that salience is an important driver of probability weighting.

of gains and losses, and to then test these predictions, both in the laboratory and in the field. Through this process, we are gradually developing better theories of how people construe these gains and losses.

One significant attempt to clarify how people think about gains and losses is the work of Kőszegi and Rabin (2006, 2007, 2009). In these papers, the authors propose a framework for applying prospect theory in economics that they argue is both disciplined and portable across different contexts. Their framework has several elements, but the most important is the idea that the reference point people use to compute gains and losses is their *expectations*, or "beliefs . . . held in the recent past about outcomes." In particular, they propose that people derive utility from the difference between consumption and *expected* consumption, where the utility function exhibits loss aversion and diminishing sensitivity. To close the model, they also assume, as a first pass, that expectations are rational, in that they match the distribution of outcomes that people will face if they follow the plan of action that is optimal, given their expectations. This framework underlies a number of the applications I describe in the next section, especially those outside the area of finance (in financial settings, a reference point such as the risk-free rate may be at least as plausible as one based on expectations).

Kőszegi and Rabin (2006) also emphasize, as do other authors, that the question at hand is not whether we should replace traditional models with models in which people derive utility *only* from gains and losses, but rather whether it is useful to consider models in which people derive utility from both gains and losses *and*, as in traditional analysis, from consumption levels. After all, even if gains and losses matter, consumption levels surely matter too, and it would be a mistake to ignore them. In some models based on prospect theory, people do derive utility only from gains and losses. However, this modeling choice simply reflects a desire for tractability, not a belief that consumption levels do not matter.

While it is widely agreed that prospect theory offers an accurate description of risk attitudes in experimental settings, some have questioned whether its predictions will retain their accuracy outside the laboratory, where the stakes are often higher and where people may have significant experience making the decision at hand. Some direct evidence bears on this issue. For example, studies using data from game shows offering large prizes and from experiments conducted in poor countries where a US researcher's budget represents a large amount of money have found that prospect theory continues to provide a good description of behavior under strong financial incentives (Kachelmeier and Shehata 1992; Post, van den Assem, Baltussen, and Thaler 2008). And while List (2003, 2004) presents evidence that prospect theory is less accurate in describing the actions of experienced traders—I return to this evidence below—Pope and Schweitzer (2011) show that prospect theory plays a role even in the behavior of highly experienced and well-incentivized professionals: in particular, professional golfers are significantly more likely to make a putt for par than a putt for scores other than par, a finding that is consistent with loss aversion relative to the reference point of par.

In the end, the best way to find out whether prospect theory can shed light on behavior in real-world settings is to derive its predictions in these settings and to confront these predictions with data. I now discuss research of this type.

Applications

Prospect theory is, first and foremost, a model of decision making under risk. As such, the most obvious places to look for applications are areas such as finance and insurance where attitudes to risk play a central role. I therefore start by discussing efforts to integrate prospect theory into these two fields and then turn to other areas of economics.[4]

Finance

Finance is the field of economics where prospect theory has been most actively applied. The research in this area applies prospect theory in three main contexts: 1) the cross section of average returns, where the goal is to understand why some financial assets have higher average returns than others; 2) the aggregate stock market; and 3) the trading of financial assets over time. I take each of these in turn.

Why do some securities have higher average returns than others? The best-known framework for thinking about this question is the famous Capital Asset Pricing Model, or CAPM. This model, which is typically derived by assuming, among other things, that investors evaluate risk according to expected utility, says that securities with higher "betas"—securities whose returns covary more with the return on the overall market—should have higher average returns. Unfortunately, this prediction has not received much empirical support (in this journal, Fama and French 2004). This raises the question: Can we do a better job explaining the cross section of average returns using a model in which investors evaluate risk in a psychologically more realistic way—specifically, according to prospect theory?

In Barberis and Huang (2008), my coauthor and I study asset prices in a one-period economy populated by investors who derive prospect theory utility from the change in the value of their portfolios over the course of the period. In this model, prospect theory leads to a new prediction, a prediction that does not emerge from the traditional analysis based on expected utility: namely, that a security's *skewness* in the distribution of its returns—even idiosyncratic skewness that is unrelated to the return on the overall market—will be priced. In particular, a positively skewed security—informally, a security whose return distribution has a right tail that is longer than its left tail—will be overpriced, relative to the price it would command in an economy with expected utility investors, and will earn a lower average return.

The intuition for this result is straightforward. By taking a significant position in a positively skewed stock, say, investors give themselves the chance—a small chance,

[4] See Camerer (2000), DellaVigna (2009), and Part IV of Kahneman (2011) for very useful, earlier discussions of prospect theory applications in economics.

admittedly—of becoming wealthy should the stock post an extraordinary right-tail performance, in other words, should it turn out to be "the next Google." Recall that under the probability weighting component of prospect theory, investors overweight the tails of the distribution they are considering—here, the distribution of potential gains and losses in wealth. This means that they overweight the unlikely state of the world in which they make a lot of money by investing in the positively skewed stock. As a result, they are willing to pay a high price for the stock, even when it means earning a low average return on it.[5]

Over the past five years, prospect theory's implications for the cross section of average returns have received significant empirical support. First, several papers, using a variety of techniques to measure skewness, have confirmed the basic prediction that more positively skewed stocks will have lower average returns (Boyer, Mitton, and Vorkink 2010; Bali, Cakici, and Whitelaw 2011; Conrad, Dittmar, and Ghysels forthcoming).

Second, several papers have argued that the skewness prediction from prospect theory can shed light on other empirical patterns. For example, a well-known puzzle is that the long-term average return of stocks that conduct an initial public offering is below that of a control group of stocks—stocks of firms that are similar to the issuing firms on important dimensions, but that happened not to do an offering. One interesting property of returns on initial public offering stocks, however, is that they are highly positively skewed: most of these stocks don't perform particularly well, but some, like Google, or Microsoft, do incredibly well. As such, prospect theory says that stocks that do an offering *should* have lower average returns. Consistent with this hypothesis, Green and Hwang (2012) find that, the higher the predicted skewness of an initial public offering stock, the lower is its long-term average return.

Researchers have used the pricing of skewness predicted by prospect theory to address several other financial phenomena: the low average return of distressed stocks, of bankrupt stocks, of stocks traded over the counter, and of out-of-the-money options (all of these assets have positively skewed returns); the low relative valuations of conglomerates as compared to single-segment firms (single-segment firms have more skewed returns); and the lack of diversification in many household portfolios (households may choose to be undiversified in positively skewed stocks so as to give themselves at least a small chance of becoming wealthy). As such, prospect theory offers a unifying way of thinking about a number of seemingly unrelated facts.[6]

[5] One attractive feature of this prediction, especially in light of the earlier discussion, is that it appears to be robust to different ways of defining what a "gain" or "loss" means to investors. In our model in Barberis and Huang (2008), investors derive prospect theory utility from changes in total wealth. The prediction that skewness will be priced continues to hold, however, if investors instead derive prospect theory utility from changes in the value of specific stocks that they own; indeed, in this case, the prediction follows even more directly. The prediction is also likely to survive the presence of expected utility investors in the economy. These investors may try to correct the overpricing of skewed securities by selling them short, but due to the risks and costs of this strategy, their efforts are unlikely to be successful.

[6] More discussion of these applications can be found in Mitton and Vorkink (2007), Eraker and Ready (2011), and Boyer and Vorkink (2011).

The *aggregate* stock market is the context for the best-known application of prospect theory in finance, namely Benartzi and Thaler's (1995) idea that prospect theory, and loss aversion in particular, can explain the famous equity premium puzzle: the fact that the average return of the US stock market has historically exceeded the average return of Treasury bills by a much greater margin than predicted by traditional consumption-based models of asset prices. According to Benartzi and Thaler, an individual who is thinking about investing in the stock market considers the historical distribution of annual stock market returns—annual because the performance of asset classes is often reported in annual terms. Since the investor is loss averse, the high dispersion of this distribution is very unappealing. To compensate for this, and thus to ensure that the investor is willing to hold his share of the supply of equity, the stock market needs to have a high *average* return, one that is significantly higher than on a safe asset like Treasury bills.[7]

Benartzi and Thaler's (1995) explanation relies not only on prospect theory, but also on an assumption known as "narrow framing," which occurs when an individual evaluates a risk separately from other concurrent risks. This manifests itself, in Benartzi and Thaler's argument, in the way investors apply prospect theory to changes in the value of one specific component of their wealth—namely, their stock market holdings. Narrow framing has been linked to many empirical findings (for example, in Barberis, Huang, and Thaler (2006), we argue that the widespread aversion to a 50:50 bet to win $110 or lose $100 is evidence not only of loss aversion but of narrow framing as well). However, we do not, as yet, have a full understanding of when and why narrow framing occurs.[8]

While Benartzi and Thaler's (1995) hypothesis is viewed by many as a plausible explanation of the equity premium puzzle, there are few direct empirical tests of it. The work that has followed their paper has instead focused on formalizing the original argument (for example, Barberis, Huang, and Santos 2001; Andries 2012; Pagel 2012a). There is, however, some evidence for the related idea that loss aversion and narrow framing can explain the nonparticipation puzzle: the fact that, historically, most households did not participate in the stock market. Dimmock and

[7] While Benartzi and Thaler (1995) focus on loss aversion, probability weighting also contributes to the high equity premium predicted by prospect theory. The reason is that the aggregate stock market is *negatively* skewed: it is subject to occasional large crashes. If investors overweight these rare events, they will require an even higher equity premium than that predicted by loss aversion alone (De Giorgi and Legg 2012). Probability weighting can therefore generate both the *high* average return on the overall stock market and the *low* average return on, for example, initial public offering stocks. In each case, the skewness of the asset, positive or negative, plays a key role.

[8] Why do we need narrow framing, rather than just loss aversion, to understand why people reject a 50:50 bet to win $110 or lose $100? Consider an individual who is loss averse but who does not engage in narrow framing. When offered the 50:50 bet, this individual does not evaluate it in isolation, but in combination with other concurrent risks—financial risk, say, or labor income risk. Loosely speaking, these other risks diversify the risk of the 50:50 bet, making it more appealing. Indeed, Barberis, Huang, and Thaler (2006) show that, unless risk aversion is implausibly high, the individual will accept the bet. This suggests that, when people turn the bet down, as they typically do, narrow framing is at work: they reject the bet because they are loss averse *and* because they evaluate it in isolation.

Kouwenberg (2010), for example, find that survey-based measures of loss aversion predict stock market participation in a cross section of households.

The third main strand of prospect theory research in finance is aimed at understanding how people trade financial assets over time. One target of interest is the "disposition effect," the empirical finding that both individual investors and mutual fund managers have a greater propensity to sell stocks that have *risen* in value since purchase, rather than stocks that have fallen in value (Odean 1998; Frazzini 2006). This behavior is puzzling because, over the horizon that these investors trade, stock returns exhibit "momentum": stocks that have recently done well continue to outperform, on average, while those that have done poorly continue to lag. As such, investors should concentrate their selling among stocks with poor past performance—but they do the opposite. This apparent unwillingness to sell stocks at a loss relative to purchase price has an important counterpart in the real estate market. Using data on Boston condominium prices from the 1990s, Genesove and Mayer (2001) find that if we take two condos, A and B, such that the two condos have the same expected selling price, but where A is expected to sell for less than its original purchase price while B is not, then the ask price that the seller posts for condo A is significantly higher than that for condo B, on average.

A long-standing idea is that this reluctance to sell assets at a loss follows naturally from prospect theory—in particular, from the convexity of the value function $v(\cdot)$ in the region of losses (Shefrin and Statman 1985). The intuition is that, if a stock (or a piece of real estate) performs poorly, this brings its owner into the loss region of the value function, where, because of the convexity, the owner becomes risk seeking. As a result, this investor holds on to the stock (or the real estate) in the hope of breaking even later on.

A number of recent papers have tried to formalize this intuition, but that task turns out to be harder than expected. In particular, some researchers have argued that, for the argument to work, the value function needs to be much more convex over losses than the experimental evidence suggests that it actually is. This issue continues to be debated (Barberis and Xiong 2009; Meng 2012).

Meanwhile, some authors have argued that the disposition effect in both the stock market and the real estate market can be better understood as a consequence of "realization utility," the idea that people derive utility *directly* from selling an asset at a gain relative to purchase price—and disutility from selling at a loss—perhaps because they think that selling assets at a gain relative to purchase price is a good recipe for long-term wealth accumulation (or conversely, that selling assets at a loss relative to purchase price is a poor recipe for wealth accumulation). In Barberis and Xiong (2012), my coauthor and I show that, if the time discount rate is sufficiently positive, even *linear* realization utility can generate a strong disposition effect, as well as other empirically observed trading patterns. While this explanation for the disposition effect differs from that based on the convexity of the prospect theory value function, it is ultimately still rooted in prospect theory, in that it relies on the investor deriving utility from gains and losses rather than from absolute wealth levels.

Insurance

Insurance is another area of economics where attitudes to risk play a central role. As such, it too is a promising place to look for applications of prospect theory. The most important consumer insurance markets are those for property and casualty insurance, mortality insurance (the main products here are life insurance and annuities), and health insurance. Thus far, prospect theory has been used to shed light on the first two of these three markets.

Sydnor (2010) studies the insurance decisions of 50,000 customers of a large home insurance company. The main decision that these households have to make is to choose a deductible from a menu of four possibilities: $100, $250, $500, and $1,000. Sydnor finds that the households that choose a $500 deductible pay an average premium of $715 per year. In choosing this policy, these households all turned down a policy with a $1,000 deductible whose average premium was just $615 per year. Given that the annual claim rate is approximately 5 percent, these households agreed to pay $100 a year to insure against a 5 percent chance of paying an additional $500 in the event of a claim! In an expected utility framework, this choice can only be rationalized by unreasonably high levels of risk aversion.

What explains this behavior? Sydnor (2010) ultimately favors an approach based on the probability weighting component of prospect theory. Under probability weighting, a household overweights tail events—in this context, the state of the world in which a claim occurs and it has to pay the deductible. Due to its extra focus on this unlikely but unpleasant outcome, the household is willing to pay a higher premium for a policy with a lower deductible. Sydnor also notes that the extent to which prospect theory can explain the data depends on the household's reference point. If the reference point is simply the household's wealth at the time it is choosing an insurance policy, then prospect theory can go some of the way, but not all the way, toward explaining the high premium the household chooses to pay. However, if, as Kőszegi and Rabin (2007) propose, the reference point is expectations about future outcomes, then prospect theory may be able to explain fully the choices we observe. The intuition is that, since a premium is a payment that a household *expects* to make, while a deductible is a payment that arises only in the unlikely event of a claim, the household doesn't experience as much loss aversion when it pays the premium as it does when it pays the deductible. As a result, it is willing to pay a higher premium.

Barseghyan, Molinari, O'Donoghue, and Teitelbaum (forthcoming) pursue this line of research further. They analyze a formal structural model of insurance choice for a prospect theory household whose reference point is its expectations about future outcomes, and estimate the model using data on home and automobile insurance choices. They, too, find evidence that probability weighting plays a role in household decisions. More precisely, their estimates suggest that, when a household chooses a policy, it significantly overweights the state of the world in which it has to file a claim. As with Sydnor's analysis, this could be because it overestimates the probability of having to file a claim; or because, as in probability weighting, it applies inflated decision weights to tail outcomes.

There are several puzzles relating to the market for mortality insurance, but the best known is the annuitization puzzle: the fact that, at the point of retirement, people allocate a much smaller fraction of their wealth to annuity products than normative models suggest they should (in this journal, Benartzi, Previtero, and Thaler 2011).

Hu and Scott (2007) argue that prospect theory offers a way of understanding why annuities are unpopular. In their framework, people think of an annuity as a risky gamble whose payoff—unknown at the moment of retirement—is the present value of the payouts to be received from the annuity before death, minus the amount initially paid for the annuity. Thus, if someone purchases an annuity at age 65 and dies at age 66, this represents a large "loss": the individual paid a lot for the annuity but received very little in return. Conversely, if this person lives until the age of 90, this represents a large "gain," in the sense that much more was received from the annuity than was initially paid in. Hu and Scott show that, if the annuity is viewed as a gamble in this way, and if it is evaluated according to prospect theory, then it will be unattractive. Loss aversion plays the largest role here: simply put, the annuity is unappealing because the individual is more sensitive to the potential loss on the annuity (if he dies soon) than to the potential gain (if he lives a long time). But probability weighting also matters: while the chance of dying very soon and hence receiving a large loss on the annuity is low, probability weighting means that this unlikely event looms large in the decision maker's thinking.[9]

The Endowment Effect

Prospect theory was originally developed as a theory of risky choice. However, in an influential paper, Thaler (1980) argues that several of the ideas in the theory may also be useful for thinking about *riskless* choice. The natural framework, formalized by Tversky and Kahneman (1991) and Kőszegi and Rabin (2006), is one where the individual derives utility from consumption relative to some reference level of consumption; and where the utility function exhibits loss aversion and diminishing sensitivity, so that, for example, the individual is more sensitive to declines in consumption relative to the reference point than to increases. A large literature, starting with Thaler (1980), has argued that some experimental findings that come under the label "endowment effect" offer strong support for this prospect theory model of riskless choice.

The term "endowment effect" actually refers to two distinct findings that may or may not be related. The first is sometimes known as "exchange asymmetries," and the second, as "WTA/WTP gaps," the gaps between willingness to accept and willingness to pay.[10]

[9] See Gottlieb (2012) for more discussion of this and other applications of prospect theory in the market for mortality insurance.

[10] The term "endowment effect" can be confusing not just because it refers to two separate empirical findings, but also because it is sometimes used to refer to evidence, and sometimes to a *theory* of that evidence, one based on prospect theory. Here, I use it to refer only to evidence.

The classic reference on exchange asymmetries is Knetsch (1989). He gives half the participants in his experiment a mug, and the other half, a candy bar. After a few minutes, during which the participants are asked to complete an unrelated questionnaire, Knetsch asks those who initially received the mug whether they would like to exchange it for the candy, and those who initially received the candy, whether they would like to exchange it for the mug. If, as in traditional economic analysis, preferences over goods do not depend on initial endowments, then whether a participant chooses to go home with a mug or with candy should not depend on the good that this participant was initially given. In fact, Knetsch finds that the initial allocation has a huge effect on subsequent choice: 89 percent of those initially given a mug opt to keep it, while only 10 percent of those initially given candy opt to exchange it for a mug.

The standard reference for willingness-to-accept/willingness-to-pay gaps is Kahneman, Knetsch, and Thaler (1990), and specifically, their Experiment 5. In this experiment, half the participants are given a mug and are asked to state, for a given list of prices, whether, for each price, they would give up the mug in exchange for that amount of money; in other words, they are asked their willingness to accept. The remaining participants are asked to state, for a given list of prices, whether, for each price, they would be willing to pay that amount of money to obtain the mug; in other words, they are asked their willingness to pay. According to traditional analysis, there should be almost no difference between these two measures. Kahneman, Knetsch, and Thaler (1990) find large differences, however: the median willingness to pay is $2.25 but the median willingness to accept is $5.75.

A leading hypothesis is that these two findings reflect the same underlying psychology of loss aversion. In the exchange asymmetry experiment, participants view an exchange as "losing" the item they were initially given and "gaining" the other item. Since they are more sensitive to losses than to gains, an exchange is unattractive, which explains why most of them stick with their initial endowment. Similarly, in the willingness-to-accept/willingness-to-pay experiment, loss aversion predicts that people will demand much more money in order to give up a mug they have previously received—here, giving up the mug is a "loss"—than they will be willing to pay in order to get one; getting a mug is the corresponding "gain."[11]

List (2003, 2004) questions the robustness of exchange asymmetries. He conducts Knetsch-type experiments at a sports card market. His participants include both nondealers and dealers; in other words, people who do not trade sports memorabilia very often, and people who do. He finds strong evidence of exchange asymmetries in the first group, but not in the second: dealers are much more willing to exchange an initial object they are given for another one of similar value. List

[11] Samuelson and Zeckhauser (1988) apply this logic more broadly. They argue that, since departing from the status quo usually entails gaining something but also losing something, and since, under loss aversion, losses loom larger than gains, people will exhibit a "status quo bias": they will cling too tightly to the status quo. They present both experimental and field evidence consistent with such a bias.

uses this evidence to suggest that prospect theory may be less useful in describing the behavior of experienced economic actors.

However, Kőszegi and Rabin (2006) argue that List's results may be fully consistent with prospect theory, albeit with an implementation of prospect theory that takes the reference point to be a person's expectations about future outcomes. Intuitively, there are fewer exchange asymmetries among dealers because dealers *expect* to exchange objects that come into their possession, and, as a result, do not experience much loss aversion when they give up the objects. This hypothesis is now being formally tested (Ericson and Fuster 2011; Heffetz and List 2011).

Plott and Zeiler (2005, 2007) show that changes in experimental conditions can significantly affect the magnitude of exchange asymmetries and willingness-to-accept/willingness-to-pay gaps, leading them to question the loss aversion interpretation of these effects. For example, they suggest that the exchange asymmetries documented by Knetsch (1989) may be due to subjects' (incorrectly) perceiving the object they were initially given as more valuable, or to them thinking of the initial object as a gift, one that it would be impolite to exchange. Plott and Zeiler's results have attracted a lot of attention, but remain controversial. For example, Kőszegi and Rabin (2006) argue, once again, that the results are consistent with loss aversion when the reference point is the decision maker's expectations. As I noted above, this hypothesis is currently being tested.

Consumption–Savings Decisions

Kőszegi and Rabin (2009) propose a way of incorporating the ideas in prospect theory into a dynamic model of consumption choice. The model builds on the authors' earlier idea that expectations are an important reference point. At each time t, the individual derives utility from two sources: 1) from the difference between actual consumption at time t and what that person recently expected consumption at that time to be, but also 2) from the difference between the individual's currently projected consumption at each future date and the consumption that person recently expected at that date. These utility terms incorporate loss aversion: the individual is more sensitive to news that consumption at some point will be lower than expected than to news that it will be higher than expected. The authors also assume that the individual is more sensitive to news that *current* consumption is different from its recently expected level than to news that future consumption will differ from its recently expected level.

This framework has some interesting implications. First, it suggests a new motive for precautionary saving: an individual facing income uncertainty will save more today so as to reduce the expected pain from finding out, later on, that it has become necessary to consume less than previously planned. Second, an individual has a tendency to overconsume, but for a reason that is quite different from the one noted in the literature on hyperbolic discounting. Specifically, in each period, the person has an incentive to surprise himself with a little extra consumption. While this comes at the cost of lower consumption later, the fact that the individual is less sensitive to news about future consumption than to news about current consumption makes the tradeoff worthwhile.

Pagel (2012b) builds on these insights to show, in a more comprehensive analysis, that the Kőszegi and Rabin (2009) framework can explain a number of facts about household consumption. For example, she finds that the precautionary saving and overconsumption motives I just described combine to produce a realistic hump-shaped pattern of consumption over the lifecycle. She also finds that the framework can shed light on the "excess sensitivity" and "excess smoothness" puzzles, whereby consumption appears to adjust insufficiently to income shocks. The intuition is that, upon receiving a negative income shock, the individual prefers to lower *future* consumption rather than current consumption. After all, news that future consumption will be lower than expected is less painful than news that current consumption is lower than expected. Moreover, when, at some future time, the individual actually lowers consumption, the pain will be limited because, by that point, expectations will have adjusted downwards.

Industrial Organization

When consumers have prospect theory preferences, firms may adopt a corresponding strategy for price setting. For example, Heidhues and Kőszegi (2012) consider a risk-neutral monopolist selling to a consumer who is loss averse, both in the dimension of the good the consumer is thinking of buying and in the dimension of money. As suggested by Kőszegi and Rabin (2006), the reference point is determined by expectations about future outcomes. In other words, the consumer derives utility from the amount of money spent relative to the amount of money he *expected* to spend; and the utility derived from obtaining the good depends on the probability with which the consumer expected to obtain it (the higher this probability, the lower the utility of obtaining the good).

It turns out that the optimal pricing strategy for this monopolist is one that supermarkets and other retailers often use in practice, namely to set a price that jumps back and forth every so often between a high "regular" price and a variety of lower sale prices. The full intuition for this conclusion has several components, but one key idea is that, by occasionally setting a low sale price at which the consumer is certain to want to buy, the firm ensures that the consumer will buy even at high prices that exceed his valuation of the good. The reason is that, because the consumer expects to obtain the good with some probability (specifically, when there is a sale on), loss aversion means that it will be painful to leave the store without the good, even if its price is high. Indeed, it turns out that, by alternating between high and low prices, the firm can induce the consumer to pay an *average* price that exceeds his valuation of the good.

Labor Supply

Prospect theory may be helpful for understanding some aspects of how labor supply reacts to wages. Research on this topic has centered on the labor supply of cab drivers. It may seem odd to focus on such a narrow segment of the labor market, but there is a reason. Models of labor supply typically assume that workers can choose the quantity of hours that they work. Driving a cab is one profession where this is literally true.

Using data on cab drivers in New York City, Camerer, Babcock, Loewenstein, and Thaler (1997) find that the number of hours that a driver works on a given day is strongly inversely related to his average hourly wage on that day. Although they do not present a formal model, the authors suggest that the data are consistent with a framework in which the driver derives prospect theory utility from the difference between his daily income and some target level, or reference level, of income. In particular, due to loss aversion, earning $20 less than the target is much more painful than earning $20 more than the target is pleasurable. It is easy to see that a driver with these preferences will typically stop work for the day after reaching the target income level. Since the driver reaches this target more quickly on days when earnings are higher, he stops working sooner on these days.[12]

A key difficulty in providing further evidence for Camerer et al.'s (1997) hypothesis is that it is not clear what determines a driver's target income. Kőszegi and Rabin (2006) break this impasse by proposing that the target is based on the driver's expectations. Specifically, they propose a model of labor supply in which the worker derives utility from the absolute levels of income and hours worked, as in traditional analysis; but in which the worker also derives prospect theory utility, on a daily basis, from the difference between income and *expected* income, and from the difference between the number of hours worked and the expected number of hours worked.

Crawford and Meng (2011) analyze this model in detail. They point out that, to a first approximation, a driver with these preferences will stop working either when he hits the income target—loss aversion means that the marginal utility of an additional dollar is much lower once he reaches this target—or when he hits the hours target (again, loss aversion means that it is much more painful to work an additional hour once this target is reached). The authors test this prediction, again using data on New York City cab drivers. As suggested by Kőszegi and Rabin, they identify a driver's targets for income and hours on the job with expected income and hours on the job, and estimate these using the driver's history of income earned and hours worked on each day of the week. The data seem to support this model. In particular, drivers appear to stop when they reach the *second* of the two targets; note that this is the income target if the driver's earnings early in the shift are lower than expected, and the hours target otherwise. These results broadly confirm Camerer et al.'s (1997) initial hypothesis, but also show the importance of identifying a driver's target with his expectations and of allowing for loss aversion both in the dimension of income

[12] This study was received skeptically in some quarters: for example, Farber (2005, 2008). The skepticism arose, in part, because Camerer et al.'s (1997) results seemed to suggest, counterintuitively, that people work less when their expected wage is high. However, Kőszegi and Rabin (2006) argue that this is not the right interpretation of the evidence. Cab drivers probably *do* work more on days when their expected earnings are higher. What Camerer et al. (1997) show is that they stop working when their earnings early in a shift have been unexpectedly high. There is no contradiction here. Intra-day wages are not significantly autocorrelated: unexpectedly high wages in the morning do not affect expected earnings in the afternoon.

and in the dimension of hours worked: the data are not consistent with a model in which the driver is loss averse *only* in the dimension of income.

Other Applications

There are other promising applications of prospect theory that I will not discuss in detail. Some recent papers study contracting between a principal and an agent when the agent has prospect theory preferences. Insights from these papers can help explain the prevalence of stock *options*, rather than just stock, in the compensation packages of both executive and nonexecutive employees (Dittman, Maug, and Spalt 2010; Spalt forthcoming).

Prospect theory has also been applied, with some success, to understanding betting markets. Snowberg and Wolfers (2010) show that probability weighting, in particular, offers a good way of thinking about one of the best-known betting anomalies, the "favorite-longshot bias," in which the market odds of longshots in horse races significantly overstate their chance of winning. In Barberis (2012), I use probability weighting to explain a broader phenomenon, namely, the popularity of casino gambling. In a dynamic setting—a casino that offers gamblers a sequence of bets, say—probability weighting predicts a time inconsistency, in the sense that the action that an individual takes in some specific state of the world may differ from the action that the individual previously planned to take in that state. I analyze this inconsistency and argue that, far from being an unattractive feature of prospect theory, it may actually be helpful for understanding observed behavior—for example, the way people often gamble longer in casinos than they were originally intending, particularly when losing.

There are areas of economics where prospect theory has not been applied very extensively, even though it has the potential to offer useful insights. Public finance, health economics, and macroeconomics are three such fields. To give just one example among many, the concept of loss aversion relative to a reference point may be a helpful way of thinking about the downward rigidity of nominal wages that plays a significant role in some models of the business cycle.

All of the applications discussed above fall under the umbrella of positive economics: we used prospect theory to make sense of observed behavior. Some applications, however, use the insights of prospect theory in a more *prescriptive* way: to nudge people toward behaviors that are viewed as more desirable. For example, Fryer, Levitt, List, and Sadoff (2012), Levitt, List, Neckerman, and Sadoff (2012), and Hossain and List (forthcoming) find that teachers, students, and factory workers, respectively, exert more effort when they are given monetary incentives framed as losses rather than gains—a finding that is consistent with loss aversion. Loss aversion is also a major influence in the design of Thaler and Benartzi's (2004) Save More Tomorrow framework for increasing saving in retirement plans: in this framework, employees' saving rates are increased only when they receive pay raises, thereby protecting them from any painful "losses" in nominal take-home pay.

The common preference for lottery-like payoffs, a preference embedded in probability weighting, has also been used to encourage a range of behaviors. In

many countries, banks offer savings' accounts that, in lieu of paying interest, enter depositors into a lottery. These products have proven popular, particularly among low-income individuals (Kearney, Tufano, Guryan, and Hurst 2010); for legal reasons, however, they are not available in the United States. In a different setting, Volpp et al. (2008a, b) try to encourage people to lose weight, or to stick to a drug regimen, by entering them into a lottery if they lose a certain number of pounds or remember to take their pills on time. This turns out to be an effective intervention.

Discussion

One might have thought that, more than 30 years after the publication of Kahneman and Tversky's (1979) paper on prospect theory, we would have a clear sense of how important a role their theory can play in economic analysis. This is not the case. Because of the difficulties inherent in applying prospect theory in economics, it is only in the last few years that we have made real progress in doing so. Indeed, this research effort is still in its early stages. While it is too soon, then, to draw any firm conclusions about prospect theory's place in economics, a few observations seem appropriate.

At this point, the fields of economics where prospect theory has been most extensively applied are finance and insurance. This emphasis is not surprising. Prospect theory came into being as a model of decision making under risk; it may therefore be best suited to settings where attitudes to risk play a crucial role. Indeed, until a few years ago, the only significant applications of prospect theory outside finance and insurance were the endowment effect and the work on the labor supply of cab drivers—a remarkably short list, and one that can be criticized: the endowment effect for being "only" an experimental finding, and the work on labor supply for being relevant to a potentially narrow segment of the working population. Nonetheless, a clear trend of the past few years is that prospect theory has extended its reach into several other areas of economics—to consumption choice, to industrial organization, to contract theory, to name just a few—and has done so in promising ways. This trend is sure to continue. Ten years from now, prospect theory's visibility in these other areas may well match or exceed its visibility in finance.

The research described in this paper also gives us a preliminary sense of the relative importance of the various components of prospect theory in economic decision making. Reference dependence is the most basic idea in prospect theory, and if any element of the theory finds a permanent place in economic analysis, it will surely be this one. Loss aversion clearly also plays a useful role in many of the applications discussed above. Diminishing sensitivity, by contrast, seems much less important. It features in one of our applications—the disposition effect—but even there, its role is unclear. Probability weighting, on the other hand, has drawn increasing interest in recent years. Indeed, within the risk-related areas of finance, insurance, and gambling, probability weighting plays a more central role than loss aversion and has attracted significantly more empirical support.

The fundamental difficulty in applying prospect theory in economics is that, even if we accept that the carriers of utility are gains and losses, it is often unclear what a gain or loss represents in any given situation. This difficulty remains unresolved; addressing it is a key challenge. Kőszegi and Rabin (2006) provide a very thoughtful analysis of this issue, but their proposal remains a hypothesis in need of more testing and, in any case, is unlikely to be completely correct. This may be particularly true in the field of finance where there are natural reference points other than expectations, and where the gains and losses that investors think about are often more likely to be the monetary gains and losses on specific investments ("narrow framing") rather than the gains and losses in consumption that Kőszegi and Rabin (2006) stress.

In this essay, I have argued that a variety of observed behaviors stem from individuals thinking about risk in the way described by prospect theory. If subsequent research confirms this claim, the natural next question is: Should anything be done about it? If people avoid annuities, "overpay" for initial public offerings, or go to casinos because they evaluate risk according to prospect theory, does that mean that these behaviors are "mistakes"? If so, should there be an effort to change people's behavior? These questions are difficult to answer because we do not, as yet, have a full understanding of whether loss aversion or probability weighting should be thought of as mistakes. One possible approach to studying this issue is to explain to people, in an appropriate way, that they may be acting the way they are because of prospect theory preferences; and to then see if, armed with this information, they change their behavior.[13]

Even prospect theory's most ardent fan would concede that economic analysis based on this theory is unlikely to replace the analysis that we currently present in our introductory textbooks. It makes sense to teach students the fundamental concepts of economics using a traditional utility function, not least because this is simpler than using prospect theory. Indeed, while Mankiw's best-selling undergraduate economics textbook devotes part of a chapter to behavioral economics, it makes no specific mention of prospect theory anywhere in its 900 pages. However, as prospect theory becomes more established in economics, a reasonable vision for future textbooks is that, once they complete the traditional coverage of some topic—of consumer behavior, say, or of consumption-savings decisions, industrial organization, or labor supply—they will follow this with a section or chapter that asks: Can we make more sense of the data using models that are based on psychologically more realistic assumptions? I expect prospect theory to figure prominently in some of these, as yet unwritten, chapters.

[13] A behavior that is closely associated with prospect theory and that *is* widely viewed as a mistake is narrow framing: evaluating a risk in isolation rather than in combination with other concurrent risks. If some phenomenon—nonparticipation in the stock market, say—is traced to narrow framing, it is easier to make a case for trying to change the pattern of thinking that underlies the phenomenon.

■ *I am very grateful to David Autor, Botond Kőszegi, John List, Ted O'Donoghue, Matthew Rabin, Andrei Shleifer, and Timothy Taylor for extensive comments on an early draft of this essay.*

References

Abdellaoui, Mohammed. 2000. "Parameter-free Elicitation of Utility and Probability Weighting Functions." *Management Science* 46(11): 1497–1512.

Andries, Marianne. 2012. "Consumption-based Asset Pricing with Loss Aversion." http://ssrn.com/abstract=2140880.

Bali, Turan G., Nusret Cakici, and Robert F. Whitelaw. 2011. "Maxing Out: Stocks as Lotteries and the Cross-section of Expected Returns." *Journal of Financial Economics* 99(2): 427–46.

Barberis, Nicholas. 2012. "A Model of Casino Gambling." *Management Science* 58(1): 35–51.

Barberis, Nicholas, and Ming Huang. 2008. "Stocks as Lotteries: The Implications of Probability Weighting for Security Prices." *American Economic Review* 98(5): 2066–2100.

Barberis, Nicholas, Ming Huang, and Tano Santos. 2001. "Prospect Theory and Asset Prices." *Quarterly Journal of Economics* 116(1): 1–53.

Barberis, Nicholas, Ming Huang, and Richard H. Thaler. 2006. "Individual Preferences, Monetary Gambles, and Stock Market Participation: A Case for Narrow Framing." *American Economic Review* 96(4): 1069–90.

Barberis, Nicholas, and Wei Xiong. 2009. "What Drives the Disposition Effect? An Analysis of a Long-standing Preference-based Explanation." *Journal of Finance* 64(2): 751–84.

Barberis, Nicholas, and Wei Xiong. 2012. "Realization Utility." *Journal of Financial Economics* 104(2): 251–71.

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C. Teitelbaum. Forthcoming. "The Nature of Risk Preferences: Evidence from Insurance Choices." *American Economic Review.*

Benartzi, Shlomo, Alessandro Previtero, and Richard H. Thaler. 2011. "Annuitization Puzzles." *Journal of Economic Perspectives* 25(4): 143–64.

Benartzi, Shlomo, and Richard H. Thaler. 1995. "Myopic Loss Aversion and the Equity Premium Puzzle." *Quarterly Journal of Economics* 110(1): 73–92.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice under Risk." *Quarterly Journal of Economics* 127(3): 1243–85.

Boyer, Brian, Todd Mitton, and Keith Vorkink. 2010. "Expected Idiosyncratic Skewness." *Review of Financial Studies* 23(1): 169–202.

Boyer, Brian H., and Keith Vorkink. 2011. "Stock Options as Lotteries." http://ssrn.com/abstract=1787365.

Bruhin, Adrian, Helga Fehr-Duda, and Thomas Epper. 2010. "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica* 78(4): 1375–1412.

Camerer, Colin F. 2000. "Prospect Theory in the Wild: Evidence from the Field." Chap. 16 in *Choices, Values and Frames,* edited by Daniel Kahneman and Amos Tversky. Cambridge University Press.

Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler. 1997. "Labor Supply of New York City Cabdrivers: One Day at a Time." *Quarterly Journal of Economics* 112(2): 407–441.

Conrad, Jennifer, Robert F. Dittmar, and Eric Ghysels. Forthcoming. "Ex Ante Skewness and Expected Stock Returns." *Journal of Finance.*

Crawford, Vincent, and Juanjuan Meng. 2011. "New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income." *American Economic Review* 101(5): 1912–32.

De Giorgi, Enrico, and Shane Legg. 2012. "Dynamic Portfolio Choice and Asset Pricing with Narrow Framing and Probability Weighting." *Journal of Economic Dynamics and Control* 36(7): 951–72.

DellaVigna, Stefano. 2009. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature* 47(2): 315–72.

Dimmock, Stephen G., and Roy Kouwenberg.

2010. "Loss-aversion and Household Portfolio Choice." *Journal of Empirical Finance* 17(3): 441–59.

Dittman, Ingolf, Ernst Maug, and Oliver Spalt. 2010. "Sticks or Carrots? Optimal CEO Compensation when Managers are Loss Averse." *Journal of Finance* 65(6): 2015–50.

Eraker, Bjorn, and Mark Ready. 2011. "Do Investors Overpay for Stocks with Lottery-like Payoffs? An Examination of the Returns on OTC Stocks." http://ssrn.com/abstract=1733225.

Ericson, Keith M. Marzilli, and Andreas Fuster. 2011. "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments." *Quarterly Journal of Economics* 126(4): 1879–1907.

Fama, Eugene F., and Kenneth R. French. 2004. "The Capital Asset Pricing Model: Theory and Evidence." *Journal of Economic Perspectives* 18(3): 25–46.

Farber, Henry S. 2005. "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers." *Journal of Political Economy* 113(1): 46–82.

Farber, Henry S. 2008. "Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers." *American Economic Review* 98(3): 1069–82.

Frazzini, Andrea. 2006. "The Disposition Effect and Underreaction to News." *Journal of Finance* 61(4): 2017–46.

Fryer, Roland G., Steven D. Levitt, John List, and Sally Sadoff. 2012. "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment." NBER Working Paper 18237.

Genesove, David, and Christopher Mayer. 2001. "Loss Aversion and Seller Behavior: Evidence from the Housing Market." *Quarterly Journal of Economics* 116(4): 1233–60.

Gonzalez, Richard, and George Wu. 1999. "On the Shape of the Probability Weighting Function." *Cognitive Psychology* 38(1): 129–66.

Gottlieb, Daniel. 2012. "Prospect Theory, Life Insurance, and Annuities." Wharton School Research Paper 44. http://ssrn.com/abstract=2119041.

Green, T. Clifton, and Byoung-Hyoun Hwang. 2012. "Initial Public Offerings as Lotteries: Skewness Preference and First-Day Returns." *Management Science* 58(2): 432–44.

Heffetz, Ori, and John A. List. 2011. "Is the Endowment Effect a Reference Effect?" NBER Working Paper 16715.

Heidhues, Paul, and Botond Kőszegi. 2012. "Regular Prices and Sales." http://emlab.berkeley.edu/~botond/rps.pdf.

Hossain, Tanjim, and John A. List. Forthcoming. "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Science.* (Published online before print, July 18, 2012, http://mansci.journal.informs.org/content/early/2012/07/18/mnsc.1120.1544.abstract.)

Hu, Wei-Yin, and Jason S. Scott. 2007. "Behavioral Obstacles in the Annuity Market." *Financial Analysts Journal* 63(6): 71–82.

Kachelmeier, Steven J., and Mohamed Shehata. 1992. "Examining Risk Preferences under High Monetary Incentives: Experimental Evidence from the People's Republic of China." *American Economic Review* 82(5): 1120–41.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* New York: Farrar, Straus, and Giroux.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98(6): 1325–48.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47(2): 263–91.

Kearney, Melissa Schettini, Peter Tufano, Jonathan Guryan, and Erik Hurst. 2010. "Making Savers Winners: An Overview of Prize-linked Savings Products." NBER Working Paper 16433.

Knetsch, Jack L. 1989. "The Endowment Effect and Evidence of Nonreversible Indifference Curves." *American Economic Review* 79(5): 1277–84.

Kőszegi, Botond, and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121(4): 1133–65.

Kőszegi, Botond, and Matthew Rabin. 2007. "Reference-Dependent Risk Attitudes." *American Economic Review* 97(4): 1047–73.

Kőszegi, Botond, and Matthew Rabin. 2009. "Reference-Dependent Consumption Plans." *American Economic Review* 99(3): 909–36.

Levitt, Steven D., John A. List, Susanne Neckerman, and Sally Sadoff. 2012. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." NBER Working Paper 18165.

List, John A. 2003. "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics* 118(1): 41–71.

List, John A. 2004. "Neoclassical Theory versus Prospect Theory: Evidence from the Marketplace." *Econometrica* 72(2): 615–25.

Meng, Juanjuan. 2012. "Can Prospect Theory Explain the Disposition Effect? A New Perspective on Reference Points." http://ssrn.com/abstract=1851883.

Mitton, Todd, and Keith Vorkink. 2007. "Equilibrium Underdiversification and the Preference for Skewness." *Review of Financial Studies* 20(4): 1255–88.

Odean, Terrance. 1998. "Are Investors Reluctant to Realize Their Losses?" *Journal of Finance* 53(5): 1775–98.

Pagel, Michaela. 2012a. "Expectations-based Reference-Dependence and Asset Pricing." Unpublished paper.

Pagel, Michaela. 2012b. "Expectations-based Reference-Dependent Life-cycle Consumption." Unpublished paper.

Plott, Charles R., and Kathryn Zeiler. 2005. "The Willingness to Pay–Willingness to Accept Gap, the 'Endowment Effect,' Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." *American Economic Review* 95(3): 530–45.

Plott, Charles R., and Kathryn Zeiler. 2007. "Exchange Asymmetries Incorrectly Interpreted as Evidence of Endowment Effect Theory and Prospect Theory?" *American Economic Review* 97(4): 1449–66.

Pope, Devin G., and Maurice E. Schweitzer. 2011. "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review* 101(1): 129–57.

Post, Thierry, Martin J. van den Assem, Guido Baltussen, and Richard H. Thaler. 2008. "Deal or No Deal? Decision Making under Risk in a Large-Payoff Game Show." *American Economic Review* 98(1): 38–71.

Rabin, Matthew. 2000. "Risk Aversion and Expected-Utility Theory: A Calibration Theorem." *Econometrica* 68(5): 1281–92.

Rottenstreich, Yuval, and Christopher K. Hsee. 2001. "Money, Kisses, and Electric Shocks: on the Affective Psychology of Risk." *Psychological Science* 12(3): 185–90.

Samuelson, William, and Richard J. Zeckhauser. 1988. "Status Quo Bias in Decision Making." *Journal of Risk and Uncertainty* 1(1): 7–59.

Shefrin, Hersh, and Meir Statman. 1985. "The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence." *Journal of Finance* 40(3): 777–90.

Snowberg, Erik, and Justin Wolfers. 2010. "Explaining the Favorite–Long Shot Bias: Is it Risk-Love or Misperceptions?" *Journal of Political Economy* 118(4): 723–46.

Spalt, Oliver. Forthcoming. "Probability Weighting and Employee Stock Options." *Journal of Financial and Quantitative Analysis.*

Sydnor, Justin. 2010. "(Over)insuring Modest Risks." *American Economic Journal: Applied Economics* 2(4): 177–99.

Thaler, Richard H. 1980. "Toward a Positive Theory of Consumer Choice." *Journal of Economic Behavior and Organization* 1(1): 39–60.

Thaler, Richard H., and Shlomo Benartzi. 2004. "Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving." *Journal of Political Economy* 112(S1): S164–S187.

Tversky, Amos, and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *Quarterly Journal of Economics* 106(4): 1039–61.

Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5(4): 297–323.

Volpp, Kevin G., Leslie K. John, Andrea B. Troxel, Laurie Norton, Jennifer Fassbender, and George Loewenstein. 2008a. "Financial Incentive-based Approaches for Weight Loss." *JAMA* 300(22): 2631–37.

Volpp, Kevin G., George Loewenstein, Andrea B. Troxel, Jalpa Dosli, Maureen Price, Mitchell Laskin, and Stephen E. Kimmel. 2008b. "A Test of Financial Incentives to Improve Warfarin Adherence." *BMC Health Services Research* 8: 272.

Wakker, Peter P. 2010. *Prospect Theory: For Risk and Ambiguity.* Cambridge University Press.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 197–222

The RAND Health Insurance Experiment, Three Decades Later[†]

Aviva Aron-Dine, Liran Einav, and Amy Finkelstein

In the voluminous academic literature and public policy discourse on how health insurance affects medical spending, the famous RAND Health Insurance Experiment stands apart. Between 1974 and 1981, the RAND experiment provided health insurance to more than 5,800 individuals from about 2,000 households in six different locations across the United States, a sample designed to be representative of families with adults under the age of 62. The experiment randomly assigned the families to health insurance plans with different levels of cost sharing, ranging from full coverage ("free care") to plans that provided almost no coverage for the first approximately $4,000 (in 2011 dollars) that were incurred during the year. The RAND investigators were pioneers in what was then relatively novel territory for the social sciences, both in the conduct and analysis of randomized experiments and in the economic analysis of moral hazard in the context of health insurance.

More than three decades later, the RAND results are still widely held to be the "gold standard" of evidence for predicting the likely impact of health insurance reforms on medical spending, as well as for designing actual insurance policies. In the light of rapid growth in health spending and the pressure this places on public sector budgets, such estimates have enormous influence as federal and state policymakers consider potential policy interventions to reduce public spending on health care.

■ *Aviva Aron-Dine received her Ph.D. in Economics in June 2012 from the Massachusetts Institute of Technology, Cambridge, Massachusetts. Liran Einav is Professor of Economics, Stanford University, Stanford, California. Amy Finkelstein is Ford Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Einav and Finkelstein are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are aviva.arondine@gmail.com, leinav@stanford.edu, and afink@mit.edu.*

[†] To access the Appendix, visit
http://dx.doi.org/10.1257/jep.27.1.197.

doi=10.1257/jep.27.1.197

On cost grounds alone, we are unlikely to see something like the RAND experiment again: the overall cost of the experiment—funded by the US Department of Health, Education, and Welfare (now the Department of Health and Human Services)—was roughly $295 million in 2011 dollars (Greenberg and Shroder 2004).[1]

In this essay, we reexamine the core findings of the RAND health insurance experiment in light of the subsequent three decades of work on the analysis of randomized experiments and the economics of moral hazard. For our ability to do so, we owe a heavy debt of gratitude to the original RAND investigators for putting their data in the public domain and carefully documenting the design and conduct of the experiment. To our knowledge, there has not been any systematic reexamination of the original data and core findings from the RAND experiment.[2]

We have three main goals. First, we re-present the main findings of the RAND experiment in a manner more similar to the way they would be presented today, with the aim of making the core experimental results more accessible to current readers. Second, we reexamine the validity of the experimental treatment effects. All real-world experiments must address the potential issues of differential study participation and differential reporting of outcomes across experimental treatments: for example, if those who expected to be sicker were more likely to participate in the experiment when the insurance offered more generous coverage, this could bias the estimated impact of more generous coverage. Finally, we reconsider the famous RAND estimate that the elasticity of medical spending with respect to its out-of-pocket price is −0.2. We draw a contrast between how this elasticity was originally estimated and how it has been subsequently applied, and more generally we caution against trying to summarize the experimental treatment effects from nonlinear health insurance contracts using a single price elasticity.

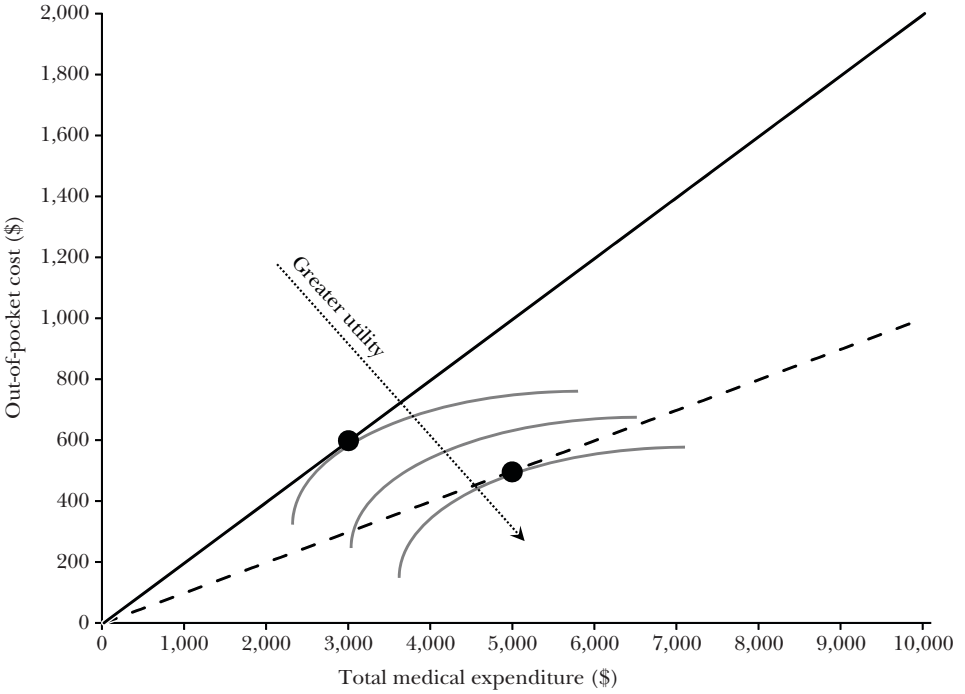
The Key Economic Object of Interest

Throughout the discussion, we focus on one of RAND's two enduring legacies—its estimates of the impact of different health insurance contracts on medical

[1] Indeed, since the RAND Health Insurance Experiment, there have been, to our knowledge, only two other randomized health insurance experiments in the United States, both using randomized variations in eligibility to examine the effect of providing public health insurance to uninsured populations: the Finkelstein et al. (2012) analysis of Oregon's recent use of a lottery to expand Medicaid access to 10,000 additional low-income adults, and the Michalopoulos et al. (2011) study funded by the Social Security Administration to see the impact of providing health insurance to new recipients of disability insurance during the two-year waiting period before they were eligible for Medicare.

[2] For many other early and influential social science experiments, researchers have gone back and reexamined the original data from the experiments in light of subsequent advances. For example, researchers have reexamined the Negative Income Tax Experiments (Greenberg and Hasley 1983; Ashenfelter and Plant 1990), the Perry preschool and other early childhood interventions experiments (Anderson 2008; Heckman, Moon, Pinto, Savelyev, and Yavitz 2010; Heckman, Pinto, Shaikh, and Yavitz 2011), the Hawthorne effect (Levitt and List 2011), Project STAR on class size (Krueger 1999; Krueger and Whitmore 2001), and the welfare-to-work experiments (Bitler, Gelbach, and Hoynes 2006).

Figure 1
The Price Elasticity of Healthcare Utilization: A Hypothetical Example



Notes: The figure presents two different budget sets arising from two different hypothetical insurance contracts: the solid line represents the budget set of an individual who has an insurance contract in which the individual has a constant 20 percent coinsurance rate, while the dashed line represents the budget set under a more generous insurance plan with a 10 percent coinsurance. The arcs are indifference curves. In this example, individuals would increase their total healthcare spending from $3,000 to $5,000 in response to a 50 percent reduction in the out-of-pocket price—that is, an elasticity of –1.33.

spending—and do not examine its influential findings regarding the health effects of greater insurance coverage. We made this choice in part because the publicly available health data are not complete (and therefore do not permit replication of the original RAND results), and in part because the original health impact estimates were already less precise than those for health spending, and our exercises below examining potential threats to validity would only add additional uncertainty.

Figure 1 illustrates the key object of interest. Healthcare utilization is summarized on the horizontal axis by the total dollar amount spent on healthcare services (regardless of whether it is paid by the insurer or out of pocket). The amount of insurance coverage is represented by how this total amount translates to out-of-pocket spending on the vertical axis. The figure presents two different budget sets arising from two different hypothetical insurance contracts: the solid line represents the budget set of an individual who has an insurance contract in which the individual pays 20 cents for any dollar of healthcare utilization—that

is a plan with a constant 20 percent coinsurance rate—while the dashed line represents the budget set under a more generous insurance plan in which the individual pays only 10 cents for any dollar of healthcare spending—that is, a 10 percent coinsurance.

Our focus in this essay is on the effect of the health insurance coverage on healthcare utilization. If utility increases in healthcare utilization and in income net of out-of-pocket medical spending, the optimal spending for an individual can be represented by the tangency point between their indifference curve and the budget set, as shown in Figure 1. The way the figure is drawn, individuals would increase their total healthcare spending from $3,000 to $5,000 in response to a 50 percent reduction in the out-of-pocket price—that is, an elasticity of –1.33.[3] A focus of the RAND experiment was to obtain estimates of this elasticity from an experiment that randomized which budget set consumers faced. This elasticity is generally known as the "moral hazard" effect of health insurance. This term was (to our knowledge) first introduced into the modern academic literature by Arrow (1963) who defined moral hazard in health insurance as the notion that "medical insurance increases the demand for medical care"; it has since come to be used more specifically to refer to the price sensitivity of demand for health care, conditional on underlying health status (Pauly 1968; Cutler and Zeckhauser 2000).

Figure 1 abstracts, of course, from many important aspects of actual health insurance contracts and healthcare consumption choices that are faced in the real world and in the RAND Health Insurance Experiment. First, summarizing healthcare utilization by its overall dollar cost does not take into account the heterogeneity in healthcare needs. One common distinction is often drawn between inpatient and outpatient spending. The former is associated with hospitalizations, while the latter is associated with visits to the doctor's office, lab tests, or procedures that do not require an overnight stay. It seems plausible that the rate at which individuals trade off healthcare spending and residual income could differ across such very different types of utilization and, therefore, that these different types of spending would respond very differently to a price reduction through insurance.

A second simplification is that Figure 1 considers two linear contracts, for which the concept of price, and price elasticity, is clearly defined. However, most health insurance contracts in the world, as well as those offered by the RAND experiment, are nonlinear, and annual healthcare utilization consists of many small and uncertain episodes that accumulate. The concept of a single elasticity, or even of a single price, is therefore not as straightforward as may be suggested by Figure 1. We return to this point later in this essay.

[3] $((P_2 - P_1)/P_1)/((Q_2 - Q_1)/Q_1) = ((5{,}000 - 3{,}000)/3{,}000)/((.1 - .2)/.2) = -1.33$. Later we will use arc elasticities, which are slightly different.

A Brief Summary of the RAND Health Insurance Experiment

In the RAND experiment, families were assigned to plans with one of six consumer coinsurance rates—that is, the share of medical expenditures paid by the enrollee—and were covered by the assigned plan for three to five years. Four of the six plans simply set different overall coinsurance rates of 95, 50, 25, or 0 percent (the last known as "free care"). A fifth plan had a "mixed coinsurance rate" of 25 percent for most services but 50 percent for dental and outpatient mental health services, and a sixth plan had a coinsurance rate of 95 percent for outpatient services but 0 percent for inpatient services (following the RAND investigators, we refer to this last plan as the "individual deductible plan"). The most common plan assignment was free care (32 percent of families), followed by the individual deductible plan (22 percent), the 95 percent coinsurance rate (19 percent), and the 25 percent coinsurance rate (11 percent).[4]

To limit the financial exposure of participants, families were also randomly assigned, within each of the six plans, to different out-of-pocket maximums, referred to as the "Maximum Dollar Expenditure." The possible Maximum Dollar Expenditure limits were 5, 10, or 15 percent of family income, up to a maximum of $750 or $1,000 (roughly $3,000 or $4,000 in 2011 dollars). On average, about one-third of the individuals who were subject to a Maximum Dollar Expenditure hit it during the year, although this of course was more likely for plans with high coinsurance rates.

The first three columns of Table 1 show the six plans, the number of individuals and families in each, and the average share of medical expenses that they paid out-of-pocket. Newhouse et al. (1993, chapter 2 and appendix B) provide considerably more detail on this and all aspects of the experiment.

Families were not assigned to plans by simple random assignment. Instead, within a site and enrollment month, the RAND investigators selected their sample and assigned families to plans using the "finite selection model" (Morris 1979; Newhouse et al. 1993, appendix B), which seeks to 1) maximize the sample variation in baseline covariates while satisfying the budget constraint for the experiment; and 2) use a form of stratified random assignment to achieve better balance across a set of baseline characteristics than would likely be achieved (given the finite sample) by chance alone.

The data come from several sources. Prior to plan assignment, a screening questionnaire collected basic demographic information and some information on health, insurance status, and past healthcare utilization from all potential enrollees. During the three-to-five year duration of the experiment, participants signed over all payments from their previous insurance policy (if any) to the RAND experiment

[4] Our analysis omits 400 additional families (1,200 individuals) who participated in the experiment but were assigned to coverage by a health maintenance organization. Due to the very different nature of this plan, it is typically excluded from analyses of the impact of cost sharing on medical spending using the RAND data (Keeler and Rolph 1988; Manning, Newhouse, Duan, Keeler, Leibowitz, and Marquis 1987; Newhouse et al. 1993).

Table 1
Plan Summary Statistics and Refusal and Attrition Rates

Plan	*Individuals (families)*	*Average out-of-pocket share*[c]	*Share refusing enrollment*	*Share attriting*	*Share refusing or attriting*
Free Care	1,894 (626)	0%	6%	5%	12%
25% Coinsurance	647 (224)	23%	20%	6%	26%
Mixed Coinsurance[a]	490 (172)	28%	19%	9%	26%
50% Coinsurance	383 (130)	44%	17%	4%	21%
Individual Deductible[b]	1,276 (451)	59%	18%	13%	28%
95% Coinsurance	1,121 (382)	76%	24%	17%	37%
All plans	5,811 (1,985)	34%	16%	10%	24%
p-value, all plans equal			< 0.0001	< 0.0001	< 0.0001
p-value, Free Care vs. 95%			< 0.0001	< 0.0001	< 0.0001
p-value, Free Care vs. 25%			0.0001	0.5590	0.0001
p-value, 25% vs. 95%			0.4100	0.0003	0.0136

Notes: "Coinsurance rate" refers to the share of the cost that is paid by the individual. In the 25 percent, mixed, 50 percent, and 95 percent coinsurance rate plans, families were assigned out-of-pocket maximums of 5 percent, 10 percent, or 15 percent of family income, up to a limit of $750 or $1,000. In the individual deductible plan, the out-of-pocket maximum was $150 per-person up to a maximum of $450 per family. The sample counts for the 95 percent coinsurance rate plans include 371 individuals who faced a 100 percent coinsurance rate in the first year of the experiment. Refusal and attrition rates are regression-adjusted for site and contact month fixed effects and interactions, because plan assignment was random only conditional on site and month of enrollment (see Newhouse et al. 1993, appendix B). "Contact month" refers to the month in which the family was first contacted by the experiment and is used in lieu of month of enrollment because month of enrollment is available only for individuals who agreed to enroll. Refusal and attrition rates exclude the experiment's Dayton site (which accounted for 1,137 enrollees) because data on Dayton refusers were lost. An individual is categorized as having attrited if he leaves the experiment at any time prior to completion.

[a] The "Mixed Coinsurance" plan had a coinsurance rate of 50 percent for dental and outpatient mental health services, and a coinsurance rate of 25 percent for all other services.

[b] The "Individual Deductible" plan had a coinsurance rate of 95 percent for outpatient services and 0 percent for inpatient services.

[c] To compute the average out-of-pocket share we compute the ratio of out-of-pocket expenses to total medical expenditure for each enrollee, and report the average ratio for each plan.

and filed claims with the experiment as if it was their insurer; to be reimbursed for incurred expenditures, participants had to file claims with the experimenters. These claim filings, which provide detailed data on health expenditures incurred during the experiment, make up the data on healthcare spending and utilization outcomes. The RAND investigators have very helpfully made all these data and detailed documentation available online, allowing us to replicate their results (almost) perfectly (see Table A1 of the online Appendix) and to conduct our own analysis of the data.[5]

[5] We accessed the RAND data via the Inter-University Consortium for Political and Social Research; the data can be downloaded at http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/6439?q=Rand+Health+Insurance+Experiment. The online Appendix and code for reproducing our results can be found at http://e-jep.org.

Experimental Analysis

As in all modern presentations of randomized experiments, we begin by reporting estimates of experimental treatment effects. We then continue by investigating potential threats to the validity of interpreting these treatment effects as causal estimates.

Empirical Framework

In our analysis, we follow the RAND investigators and use the individual-year as the primary unit of analysis. We denote an individual by i, the plan the individual's family was assigned to by p, the calendar year by t, and the location and start month by l and m, respectively. The baseline regression takes the form of

$$y_{i,t} = \lambda_p + \tau_t + \alpha_{l,m} + \varepsilon_{i,t}$$

where an outcome $y_{i,t}$ (for example, medical expenditure) is used as the dependent variable, and the explanatory variables are plan, year, and location-by-start-month fixed effects. The key coefficients of interest are the six plan fixed effects, λ_p. Because, as described earlier, there was an additional randomization of Maximum Dollar Expenditure limits, the estimated coefficients represent the average effect of each plan, averaging over the different limits that families were assigned to within the plan. Because plan assignment was only random conditional on location and start (that is, enrollment) month, we include a full set of location by start month interactions, $\alpha_{l,m}$. We also include year fixed effects, τ_t, to account for any underlying time trend in the cost of medical care. Because plans were assigned at the family rather than individual level, all regression results cluster the standard errors on the family.

Treatment Effects

Table 2 reports the treatment effects of the different plans based on estimating the basic regression for various measures of healthcare utilization. The reported coefficients (the λ_p's from the above regression) indicate the effect of the various plans on that measure of utilization relative to the free care plan (whose mean is given by the constant term). Column 1 reports results for a linear probability model in which the dependent variable takes the value of one when spending is positive and zero otherwise. In column 2, the dependent variable is the amount of annual medical spending (in 2011 dollars).

The point estimates of both specifications indicate a consistent pattern of lower spending in higher cost-sharing plans. For example, comparing the highest cost-sharing plan (the 95 percent coinsurance plan) with the free care plan, the results indicate a 17 percentage point (18 percent) decline in the fraction of individuals with zero annual medical spending and a $845 (39 percent) decline in average annual medical spending. As the last row shows, we can reject the null hypothesis that spending in the positive cost-sharing plans is equal to that in the free care plan.

Table 2
Plans' Effects on Utilization

	Total spending[a]		*Inpatient spending*		*Outpatient spending*	
	Share with any (1)	*Spending in $ (2)*	*Share with any (3)*	*Spending in $ (4)*	*Share with any (5)*	*Spending in $ (6)*
Constant (Free Care Plan, $N = 6{,}840$)	0.931 (0.006)	2,170 (78)	0.103 (0.004)	827 (60)	0.930 (0.006)	1,343 (35)
25% Coinsurance ($N = 2{,}361$)	–0.079 (0.015)	–648 (152)	–0.022 (0.009)	–229 (116)	–0.078 (0.015)	–420 (62)
Mixed Coinsurance ($N = 1{,}702$)	–0.053 (0.015)	–377 (178)	–0.018 (0.009)	21 (141)	–0.053 (0.016)	–398 (70)
50% Coinsurance ($N = 1{,}401$)	–0.100 (0.019)	–535 (283)	–0.031 (0.009)	4 (265)	–0.100 (0.019)	–539 (77)
Individual Deductible ($N = 4{,}175$)	–0.124 (0.012)	–473 (121)	–0.006 (0.007)	–67 (98)	–0.125 (0.012)	–406 (52)
95% Coinsurance ($N = 3{,}724$)	–0.170 (0.015)	–845 (119)	–0.024 (0.007)	–217 (91)	–0.171 (0.016)	–629 (50)
p-value: all differences from Free Care = 0	< 0.0001	< 0.0001	0.0008	0.1540	< 0.0001	< 0.0001

Notes: Table 2 reports the treatment effects of the different plans based on estimating the basic regression for various measures of healthcare utilization. The reported coefficients are from an ordinary least squares regression and indicate the effect of the various plans on that measure of utilization relative to the free care plan (whose mean is given by the constant term). Column 1 reports results for a linear probability model in which the dependent variable takes the value of one when spending is positive, and zero otherwise. In column 2, the dependent variable is the amount of annual medical spending (in 2011 dollars). The other columns of Table 2 break out results separately for inpatient spending and outpatient spending. Standard errors, clustered on family, are in parentheses below the coefficients. Because assignment to plans was random only conditional on site and start month (Newhouse et al. 1993), all regressions include site by start month dummy variables, as well as year fixed effects. All spending variables are inflation adjusted to 2011 dollars (adjusted using the CPI-U). Site by start month and year dummy variables are de-meaned so that the coefficients reflect estimates for the "average" site-month-year mix.

[a] Total spending is the sum of inpatient and outpatient spending (where outpatient spending includes dental and outpatient mental health spending).

The other columns of Table 2 break out results separately for inpatient spending, which accounted for 42 percent of total spending, and outpatient spending, which accounted for the other 58 percent. Once again the patterns suggest less spending in plans with higher cost-sharing. We are able to reject the null of no differences in spending across plans for "any inpatient" and for both measures of outpatient spending. The effect of cost sharing on the level of inpatient spending is consistently small and generally insignificant, suggesting that more serious medical episodes may be less price sensitive, which seems plausible.

Another way to approach the data is to look at the extent to which the effect of cost sharing might vary for those with higher levels of medical spending. To

explore this, we use quantile regressions to estimate the above equation, and then assess the way by which the estimated plan effects vary across the quantiles of medical spending. Detailed results for these specifications are available in Table A2 of the online Appendix available with this article at http://e-jep.org. The results are consistent with a lower percentage treatment effect for higher-spending individuals. This pattern is likely to arise from a combination of two effects. First, consistent with the results for inpatient spending, more serious and costly medical episodes may be less responsive to price. Second, individuals with high utilization typically hit the Maximum Dollar Expenditure limit early in the coverage year, and so for much of their coverage period they face a coinsurance rate of zero percent regardless of plan assignment.

Threats to Validity

The great strength of a randomized experimental approach, of course, is that a straight comparison of those receiving the treatment and those not receiving the treatment, like the regression coefficients reported in Table 2, can plausibly be interpreted as a causal effect of the treatment. However, this interpretation requires that no systematic differences exist across individuals who participate in the different plans that could be correlated with measured utilization. In this section, we consider in turn three possible sources of systematic differences that need to be considered in any real-world experimental context: 1) nonrandom assignment to plans, 2) differential participation in the experiment across treatment arms, and 3) differential reporting (in this case, of medical care utilization) across treatment arms.

The first potential threat to validity concerns whether the stratified random assignment to plans, described earlier, was successfully implemented. To investigate, we estimated a version of the earlier equation but, instead of using healthcare spending as the dependent variable, we used as outcomes various personal characteristics, such as age or education, of people assigned to different plans. In effect, such regressions show whether there is a statistically significant correlation between any particular characteristic of a person and the plan to which that person was assigned—which would be a warning sign for concern about the randomization process. We first focused on characteristics used by the investigators in the finite selection model that determined the randomization, including, for example, variables for size of family, age categories, education level, income, self-reported health status, and use of medical care in the year prior to the start of the experiment. Unsurprisingly, given that the assignment algorithm was explicitly designed to achieve balances across plan assignment on these characteristics, our statistical tests are unable to reject the null that the characteristics used in stratification are balanced across plans. (More specifically, we used a joint *F*-test, as reported in panel A of Table A3 of the online Appendix available with this paper at http://e-jep.org.)

We next estimated these same types of regressions, but now using as the dependent variable individual characteristics not used by the original researchers in plan assignment. These include, for example, the kind of insurance (if any) the person

had prior to the experiment, whether family members grew up in a city, suburb, or town, and spending on medical care and dental care prior to the experiment. Using these statistics, people's characteristics did not appear to be randomly distributed across the plans (as shown by the joint *F*-test results in panel B of Table A3 of the online Appendix). However, as we looked more closely, this result appeared to be driven only by assignment in the 50 percent coinsurance plan, which has relatively few people assigned to it. While these imbalances may have been due to sampling variation, there may also have been some problem with the assignment of families to the 50 percent plan; indeed, midway through the assignment process the RAND investigators stopped assigning families to this plan. With this (small) plan deleted, our statistical tests are unable to reject the null hypothesis that covariates that were not used in stratification are also balanced across plans. We proceed below on the assumption that the initial randomization was in fact valid—at least for all plans except for the 50 percent coinsurance plan. However, we also assess the sensitivity of the results to the inclusion of baseline covariates as controls.

To examine the second threat to validity—the concern that differential participation across plans might affect the findings—we begin with the observation that individuals assigned to more comprehensive insurance will have greater incentive to participate in the experiment. Indeed, the RAND investigators anticipated this issue, and attempted to offset these differential incentives by offering a higher lump sum payment for those randomized into less-comprehensive plans. While this differential payment may make participation incentives more similar across plans, it can do so only on average. Unless the participation incentive varies with a family's pre-experiment expectation of medical spending (and it did not), the incremental benefit from more comprehensive coverage remains greater for individuals who anticipate greater medical spending.

Thus, differential participation (or attrition) could bias the estimates of the spending response to coverage. For example, if individuals incur a fixed cost of participating in the experiment, high-expected-spending individuals might participate regardless of plan assignment, but lower-expected-spending individuals might be inclined to drop out if not randomized into a comprehensive plan, which could bias downward the estimated effect of insurance coverage on medical utilization. Alternatively, if high-expected-spending and low-expected-spending families were about equally likely to participate in the experiment when assigned to the free care plan, but high-expected-spending families were less likely than low-expected-spending families to participate when assigned to less-comprehensive plans, this differential selection would bias upward the estimated effect of insurance coverage on medical utilization.

Columns 4–6 of Table 1 presented earlier suggest scope for bias from differential participation across plans. Overall, 76 percent of the individuals offered enrollment ended up completing the experiment. Completion rates were substantially and systematically higher in more-comprehensive insurance plans, ranging from 88 percent in the (most comprehensive) free care plan to 63 percent in the (least comprehensive) 95 percent coinsurance plan. Most of the difference in

completion rates across plans was due to differences in initial enrollment rates—that is, the share of families refusing coverage from the experiment—although subsequent attrition from the experiment also plays a nontrivial role. As shown in the bottom rows of Table 1, neither the initial refusal nor the subsequent attrition differentials can be attributed to sampling variation alone.

The differential participation by plan assignment was noted and investigated by the original RAND investigators (Newhouse et al. 1993, Chapter 2). The RAND investigators primarily investigated attrition (rather than refusal), and focused on testing particular mechanisms by which bias might have arisen. We took a more agnostic view and implemented an omnibus test for differences in available observable pre-randomization characteristics among those completing the experiment in the different plans—and we reach somewhat different conclusions. First, we divided up all the pre-randomization measures into two groups: those that directly measure prior healthcare utilization—which are closely related to the primary post-randomization outcomes—and all other baseline demographic information. For either set of covariates (or for both combined) we are able to reject at the 1 percent level that these pre-randomization covariates are balanced across plans for those completing the experiment (using a joint *F*-test; see Table A4 in the online Appendix for additional details). These differentials mostly reflect imbalances that arise after assignment.[6] Of particular note, by the end of the experiment, there are imbalances across plans in participants' average number of doctors' visits in the year before the experiment and in the share of participants who had a medical exam in the year before the experiment.

The potential bias from differential nonresponse or attrition across experimental treatments is now a well-known concern for analysis of randomized social experiments. For example, Ashenfelter and Plant (1990) document the contamination to estimates arising from nonrandom attrition in the Negative Income Tax experiments from the 1970s, which were implemented around the same time. We discuss below possible ways of trying to account for this potential bias.

Finally, the third potential threat to validity is the extent to which participants in more comprehensive plans had differential incentives to report their medical spending. Data on medical utilization and expenditures from experimental participants were obtained from Medical Expense Report ("claims") forms which required a provider's signature and which the participant (or the healthcare provider) had to file with the experiment in order to be reimbursed for the expenditure. The incentive for filing claims was to get reimbursed, and so the filing incentive was weaker for participants enrolled in higher coinsurance rate plans (or their providers) than for those enrolled in lower coinsurance rate plans or the free care plan. For example, a participant assigned to the 95 percent coinsurance plan, who had yet to satisfy the Maximum Dollar Expenditure, would have had little to gain from filing a claim toward the end of the coverage year. This differential reporting

[6] This can be seen by comparing the balance at completion rates in Table A4 to the balance at assignment results in Table A3; both tables are in the online Appendix.

would therefore be expected to bias the estimates in the direction of overstating the spending response to coverage.[7]

Again, the original RAND investigators anticipated this potential problem and conducted a contemporaneous survey to try to determine the extent of the reporting bias (Rogers and Newhouse 1985). In this study of roughly one-third of all enrollees, the investigators contacted the providers for whom claims were filed by the participant or his family members, as well as a random subset of providers mentioned by other participants. From these providers, they requested all outpatient billing records for the participants and family members. For the 57 percent of providers who responded, the investigators matched the outpatient billing records to the experiments' outpatient claims data and computed the amounts corresponding to matched and unmatched billing records. The results indicate that, on average, participants in the free care plan failed to file claims for 4 percent of their total outpatient spending, while those in the 95 percent coinsurance plan failed to file claims for 12 percent of their total outpatient spending. Underreporting by participants in the other plans fell in between these two extremes (Rogers and Newhouse 1985, Table 7.3). Once again, in what follows we will attempt to adjust the estimates to address the bias that may arise from this greater underreporting of expenditures in the higher cost-sharing plans.

Robustness of Treatment Effects

The potential for bias in the RAND experiment has been a source of some recent controversy: for example, Nyman (2007, 2008) raises concerns about bias stemming from differential participation across plans, and the RAND investigators offer a rebuttal in Newhouse et al. (2008). To our knowledge, however, there has been no attempt to quantify the potential magnitude of the bias. Nor, to our knowledge, has there been a formal attempt to quantify the potential bias arising from the differential reporting documented by Rogers and Newhouse (1985).

Table 3 reports our results from such attempts. The different columns report results for different measures of spending, while the different panels show results for different pairwise plan combinations: free care versus 95 percent coinsurance; free care versus 25 percent coinsurance; and 25 percent versus 95 percent coinsurance. For each, we report results from four different specifications. Row 1 of each panel replicates the baseline results from Table 2, where here we also show estimates from log specifications due to the extreme sensitivity of the levels estimates to some of our adjustments.

We begin in row 2, by trying to adjust the estimates for the differential filing of claims by plan detected by Rogers and Newhouse (1985). Specifically, we proportionally scale up outpatient spending for participants in each plan based on the plan-specific underreporting percentages they report (Rogers and Newhouse 1985,

[7] Once again, this issue of differential reporting incentives by experimental assignment also plagued the Negative Income Tax experiments in the 1970s (Greenberg and Hasley 1983).

Table 3

Sensitivity of Results to Additional Covariates and Bounding Exercises

	Total spending			*Inpatient spending*		*Outpatient spending*		
	Share with any (1)	*Spending (in $) (2)*	*Spending (in logs) (3)*	*Share with any (4)*	*Spending (in $) (5)*	*Share with any (6)*	*Spending (in $) (7)*	*Spending (in logs) (8)*
Panel A: 95% Coinsurance plan vs. Free Care ($N = 10,564$)								
(1) Baseline specification (from Table 2)	−0.170	−845	−1.381	−0.024	−217	−0.171	−629	−1.361
	(0.015)	(119)	(0.096)	(0.007)	(91)	(0.016)	(50)	(0.093)
(2) Adjusted for underreporting	−0.100	−786	−1.313	−0.024	−217	−0.102	−582	−1.299
	(0.017)	(123)	(0.097)	(0.007)	(91)	(0.018)	(55)	(0.095)
(3) Adjusted for underreporting + controlling for pre-randomization covariates	−0.095	−728	−1.276	−0.023	−183	−0.096	−558	−1.261
	(0.016)	(111)	(0.087)	(0.007)	(85)	(0.016)	(50)	(0.084)
(4) Lee bounds + adjusted for underreporting	−0.080	745	−0.672	0.079	592	−0.081	151	−0.751
	(0.018)	(96)	(0.098)	(0.005)	(71)	(0.018)	(38)	(0.095)
Panel B: 25% Coinsurance plan vs. Free Care ($N = 9,201$)								
(1) Baseline specification (from Table 2)	−0.079	−648	−0.747	−0.022	−229	−0.078	−420	−0.719
	(0.015)	(152)	(0.095)	(0.009)	(116)	(0.015)	(62)	(0.093)
(2) Adjusted for underreporting	−0.065	−645	−0.734	−0.022	−229	−0.065	−418	−0.706
	(0.016)	(155)	(0.096)	(0.009)	(116)	(0.016)	(65)	(0.094)
(3) Adjusted for underreporting + controlling for pre-randomization covariates	−0.069	−585	−0.748	−0.022	−181	−0.068	−405	−0.718
	(0.014)	(137)	(0.084)	(0.008)	(107)	(0.014)	(59)	(0.082)
(4) Lee bounds + adjusted for underreporting	−0.055	639	−0.335	0.081	581	−0.054	205	−0.369
	(0.016)	(133)	(0.096)	(0.008)	(99)	(0.016)	(52)	(0.093)
Panel C: 95% Coinsurance plan vs. 25% Coinsurance plan ($N = 6,085$)								
(1) Baseline specification (from Table 2)	−0.091	−197	−0.633	−0.002	12	−0.093	−209	−0.641
	(0.020)	(160)	(0.120)	(0.009)	(122)	(0.020)	(61)	(0.117)
(2) Adjusted for underreporting	−0.035	−141	−0.579	−0.002	12	−0.037	−164	−0.592
	(0.022)	(164)	(0.122)	(0.009)	(122)	(0.022)	(66)	(0.118)
(3) Adjusted for underreporting + controlling for pre-randomization covariates	−0.026	−143	−0.529	−0.001	−2	−0.028	−153	−0.543
	(0.019)	(141)	(0.106)	(0.009)	(108)	(0.019)	(60)	(0.103)
(4) Lee bounds + adjusted for underreporting	−0.020	764	−0.248	0.078	657	−0.021	185	−0.313
	(0.022)	(105)	(0.120)	(0.006)	(78)	(0.023)	(42)	(0.117)

Notes: The table reports coefficients on plan dummies from an ordinary least squares regression; the ommitted category is the free care plan. The dependent variable is given in the column headings. Standard errors are in parentheses below the coefficients. Standard errors are clustered on familiy. Because assignment to plans was random only conditional on site and start month (Newhouse et al. 1993), all regressions include site by start month dummy variables, as well as year fixed effects to adjust for inflation; level regressions use inflation-adjusted spending variables (in 2011 dollars, adjusted using the CPI-U). Log variables are defined as log(var + 1) to accommodate zero values. The regressions adding pre-randomization covariates as controls (row 3) include the full set of covariates shown in Table A4 of the online Appendix. Adjustment for underreporting and bounding procedures are explained in the main text.

Table 7.3).[8] We do not make any adjustment to inpatient spending because there is no study on underreporting of inpatient spending and because we think inpatient spending is less likely to be subject to reporting bias. Most inpatient episodes were costly enough that even participants in the 95 percent coinsurance plan should have had strong incentives to file claims, because doing so would put them close to or over their Maximum Dollar Expenditure limit. Moreover, claims for inpatient episodes were generally filed by hospitals, which had large billing departments and systematic billing procedures and so were presumably less likely than individuals to fail to file claims. As shown in row 2, the adjustment reduces the estimated effects, but not by much.

The remaining rows highlight the impact of differential participation across plans on the estimates from row 2 that account for differential filing. We first consider the potential effect of observable differences across those who choose to participate in different plans. Row 3 quantifies the effect of the observable differences in participant characteristics across plans by reestimating the regression from row 2 but now controlling for the full set of pre-randomization covariates. These controls reduce further the estimated plan treatment effects but, again, not by much. Of course, this is only reassuring in so far as we believe we have a very rich set of observables that capture much of the potential differences across participants in the different plans.

A trickier issue is how to account for potential unobservable differences across individuals who select into participation in different experimental arms. There are, broadly speaking, three main approaches to this problem. Probably the most direct way to address potential bias stemming from differential nonparticipation across plans would be to collect data on outcomes (in this case, healthcare utilization) for all individuals, including those who failed to complete the experiment. Such data would allow comparison of outcomes for individuals based on initial plan assignment, regardless of participation, and then could be used for unbiased two-stage least squares estimates of the effects of cost sharing on utilization. Unfortunately, we know of no potential source of such data—individual-level hospital discharge records do not, to our knowledge, exist from this time period, and even if the records existed, there is no legal permission to match RAND participants (or nonparticipants) to administrative data.

A second approach is to make assumptions about the likely economic model of selection and use these to adjust the point estimates accordingly. (Angrist, Bettinger, and Kremer 2006, formalize one such approach in a very different experimental setting.) Then, depending on the economic model assumed, one might conclude

[8] Rogers and Newhouse (1985) have no estimates of underreporting for those individuals with zero claims. In the regressions with binary outcomes ("any spending") we somewhat arbitrarily scale up the shares of individuals by the same percentage as we scaled up spending among those who have positive spending amounts. When we analyze spending continuously, however, those who report no spending remain at zero.

that the existing point estimates are under- or overestimates of the true experimental treatment effects.

A final approach, and the one we take here, is to remain agnostic about the underlying economic mechanism generating the differential selection and instead perform a statistical exercise designed to find a lower bound for the treatment effect. In other words, this approach is designed to ask the statistical question of how bad the bias from differential participation could be. Specifically, in row 4, we follow Lee's (2009) bounding procedure by dropping the top group of spenders in the lower cost-sharing plan. The fraction of people dropped is chosen so that with these individuals dropped, participation rates are equalized between the lower cost-sharing plan and the higher cost-sharing plan to which it is being compared. As derived by Lee, these results provide worst-case lower bounds for the treatment effect under the assumption that any participant who refused participation in a given plan would also have refused participation in any plan with a higher coinsurance rate. For example, since 88 percent of those assigned to the free care plan completed the experiment compared to only 63 percent of those assigned to the 95 percent coinsurance (Table 1, column 6), for a comparison of these two plans, we drop the highest 28 percent ((88 – 63)/88) of spenders in the original free care sample, thus obtaining equal participation rates across the two samples.

Our primary conclusion from Table 3 is that after trying to adjust for differential selection and differential reporting by plan, the RAND data still reject the null hypothesis of no utilization response to cost sharing.[9] In particular, when the outcome is total spending, our ability to reject the null that utilization does not respond to consumer cost sharing survives all of our adjustments in two of the three specifications: any spending and log spending.[10]

The sensitivity analysis does, however, reveal considerable uncertainty about the magnitude of the response to cost sharing. The combination of adjusting for differential reporting and the Lee (2009) bounding exercise in row 4 opens up scope for the possibility that the treatment effects could be substantially lower than what is implied by the unadjusted point estimates. For example, focusing on column 3, our point estimate in row 1 indicates that spending under the 95 percent coinsurance

[9] Perhaps not surprisingly, there are statistical assumptions under which one cannot still reject this null. For example, we show in Table A5 of the online Appendix what we believe are (too) extreme worst-case bounds under which we can no longer reject the null. Specifically, following Manski (1990), for each year in which an individual should have been but was not present in the experiment (due to refusal or attrition), we impute the values that would minimize the treatment effect, and then further adjust the data for differential claim filing by plan, as before.

[10] In all cases, the statistically significant decline in the mean level of spending (column 2) is not robust to the bounding exercises in row 4. We think that this result is driven by the skewness of medical spending, which makes the results extremely sensitive to dropping the top 10–30 percent of spenders. In addition, we note that in some cases, the lower bounds appear to be statistically significant but with the "wrong" sign. Given strong a priori reasons to think that higher cost-sharing will not raise medical utilization, we interpret these results as simply showing that we cannot reject the null.

plan is 75 percent lower than under the free care plan, but the adjusted lower bound estimate in row 4 suggests that spending may only be 49 percent lower.[11]

Table 3 also shows that we can continue to reject the null of no response of outpatient spending for either the "any spending" specification or the log specification but are no longer able to reject the null of no response of *inpatient* utilization to higher cost sharing. The bounding exercise indicates that the response of inpatient spending is not robust to plausible adjustments for nonparticipation bias, and thus the RAND data do not necessarily reject (although they also do not confirm) the hypothesis of no price responsiveness of inpatient spending.

Finally, it is worth reemphasizing that the results in row 4 of Table 3 represent *lower bounds*, rather than alternative point estimates. We interpret the exercise as indicating that the unadjusted point estimates could substantially overstate the causal effect of cost sharing on healthcare utilization.

Estimating the Effect of Cost Sharing on Medical Spending

The most enduring legacy of the RAND experiment is not merely the rejection of the null hypothesis that price does not affect medical utilization, but rather the use of the RAND results to forecast the spending effects of other health insurance contracts. In extrapolating the RAND results out of sample, analysts have generally relied on the RAND estimate of a price elasticity of demand for medical spending of –0.2 (for which Manning, Newhouse, Duan, Keeler, Leibowitz, and Marquis 1987, is widely cited, but Keeler and Rolph 1988, is the underlying source).

This –0.2 elasticity estimate is usually treated as if it emerged directly from the randomized experiment, and is often ascribed the kind of reverence that might be more appropriately reserved for universal constants like π. Despite this treatment, the famous elasticity estimate is in fact derived from a combination of experimental data and additional modeling and statistical assumptions, as any out-of-sample extrapolation of experimental treatment effects must be. In using it out of sample, one necessarily confronts a number of statistical as well as economic issues.

Some Simple Attempts to Arrive at Estimates of the Price Elasticity

A major challenge for any researcher attempting to transform the findings from experimental treatment effects of health insurance contracts into an estimate of the price elasticity of demand for medical care is that health insurance contracts—both in the real world and in the RAND experiment—are highly nonlinear, with the price faced by the consumer typically falling as total medical spending accumulates during the year. The RAND contracts, for example, required some initial positive cost sharing, but out-of-pocket spending falls to zero after the Maximum Dollar Expenditure is reached. More generally, pricing under a typical health insurance

[11] We translate the coefficients in column 3 into percentages by exponentiating and subtracting from one.

contract might begin with a consumer facing an out-of-pocket price of 100 percent of his medical expenditure until a deductible is reached, at which point the marginal price falls sharply to the coinsurance rate that is typically around 10–20 percent, and then falls to zero once an out-of-pocket limit has been reached.

Due to the nonlinear form of the health insurance contracts, any researcher who attempts to summarize the experiment with a single price elasticity must make several decisions. One question is how to analyze medical expenditures that occur at different times, and therefore under potentially different cost-sharing rules, but which stem from the same underlying health event. Another issue is that the researcher has to make an assumption as to which price individuals respond to in making their medical spending decision. It is not obvious what single price to use. One might use 1) the current "spot" price of care paid at the time healthcare services are received (on the assumption that individuals are fully myopic), 2) the expected end-of-year price (based on the assumption that individuals are fully forward looking and with an explicit model of expectation formation), 3) the realized end-of-year price (on the assumption that changes in healthcare consumption happen at that margin), or perhaps 4) some weighted-average of the prices paid over a year. These types of modeling challenges—which were thoroughly studied and thought through by the original RAND investigators (Keeler, Newhouse, and Phelps 1977)—are inherent to the problem of extrapolating from estimates of the spending impact of particular health insurance plans and in this sense are not unique to the RAND experiment.

To get some idea of the challenges involved in translating the experimental treatment effects into an estimate of the price elasticity of demand, Table 4 reports a series of elasticity estimates that can be obtained from different, relatively simple and transparent ad-hoc manipulations of the basic experimental treatment effects. In panel A of Table 4 we convert—separately for each pair of plans—the experimental treatment effects from column 2 of Table 2 to arc elasticities with respect to the coinsurance rate. (These pairwise arc elasticities are calculated as the change in total spending as a percentage of the average spending, divided by the change in price as a percentage of the average price; in panel A we define the price as the coinsurance rate of the plan).[12] We obtain pairwise elasticities that are for the most part negative, ranging from about –0.1 to –0.5; the few positive estimates are associated with coinsurance rates that are similar and plans that are small.

We use panel B of Table 4 to report weighted averages of pairwise estimates under alternative assumptions regarding 1) the definition of the price, and 2) the definition of the elasticity. In terms of the definition of the price, in computing the elasticities in panel A we used the plan's coinsurance rate as the price and

[12] The arc elasticity of x with respect to y is defined as the ratio of the percent change in x to the percent change in y, where the percent change is computed relative to the average, namely $(x_2 - x_1)/((x_2 + x_1)/2)$. As x_2 and x_1 gets closer to each other, the arc elasticity converges to the standard elasticity. Although not commonly used elsewhere, it was heavily used by the RAND researchers because the largest plan in RAND was the free care plan. Starting with a price of zero, a percent change is not well defined, so arc elasticities are easier to work with.

Table 4
Sensitivity of Elasticity Estimates to Choice of Plan Comparisons and Price Measures

Panel A: Arc elasticities of total spending with regard to coinsurance rate, for different plan pairs[a]

	25% Coinsurance	*Mixed Coinsurance*[c]	*50% Coinsurance*	*Individual Deductible*[c]	*95% Coinsurance*
Free Care	–0.180	–0.091	–0.149	–0.119	–0.234
	(0.044)	(0.051)	(0.080)	(0.031)	(0.039)
25% Coinsurance		0.749	0.097	0.159	–0.097
		(0.533)	(0.281)	(0.128)	(0.101)
Mixed Coinsurance			–0.266	–0.101	–0.295
			(0.422)	(0.195)	(0.126)
50% Coinsurance				0.429	–0.286
				(1.176)	(0.280)
Individual Deductible					–0.487
					(0.187)

Panel B: Elasticities of total spending with regard to various price measures

	Coinsurance rate		*Average out-of-pocket price*	
	Arc elasticity[a]	*Elasticity*[b]	*Arc elasticity*[a]	*Elasticity*[b]
All plans	–0.094	NA	–0.144	NA
	(0.066)		(0.051)	
All plans except Free Care	–0.039	–0.523	–0.133	–0.524
	(0.131)	(0.082)	(0.097)	(0.085)
All plans except Free Care and Individual Deductible	–0.039	–0.537	–0.038	–0.600
	(0.108)	(0.084)	(0.108)	(0.094)

Notes: Panel A reports the pairwise arc elasticities calculated based on Table 2, column 2. Panel B reports the sample-size weighted average of various pairwise elasticities, calculated as detailed in the column-specific notes. Standard errors are in parentheses below the coefficient values. Standard errors are clustered on family. Arc elasticity standard errors are bootstrapped standard errors based on 500 replications, clustered on family.

[a] Pairwise arc elasticities are calculated as the change in total spending as a percentage of the average, divided by the change in price as a percentage of the average price, where the price is either the coinsurance rate of the plan (in panel A) or (in panel B) either (depending on the column) the coinsurance rate or the average out-of-pocket price paid by people assigned to that plan (the average out-of-pocket price of each plan is shown in Table 1).

[b] Elasticities are calculated based on pairwise regressions of log(total spending + 1) on log(price), where price is either the coinsurance rate of the plan or the average out-of-pocket price paid by people assigned to that plan.

[c] For the mixed coinsurance plan and the individual deductible plan, we take the initial price to be the average of the two coinsurance rates, weighted by the shares of initial claims that fall into each category. For the mixed coinsurance rate plans, this gives an initial price of 32 percent. For the individual deductible plan, it gives an initial price of 58 percent.

ignored the fact that once the Maximum Dollar Expenditure is reached the price drops to zero in all plans. In panel B, we consider both this elasticity with respect to the plan's coinsurance rate, but also report the elasticity with respect to the

average, plan-specific (but not individual-specific) out-of-pocket price. The plan's average out-of-pocket price (reported in Table 1, column 3) will be lower than the plan's coinsurance rate since it is a weighted average of the coinsurance rate and zero, which would be the "spot" price after the Maximum Dollar Expenditure is reached. For each price definition, we also consider two definitions of the elasticity; specifically, we calculate both arc elasticities as in panel A and more standard elasticities that are based on regression estimates of the logarithm of spending on the logarithm of price.[13] We also report results excluding the individual deductible plan, which has a different coinsurance rate for inpatient and outpatient care. Across these various simple manipulations of the experimental treatment effects in panel B, we find price elasticities that range between –0.04 and –0.6. (This exercise does not consider the additional adjustments for differential participation and reporting discussed in Table 3).

The RAND Elasticity: A Brief Review of Where It Came From

We now review the particular assumptions made by the original RAND investigators that allowed them to arrive at their famous estimate of a price elasticity of demand for medical care of –0.2; Keeler and Rolph (1988) provide considerably more detail.

To transform the experimental treatment effects into a single estimate of the single price elasticity of demand for health care, the RAND investigators grouped individual claims into "episodes." Each episode—once occurring—is thought of as an unbreakable and perfectly forecastable "bundle" of individual claims. The precise grouping relies on detailed clinical input and depends on the specific diagnosis. For example, each hospitalization constitutes a separate single episode. Routine spending on diabetes care over the entire year is considered a single episode and is fully anticipated at the start of the year, while "flare-ups" are not. Each cold or accident is a separate episode, but these could run concurrently. Once claims are grouped into episodes, the RAND investigators regress average costs per episode on plan fixed effects (and various controls) and find that plan assignment has virtually no effect on costs per episode. From this they conclude that spending on the intensive margin—that is, spending conditional on an episode occurring—does not respond to price, and focus their analysis on the price responsiveness of the extensive margin only—that is, on the occurrence rate of episodes.

To investigate the price to which individuals respond, the RAND investigators looked at whether the occurrence rate of episodes differs between individuals who face similar current prices for medical care but different future prices. Specifically, they look at whether spending is higher within a plan for individuals who are closer

[13] The latter calculations require that we exclude the free care plan, with a price of zero; as mentioned in an earlier footnote, this is the primary reason that the RAND investigators worked with arc elasticities. Because the arc elasticity estimates are based on treatment effects estimated in levels, and because we estimated smaller treatment effects (in percentage terms) for high-spending individuals (see Table A2), the arc elasticities are generally smaller than the more standard elasticities.

to hitting their Maximum Dollar Expenditures, and whether it is higher among people in cost-sharing plans who have exceeded their Maximum Dollar Expenditures compared to people in the free care plan. Of course, a concern with this comparison is that families with higher underlying propensities to spend are more likely to come close to hitting their Maximum Dollar Expenditures; the RAND investigators address this via various modeling assumptions. Finding no evidence in support of higher episode rates among individuals who are closer to hitting their Maximum Dollar Expenditure limits, the RAND investigators conclude that participants' extensive margin decisions about care utilization appear to be based entirely on the current "spot" price of care.

Given these findings, in the final step of the analysis the RAND investigators limit the sample to individuals in periods of the year when they are sufficiently far from hitting the Maximum Dollar Expenditure (by at least $400 in current dollars) so that they can assume that the coinsurance rate (or "spot" price) is the only relevant price. They then compute the elasticity of medical spending with respect to the experimentally assigned coinsurance rate. Specifically, for each category of medical spending—hospital, acute outpatient, and so on—they compute arc elasticities of spending in a particular category in the free care versus 25 percent coinsurance plan and in the free care versus 95 percent coinsurance plan. To compute these arc elasticities, they estimate spending changes for these individuals across contracts by combining their estimates of the responsiveness of the episode rate to the coinsurance rate with data on average costs per episode (which is assumed to be unresponsive to the coinsurance rate). The enduring elasticity estimate of –0.2 comes from noting that most of these arc elasticities—summarized in Keeler and Rolph (1988, Table 11)—are close to –0.2.

Using the RAND Elasticity: The Need to Summarize Plans with a Single Price

Application of the –0.2 estimate in a manner that is fully consistent with the way the estimate was generated is a nontrivial task. The RAND elasticity was estimated based on the assumption that in deciding whether to consume medical care, individuals fully anticipate spending within an "episode of care" but make their decision myopically—that is, only with regard to the current "spot" price of medical care—with respect to the potential for spending during the year on other episodes. Therefore a researcher who wanted to apply this estimate to forecasting the impact of an out-of-sample change in cost sharing would need to obtain micro data on medical claims, group these claims into "episodes" as described earlier, and calculate the "spot" price that each individual would face in each episode. Although there exist notable exceptions of studies that do precisely this (Buchanan, Keeler, Rolph, and Holmer 1991; Keeler, Malkin, Goldman, and Buchanan 1996), many subsequent researchers have applied the RAND estimates in a much simpler fashion. In doing so, arguably the key decision a researcher faces is how to summarize the nonlinear coverage with a single price. This is because the RAND elasticity is a single elasticity estimate, so it has to be applied to a single price.

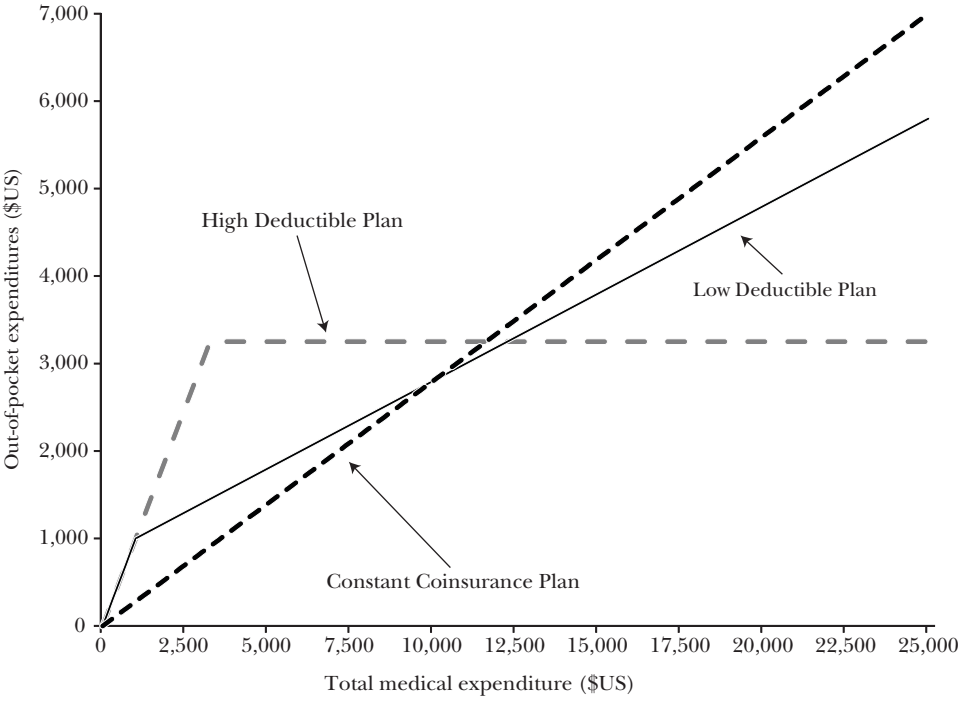
Researchers have taken a variety of different approaches to summarizing the price of medical care under a nonlinear insurance contract by a single number. For example, in predicting how medical spending will respond to high-deductible health savings accounts, Cogan, Hubbard, and Kessler (2005) applied the –0.2 elasticity estimate to the change in the average price that was paid out of pocket, where the average was taken over claims that were made at different parts of the nonlinear coverage. In extrapolating from the RAND experiment to the impact of the spread of insurance on the growth of medical spending, researchers have also used an "average price approach," summarizing the changes in the price of medical care by changes in the overall ratio between out-of-pocket medical spending and total spending (Newhouse 1992; Cutler 1995; Finkelstein 2007). Other work on the price elasticity of demand for medical care has summarized the price associated with a nonlinear coverage using the actual, realized price paid by each individual for his last claim in the coverage year (Eichner 1998; Kowalski 2009) or the expected end-of-year price (Eichner 1997).

These different methods for summarizing a nonlinear coverage with a single price can have an important effect on the estimated spending effects of alternative contracts. To illustrate this point, consider three "budget neutral" alternative coverage designs, depicted in Figure 2: a "high deductible" plan with a $3,250 per-family deductible and full insurance above the deductible; a "low deductible" plan with a $1,000 per-family deductible and a 20 percent coinsurance rate above the deductible; and a "no deductible" plan with a constant coinsurance rate of 28 percent. In describing these plans as "budget neutral," we mean that we picked them so that they would all have the same predicted cost (for the insurer) when we ignore potential behavioral responses to the different contracts and apply to each of them the same distribution of annual medical expenditures from RAND's free care plan (in 2011 dollars). The "no deductible" plan always has the same single price: that is, the buyer always pays 28 percent of the cost of health services. However, in the two nonlinear plans, the price paid by the individual will change from 100 percent of healthcare cost before the deductible is reached, to the coinsurance rate above that level.

As we described, in summarizing such a plan by a single number, one might look at a variety of "price" definitions, including the "spot" price paid at the time healthcare services are received, the realized end-of-year price, the expected end-of-year price, or at some weighted-average of the prices paid over a year. The concern is that when evaluating how changing from one insurance contract to another (or from no insurance to having insurance) would affect healthcare utilization, the method that is used to boil down the insurance contract into a single price—to which the –0.2 elasticity estimate is then applied—can yield very different conclusions about how the change in insurance contracts would increase the amount of health care consumed.

To illustrate the potential magnitudes at stake, consider an exercise in which we try to forecast the effect of reducing coverage from RAND's 25 percent coinsurance plan to a plan with a constant coinsurance rate of 28 percent, which is one of the options depicted in Figure 2. Because the new coverage has a constant coinsurance rate, the price of medical care under this coverage is clear and well defined: it

Figure 2
Nonlinear Health Insurance Coverage



Note: Consider three "budget neutral" alternative health insurance coverage designs: a "high deductible" plan with a $3,250 per-family deductible and full insurance above the deductible; a "low deductible" plan with a $1,000 per-family deductible and a 20 percent coinsurance rate above the deductible; and a "no deductible" plan with a constant coinsurance rate of 28 percent. See text for details.

is 28 cents for every dollar of healthcare spending. But in order to apply the RAND estimate of –0.2, we also need to summarize RAND's 25 percent coinsurance with a single price. Recall that the RAND plan had a Maximum Dollar Expenditure limit, so the price starts at 25 cents for every dollar, but then becomes zero once the limit is reached, so summarizing the RAND plan with a single price essentially means a choice of weights in the construction of an average price. We use three different ways to summarize the RAND 25 percent coinsurance plan with a single price: a dollar-weighted average price, a person-weighted average price, and a person-weighted average end-of-year price. Applying the distribution of spending under the free care plan, these result in three different summary prices, of 10, 17, and 13 cents for every dollar of medical spending, respectively. Applying the –0.2 estimate to changing from each of these prices to 28 cents, which is the constant price in the alternative coverage, we obtain a reduction in healthcare spending of 18, 9, and 14 percent, respectively. Thus, in this example, the decision of how to define the price leads to differences in the predicted reduction of spending that vary by a factor of 2.

The Dangers of Summarizing Nonlinear Coverage by a Single Price

The preceding exercise illustrated how the manner by which a nonlinear coverage is summarized by a single price could be important. In general, there is no "right" way to summarize a nonlinear budget set with a single price. The differing implications of alternative reasonable, yet ad hoc "fixes" to this problem should give us pause when considering many of the subsequent applications of the RAND experimental results. It also suggests that, going forward, attempts to estimate the impact of health insurance contracts on healthcare spending would benefit from more attention to how the nonlinearities in the health insurance contracts may affect the spending response.

Fortunately, just as there has been intellectual progress in the design and analysis of experimental treatment effects in the decades since RAND, there has similarly been progress on the analysis of the behavioral response to nonlinear budget sets (for example, Hausman 1985). Much of the initial work in this area focused on analyzing the labor supply response to progressive taxation. Recently, however, researchers have begun to apply the techniques of nonlinear budget set estimation to the analysis of the effect of (nonlinear) health insurance contracts (Marsh 2012; Kowalski 2012), and further work in this area could be of great value.

Of course, even equipped with these techniques, current researchers must grapple with many of the same issues that the original RAND investigators faced. In particular, they must model the distribution of medical shocks throughout the year in the population under analysis, as well as the evolution of individuals' beliefs about these shocks. Another key issue is whether individuals take into account the entire nonlinear budget set induced by the health insurance contract in making their spending decision, or whether they respond only to the current "spot" price, or to something in between. Although fully forward-looking rational individuals should only respond to the expected end-of-year price, if individuals are myopic, liquidity constrained, or unsure of the details of their contract, they might also respond, at least to some extent, to the "spot" price. In recent empirical work, we investigate this question using data on medical spending by people covered by employer-provided health insurance (Aron-Dine, Einav, Finkelstein, and Cullen 2012). We concluded that, in our specific setting, individuals do appear to take into account the nonlinear budget set in making medical spending decisions but that they are not fully forward looking as they also take account of the spot price. In our calibration results, the predicted spending change associated with introducing a nonlinear health insurance contract can vary greatly depending on what one assumes about the degree of forward-looking behavior, suggesting that more evidence on this question would be useful.

More generally, any transformation of the experimental treatment effects into estimates that can be used out-of-sample will require more assumptions than required to obtain those treatment effects in the first place. More than three decades after the RAND experiment, the development and use of new approaches to doing such out-of-sample extrapolation remains an active and interesting area for research.

Concluding Remarks

At the time of the RAND Health Insurance Experiment, it was vigorously argued that medical care was determined by "needs," and therefore was not sensitive to price. As Cutler and Zeckhauser (2000) wrote, the RAND experiment was instrumental in rejecting this view: "Sound methodology, supported by generous funding, carried the day. The demand elasticities in the Rand Experiment have become the standard in the literature, and essentially all economists accept that traditional health insurance leads to moderate moral hazard in demand."

But as this core lesson of the RAND experiment has become solidified in the minds of a generation of health economists and policymakers, there has been a concomitant fading from memory of the original experimental design and analytical framework. While this progression may be natural in the lifecycle of transformative research, it seems useful to remind a younger generation of economists of the details and limitations of the original work.

In this essay, we re-presented and reexamined the findings of the RAND experiment from the perspective of three subsequent decades of progress in empirical work on the design and analysis of randomized experiments, as well as on the analysis of moral hazard effects of health insurance—much of it inspired, no doubt, to a large degree by the enduring influence of the RAND results. This landmark and pioneering study was uniquely ambitious, remarkably sophisticated for its time, and entrepreneurial in the design and implementation of the then-new science of randomized experiments in the social sciences.

Our reexamination concludes that despite the potential for substantial bias in the original estimates stemming from systematically differential participation and reporting across experimental arms, one of the central contributions of the RAND experiment is robust: the rejection of the null hypothesis that health spending does not respond to the out-of-pocket price. Naturally, however, these potential biases introduce uncertainty about the magnitude of the impact of the different insurance plans on medical spending. Moreover, the translation of these experimental estimates into economic objects of interest—such as a price elasticity of demand for medical care—requires further assumptions and machinery, which go beyond the "raw" experimental results. While economic analysis has made progress in the intervening decades in developing techniques that may offer new approaches to the economic analysis of moral hazard effects of health insurance, it will always be the case that, like the famous –0.2 price elasticity of demand estimate produced by the original RAND investigators, any attempt by researchers to apply the experimental estimates out of sample will involve more assumptions—and hence scope for uncertainty—than the direct experimental estimates themselves. This point, while straightforward and uncontroversial (we'd think), may have become somewhat lost in the intervening decades of use of the RAND estimates. Our hope is that this essay may help put both the famous experiment and its results back in context.

■ *We are grateful to the JEP editors (David Autor, John List, and the indefatigable Timothy Taylor), as well as to Ran Abramitzky, Tim Bresnahan, Dan Fetter, Emmett Keeler, Will Manning, Joe Newhouse, Matt Notowidigdo, Sarah Taubman, and Heidi Williams for helpful comments, and to the National Institute of Aging (Grant No. R01 AG032449) for financial support; Aron-Dine also acknowledges support from the National Science Foundation Graduate Research Fellowship program. We thank the original RAND investigators for making their data so easily available and accessible. Aron-Dine contributed to this paper while she was a graduate student in Economics at MIT. She is currently employed as a Special Assistant to the President for Economic Policy at the National Economic Council. The views in this paper do not represent those of the National Economic Council or the White House.*

References

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103(484): 1481–95.

Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Columbia." *American Economic Review* 96(3): 847–62.

Aron-Dine, Aviva, Liran Einav, Amy Finkelstein, and Mark R. Cullen. 2012. "Moral Hazard in Health Insurance: How Important is Forward Looking Behavior?" NBER Working Paper 17802.

Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* 53(5): 941–73.

Ashenfelter, Orley, and Mark W. Plant. 1990. "Non-Parametric Estimates of the Labor-Supply Effects of Negative Income Tax Programs." *Journal of Labor Economics* 8(1): S397–S415.

Bitler, Marianne, Jonah Gelbach, and Hilary Hoynes. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96(4): 988–1012.

Buchanan, Joan L., Emmett B. Keeler, John E. Rolph, and Martin R. Holmer. 1991. "Simulating Health Expenditures under Alternative Insurance Plans." *Management Science* 37(9): 1067–90.

Cogan, John F., R. Glenn Hubbard, and Daniel P. Kessler. 2005. *Healthy, Wealthy, and Wise: Five Steps to a Better Healthcare System*, 1st ed. Washington, DC: AEI Press.

Cutler, David M. 1995. "Technology, Health Costs, and the NIH." Paper prepared for the National Institutes of Health Economics Roundtable on Biomedical Research. http://www.economics.harvard.edu/files/faculty/13_Technology,%20Health%20Costs%20and%20the%20NIH.pdf.

Cutler, David M., and Richard J. Zeckhauser. 2000. "The Anatomy of Health Insurance." In *Handbook of Health Economics*, edited by A. J. Culyer and J. P. Newhouse, volume 1, 563–643. Amsterdam: Elsevier.

Eichner, Matthew J. 1997. "Medical Expenditures and Major Risk Health Insurance." MIT, PhD Dissertation, Chapter 1.

Eichner, Matthew J. 1998. "The Demand for Medical Care: What People Pay Does Matter." *American Economic Review* 88(2): 117–21.

Finkelstein, Amy. 2007. "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare." *Quarterly Journal of Economics* 122(3): 1–37.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics* 127(3): 1057–1106.

Greenberg, David, and Harlan Halsey. 1983. "Systematic Misreporting and Effects of Income Maintenance Experiments on Work Effort: Evidence from the Seattle-Denver Experiment." *Journal of Labor Economics* 1(4): 380–407.

Greenberg, David, and Mark Shroder. 2004. *The Digest of Social Experiments*, 3rd ed. Washington, DC: Urban Institute Press.

Hausman, Jerry A. 1985. "The Econometrics of Nonlinear Budget Sets." *Econometrica* 53(6): 1255–82.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Q. Yavitz. 2010. "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the High Scope Perry Preschool Program." *Quantitative Economics* 1(1): 1–46.

Heckman, James J., Rodrigo Pinto, Azeem M. Shaikh, and Adam Q. Yavitz. 2011. "Inference with Imperfect Randomization: The Case of the Perry Preschool Program." http://home.uchicago.edu/~amshaikh/webfiles/perry.pdf.

Keeler, Emmett B., Joseph P. Newhouse, and Charles E. Phelps. 1977. "Deductibles and the Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule under Uncertainty." *Econometrica* 45(3): 641–56.

Keeler, Emmett B., and John E. Rolph. 1988. "The Demand for Episodes of Treatment in the Health Insurance Experiment." *Journal of Health Economics* 7(4): 337–67.

Keeler, Emmett B., Jesse D. Malkin, Dana P. Goldman, and Joan L. Buchanan. 1996. "Can Medical Savings Accounts for the Nonelderly Reduce Health Care Costs?" *JAMA* 275(21): 1666–71.

Kowalski, Amanda E. 2009. "Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care." NBER Working Paper 15085.

Kowalski, Amanda E. 2012. "Estimating the Tradeoff between Risk Protection and Moral Hazard with a Nonlinear Budget Set Model of Health Insurance." NBER Working Paper 18108.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2): 497–532.

Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *Economic Journal* 111(468): 1–28.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76(3): 1071–1102.

Levitt, Steven D., and John A. List. 2011. "Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments." *American Economic Journal: Applied Economics* 3(1): 224–39.

Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, Arleen Leibowitz, and M. Susan Marquis. 1987. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." *American Economic Review* 77(3): 251–77.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80(2): 319–23.

Marsh, Christina. 2012. "Estimating Demand Elasticities using Nonlinear Pricing." http://clmarsh.myweb.uga.edu/docs/Demand_Elasticities_Marsh.pdf.

Michalopoulos, Charles, David Wittenburg, Dina A. R. Israel, Jennifer Schore, Anne Warren, Aparajita Zutshi, Stephen Freedman, and Lisa Schwartz. 2011. *The Accelerated Benefits Demonstration and Evaluation Project: Impacts on Health and Employment at Twelve Months.* New York: MDRC. http://www.mdrc.org/sites/default/files/full_528.pdf.

Morris, Carl. 1979. "A Finite Selection Model for Experimental Design of the Health Insurance Study." *Journal of Econometrics* 11(1): 43–61.

Newhouse, Joseph P. 1992. "Medical Care Costs: How Much Welfare Loss?" *Journal of Economic Perspectives* 6(3): 3–21.

Newhouse, Joseph P., Robert H. Brook, Naihua Duan, Emmett B. Keeler, Arleen Leibowitz, Willard G. Manning, M. Susan Marquis, Carl N. Morris, Charles E. Phelps, and John E. Rolph. 2008. "Attrition in the RAND Health Insurance Experiment: A Response to Nyman." *Journal of Health Politics, Policy and Law* 33(2): 295–308.

Newhouse, Joseph P., and the Insurance Experiment Group. 1993. *Free for All.* Cambridge: Harvard University Press.

Nyman, John A. 2007. "American Health Policy: Cracks in the Foundation." *Journal of Health Politics, Policy and Law* 32(5): 759–83.

Nyman, John A. 2008. "Health Plan Switching and Attrition Bias in the RAND Health Insurance Experiment." *Journal of Health Politics, Policy and Law* 33(2): 309–17.

Pauly, Mark. 1968. "The Economics of Moral Hazard: Comment." *American Economic Review* 58(3): 531–37.

Rogers, William H., and Joseph P. Newhouse. 1985. "Measuring Unfiled Claims in the Health Insurance Experiment." In *Collecting Evaluation Data: Problems and Solutions*, edited by Leigh Burstein, Howard E. Freeman, and Peter H. Rossi, 121–133. Beverly Hills: Sage.

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 223–230

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., Saint Paul, Minnesota, 55105.

Potpourri

The theme for the *2013 World Development Report* from the World Bank is one word: "Jobs." "To many, a 'job' brings to mind a worker with an employer and a regular paycheck. Yet, the majority of workers in the poorest countries are outside the scope of an employer–employee relationship. Worldwide, more than 3 billion people are working, but their jobs vary greatly. Some 1.65 billion are employed and receive regular wages or salaries. Another 1.5 billion work in farming and small household enterprises, or in casual or seasonal day labor. Meanwhile, 200 million people, a disproportionate share of them youth, are unemployed and actively looking for work. Almost 2 billion working-age adults, the majority of them women, are neither working nor looking for work, but an unknown number of them are eager to have a job. . . . The problem for most poor people in these

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

http://dx.doi.org/10.1257/jep.27.1.223. doi=10.1257/jep.27.1.223

countries is not the lack of a job or too few hours of work; many hold more than one job and work long hours. Yet, too often, they are not earning enough to secure a better future for themselves and their children, and at times they are working in unsafe conditions and without the protection of their basic rights. Jobs are instrumental to achieving economic and social development. Beyond their critical importance for individual well-being, they lie at the heart of many broader societal objectives, such as poverty reduction, economy-wide productivity growth, and social cohesion. The development payoffs from jobs include acquiring skills, empowering women, and stabilizing post-conflict societies." October 2012. At http://econ.worldbank.org/wdr.

The Congressional Budget Office has calculated "Effective Marginal Tax Rates for Low- and Moderate-Income Workers." "The effective marginal tax rate is the percentage of an additional dollar of earnings that is unavailable to a worker because it is paid in taxes or offset by reductions in benefits from government programs. . . . When lawmakers target assistance to people of limited means, that assistance declines as income rises. . . . The Congressional Budget Office (CBO) finds that working taxpayers with income below 450 percent of federal poverty guidelines (commonly known as the federal poverty level, so abbreviated as FPL) face a marginal tax rate of 30 percent, on average, under the provisions of law in effect in 2012. . . . Over the next two years, CBO estimates, various provisions of current law will cause marginal tax rates among this population to rise, on average, to 32 percent in 2013 and to 35 percent in 2014. CBO also finds that under provisions of law in effect between 2012 and 2014, marginal tax rates vary greatly across earnings ranges and among individuals within the same earnings range." November 2012. At http://www.cbo.gov/sites/default/files/cbofiles/attachments/11-15-2012-MarginalTaxRates.pdf.

The World Economic Forum has published *The Global Enabling Trade Report 2012*, a group of essays with the overall theme of "Reducing Supply Chain Barriers." In chapter 1.1, Robert Z. Lawrence, Sean Doherty, and Margareta Drzeniek Hanouz (who are also the editors of the report) sum up the growing importance of global supply chains in this way: "Traded commodities are increasingly composed of intermediate products. Reductions in transportation and communication costs and innovations in policies and management have allowed firms to operate global supply chains that benefit from differences in comparative advantage among nations, both through international intra-firm trade and through networks that link teams of producers located in different countries. Trade and foreign investment have become increasingly complementary activities. . . . Increasingly, countries specialize in tasks rather than products. Value is now added in many countries before particular goods and services reach their final destination, and the traditional notion of trade as production in one country and consumption in another is increasingly inaccurate." May 2012. At http://www3.weforum.org/docs/GETR/2012/GlobalEnablingTrade_Report.pdf.

Most economists occasionally face the existential question: Is more GDP good? In response, Nicholas Oulton has written "Hooray for GDP!" Oulton addresses, and critiques, four arguments against focusing on GDP: "1) GDP is hopelessly flawed

as a measure of welfare. It ignores leisure and women's work in the home. It takes no account of pollution and carbon emissions. 2) GDP ignores distribution. In the richest country in the world, the United States, the typical person or family has seen little or no benefit from economic growth since the 1970s. But over the same period inequality has risen sharply. 3) Happiness should be the grand aim of policy. But the evidence is that, above a certain level, a higher material standard of living does not make people any happier. . . . 4) Even if higher GDP were a good idea on other grounds, it's not feasible because the environmental damage would be too great." Oulton does not attempt an exhaustive review, but provides a selective sampling of the arguments and evidence. Occasional paper 30, Centre for Economic Performance, London School of Economics and Political Science, August 2012. At http://cep.lse.ac.uk/pubs/download/occasional/op030.pdf.

The October 2012 issue of *Nature Biotechnology* offers several articles that "Focus on Commercializing Biomedical Innovations." From the opening "Editorial": "Investment in biomedical innovation is not what it once was. Millions of dollars have fled the life sciences risk capital pool. . . . Never has there been a more pressing need to look beyond the existing pools of funding and talent to galvanize biomedical innovation." In one of the papers, Jose-Maria Fernandez, Roger M. Stein, and Andrew W. Lo offer a proposal for "Commercializing Biomedical Research through Securitization Techniques." "Industry professionals cite the existence of a 'valley of death'—a funding gap between basic biomedical research and clinical development. For example, in 2010, only $6–7 billion was spent on translational efforts, whereas $48 billion was spent on basic research and $127 billion was spent on clinical development that same year. . . . We propose an alternative for funding biomedical innovation that addresses these issues through the use of 'financial engineering'. . . Our approach involves two components: (i) creating large diversified portfolios—'megafunds' on the order of $5–30 billion—of biomedical projects at all stages of development; and (ii) structuring the financing for these portfolios as combinations of equity and securitized debt so as to access much larger sources of investment capital. These two components are inextricably intertwined: diversification within a single entity reduces risk to such an extent that the entity can raise assets by issuing both debt and equity, and the much larger capacity of debt markets makes this diversification possible for multi-billion-dollar portfolios of many expensive and highly risky projects." The issue is available at http://www.nature.com/nbt/journal/v30/n10/full/nbt.2400.html. The Fernandez, Stein, and Lo paper is at http://www.nature.com/nbt/journal/v30/n10/full/nbt.2374.html.

Sri Wening Handayani and Babken Babajanian have edited a collection of essays for the Asian Development Bank on the topic *Social Protection for Older Persons: Social Pensions in Asia.* As one example, Sharifa Begum and Dharmapriya Wesumperuma provide an "Overview of the Old Age Allowance Programme in Bangladesh." "[E]ven a low pension level can have a meaningful impact on the lives of older people and their families. The pension level in Bangladesh is very low (at $4.50 per month) and there is a strong argument for increasing it. Nevertheless, the impacts of the pension so far have been far from negligible. . . . [T]he

social pension appears to benefit women more than men. This particularly relates to impacts on health and psychosocial well-being. . . . In theory, the relatively high coverage of the social pension means that it should be able to cover the most-poor elderly. However, a large portion of the beneficiaries (20%–40%) do not actually meet the eligibility criteria. Meanwhile, some of the most vulnerable older people miss out. . . . [W]orking with the private sector for pension payments appears to have some benefits, though . . . a number of issues have arisen through this delivery mechanism. These include . . . issues of long queues on payday and some cases of malpractice of banking staff. Other countries in Asia considering the use of banks in delivering social pensions and other cash transfers would do well to assess how these issues can be overcome." July 2012. http://www.adb.org/publications/social-protection-older-persons-social-pensions-asia.

As world population climbs toward a projected nine billion or so by mid-century, can agricultural productivity keep up? Keith Fuglie and Sun Ling Wang offer some thoughts in "New Evidence Points to Robust But Uneven Productivity Growth in Global Agriculture." "Improving agricultural productivity has been the world's primary safeguard against a recurring Malthusian crisis—where the needs of a growing population outstrip the ability of man and resources to supply food. Over the past 50 years, global gross agricultural output has more than tripled in volume, and productivity growth in agriculture has enabled food to become more abundant and cheaper. In inflation-adjusted dollars, agricultural prices fell by an average of 1 percent per year between 1900 and 2010, despite an increase in the world's population from 1.7 billion to nearly 7.0 billion over the same period. Nonetheless, food prices have been rising since around 2001. This has renewed concerns about the pace of agricultural productivity growth. . . . Perhaps the single, most important factor separating countries that have successfully sustained long-term productivity growth in agriculture from those that have not is their capacity for agricultural R&D. . . . Recent research has identified a number of other factors that account for cross-country differences in agricultural TFP [total factor productivity]. Improvements in what can broadly be characterized as the 'enabling environment' have encouraged the adoption of new technologies and practices by some countries; these include policies that improve economic incentives for producers, strengthen rural education and agricultural extension services, and improve rural infrastructure and access to markets." *Amber Waves*, September 2012, published by the US Department of Agriculture, Economic Research Service, http://www.ers.usda.gov/amber-waves/2012-september/global-agriculture.aspx.

As low-income countries become better off, they typically raise their poverty lines, as Martin Ravallion points out in "A Relative Question." "For example, China recently doubled its national poverty line from 90 cents a day to $1.80 (adjusted to reflect constant 2005 purchasing power). Other countries—including Colombia, India, Mexico, Peru, and Vietnam—have also recently revised their poverty lines upward. . . . It would not be fair to the more than 1 billion people who still live on less than $1.25 a day to abandon the emphasis on fighting absolute poverty. Eliminating such extreme poverty must remain the global development community's number one priority. But the world is changing rapidly. The convergence in living

standards across the globe is accompanied by emerging convergence in our ideas about what poverty means . . . New poverty targets will undoubtedly emerge that reflect these new perceptions." *Finance & Development*, December 2012, pp. 40–42. http://www.imf.org/external/pubs/ft/fandd/2012/12/ravallion.htm.

International Reserves

Edwin Truman offers "Reflections on Reserve Management and International Monetary Cooperation." "At the end of 2011, international reserve assets alone amounted to 17 percent of world GDP and an average of 29 percent of the national GDP of emerging market and developing countries. . . . Including the international assets of SWFs [sovereign wealth funds] and similar entities would boost those percentages substantially above 20 percent and close to 40 percent respectively. . . . [E]nhancement of cooperative arrangements in this area is falling behind the need for them in the face of the explosion of the size and number of significant public investors, bringing in many non-traditional investors. This is a global issue. The notion that a country's public investments are the exclusive concern of the country itself is analytically wrong and fundamentally dangerous. Two countries (at least) share an exchange rate. Similarly, two countries (at least) share the effects of cross-border public investments. . . . The alternative to increased cooperation on public sector investment policies is a currency war. . . . The greater risk is that restrictions and barriers will increase affecting not only cross-border official investments, but all cross-border financial transactions. Once we start down that path, a trade war would not be difficult to envisage, and the consequences for global growth and stability could be severe." Remarks delivered at the World Bank/Bank for International Settlements Joint Fourth Public Investors' Conference, December 3, 2012. http://www.iie.com/publications/papers/truman20121203.pdf.

The Independent Evaluation Office of the International Monetary Fund offers a contrasting view in "International Reserves: IMF Concerns and Country Perspectives." "There was a common view among country authorities that the IMF tended to underestimate the benefits of reserves. Thinking about the tradeoff between costs and benefits of reserves, country officials often mentioned a range of benefits that they considered important but were not easily incorporated into either single indicators or formal models. In addition to precautionary self-insurance (also emphasized by the Fund), they mentioned other important advantages: reserves provide a country with reliability of access and the policy autonomy to act quickly, flexibly, and counter cyclically, and, as was evident during the global crisis, they inspire confidence. Reserves have also allowed authorities to avoid the stigma associated with approaching the Fund for resources—an issue that is very much alive in a number of countries. . . . Moreover, factors other than reserve accumulation—notably the leverage-induced fluctuations in global liquidity, inadequate financial sector regulation, and capital flow volatility—are more pertinent sources of concern for systemic resiliency." August 13, 2012. At http://www.ieo-imf.org/ieo/files/completedevaluations/IR_Main_Report.pdf.

Issues in Manufacturing

The McKinsey Global Institute has published, *Manufacturing the Future: The Next Era of Global Growth and Innovation*. "The role of manufacturing in the economy changes over time. Empirical evidence shows that as economies become wealthier and reach middle-income status, manufacturing's share of GDP peaks (at about 20 to 35 percent of GDP). Beyond that point, consumption shifts toward services, hiring in services outpaces job creation in manufacturing, and manufacturing's share of GDP begins to fall along an inverted U curve. Employment follows a similar pattern: manufacturing's share of US employment declined from 25 percent in 1950 to 9 percent in 2008. In Germany, manufacturing jobs fell from 35 percent of employment in 1970 to 18 percent in 2008, and South Korean manufacturing went from 28 percent of employment in 1989 to 17 percent in 2008. As economies mature, manufacturing becomes more important for other attributes, such as its ability to drive productivity growth, innovation, and trade. Manufacturing also plays a critical role in tackling societal challenges, such as reducing energy and resource consumption and limiting greenhouse gas emissions. . . . Manufacturing continues to make outsize contributions to research and development, accounting for up to 90 percent of private R&D spending in major manufacturing nations. The sector contributes twice as much to productivity growth as its employment share, and it typically accounts for the largest share of an economy's foreign trade; across major advanced and developing economies, manufacturing generates 70 percent of exports." November 2012. At http://www.mckinsey.com/insights/mgi/research/productivity_competitiveness_and_growth/the_future_of_manufacturing.

Joe Mahon presents an "Interview: Susan Houseman on Measuring Manufacturing Productivity." "I still see a lot of analysts who say, 'Look at how fast manufacturing is growing; manufacturing output is growing faster than GDP. There's nothing wrong; manufacturing is doing great.' But we find that without the computer industry, growth in manufacturing real value added falls by two-thirds and productivity growth falls by almost half. It doesn't look like a strong sector without computers. That's the first point. The second point . . . is that there's been a lot of growth in manufacturers' use of foreign intermediate inputs since the 1990s, and most of those inputs come from developing and low-wage countries where costs are lower. We point out that those lower costs aren't being captured by statistical agencies, and so, as a result, the growth of those imported inputs is being undercounted." *fedgazette*, Federal Reserve Bank of Minneapolis, October 2012. At http://www.minneapolisfed.org/publications_papers/pub_display.cfm?id=4982. Houseman presented a more detailed version of these arguments in a paper with Christopher Kurz, Paul Lengermann, and Benjamin Mandel, "Offshoring Bias in U.S. Manufacturing," which appeared in the Spring 2011 issue of this journal.

Antonio Regalado interviews Ricardo Hausman in "You Must Make the New Machines." "The step that makes the most sense for the U.S. is to become the producer of the machinery that will power the next global manufacturing revolution. That is where the most complex and sophisticated products are, and that is the work that can pay higher wages. . . . My guess is that developments around

information technology, 3-D printing, and networks will allow for a redesign of manufacturing. The world will be massively investing in it. The U.S. is well positioned to be the source of those machines. It can only be rivaled by Germany and Japan." *MIT Technology Review,* January 4, 2013. At http://www.technologyreview.com/news/509281/you-must-make-the-new-machines/.

Some Nobel Laureates

Each year when the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel is awarded, the Prize Committee posts useful background information. This year's "Information for the Public" is titled "Stable Matching: Theory, Evidence, and Practical Design." It begins: "This year's Prize to Lloyd Shapley and Alvin Roth extends from abstract theory developed in the 1960s, over empirical work in the 1980s, to ongoing efforts to find practical solutions to real-world problems. Examples include the assignment of new doctors to hospitals, students to schools, and human organs for transplant to recipients. Lloyd Shapley made the early theoretical contributions, which were unexpectedly adopted two decades later when Alvin Roth investigated the market for U.S. doctors. His findings generated further analytical developments, as well as practical design of market institutions." Available at http://www.nobelprize.org/nobel_prizes/economics/laureates/2012/popular-economicsciences2012.pdf. A more detailed "Scientific Background" paper titled "Stable Allocations and the Practice of Market Design" is available at http://www.nobelprize.org/nobel_prizes/economics/laureates/2012/advanced-economicsciences2012.pdf, October 15, 2012.

Elinor Ostrom gave the Hayek Memorial Lecture at the Institute for Economic Affairs on the topic: "The Future of the Commons: Beyond Market Failure and Government Regulation." Here's Ostrom: "Challenge one, as I mentioned, is the panacea problem. A very large number of policymakers and policy articles talk about 'the best' way of doing something. For many purposes, if the market was not the best way people used to think that it meant that the government was the best way. We need to get away from thinking about very broad terms that do not give us the specific detail that is needed to really know what we are talking about. We need to recognise that the governance systems that *actually have worked in practice* fit the diversity of ecological conditions that exist in a fishery, irrigation system or pasture, as well as the social systems. There is a huge diversity out there, and the range of governance systems that work reflects that diversity. We have found that government, private and community-based mechanisms all work in some settings. People want to make me argue that community systems of governance are always the best: I will not walk into that trap. There are certainly very important situations where people can self-organise to manage environmental resources, but we cannot simply say that the community is, or is not, the best; that the government is, or is not, the best; or that the market is, or is not, the best. It all depends on the nature of the problem that we are trying to solve." The Institute of Economic Affairs has published her lecture as part of a short e-book, together with several useful supporting essays. Chapter 3

2013 Application/Renewal for Membership

AMERICAN ECONOMIC ASSOCIATION
2014 Broadway, Suite 305
Nashville, TN 37203
Ph. 615-322-2595 fax: 615-343-7590
Federal ID No. 36-2166945
www.vanderbilt.edu/AEA

RENEWING MEMBERS, ENTER ACCT. NUMBER & EXP. DATE	**IF PAYING BY CREDIT CARD, PLEASE FILL OUT BELOW**
ACCOUNT NUMBER:	CARD NUMBER:
EXPIRATION DATE:	EXP DATE: CSC CODE:

FIRST NAME:	MI:	LAST NAME:
ADDRESS:		
CITY:		STATE/PROVINCE: ZIP:
COUNTRY:		☐ Check here if non-US
PHONE:		FAX:
PRIMARY FIELD OF SPECIALIZATION:		
SECONDARY FIELD OF SPECIALIZATION:		
EMAIL:		☐ Check here to exclude your email address from the public directory

Please include my email address to receive:
☐ Announcements about public policy affecting economists or the economics profession
☐ Surveys of economists for research purposes
☐ Commercial advertising

MEMBERSHIP DUES — Based on annual income. Please select one below.

☐ Annual income of $70,000 or less	$20	$
☐ Annual income of $70,000 to $105,000	$30	$
☐ Annual income over $105,000	$40	$

The AEA dues above include online access to all seven AEA journals.
For print or CD subscription(s) indicate preference below and add appropriate charge(s).

Journal		**Print**	**Int'l Postage***	**CD***	
AER	(7 issues, incl. P&P)	☐ Add $20	☐ Add $25	☐ Add $15	$
AER Papers & Proceedings Only*		☐ Add $10	n/a	n/a	$
JEL	(4 quarterly issues)	☐ Add $15	☐ Add $15	☐ Add $15	$
JEP	(4 quarterly issues)	☐ Add $15	☐ Add $15	☐ Add $15	$
AEJ: Applied	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$
AEJ: Policy	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$
AEJ: Macro	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$
AEJ: Micro	(4 quarterly issues)	☐ Add $15	☐ Add $15	n/a	$

* Int'l postage applies only to print journals mailed outside of the U.S. No additional postage is required for CDs or the AER Papers and Proceedings.

AEA Journals via JSTOR online

JSTOR	☐ Add $16	$

	Sub Total	$
Check One: ☐ 1 Year ☐ 2 Years ☐ 3 Years	TOTAL AMOUNT	$

Make checks payable to: American Economic Association.
Must be drawn on a US bank.
Apply online at http://www.aeaweb.org/membership.php

Payments must be made in advance. We accept checks (in US dollars only, with correct coding for processing in US banks) and credit cards; online or by faxing or mailing the application. Please choose one method;
it is the Association's policy NOT TO REFUND dues.

The *Journal of Economic Perspectives*: Proposal Guidelines

Considerations for Those Proposing Topics and Papers for *JEP*

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.

Philosophy and Style

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of *JEP* articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author.** The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

JEP does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-U.S. country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives.* Our stock in trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship. If you are not familiar with this journal, it is freely available on-line at <http://e-*JEP*.org>.

Guidelines for Preparing *JEP* Proposals

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP.* However, given

the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and authors an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future articles for *JEP* may be sent (electronically please) to the assistant editor, Ann Norman, at <anorman@JEPjournal.org>. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

Guidelines for Empirical Papers Submitted to *JEP*

The *JEP* is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1) The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.

2) In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.

3) The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.

AEA
AMERICAN ECONOMIC ASSOCIATION

Membership Information

Our publications include:

American Economic Review (*AER*)	*American Economic Journal*: *Applied Economics* (*AEJ*: *AE*)
and the *AER Papers and Proceedings* (May issue)	*American Economic Journal*: *Economic Policy* (*AEJ*: *EP*)
Journal of Economic Literature (*JEL*)	*American Economic Journal*: *Macroeconomics* (*AEJ*: *MAC*)
Journal of Economic Perspectives (*JEP*)	*American Economic Journal*: *Microeconomics* (*AEJ*: *MIC*)

In addition to having access to all seven academic journals of the Association, your benefits include:

- Receiving the *AER*, *JEL*, and *JEP* online and in print or on CD
- Receiving online and in print the *American Economic Journals* (*AE*, *EP*, *MAC*, and *MIC*)
- Access to pre-publication accepted articles for AEA journals
- Discounts on submission fees for the *AER* and the *AEJ*s
- Quarterly AEA Virtual Field Journals: Notification of articles in all of the AEA journals in the subject classifications of your choice
- Access to AEA journals in JSTOR for an additional $16 annually
- *EconLit for Members*: The *EconLit* online bibliography (without all the search features and full-text links provided in institutional settings)
- *EconLit for Members* update alerts by *JEL* code(s) of your choice

as well as

Continuing Education Program discounts,
Listing in the AEA Directory of Members,
and
Group Term Life Insurance & Short Term Recovery
Health Care through Marsh US Consumer

Only AEA members may:

- Vote in the annual election of officers and at the Annual Business Meeting
- Contribute to *AEJ* and *JEP* online "Discussion Forums"
- Submit papers to be considered for presentation at the AEA Annual Meeting Program
- View webcasts of certain AEA Annual Meeting and Continuing Education sessions online

Regular membership dues are based on annual income.

Online member benefits begin immediately. Requested journals in print or CD begin with the issue following posting of your payment. Membership may not be back-started. Journals are mailed second class; please allow 6 to 8 weeks for arrival of journals shipped outside the US. CDs are mailed First Class.

Contact Information

American Economic Association
2014 Broadway, Suite 305
Nashville, TN 37203

Phone: (615) 322–2595
Fax: (615) 343–7590
www.vanderbilt.edu/AEA/

aeainfo@vanderbilt.edu

It is important to include your e-mail address and to keep it up to date. It often is used for verification of services. In addition, we notify members of important dates and new services by e-mail.

in *The Future of the Commons: Beyond Market Failure and Government Regulation,* by Elinor Ostrom, with contributions from Christina Chang, Mark Pennington, and Vlad Tarko. 2012. At http://www.iea.org.uk/sites/default/files/publications/files/IEA%20Future%20of%20the%20Commons%20web%2029-1.10.12.pdf.

The Society for Economic Dynamics has a short, delightful interview with Robert Lucas in its newsletter, *Economic Dynamics*. Lucas says: "My paper, 'Econometric Policy Evaluation: A Critique' was written in the early 70s. Its main content was a criticism of specific econometric models—models that I had grown up with and had used in my own work. These models implied an operational way of extrapolating into the future to see what the 'long run' would look like. . . . Of course every economist, then as now, knows that expectations matter but in those days it wasn't clear how to embody this knowledge in operational models. . . . But the term 'Lucas critique' has survived, long after that original context has disappeared. It has a life of its own and means different things to different people. Sometimes it is used like a cross you are supposed to use to hold off vampires: Just waving it at an opponent defeats him. Too much of this, no matter what side you are on, becomes just name calling." November 2012. At http://www.economicdynamics.org/News271.htm#interview.

Discussion Starters

Like many economists, I'm always on the lookout for analysis of the benefits, costs, and tradeoffs of life's difficult questions. Thus, I was delighted to run across "The Hygienic Efficacy of Different Hand-Drying Methods: A Review of the Evidence," by Cunrui Huang, Wenjun Ma, and Susan Stack. Based on a review of 12 studies, paper towels clearly win out over regular air dryers, jet air dryers, and cloth rollers for preventing the spread of germs, for consumer preference, and for noise prevention, while other issues like environmental impact are essentially a wash between the alternatives. August 2012 issue of the *Mayo Clinic Proceedings,* 87(8): 791–98. At http://www.mayoclinicproceedings.org/article/S0025-6196%2812%2900393-X/fulltext.

Bartow J. Elmore traces "The American Beverage Industry and the Development of Curbside Recycling Programs, 1950–2000." From the abstract: "Many people today consider curbside recycling the quintessential model of eco-stewardship, yet this waste-management system in the United States was in many ways a polluter-sponsored initiative that allowed corporations to expand their productive capacity without fixing fundamental flaws in their packaging technology. For the soft-drink, brewing, and canning industries, the promise of recycling became a powerful weapon for combating mandatory deposit bills and other source-reduction measures in the 1970s and 1980s." *Business History Review,* Autumn 2012, 86(3): 477–501.

■ *Thanks to Larry Willmore for his suggestions.*

Journal of Economic Perspectives—Volume 27, Number 1—Winter 2013—Pages 231–232

Notes

For additional announcements, check out the continuously updated JEP online Bulletin Board, http://www.aeaweb.org/bulletinboard.php. Calls for papers, notices of professional meetings, and other announcements of interest to economists should be submitted to Ann Norman at jep@jepjournal.org in one or two paragraphs containing the relevant information. These will be posted at the JEP online Bulletin Board. Given sufficient lead time (at least one month before an issue goes online), we will also print a shorter, one-paragraph version of your notice in the "Notes" section of the Journal of Economic Perspectives. *We reserve the right to edit material received.*

Call for sessions and papers for the January 2014 American Economic Association Annual Meeting. Members wishing to give papers or organize complete sessions for the program for the meetings in Philadelphia, January 3–5, 2014, are invited to submit proposals electronically to Professor William Nordhaus via the AEA website http://www.aeaweb.org/Annual_Meeting/submissions.php. While papers covering a wide array of topics in economics will be included on the 2014 program, Professor Nordhaus especially encourages interdisciplinary proposals.

To be considered, individual paper proposals (with abstracts) and up to two *Journal of Economic Literature* bibliographic codes in rank order should be submitted **by April 1, 2013. The deadline for complete session proposals is April 15, 2013**. At least one author of each paper must be an AEA member. All authors of papers on a complete session must join the AEA if the session is selected for the program.

Econ-Harmony allows AEA members with an individual paper to submit to the 2014 AEA Meetings to post information about their paper and search for others with similar interests who might join them to form a complete session submission, and it provides an opportunity to volunteer as a session chair. Proposals for complete sessions have historically had a higher probability of inclusion (35–40%) than papers submitted individually (10–15%). Individual paper contributors are strongly encouraged to use the AEA's Econ-Harmony website to form integrated sessions. Proposals for a complete session should be submitted only by the session organizer. Sessions normally contain three or four papers.

Please make certain your information is complete before submission. No changes will be accepted until a decision is made about inclusion on the program (usually in July). Econometric studies or highly mathematical papers are not appropriate for sessions sponsored by the AEA: such papers should be submitted to the Econometric Society. Do not send a complete paper. The Association discourages multiple proposals from the same person, and under no circumstances should the same person submit more than two proposals.

Some of the papers presented at the annual meeting are published in the May *American Economic Review* (the Papers & Proceedings). The President-elect includes at least three contributed sessions (12 papers) from among those submitted in response to the Call for Sessions and Papers. Econ-Harmony will open February 12, 2013. Go to http://www.aeaweb.org/econ-harmony/.

Call for poster proposals. The **Committee on Economic Education will sponsor a poster session** at the **2014 ASSA Meetings** in Philadelphia, January 3–5, 2014, devoted to active learning strategies across the economics curriculum. Instead of papers, session presenters will prepare large visual poster summaries of their work, which will be mounted in an exhibition room to allow presenters to talk directly with session participants. Although we encourage presenters to include evidence that their strategy enhances learning, we do not require quantifiable evidence. Presenters should emphasize the originality of their strategy and provide sufficient information so that session participants may apply the technique in their own classrooms.

Proposals should describe the teaching strategy and explain how it will be presented in the poster. Posters marketing textbooks, commercial software, or similar materials will not be considered for the session. Proposals are limited to two pages, should include full contact information for all authors and

are **due by April 1, 2013**. Send proposals to: Steven Cobb at scobb@unt.edu.

ASSA 2013 Webcasts are open-access, available online compliments of the AEA: "Sovereign Debt Crises and Policies: History and Future Prospects," "Health Insurance and Government Mandates: A Session to Honor Amy Finkelstein, John Bates Clark Medalist for 2012," "Models or Muddles: How the Press Covers Economics and the Economy," "Reflections on the 100th Anniversary of the Federal Reserve," and "What Do Economists Think About Major Public Policy Issues?" Go to http://www.aeaweb.org/webcasts/2013/index.php. Webcasts from 2009 through 2012 and Continuing Education Programs are available at http://www.aeaweb.org/webcasts/index.php.

The Committee on Economic Education (CEE) announces that its Third Annual AEA Conference on Teaching (at the undergraduate and graduate levels) and Research in Economic Education (all levels, including precollege) will be held May 29 to May 31, 2013 in Chicago, hosted by the Committee on Economic Education, in cooperation with the *Journal of Economic Education* and the Federal Reserve Bank of Chicago. The conference is at the Renaissance Chicago Downtown Hotel. The first night of the conference will include a tour of the Money Museum and dinner at the Federal Reserve Bank of Chicago. Plenary talks will be given by John List, Steven Levitt, Derek Neal, and Dan Sullivan (Director of Research of the Chicago Fed).

Submissions are now closed. Conference registration will open February 15, and registration will cost $125. Conference rates for rooms at the Renaissance will be $209/night. For more details go to http://www.aeaweb.org/committees/AEACEE/index.php.

The Committee on the Status of Women in the Economic Profession (CSWEP) will sponsor sessions at the January 2014 ASSA Annual Meeting in Philadelphia, PA. We will be organizing three sessions on gender-related topics and three sessions on econometrics topics. Accepted papers will be considered for publication in the Papers and Proceedings issue of the *American Economic Review*. Abstracts of individual papers and complete session proposals will be considered. Email a cover letter (specifying to which set of sessions the paper is being submitted) and a copy of a one- to two-page abstract (250–1,000 words), clearly labeled with the paper title, authors, names, affiliation, and contact information for all the authors **by March 1, 2013**, to cswep@econ.duke.edu.

The Committee on Economic Statistics of the American Economic Association (AEAStat) will sponsor three sessions on economic measurement at the January 2014 AEA meetings to be held in Philadelphia, PA. The Committee welcomes both submissions of individual papers and proposals for sessions of three or four papers. The Committee is interested in receiving submissions in any area of economics, especially topics making use of new and unexplored datasets. Examples of topics that would be of interest include (but are not restricted to): the duration of unemployment and its causes; the mortgage and credit markets; financial risk in the household, business, and banking sectors; how the diffusion of innovation across firms contributes to productivity growth; and how income inequality, economic mobility, and economic opportunities in the US today compare with other countries and with earlier times. One of the three sessions organized by the Committee will be published in the *AER* Papers and Proceedings volume. Abstracts for individual papers or for the papers to be included in a proposed session should be submitted no later than **April 15, 2013**, to Robert Feenstra, Chair, Committee on Economic Statistics, aea-stat@umich.edu. Submissions should be PDF files and include name, institutional affiliation, and email address of all participants.

Job Openings for Economists (JOE). The AEA is initiating a listing of retired economists who may be interested in teaching on either a part-time or temporary basis. Individuals can add or delete their name at any time during the period that the listing is open. The listing will be active from February 1 through November 30 each year. Listings will be deleted on November 30; the service will be closed during December and January, re-opening on February 1. Go to http://www.aeaweb.org/joe/.

Call for proposals for special issues. The *Journal of Benefit-Cost Analysis* is soliciting proposals for occasional special issues. The short proposal should include the thematic topic, a preliminary set of papers and authors, and reasoning for its suitability for the *JBCA*. The author of the proposal will be a guest editor. Submissions and queries received at jbca@umbc.edu.

The 2013 LIS Introductory Summer Workshop will be held June 30–July 7, at the University of Luxembourg campus. **Applications are due March, 15 2013.** This one-week workshop introduces researchers in the social sciences to comparative research in income distribution, employment, and social policy, using the *Luxembourg Income Study Database* (LIS)—the largest available database of harmonized income microdata collected over a period of decades. It also introduces them to the *Luxembourg Wealth Study* (LWS)—the first cross-national database of harmonized wealth microdata in existence. Tuition is €1,400, including room and board. Instruction is in English. Go to http://www.lisdatacenter.org/news-and-events/lis-introductory-summer-workshop/.

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary *JEP* for your e-reader, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $20.00, $30.00, or $40.00, depending on income; for an additional $15.00, you can receive this journal in print. Change of address notice must be received at least six weeks prior to the publication month.

The Journal of

Economic Perspectives

Winter 2013, Volume 27, Number 1

Symposia

Patents

Trading Polution Permits

Articles

AMERICAN ECONOMIC ASSOCIATION