VOLUME 5 | NUMBER 2 | AUTUMN 2009

Competition Policy International

Edited by David Evans

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Letter From the Editor

David S. Evans

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Table of Contents

Table of Contents *Continued*

From the Editor

This Autumn 2009 issue marks several anniversaries; it is the tenth volume of CPI, the end of our fifth year, and the last issue we will publish in the first decade of the 21st century. Since our first issue, we've published 134 articles from many of the leading thinkers, doers, and judges of antitrust from around the world. As the global competition policy community has grown, so has this publication. Over the course of the year the CPI website attracts visitors from more than 150 countries. We extend our thanks to this vibrant community.

Our tenth issue follows a very difficult year for the economies in many countries. Looking back, the September 2007 run on the Northern Rock Bank in Britain was a warning shot of what was to come. After an initial injection of liquidity it was soon nationalized. A year later Lehman Brothers collapsed and a global financial meltdown appeared imminent. Governments came to the rescue of many financial institutions as well as other industries, such as automobiles, that were subject to collateral damage as lending and spending cratered. Forced mergers and bailouts occurred with seeming abandon. Financial regulators talked much about firms being too big to fail but less about whether firms were too big and why.

The first collection of articles in this issue deals with several antitrust aspects of the financial crisis. Philip Lowe kicks off the discussion with an article on DG Competition's views. Bruce Lyons argues that it makes sense to bail out banks under the circumstances but that one should be circumspect about helping other sectors. John Kwoka then argues in favor of the help that the U.S. government gave to its beleaguered domestic automobile industry. Lorenzo Coppi and Jenny Haydoc review the European Commission's policies on state aid and the financial crisis. The symposium concludes with an article by Ken Heyer and Sheldon Kimmel who argue that there is no reason for competition authorities to relax their examination of failing firm defenses given the crisis.

We then turn to the controversial issue of reverse payment settlements—cases in which branded pharmaceutical companies sue generic entrants for patent infringement and settle the litigation by paying the generic entrant some money in return for delaying entry. The U.S. Federal Trade Commission has challenged these types of settlements vigorously but the courts have not seen things the

same way. William Rooney and Elai Katz provide an overview, Michael Kades describes the problem with the *per se* legal treatment of some reverse payment settlements, and Anne Layne Farrar argues for a moderate approach. The European competition authorities and courts have not yet addressed the issue although it has been raised in the pharmaceutical sector inquiry. Marc van der Woude explains the approach he believes is required under EU law.

The past several years have seen considerable debate over single-firm conduct. An important issue is whether the single-monopoly profit theorem—which is closely identified with the Chicago School—convincingly demonstrates that firms usually lack the incentives to use tying, bundling, and other devices for purposes that reduce consumer welfare. In a widely circulating and influential working paper, soon to be published in the *Harvard Law Review*, Einer Elhauge provocatively asserts the theorem is dead and argues that many forms of tying and bundling should be considered highly suspect. CPI recruited four commentators who we thought would have diverse views on Elhauge's paper, and indeed they did. The commentary begins with Harry First who provides a supportive summary of Elhauge's argument, is followed by Daniel Crane and Joshua Wright who dispute Elhauge's conclusions on bundled discounts, continues with Barry Nalebuff who clarifies issues surrounding the welfare-effects of price discrimination (which is key to Elhauge's analysis) and agrees and disagrees with various aspects of Elhauge's piece, and concludes with Paul Seabright who argues that the single-monopoly profit theorem may have some life left in it.

Continuing our anniversary theme, Thomas Kauper provides his perspective on the government antitrust case that led to the breakup of the American Telephone & Telegraph Company 25 years ago. Kauper was the Assistant Attorney General for Antitrust who brought the case against AT&T in 1974.

This issue concludes with a classic piece by Arnold Harberger on the social cost of monopoly. As Hill Wellford, who introduces the article explains, Harberger's piece was revolutionary both because it documented that the social costs of monopoly were surprisingly small and because it pioneered the use of empirical methods in antitrust.

The classic has been a feature of CPI from the beginning. We believe that there is a tendency to forget some of the lessons from leading thinkers on antitrust over the years and that it is helpful to go back and read originals or at least be reminded of them. The first classic we reprinted was Oliver Williamson's

Economies as an Antitrust Defense: The Welfare Tradeoffs. We extend our congratulations to Professor Williamson who was awarded the 2009 Nobel Prize in Economics for work that has had a profound influence on our theoretical and empirical understanding of firm governance, transactions costs, and contractual relationships.

On behalf of CPI's readers and its editorial team, I am delighted to extend my thanks to all the contributors of this issue.

David S. Evans
University College London and University of Chicago

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Competition Policy and the Global Economic Crisis

Philip Lowe

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Competition Policy and the Economic Crisis

*Philip Lowe**

The Commission stands firm on the importance of maintaining the competition rules and a policy of robust competition policy enforcement. I propose to discuss in this article first why we believe that competition policy is one of the tools we need to deploy to help maintain the integrity of the EU single market and to help our economies out of the crisis, and then to examine, concretely, how the crisis has affected and is affecting our approach to enforcing the EC State aid rules, as well as the EC antitrust and merger control rules.

*Director General, DG Competition, European Commission. The author would like to thank Mercedes Campo Mozo, Monica Cunningham, Anna Emanuelson, Andras Inotai, Philip Kiena, Peter Ohrlander, Sam Pieters, Koen Van de Casteele, and Christoph Walkner for their assistance in preparing this article.

I. Introduction

The past year has been challenging for the economy and for business, but also for policy makers. Governments, central banks, and financial regulators are all working hard to stabilize the world financial system and to introduce the regulations and institutions necessary to try to ensure that the current crisis cannot recur. At the same time, policy-makers are working on policies to help minimize the impact of the crisis on the real economy.

Within the European Union, the European Commission, and, in particular, my directorate general, has the task of scrutinizing government aid to financial institutions and to the real economy, under the competition rules laid down in the EC Treaty.[1] More widely, the Commission, in the area of internal market and economic and financial policy, has also put in place measures[2] to help restore consumer and business confidence, restart lending, and stimulate investment in the EU economies, and is working on proposals for a new regulatory and supervisory framework for financial services.

At the outset of the crisis there was pressure on the Commission to set aside the competition rules on State aid, in order to allow EU Member States freedom to implement financial sector rescue measures as they saw fit. However, it was very quickly recognized that there was a need to enforce common rules so as to help maintain a level playing field in the EU and avoid large scale movements of funds between Member States by investors in search of the highest level of protection. Under the EU state aid rules mechanisms were put in place to minimize the distortions of competition that might result from the large-scale award of rescue aid, so as to avoid disrupting the European Single Market and to prepare for the return to normal market functioning.

THE COMMISSION STANDS FIRM ON THE IMPORTANCE OF MAINTAINING THE COMPETITION RULES AND A POLICY OF ROBUST COMPETITION POLICY ENFORCEMENT.

As the crisis has spread into and deepened in the real economy, our mergers and antitrust policies have also come under pressure.

The Commission stands firm on the importance of maintaining the competition rules and a policy of robust competition policy enforcement. I propose to discuss in this article first why we believe that competition policy is one of the tools we need to deploy to help maintain the integrity of the EU single market and to help our economies out of the crisis, and then to examine, concretely, how the crisis has affected and is affecting our approach to enforcing the EC State aid rules, as well as the EC antitrust and merger control rules.

II. Competition Policy in General and the Crisis

The crisis has not undermined the economic principle that competition breeds competitiveness: it enables an efficient allocation of resources and stimulates technological development and innovation. This, in turn, leads to a wider choice of products and services, lower prices, better quality, and higher productivity. The benefits of pursuing a competition policy based on these principles are clear. For example, the opening up of telecommunications and air transport services to competition means that we now have lower prices and a wider choice of telecommunications and air transport services in Europe than previously. In 2008, the Commission's application of competition policy tools resulted in estimated consumer benefits of more than 11 billion EUROs.[3]

The benefits of competition are particularly relevant at times of economic crisis. By producing consumer savings through the breakup of cartels or by prohibiting anticompetitive mergers, competition policy stimulates demand and leads to concrete improvements in consumers' purchasing power. At the same time, competition not only leads to lower prices for consumers (and thereby lowers inflation) but it also reduces price levels in wholesale and intermediary markets. This, in turn, has a beneficial effect on the competitiveness of those undertakings that act as customers on these markets. For example, introducing more competition to telecommunications markets led to an average decrease of 45 percent of the price businesses paid for international calls between 1998 and 2003.[4]

The link between effective competition and economic growth is particularly important in times of economic recession. As markets characterized by effective competition make companies innovate more, they drive economic growth through the improvement of total factor productivity. Total factor productivity growth can be several percentage points higher in sectors where the intensity of competition is higher. This can make the difference when markets cannot rely on large amounts of capital to stimulate growth.

And at a time when people are concerned with growing unemployment, it is important to emphasize that there is absolutely no evidence to suggest that more competition leads to net employment losses.

Markets subject to external competitive pressures also grow faster. It is estimated that if trade between EU Member States was eliminated (for example, as a result of market-sharing agreements or State restrictions on external competition) average productivity would fall by 13 percent.[5] Sealing off markets from outside competition allows companies to raise prices and to restrict output which, in turn, further deepens recession.

And at a time when people are concerned with growing unemployment, it is important to emphasize that there is absolutely no evidence to suggest that more competition leads to net employment losses. For example, in the wake of open-

ing the air transport sector to competition, direct airline employment in Europe rose by 6 percent between 1992 and 2001.[6]

It follows that alongside fiscal (and in some countries monetary) policy, competition policy should be an integral part of the toolbox on which governments rely for responses to the economic crisis. According to widely quoted research from the University of California, the relaxation of antitrust rules in the United States in the 1930s probably helped prolong the economic crisis by seven years. The relaxation of the antitrust rules—which included exempting certain industries from competition law—was partly to blame for the slowing down of the economy and for an unemployment rate of around 20 percent.[7]

This does not mean that competition policy (and competition enforcement agencies) do not face particular challenges arising from the crisis. However, a well-established competition regime should not require a lot of adjustment to cope with these challenges. And there should be no need to compromise on the principles of competition policy.

The types of adjustments that may be required are:

1. In order to be able to respond to urgent situations (e.g. the need to ensure that rescue measures for banks could go ahead quickly, in the interest of financial stability) processes may need to be streamlined and timelines adjusted to take account of the market situation so as to be able to respond accordingly.

2. In contributing to an effective response to the crisis, where we have discretionary powers, competition policy should arguably focus on those sectors that either directly or indirectly affect household expenditure to the greatest extent in order to ease the burden on consumers, as well as on sectors that are the most important for productivity growth. In the EU, network industries such as energy and telecommunications meet both criteria and therefore arguably should be the focus of attention. More generally, prioritization is increasingly important so as to ensure that enforcement action is targeted towards those infringements that have the greatest impact on consumers.

3. In an environment where confidence in markets may have decreased and where there is a greater chance of government intervention, competition advocacy will have a greater role to play in ensuring that State measures take on board competition principles and do not create disproportionate restrictions of competition, which will ultimately harm the economy and make things worse for consumers.

Finally, as the economic crisis puts pressure on State budgets and public sector expenditure may need to be cut back, authorities in charge of competition policy must also justify their resources to taxpayers. This requires them to constant-

ly improve their efficiency and effectiveness and to demonstrate to society that they deliver real benefits.

III. State Aid to the Financial Sector

Early on in the crisis EU Member States decided it was necessary to inject large amounts of State aid into the financial sector. The European Commission became involved, because of our powers to scrutinize State aid under the EU competition rules.

The State aid provided to EU banks and insurance companies have had clear benefits. They have helped avoid the meltdown of the financial system and helped re-open markets, re-establish lending to the real economy, and put financial markets back on the path towards normal market functioning (that is to say, without state support). Financial stability and protecting and preserving competitive markets are complementary objectives. Competition policy is there to support financial stability and create the right conditions for stable financial markets in both the short- and the longer-term—which is why it is crucial to ensure that bail-outs in the banking and insurance sector respect fundamental competition principles.

FROM THE START, OUR OBJECTIVE IN APPLYING THE STATE AID RULES WAS TO PRESERVE THE LEVEL PLAYING FIELD FOR EUROPEAN BANKS.

From the start, our objective in applying the State aid rules was to preserve the level playing field for European banks, by preserving competition between banks in different Member States and between banks throughout Europe which are competing on the same markets, taking into account their different risk profiles. We try to ensure that State aid measures do not undo all the benefits of the Single Market, and do not have the effect of delaying the return to normal competitive market functioning.

At the same time, it has been crucial to provide a clear and predictable framework for rapid approval of Member State rescue measures for individual banks and national schemes to support the banking sector. In the interests of speed and efficiency we have been flexible on process—but firm on the principles underpinning the state aid rules.

In order to assist Member states to take urgent and effective measures to preserve stability and to provide legal certainty, between October 2008 and July 2009 the Commission adopted four Communications indicating how we would apply the State aid rules to government measures to support the financial sector in the context of the current crisis. On October 13, 2008 the Commission adopted guidance indicating how we would apply State aid rules to state support schemes and individual assistance for financial institutions.[8]

Essentially the conditions it insisted on are:

- Non-discriminatory access to the schemes in order to protect the functioning of the Single Market by making sure that eligibility for a support scheme is not based on nationality;
- State commitments should be limited in time—and reviewed at least every six months—so that support can be provided as long as necessary but that it will be reviewed and adjusted or terminated as soon as improved market conditions permit;
- State support should be clearly defined and limited in scope to what is necessary to address the acute crisis in financial markets, while excluding unjustified benefits for shareholders of financial institutions at the taxpayer's expense;
- The private sector should contribute by way of an adequate remuneration for the introduction of general support schemes (such as a guarantee scheme) and it should also cover at least a significant part of the cost of assistance, so as to ensure that there are incentives to return state money;
- Beneficiaries should be subject to constraints on their behavior so as to prevent an abuse of state support by means of, for example, expansion and aggressive market strategies on the back of a state guarantee; and
- There should be an appropriate follow-up in the form of structural adjustment measures for the financial sector as a whole and/or restructuring by individual financial institutions that benefited from state intervention.

BY THE END OF 2008 THE SOLUTIONS BEING DEVISED BY MEMBER STATES EVOLVED FROM LARGELY GUARANTEE-BASED SCHEMES TO OTHER MEASURES SUCH AS RECAPITALIZATIONS.

The principles set out in the *Banking Communication* are based on our pre-existing Guidelines on rescue and restructuring aid.[9] As a rule, rescue and restructuring aid is assessed under Article 87(3)(c), which allows the Commission to authorize "aid to facilitate the development of certain economic activities [..] where such aid does not adversely affect trading conditions to an extent contrary to the common interest." The Commission relies on this provision to authorize aid to correct disparities caused by market failures or to ensure economic and social cohesion—but makes such aid subject to strict conditions. However, the Commission has recognized that the severity of the crisis justifies the award of aid on the basis of Article 87(3)(b) of the EC Treaty, under which aid can be allowed in order to "remedy a serious disturbance to the economy of a Member State."

By the end of 2008 the solutions being devised by Member States evolved from largely guarantee-based schemes to other measures such as recapitalizations. On December 5, 2008, following detailed discussions with the European Central Bank and the Member States, the Commission adopted detailed guidance on how it would assess these bank recapitalization schemes,[10] complementing the October 13 guidelines.

The *Recapitalisation Communication* distinguishes between banks that are fundamentally sound and receive temporary support to enhance the stability of financial markets and restore lending to businesses and consumers, and distressed banks whose business model has brought about a risk of insolvency and which pose a greater risk of distortions to competition.

In particular, the *Recapitalisation Communication* establishes principles for pricing the injections of capital made by States into banks. For fundamentally sound banks, the price of capital injections should be linked to: the base rates set by central banks to which a risk premium is added to reflect the risk profile of the beneficiary bank; the type of capital used; and the nature of the safeguards against abuse of public funding that accompany the recapitalization measure. This pricing mechanism needs to carry sufficient incentives to keep the duration of state involvement to a minimum, for instance by having a rate of remuneration that increases over time.

BANKS IN DISTRESS WHICH ARE AT RISK OF INSOLVENCY SHOULD, IN PRINCIPLE, BE REQUIRED TO PAY MORE FOR STATE SUPPORT AND SHOULD BE SUBJECT TO STRICTER SAFEGUARDS.

Banks in distress which are at risk of insolvency should, in principle, be required to pay more for state support and should be subject to stricter safeguards. Injections of state capital into these banks are acceptable only on condition that they are followed by far-reaching restructuring to restore long-term viability, which may include changes to management and corporate governance.

By way of these first two Communications, the Commission introduced some necessary flexibility into our handling of national financial sector rescue schemes and individual financial institution rescue measures, without losing sight of key state aid principles. While giving Member States clear guidelines on what would or would not be acceptable, we aimed to achieve a degree of consistency in Member State responses across Europe.

Flexibility in process as well as in substance has also been very important. Support schemes such as guarantees or re-capitalization schemes have been cleared by the Commission very quickly as long as the schemes fulfill conditions, which guarantee that they are well-targeted and proportionate and contain safeguards against unnecessary negative effects on competition.

While it seems clear that the financial sector rescue packages adopted by Member States since October 2008 averted the risk of financial meltdown, by early 2009 it also seemed clear that further measures were needed to restore trust and to return the financial sector to normal functioning.

One reason why credit remained squeezed seemed to be uncertainty about the value and location of impaired assets held by banks. On February 25, 2009, after detailed discussions with the Member States, the Commission adopted a Communication on the treatment of impaired assets.[11] This Communication discusses the budgetary and regulatory implications of asset relief measures that could be adopted by Member States to remove impaired or toxic assets from the balance sheets of banks, and provides guidance on the application of the State aid rules to such measures.

The *Impaired Assets Communication* stipulates that:

- Member States must make asset relief measures conditional on full transparency and disclosure of impaired assets and must ensure that the costs of the impaired assets are shared among the Member States, shareholders, and creditors of the financial institutions.
- Member States should take a coordinated approach to identifying assets eligible for asset relief measures and to valuing assets. The primary task of carrying out asset valuation is performed at the national level, and validated by the appropriate supervisory authority. However, each individual case is checked by the Commission with the help of external experts.
- Finally, restructuring measures should follow, so as to ensure the return to viability of the banks in question, and the return to normal market conditions.

The measures in question could involve asset purchases (including "bad" bank scenarios), asset swaps, state guarantees, or hybrid systems—the choice is, of course, up to the Member States who are responsible for the methods and design of asset relief measures. The complexity of asset eligibility and valuation is illustrated by the fact that, to date, the Commission has given final approval for very few impaired asset measures, and is still investigating others.

THE COMPLEXITY OF ASSET ELIGIBILITY AND VALUATION IS ILLUSTRATED BY THE FACT THAT, TO DATE, THE COMMISSION HAS GIVEN FINAL APPROVAL FOR VERY FEW IMPAIRED ASSET MEASURES.

Finally, on July 23, 2009 the Commission published guidelines setting out its approach to assessing restructuring aid given by Member States to banks.[12] Essentially, those banks that have received large amounts of aid and that have unsustainable business models will have to restructure in order to return to long-term viability without relying on State support.

The *Restructuring Communication* stipulates that banks in need of restructuring have to demonstrate strategies to achieve long-term viability under adverse economic conditions; this involves rigorous stress testing of the businesses. In some cases, divestments will not be needed but in many cases they will be essential, either to ensure viability of core businesses or to reflect the negative competitive impact of aid on key market segments. However, the Commission also needs to be realistic about divestments, for example with respect to the likelihood of finding buyers and the time period for divestiture.

Additionally, banks that have received large amounts of aid and that have unsustainable business models should, along with their capital holders, contribute to the cost of restructuring as much as possible with their own resources. This creates appropriate incentives for future behavior. An appropriate price for State support ensures that the aid cannot be used to finance activities such as acquisitions which are not linked to the restructuring process. Similarly, aid should not be used to pay interest to holders of hybrid capital instruments when a bank receiving aid is making losses, unless this remuneration is essential to attract new capital.

Finally, the Commission needs to create conditions which foster the development of competitive markets after the crisis. Where restructuring is necessary, decisions need to be taken now, in order to chart the road map of the bank to viability without state support. This may be achievable over two to three years, but restructuring may even take up to five years. Banks which do not need fundamental restructuring, because their basic business models are sound, also need to plan their return to normal market operation without state support. Essentially, exit strategies from national support schemes for all banks now need to be developed providing the conditions for a sustainable recovery of private markets as a whole are met. This requires detailed discussions among the European Commission and the Member States, national central banks and regulators, the European Central Bank, and coordination across all policy areas.

In addition to these Communications, in the past year the Commission has taken around 70 decisions approving national schemes for aid to the financial sector.

Taken as a whole, the four Communications from the Commission provide guidance as to what we see as the key principles that Member States need to comply with, in order to: 1) reduce the risk that national measures to support the financial sector might fragment the Single Market; 2) minimize any distortions of competition that might result from the state intervention; and 3) avoid distorting the incentives of market players in the financial sector going forward.

In addition to these Communications, in the past year the Commission has taken around 70 decisions approving national schemes for aid to the financial sector—taking the form of guarantee schemes, bank recapitalization schemes,

and asset relief schemes—as well as individual rescue aid measures and some restructuring aid decisions.[13]

An example of a complex, ongoing investigation is the ING "illiquid assets" case. On March 31, 2009, the Commission approved for 6 months the illiquid asset back-up facility provided by the Dutch State to the financial group ING. At the same time, the Commission initiated the formal investigation procedure laid down in Article 88 (2) of the EC Treaty to verify that the conditions laid down in the *Impaired Assets Communication* regarding valuation (including the valuation methodology) and burden sharing of the measure are met.

In January 2009, the Dutch State and ING agreed on a so-called illiquid assets back-up facility for a portfolio of U.S. $39 billion par value worth of securitized U.S. mortgage loans, mostly consisting of so-called Alt-A mortgages. Alt-A loans are the category of U.S. loans between prime and sub-prime, often granted on the basis of a simple declaration by the borrower about his income with no other proof required.

Under the transaction, the Dutch State will buy the right to receive the cash flows on 80 percent of this U.S. $39 billion portfolio by paying ING about U.S. $28 billion. That amount will be paid by the Dutch State in accordance with a pre-agreed payment scheduled.

Following an initial assessment of the measure, the Commission decided for reasons of financial stability, similar to those governing the assessment of rescue aid, not to raise objections for a period of six months. The Commission found that the measure complies with the conditions on eligibility of assets, asset management arrangement, transparency and disclosure, and a guarantee fee as stipulated in the *Impaired Assets Communication*. However, some conditions like valuation and burden sharing require further in-depth analysis, which is why the Commission opened an in-depth investigation.[14]

ING had already benefited from an emergency recapitalization of 10 billion Euros, which the Commission approved in November 2008.[15]

In essence these measures are all part of the process undertaken by Member States to restore stability to the banking sector and put it on the path back to normal market functioning, without State support. In parallel, a move toward regulatory reform of the financial sector is underway. The Commission has put forward a number of proposals to improve regulation and supervision of the financial sector.[16]

This regulatory program and the restructuring of banks are complementary routes to the same goal of the return to viability of individual banks and of the European banking sector as a whole. Banks must operate on the basis of sound business models in a regulatory framework in which they can compete on the merits with balanced incentives without state aid. They must be able to exit the

market or restructure when they are no longer competitive, without triggering the systemic consequences that have characterized the current crisis.

IV. State Aid to the "Real" Economy

State aid issues are, of course, not confined to the financial sector. Before the end of 2008, the effects of the credit crisis were being felt in the "real" economy and Member States began to consider what measures they could take to tackle that crisis too.

As stated, relaxing or suspending the State aid rules for the duration of a financial and economic crisis has never been an option—the effect would be that some companies would have enjoyed State subsidies, giving them a competitive advantage over their competitors. A subsidy race between Member States would not only be financially unsustainable, it would also delay the necessary restructuring of the economy and thus deepen the recession and its long-term effects.

ALTHOUGH PUBLIC INTERVENTION HAS TO BE DECIDED AT NATIONAL LEVEL, IT NEEDS TO BE IMPLEMENTED WITHIN A COORDINATED FRAMEWORK AND ON THE BASIS OF PRINCIPLES COMMON TO THE WHOLE OF THE EU.

Although public intervention has to be decided at national level, it needs to be implemented within a coordinated framework and on the basis of principles common to the whole of the EU.[17]

The Commission's policy has been to encourage a horizontal approach that benefits the whole economy, rather than specific industrial sectors. However, this does not mean that Member States do not have flexibility to target specific problems within their territory.

For the real economy, on December 17, 2008 the Commission adopted a Temporary State Aid Framework which provides additional possibilities for Member States to grant State aid until the end of 2010. Some technical adjustments to this framework, mainly on guarantees, were introduced on February 25, 2009.

The main objective of the Temporary Framework is to reduce the negative effects of the crisis in the real economy by facilitating companies' access to finance. Sufficient and affordable access to finance is clearly a pre-condition for investment, growth, and job creation by the private sector. In the short-term, the economic crisis has negative consequences on the viability of European companies. In the long-term, it could delay investments in sustainable growth and other Lisbon Strategy objectives.

The Temporary Framework has additional objectives: 1) to contribute to the immediate unblocking of bank lending and continuity in companies' access to

finance; 2) to ensure that limited amounts of the necessary aid reach the recipients in the most rapid and effective way; and 3) to encourage companies to continue investing into a sustainable future, including the development of green products.

Although Member States can already grant State aid for a range of different objectives (environmental aid, rescue and restructuring aid, etc.), there was a need for additional measures targeted to the exceptional difficulties in obtaining finance.

The measures contained in the Temporary Framework are—like the crisis measures adopted in the banking sector—based on Article 87 (3) (b) of the Treaty. This is the reason why the new measures are limited in time, until the end of 2010.

On the basis of the Temporary Framework Member States may:

- Give 500,000 EUROs per undertaking to cover investments and/or working capital over a period of two years.
- Offer State guarantees for loans at a reduced premium. The guarantee may relate to both investment and working capital loans and it may cover up to 90 percent of the loan.
- Offer aid in the form of subsidized interest rate applicable to all type of loans. This reduced interest rate can be applied for interest payments until the end of 2012.
- Offer subsidized loans for the production of green products involving the early adaptation to or going beyond future Community product standards.

The Commission considers that environmental goals should remain a priority despite the crisis—and, for this reason, it sought to give support to companies investing in environmental projects.

THE COMMISSION CONSIDERS THAT ENVIRONMENTAL GOALS SHOULD REMAIN A PRIORITY DESPITE THE CRISIS.

Furthermore, the Temporary Framework also allows for:

1. A temporary derogation from the Community guidelines on Risk Capital[18] guidelines in order to allow 2.5 million of risk capital injection in small- and medium- sized enterprises ("SMEs") per year (instead of 1.5 million EUROs) and a reduction of the minimum level of private participation (from 50 percent to 30 percent).
2. A simplification of the Communication on short-term export credit insurance.[19] This makes it easier for Member States to demonstrate that certain risks are temporarily non-marketable and can thus be covered by the State.

Member States do need to notify all the measures contained in the Temporary Framework—but special procedures have been put in place to ensure that the Commission is in a position to very quickly adopt decisions allowing State aid under the Temporary Framework. To date, over 65 aid scheme decisions have been adopted under the Temporary Framework.

To give some examples of decisions under the Temporary Framework:

On December 30, 2008 the European Commission approved two German measures to support the real economy, the first under the Temporary Framework. The first measure was intended to provide liquidity for companies affected by the credit squeeze, and allows interest rate reductions on loans to finance investments and working capital of up to 50 million EUROs to be granted to companies with a turnover of less than 500 million EUROs. The second measure is a framework scheme which allows federal, regional, and local bodies to provide aid of up to 500,000 EUROs to firms in need. It only applies to companies that were not in financial difficulties on July 1, 2008.[20]

On June 12, 2009 the European Commission authorized a Finnish guarantee scheme aimed at providing relief to companies encountering financing difficulties as a result of the credit squeeze. The scheme allows authorities to grant aid in the form of subsidized guarantees for investment and working capital loans concluded by December 31, 2010. The scheme meets the conditions laid down in the Temporary Framework because it is limited in time, respects the relevant thresholds, and applies only to companies that were not in difficulty on July 1, 2008.[21]

In adopting the Temporary Framework, the Commission sought to react in a pragmatic and responsible way to the evolving market circumstances, so as to enable Member States to react to market circumstances, but without compromising the State aid rules and the EU Single Market.

The Commission is also thinking ahead and preparing also for the review process. We are closely monitoring the aid schemes put in place by Member States under the Temporary Framework—a report on these measures should be provided to the Commission by Member States by October 31, 2009.

As with financial sector measures, the Commission's aim has been to be flexible on process—by facilitating national umbrella schemes—but firm on the underlying principles. It is important the Commission responds to market conditions while, at the same time, resisting pressures to allow Member States to adopt protectionist measures and provide long term support to ailing national companies, contrary to the principles of fair competition among EU companies. EU State aid policy provides a framework for ensuring that restructuring is based on a feasible, coherent, and far-reaching plan to restore long term viability of companies, which also helps safeguard employment.

V. Mergers and the Crisis

The picture under the EC merger control rules is quite different. In contrast with the wholesale government interventions providing financial support to the banking and insurance sectors, there has been relatively little merger activity directly related to banking rescue or restructuring (or other financial firms) that has been subject to review by the Commission. Some cases—such as the Lloyds/HBOS merger in the United Kingdom and the Commerzbank/Dresdner merger in Germany—have been dealt with by National Competition Authorities in the relevant EU Member States.

It is, however, likely that as the worst of the financial sector turbulence calms down, there will be further mergers in the banking sector. The same applies to other areas of the real economy where the effects of the economic downturn may result in some consolidation.

In assessing mergers that occur against the backdrop of the financial and economic crisis, the Commission's priority is to ensure that we maintain effective scrutiny under the competition test laid down in the EC Merger Regulation.[22] The purpose of the test is to ensure that consumer welfare is preserved. In the shorter term, this will be achieved by maintaining financial and economic stability; but, in the mid- to long-term, it will be achieved by preserving competitive market structures.

We believe that the EC Merger Regulation is an appropriate and sufficiently flexible tool for merger control enforcement in times of crisis as well as in normal times. There is no need for special procedures to be adopted for the review of mergers in time of crisis, nor is there a need to amend our substantive test for approving mergers. But, of course, the crisis has thrown up procedural and substantive challenges, some of which are directly linked to Member State intervention in the economy as a result of the crisis. I will deal with these in some detail.

> Consequently, any nationalization measure has to be assessed under the competition rules in the same way as any other change of ownership.

In terms of procedure, one issue that arises is how to deal with nationalizations. The EC Treaty is neutral on the question of private or public ownership. Consequently, any nationalization measure has to be assessed under the competition rules in the same way as any other change of ownership. The first step would be to determine whether a nationalization measure is a merger within the meaning of our merger rules—which is something we assess very carefully, on a case-by-case basis. This is a particularly sensitive issue where a government takes control of two or more companies or banks which are competitors on the same markets.

Another issue that arises is whether the time limits for the approval process laid down under the EC Merger Regulation and its implementing provisions need to be adjusted in a crisis situation. Timing of the review process is always important to the merging parties and may be even more pressing in case of rescue mergers. However, in order to carry out an effective and thorough review of whether any particular merger is likely to give rise to competition concerns, it is important that the Commission has sufficient time. The rules, as they stand, give a certain degree of flexibility. For instance, the Commission can give the parties permission to derogate from the normal standstill obligation and implement a merger immediately, pending the outcome of the review.

In exceptional cases, we may also need to work faster than usual. In the BNP Paribas/ Fortis case, from December 2008, the Commission adopted its authorization decision two weeks before the normal deadline. The case, which concerned the acquisition of Fortis' Belgian and Luxembourg assets by BNP Paribas, was only cleared subject to conditions relating to the credit card market, so as to avoid narrowing consumer choice for credit cards.

Remedies is another area where we may need to show some flexibility on timing. Where we are considering proposing that a merger be cleared subject to, for instance, a commitment to divest a business, it may be necessary to take into account the difficulty in finding buyers given the current economic climate. This can addressed, depending on the circumstances, either by requiring upfront buyers, in order to guarantee the effectiveness of the proposed remedy, or by extending the divestment period. However, both of these possibilities are already covered by our revised Notice on Remedies, adopted in October 2008.[23]

The Remedies Notice reflects the Commission's experience of remedies in a large number of cases, a study on remedies in past cases that we carried out in 2005, as well as recent judgments by the European courts. It also takes into account amendments brought to the EC Merger Regulation in 2004, such as the possibility of extending the compulsory merger deadlines in order to discuss and assess remedies.

In terms of the Commission's substantive assessment the competition test under the EC Merger Regulation already allows the Commission to take into account rapidly evolving market conditions in its competition assessment.

In terms of the Commission's substantive assessment the competition test under the EC Merger Regulation already allows the Commission to take into account rapidly evolving market conditions in its competition assessment. Even in sectors suffering particularly from the current economic crisis, the Commission takes the view that it is important to ensure that markets remain competitive. In the European airline sector, for instance, the Commission takes great care that the interests of consumers in having a competitive choice of airline services in Europe are safeguarded, particularly in view of the current consolidation process.

In the Lufthansa/ SN Brussels Airlines case, on June 22, 2009 the Commission approved the acquisition by Lufthansa of SN Brussels Airlines. The Commission's decision is conditional upon the implementation of a set of remedies offered by Lufthansa to alleviate the Commission's competition concerns, in particular on a number of routes between Belgium and Germany and Belgium and Switzerland. Taking into account past experience with remedies in the airline sector, these commitments aim at generally enhancing the attractiveness of the route for new entrants. They provide for an efficient and timely slot allocation mechanism. Furthermore, any new entrant will obtain grandfathering rights over the relevant slots, once it has operated a route for a certain pre-determined period of time. This specifically targets the problem of slot congestion, which is an important entry barrier on the problematic routes. Ancillary remedies, such as interlining, special pro-rate or code-share agreements, and the participation in Frequent Flyer Programs are also foreseen.

> In the event of a rescue merger, the Commission's policy and practice provide for consideration of the so-called "failing firm defense."

In the event of a rescue merger, the Commission's policy and practice provide for consideration of the so-called "failing firm defense." However, the conditions set out in the Guidelines on horizontal mergers would need to be met.[24] These guidelines suggest that the Commission may decide that an otherwise problematic merger can nonetheless be allowed if one of the merging parties is a failing firm, as long as the deterioration of the competitive structure of the market that follows the merger cannot be said to be caused by the merger.

The Guidelines identify the following three criteria as being especially relevant to the Commission's assessment of a failing firm defense:

1. First, the allegedly failing firm would, in the near future, be forced out of the market because of financial difficulties if not taken over by another undertaking.

2. Second, there is no less anticompetitive alternative purchase than the notified merger.

3. Third, in the absence of a merger, the assets of the failing firm would inevitably exit the market.

In a period of financial crisis and market collapse, it may often be difficult to obtain reliable information to test the merger against these criteria, for example the criterion of an alternative purchaser. However, this does not absolve the Commission from carrying out as thorough an investigation of the arguments as possible.

Under the EC Merger Regulation[25] the EU Member States can also intervene in order to prohibit, on public policy grounds, a merger that the Commission

might otherwise approve. But they do not have the right to clear mergers that the Commission would prohibit on competition grounds.

It is sometimes argued that in times of crisis, it would be appropriate for the Commission to be able to take into account other wider considerations, such as employment. However, experience has shown that a legal instrument such as the EC Merger Regulation is most effective when it is directed to one single objective. Employment concerns need to be addressed through other instruments. It is hard to see how it would be possible to agree on the wider objectives that should be taken into account in our assessment or, indeed, how it would be possible to agree on how these objectives should be implemented.

VI. Antitrust Policy and the Crisis

The current financial and economic crisis has not—at least to date—resulted in wholesale government intervention in company behavior, such as promoting or encouraging collective action or measures by companies to combat the effects of the crisis. Nor have companies brought to our attention many such initiatives of their own. However, we have come under some pressure from both governments and companies to suggest that we might relax our application of the EU antitrust rules, namely Articles 81 and 82 of the EC Treaty which respectively prohibit anticompetitive agreements between undertakings and abuses of dominance, in the event that such schemes might be thought necessary.

But in our view crisis conditions cannot justify collective or concerted action through so-called "crisis cartels" aiming to reduce capacity or production.

It is probably unavoidable that in times of recession many companies will suffer. There is a risk of reduced profits and overcapacity—but in our view crisis conditions cannot justify collective or concerted action through so-called "crisis cartels" aiming to reduce capacity or production.

In recent years the Commission has made cartels—arguably the most harmful type of competition infringement—a priority. We have implemented a comprehensive policy framework for cartels, including a very successful leniency program[26] and an effective fining policy.[27]

In the interest of maintaining competitive markets in the EU, which are fundamental to ensuring the economy finds its way out of the crisis, we believe it would be very unwise to relax our rules on cartels or indeed to pursue cartels any less vigorously. Of course, collective action can take other forms, some of which may be less harmful than cartels. However, any such cooperation between companies would have to satisfy the criteria laid down in Article 81(3)—that the companies concerned would have to show that the agreement contributed to improving production or distribution, or to promoting technical or economic

progress, while allowing consumers a fair share of the resulting benefit, but without imposing unnecessary restrictions or eliminating competition. The Commission would view any argument related to the economic crisis with considerable skepticism—and it would seem extremely unlikely that any agreement on prices or output could be justified. Nonetheless, the point is that under the rules certain types of cooperation are allowed, if they are truly necessary and proportionate.

In many ways, the focus of our enforcement policy in recent years is also suitable to meet the challenges posed by the current financial and economic crisis. The Commission has pursued a policy of targeting its antitrust enforcement efforts on those infringements that cause the most harm to consumers. It has consolidated an economics-centered, effects-based approach across the board—except with respect to naked cartels—and improved prioritization.

One tool we have used to this end is the sector inquiry; the Commission has carried out major inquiries in recent years into energy, financial services, and pharmaceuticals.[28] Our final report on competition in pharmaceuticals in Europe was published in July 2009. These inquiries were launched in sectors of the economy where there were indications that competition was not working as well as it might. They have helped us understand the sectors, identify where the obstacles to competition lie, and decide on the best course of action. For instance, in energy our sector inquiry resulted in both regulatory changes—the Third Energy Package[29]—and antitrust enforcement action. One lesson it has taught us is that competition enforcement action is not always the only solution to a competition problem—sometimes regulatory action is an option.

Decisions taken by the Commission following the energy sector inquiry have had a clear impact on improving competitor access to the market and potentially improving consumer choice. On March 18, 2009, the Commission opened the German gas market to competition by accepting commitments from RWE to divest its transmission network. The Commission had concerns that RWE may have abused its dominant position on its gas transmission network to restrict its competitors' access to the network. In order to alleviate these concerns, RWE offered to divest its entire Western German high-pressure gas transmission network.

In a separate case, the Commission imposed the first fines in the energy sector, amounting to 553 million EUROs on GDF Suez, as well as on the German E.ON Group for participating in a market-sharing agreement in the French and German gas markets. The Commission found that in 1975, when E.ON/Ruhrgas and GDF decided to jointly build the MEGAL pipeline across Germany to import Russian gas into Germany and France, they agreed not to sell gas transported over this pipeline in each other's home markets. They maintained the market-sharing agreement in place after European gas markets were liberalized thus denying French and German gas consumers the benefits of the 1998 liberalization, including more price competition and choice of suppliers.

The other focus of the Commission's enforcement action under the antitrust rules is against unilateral conduct such as abuses of dominance where we are again targeting our enforcement action against those infringements that cause the most harm to consumers. In December 2008 we adopted our Guidance on enforcement priorities in relation to exclusionary abuses of dominance,[30] but we have, in essence, been applying the principles underlying the Guidance for some time, notably in IT cases such as the Telefonica margin squeeze case, in Microsoft, and in the recent Intel decision. We are also focusing on the energy sector, with the E.On and RWE commitments decisions and other ongoing cases.

On May 13, 2009, the Commission adopted a prohibition decision in the Intel case finding that Intel infringed Article 82 of the EC Treaty. The decision orders Intel to cease its anticompetitive practices to the extent that they are ongoing and refrain from engaging in similar or equivalent practices, and imposes a fine of 1.06 billion EUROs. The Commission found that Intel engaged in two specific forms of illegal practice. First, Intel gave wholly or partially hidden rebates to computer manufacturers conditional upon (near) exclusivity for its x86 Central Processing Unit ("CPU"). Intel also made direct payments to a major retailer to stock only computers with its x86 CPUs. Second, Intel made direct payments to computer manufacturers to halt or delay the launch of specific products containing competitors' x86 CPUs and to limit the sales channels available to these products.

In the context of the financial and economic crisis, we have faced criticism over the level of our fines. In 2006 we adapted our fining policy to ensure that our fines would act as an effective deterrent and would better reflect the economic harm caused by cartels and other anticompetitive behavior. In the absence of criminal sanctions at EU level and taking into account the fact that there is little civil litigation, fines are the only instrument the Commission has to sanction and deter companies from engaging in the most serious violations of the antitrust rules.

WHILE OUR ANTITRUST FINES MAY NOW BE, ON AVERAGE, HIGHER THAN IN PREVIOUS YEARS, WE DO NOT BELIEVE THAT THEY ARE TOO HIGH NOW—RATHER, PREVIOUSLY THEY WERE TOO LOW TO BE A DETERRENT.

While our antitrust fines may now be, on average, higher than in previous years, we do not believe that they are too high now—rather, previously they were too low to be a deterrent. The Commission enforces EU competition rules across the largest integrated economic area in the world, and we target the most serious infringements, so the size of the Commission's fines also reflects the size and importance of the companies that we are investigating. Our fines are based on sound economic principles and are directly related to the economic harm likely to have occurred on the market, and to the duration of the infringement. And, at any event, the Commission is always bound by the threshold of 10 percent of the undertaking's worldwide turnover, which has remained unchanged since 1962. Most of our fines remain well below this legal maximum.

The Commission does have the option of reducing the cartel fine it would impose if the company in question is unable to pay. A reduction of this kind could only be granted if paying the fine would seriously endanger the economic viability of the company. While this situation might occur in the context of the crisis, the Commission would make an extremely careful assessment before granting any such reduction.

I BELIEVE THAT OUR FOCUS ON ELIMINATING CONSUMER HARM—RATHER THAN PROTECTING INEFFICIENT COMPETITORS—WILL STAND US IN GOOD STEAD IN THE CURRENT CRISIS.

I believe that our focus on eliminating consumer harm—rather than protecting inefficient competitors—will stand us in good stead in the current crisis. In times of economic recession, allowing consumers to make the best use of their buying power is essential. The recession cannot be an excuse for the burden of the downturn to be transferred, through cartels and abusive practices from companies which are doing badly, to consumers in general.

VII. Conclusions—Lessons from the Crisis

The best strategy to get out of the current crisis must include a robust and rigorous competition policy. However, the crisis naturally has and continues to have an effect on the way the Commission enforces competition policy. Governments and companies alike are faced with very real constraints as a result of the crisis, and the Commission has to make sure that it does not put procedural obstacles in the way of necessary and urgent rescue measures which aim to stabilize our economies. But, equally, we would be failing at our job, and failing the European consumers and the economy as a whole, if we did not ensure that these measures comply with competition principles. The route to recovery lies with competitive markets, not markets where inefficient and ailing companies are propped up by state support, illegal cartels, or abuses of market power, nor with markets where consumers pay to support structures which are not sustainable.

In order to ensure competitive markets, we also need competition-friendly regulation. We need to ensure that regulatory initiatives take account of competition principles, in the financial sector and in other sectors of the economy, as well as horizontal measures such as consumer protection initiatives that cut across many areas.

1 Articles 87, 88, and 89 of the EC Treaty. The scrutiny of State aid is a task that falls to the Commission's Directorate General for Competition (DG Competition)—although formal decisions on State aid are in principle the responsibility of the full College of Commissioners. At the height of the crisis, the Commission delegated powers to Mrs. Kroes, the Commissioner in charge of competition, to adopt decisions authorizing rescue aid under an emergency procedure.

2 Commission Communication, *A European Economic Recovery Plan*, COM (2008) 800 final, (November 26, 2008).

3 Based on the methodology applied for calculating customer benefits as set out in DG Competition's Annual Management Plan 2009, *available at* http://ec.europa.eu/competition/publications/annual_management_plan/amp_2009_en.pdf.

4 Commission Communication, *European Electronic Communications Regulation and Markets 2003—Report on the Implementation of the EU Electronic Communications Regulatory Package*. COM(2003) 715 final.

5 Commission Communication, *European Competitiveness Report 2008*, COM(2008) 774 final.

6 Study by U.K. Civil Aviation Authority on *The Effect of Liberalisation on Aviation Employment* (2004), available at http://www.caa.co.uk/docs/33/cap749.pdf.

7 H. L. Cole & L. E. Ohanianm, *New Deal Policies and the Persistence of the Great Depression: A General Equilibrium Analysis*, 112 J. Pol. Econ. 4, pp. 779-816, (2004).

8 Commission Communication, *The application of the State aid rules to measures taken in relation to financial institutions in the context of the current global financial crisis*, OJEC [2008] C 270/8 (the "Banking Communication").

9 Commission Communication, *Community guidelines on State aid for rescuing and restructuring firms in difficulty*, OJEC [2004] C 244/2; *Prolongation of Community guidelines on State aid for rescuing and restructuring firms in difficulty*, OJEC [2009] C 156/2.

10 Commission Communication, *The recapitalisation of financial institutions in the current financial crisis: limitation of aid to the minimum necessary and safeguards against undue distortions of competition*, OJEC [2009] C 10/2 (the "Recapitalisation Communication").

11 Commission Communication, *Treatment of Impaired Assets in the Community Banking Sector*, OJEC [2009] C 72 (the "Impaired Assets Communication").

12 Commission communication on the return to viability and the assessment of restructuring measures in the financial sector in the current crisis under the State aid OJEC [2009] C 195/9 (the "Restructuring Communication").

13 The European Commission publishes an overview of national measures adopted as a response to the financial/economic crisis, which is regularly updated http://ec.europa.eu/competition/sectors/financial_services/financial_crisis_news_en.html.

14 On September 15, the Commission announced that it was extending the temporary clearance of the measure because the investigation is not yet complete.

15 Case N 528/2008.

16 Charlie McCreevy, Towards an integrated approach to regulation across the EU, Speech/09/398, European Commissioner for Internal Market and Services (September 18, 2009).

17 Conclusions of the ECOFIN Council of 7 October 2008.

18 Community guidelines on state aid to promote risk capital investments in small and medium-sized enterprises OJEC [2006] C 194/2.

19 Communication of the Commission to Member States amending the communication pursuant to Article 93(1) of the EC Treaty applying Articles 92 and 93 of the Treaty to short-term export-credit insurance, OJEC[2005] C 325/22.

20 Case numbers N 661/2008 and N 668/2008 (the latter was amended on 5 June 2009 and 16 July 2009).

21 Case number N82b/2009.

22 Article 2 of the EC Merger Regulation (Council Regulation 139/2004 on the control of concentrations between undertakings, OJEC [2004] L 24/1) stipulates that "a concentration which would significantly impede effective competition in the common market or a substantial part of it, in particular as a result of the creation or strengthening of a dominant position, shall be declared incompatible with the common market."

23 Commission Notice on remedies acceptable under Council Regulation 139/2004 and under Commission Regulation 802/2004, OJEC [2008] C267/1.

24 Guidelines on the assessment of horizontal mergers under the Council Regulation on the control of concentrations between undertakings, OJEC [2004] C 31/5, ¶¶.88-91.

25 Article 21 of the EC Merger Regulation.

26 Commission Notice on Immunity from fines and reduction of fines in cartel cases, OJEC [2006] C298/17.

27 Commission Guidelines on the method of setting fines imposed pursuant to Article 23(2)(a) of Regulation No 1/2003, OJEC [2006] C 210/2.

28 *See* http://ec.europa.eu/competition/sectors/pharmaceuticals/inquiry/index.html.

29 The Third Energy Package consists of two directives and three regulations adopted by the European Parliament and the Council on July 13, 2009; *see* http://ec.europa.eu/energy/gas_electricity/third_legislative_package_en.htm.

30 Communication from the Commission—Guidance on the Commission's enforcement priorities in applying Article 82 of the EC Treaty to abusive exclusionary conduct by dominant undertakings OJEC [2009] C 45/7.

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Competition Policy, Bailouts, and the Economic Crisis

Bruce Lyons

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Competition Policy, Bailouts, and the Economic Crisis

*Bruce Lyons**

The aims of this paper[1] are twofold. First, I explain the economics of bank bailouts as distinct from bailouts for other sectors of the economy. Why do all the rules of good competition policy appear to fly out of the window when the banks get into trouble? Does this mean that we should abandon the rules equally for car manufacturers and other industries in trouble? I argue that a unique combination of two characteristics made it essential to bailout or nationalize the banks in the current crisis. No other sector of the economy can claim the same justification. Second, I review the threat of a retreat to politically-determined industrial policy and the need for vigilant implementation of economic effects-based competition policy.

*Bruce Lyons is Professor of Economics in the School of Economics at UEA and Deputy Director of the ESRC Centre for Competition Policy, University of East Anglia.

I. The Credit Crunch

The current global economic crisis had its roots in slack economic policy and huge strategic errors by the banks.[2] Permitted by weak regulation and driven by biased incentives, the banks borrowed (and lent) far too much given their low capital bases, and were caught out when the housing price bubble began to burst, heralding large-scale defaults. The global reach of this behavior was compounded by the sale and purchase of opaque mortgage-backed securities and their derivatives between financial institutions. The banks' recklessness was facilitated by weak corporate governance, ineffective regulation, permissive monetary policy, and massive international flows of funds.[3] Like unlimited supply of food in the animal kingdom, huge flows of funds into western banks suppressed the power of competition to select only the fittest to survive. Similarly, rapid recession, like periods of limited food, soon picks off the unfit and, if the drought is severe, many of the fit as well.

THE BANKS' RECKLESSNESS WAS FACILITATED BY WEAK CORPORATE GOVERNANCE, INEFFECTIVE REGULATION, PERMISSIVE MONETARY POLICY, AND MASSIVE INTERNATIONAL FLOWS OF FUNDS.

There have been two enormous market consequences of these events and a third may be round the corner. The first was that many of the world's most renowned banks have been pushed close to bankruptcy. For some, this was the direct result of their own recklessness, but others have been sucked into the whirlpool. Governments across the world have stepped in to bail them out by guaranteeing loans, injecting capital, underwriting toxic assets, and acquiring their shares. Such has been this commitment that only one bank of major significance has so far gone bust (Lehman Bros). This "success" has been achieved only at huge cost to current and future taxpayers.

The second consequence was contagion into the non-financial sectors of the economy. The banks cut lending in every way they could in order to rebuild their reserves.[4] This created severe financial constraints for their business and private customers, resulting in investment cuts, reduced demand, and a powerful negative multiplier across the global economy. Beyond financial constraints, consumer and investor confidence were shattered creating a further squeeze on demand. Fearing a Japanese style "lost decade" of deflation and stagnation, governments and monetary authorities have been trying to reverse this by slashing interest rates, buying securities, increasing public spending, and temporarily reducing taxes. Much of this may have been necessary as an emergency measure, even though the haste, panic, and haggling with which such packages were put together suggests many initiatives will have been substantially wasteful.

The third potential consequence could be an interventionist industrial policy in the wider economy and the emasculation of competition policy. Currently, this has happened only to a minor extent, but aspects of rescue packages promoted by governments across the globe point to the danger. In the last decade, com-

petition policy has been reinvented across Europe[5] and introduced in many emerging economies with vigor and new focus on economic foundations. This has been a huge success in protecting consumer-responsive markets and efficient business practices. The discipline of competition policy has also allowed the reduction of inefficient forms of regulation and public ownership. While modern competition policy is economically robust, it remains politically fragile and thus vulnerable to crude, populist, deeply-flawed claims that it is an unnecessary luxury in times of recession—or even that the crisis itself is due to "too much competition." A more considered analysis shows this to be untrue, but it is all too easy to see why the mistaken view might take root.

The aims of this paper are twofold. First, I explain the economics of bank bailouts and why they are different from bailouts for other sectors of the economy. Second, I review the threat of a retreat to politically-determined industrial policy and opportunities for the implementation of an active competition policy. Section 2 highlights a unique combination of two characteristics that made it essential to bailout or nationalize banks in the recent crisis. In section 3, I assess the dangers of bailing out failing firms in sectors that do not exhibit both these characteristics. The recent trend in interventions and the positive role of competition policy during the recession are reviewed in section 4. Section 5 presents a brief conclusion.

II. Bailouts, Nationalization, and Regulation for Banks[6]

A. CAUSES

After years of lecturing and lobbying from the West, China adopted its new Anti-Monopoly Law only last year (2008). China may, therefore, be puzzled to see so much government intervention in banks in recent months, including: massive individualized subsidies, direct "interference" in business decisions, politicians promoting mergers, and nationalization.[7] Why do all the rules of good competition policy appear to fly out of the window when the banks get into trouble? Does this mean that we should abandon the rules equally for car manufacturers and other firms or industries in trouble? I address the first question in the remainder of this section and the second in section 3.

WHY DO ALL THE RULES OF GOOD COMPETITION POLICY APPEAR TO FLY OUT OF THE WINDOW WHEN THE BANKS GET INTO TROUBLE?

All markets have their own idiosyncrasies but each works fundamentally in the same way. Only rarely are the idiosyncrasies so substantial that they warrant special treatment. It is an unfortunate truth that banking is different to other industries due to a unique combination of two essential characteristics that cre-

ate the potential for systemic economic collapse: contagion within the banking sector and contagion from banks to the entire real economy. Before getting to these twin contagions, note the importance of confidence and potential for panic in banking.

A bank can only survive if everyone is pretty certain that it will survive. It cannot survive a loss of confidence.[8] Banks necessarily borrow short (i.e. customers can withdraw their money at short notice) and lend long, which means they must rely on funder confidence to keep funds flowing in to support their loan book. Banks lend a multiple of what has been deposited and can do this in normal times because most people leave much of their money in the bank. However, in the absence of full guarantees, individual savers have a great deal to lose if a bank goes bust and very little to gain by keeping their money in a particular bank. Even a rumor of potential failure can result in massive withdrawals and, in the absence of intervention, failure is a self-fulfilling prophesy. This can happen even if a bank's loan book is sound because the bank will not have the liquidity to pay all depositors their money.

The problem moves from liquidity crisis to a much more serious insolvency crisis when loans go bad and the bank has insufficient capital to absorb losses. Depositors could not be paid out even if all the good loans could be called in. The loss of confidence cannot then be soothed. The queues outside Northern Rock in the United Kingdom in September 2007 were an early sign of the fragility of the banking system even when most retail depositors were covered by an explicit government guarantee. Wholesale funds from other banks and international lenders were quantitatively much more important and unguaranteed, and it was these that hemorrhaged from Northern Rock to bring it down. Few other products are so sensitive to confidence.[9] Nevertheless, banks would not warrant special treatment if this was the end of the story because creditors could simply move their deposits to a rival bank which could consequently increase its loans.

The first truly distinctive characteristic of banking from the competition perspective is that the balance sheet of banks are so interconnected that the collapse of a large bank is contagious and contaminates the whole banking system. To a small extent this is because funders (from small retail depositors to international wholesale funds) wonder which will be the next troubled bank from which they should withdraw their funds. But if the crisis was merely one of confidence, that worry could easily be addressed by the central bank providing liquidity to a bank subject to a run. For relatively small bank failures, when banks have adequate capital and when specific risks and reasons for failure are understood, the banking system is typically quite stable.[10]

Banks in highly developed economies do not fail due to liquidity problems alone, but they are interconnected in more significant ways. Banks lend to each other so if one is unable to repay its debts, that failure creates bad debts the lending bank which, in turn, undermines its solvency (counterparty risk). Before the

current crisis, most banks had shared a similar belief about continuing asset price rises and they did not diversify the associated risk sufficiently outside the banking system. Instead, they exchanged ever more complex and opaque collateralized debt obligations, most importantly those based on mortgages. The risks stayed within the system.

In August 2007, the banks apparently suddenly noticed the rising mortgage repayment delinquencies and foreclosures as house price inflation tumbled. They stopped lending to each other, justifiably concerned that they could not calculate the risks in their own balance sheets, let alone those of counterparties. The self-inflicted wounds of inadequate capital, bad loans notably in U.S. subprime mortgages, and foolish trading in derivatives spread the damage and destroyed the already limited capital of many banks and related financial institutions.[11] Like firms in all industries, banks go bust when their capital is exhausted by bad trading but, because of the interconnectivity between banks, bad loans and bad assets quickly spread through the global banking system.

LIKE FIRMS IN ALL INDUSTRIES, BANKS GO BUST WHEN THEIR CAPITAL IS EXHAUSTED BY BAD TRADING BUT, BECAUSE OF THE INTERCONNECTIVITY BETWEEN BANKS, BAD LOANS AND BAD ASSETS QUICKLY SPREAD THROUGH THE GLOBAL BANKING SYSTEM.

The banking crisis lurched towards potential catastrophe a year to the day after those Northern Rock queues on U.K. high streets, when the major U.S. bank Lehman Bros was allowed to collapse and the global financial system nearly followed. In simple economic terms, this first distinctive characteristic is that a large bank with substantial trading activities has a negative externality on its rivals—if it collapses, the stability of its rivals is undermined.[12] This is in sharp contrast to, say, a grocer or a car manufacturer where others in the industry can usually benefit from the collapse of a rival.[13]

The second distinctive characteristic is that bank finance provides the essential oil in the entire economic system, allowing firms to make investments and to absorb the bumps of fluctuating revenues and payments. In normal times, banks lend to each other for exactly the same reason. Additionally, traditional investment banking puts together funding for bigger projects. Without this oil provided by the banks, the economy seizes up. The product of no other industry is as essential to every other market in the system. Banks are particularly important for smaller firms which do not have the scale to issue corporate bonds and do not have access to the internal capital markets of large business groups.[14] They are also important for financing large purchases by consumers (e.g. housing, cars). Unfortunately, during a banking crisis, the first reaction of a bank is to stop making loans in order to compensate for its loss of deposits and asset write-offs. If the banks thus fail to fulfill their crucial lending function, this leads to a fall in demand and macroeconomic recession. This is the second dimension of contagion.[15]

Thus, the deposit-side of banks is vulnerable to contagion in the withdrawal of funds and especially asset write-downs, and the consequent loan-side collapse contaminates the whole economy as banks try to rebuild their balance sheets. These two characteristics combined into a compelling argument for treating the banks as a special case in the current crisis. The prospect of contagious bank failures justifies intervention both to provide them with liquidity and to keep them solvent. However bitter the taste to taxpayers, this applies even when the banks' plight is their own fault.

This double contagion is unique. A food product may be vulnerable to a health scare and a contagious loss of confidence for that particular product, but this would not result in global recession if it was taken off the supermarket shelves.[16] Electricity may be required for the production of practically every other product in the economy, but it does not suffer from within-sector contagion—electricity supply did not collapse with Enron and would be little affected by the bankruptcy of a major supplier. Only the banking system combines both of these characteristics to create the potential for genuinely systemic contagion. A detonator alone makes only a small bang, and TNT alone is a relatively stable material, but put the two together and you have a truly dangerous bomb. As it is, the banking crisis detonated a huge bomb under the global economy. The collapse of another major bank could have been nuclear. There was no sensible alternative but to bail out or nationalize failing major banks.[17]

AS IT IS, THE BANKING CRISIS DETONATED A HUGE BOMB UNDER THE GLOBAL ECONOMY.

There is one more twist to the story. This specialness of banks has been a substantial cause of the crisis. The major banks are now sure of what they already thought they knew: they will always be bailed out. The shock of the Lehman collapse was the exception that only served to prove the government guarantee. The consequences of collapse were seen to be so awful that governments have bailed out the banks ever since.

The anticipation of this bailout had created a moral hazard that biased decisions towards risk taking. The upside for banks was huge potential profits and the downside was a bailout. This asymmetry was reflected in the bonus structure for executives and the traders they employed. The reward for short-term trading success was huge, while there was no equivalent sacrifice for having made losses and no claw-back for short-term profitable trades that turn sour. This system allowed banks to share the same bullish beliefs in asset prices without diversifying the risk outside the banking system. It also encouraged heavy duty lobbying to reduce the effectiveness of regulation. Some banks did remain prudent, but others competed on upside alone.[18]

B. SOLUTIONS

Having identified some of the problems, what should have been done to solve them? In the short term, the urgency should have been to get banks lending

again and so to limit the contagion of the banking crisis to the rest of the economy. Most governments tried to do this indirectly by recapitalizing the banks, often in return for some form of preferred stock (i.e. something between a standard loan and common equity). This allowed them to say that a bank was not being nationalized even when the taxpayer became the majority stock holder and took a high risk of not being repaid.

Governments have also provided credit insurance and toxic asset underwriting (ex post i.e. after the assets had turned toxic!) and central banks have purchased large quantities of bonds from the banks.[19] While this bailout has saved many banks from collapse, it did not get them lending again on a sufficient scale and urgency. These banks have instead used this funding to rebuild their own capital while they operated in the shadow of collapse. This created "zombie banks" which drain funds while failing to fulfill their raison d'être.

THIS CREATED "ZOMBIE BANKS" WHICH DRAIN FUNDS WHILE FAILING TO FULFILL THEIR RAISON D'ÊTRE.

Government loan guarantees have also failed to stimulate lending on a significant scale. Unfortunately, against this limited success, the bailouts will further reinforce the asymmetry in risk-taking by banks once more normal times return. Meanwhile, the banks' self-preservation measures made the recession bite harder, thus "justifying" their failure to lend to businesses by claiming that those businesses have become too risky.

There would have been less contagion into the real economy if a form of temporary nationalization (beyond passive ownership of preference shares) had been adopted early in the crisis. The idea would be for those banks which were nationalized to be run by trustees and concentrate on traditional lending based on investment and repayment prospects. It would draw on the traditional skills and expertise of bankers in assessing loans and creditworthiness, but importantly should not undercut terms provided by private lenders in normal times. Their loans would be made on full commercial terms and such banks would be privatized as soon as economic conditions permitted.

Had such nationalization been adopted in late 2008, this would have limited the contagion into the real sector.[20] Competition authorities could have been instructed to monitor that each nationalized bank was indeed operating on genuine commercial terms both in attracting funds and in lending activities.[21]

There are major problems with such a strategy both in the process of nationalization and in the State running a commercial bank. Nationalization of a bank that would be bankrupt in the absence of government help would, quite fairly and efficiently, wipe out the common shareholders and reduce the payout for junior creditors. It would probably also cause shareholders and subordinated creditors of some other banks to flee in anticipation of nationalization. This means that several major but weak banks would have to be nationalized simulta-

neously. This would undoubtedly be politically uncomfortable. Since pension and insurance funds also invest in banks, the spillover could be far-reaching and the state may have to absorb some of the creditor losses to keep otherwise well-managed insurance companies afloat. However, there is no reason to provide such insurance to shareholders in general.

It has to be acknowledged that the aim of a nationalized bank to make loans only on commercial terms has limited credibility because politicians are genetically prone to fiddling with any high profile asset they own. This certainly happens under long-term state ownership but not necessarily over the short term. This problem must be balanced against the prospect of "standard" bailouts creating zombie banks that are not lending and so causing a protracted recession.[22] As soon as the economy recovers and an appropriate regulatory regime has been established, these banks should be privatized, though in a restructured form to minimize the risk of future contagious bank failures.

IT HAS TO BE ACKNOWLEDGED THAT THE AIM OF A NATIONALIZED BANK TO MAKE LOANS ONLY ON COMMERCIAL TERMS HAS LIMITED CREDIBILITY BECAUSE POLITICIANS ARE GENETICALLY PRONE TO FIDDLING WITH ANY HIGH PROFILE ASSET THEY OWN.

These rapidly privatized banks should probably be much smaller than the ones that failed, and so less prone to causing future systemic collapse. This would help to balance the sharp increase in bank concentration that has been a consequence of the crisis. For example, in the United States, we have seen the consolidation of: Bank of America, Countrywide, and Merrill Lynch; JP Morgan, Washington Mutual and Bear Stearns; Wells Fargo and Wachovia. In the United Kingdom: Lloyds TSB and HBoS; Santander and Bradford & Bingley; Nationwide and Dunfermline;[23] while Northern Rock has been the only conventional nationalization. Internationally, Lehman assets were picked up by Barclays (United Kingdom and United States) and Nomura (Asia). No one can seriously claim that this change in banking market structure has been due to the natural market forces that should rightly shape an efficient market structure.

In the medium term, major revisions of bank regulation are necessary so that banks can compete as private firms with balanced incentives. Financial markets are not unique in having special features that require a specific regulatory framework to align competition and welfare. For example, some industries (e.g. infrastructure networks distributing electricity, water, or rail services) are subject to such strong economies of scale that they are natural monopolies and so require a specialist regulator to control maximum prices; but banks do not have such strong scale or network economies to make them anywhere near natural monopolies.

A more relevant example is pharmaceuticals, for which there are powerful health and safety reasons to regulate new drugs. In late 1950s Europe, this regulation was entirely insufficient, with the result that thalidomide was prescribed

to pregnant women. The resultant tragedy brought about a new and necessary regulatory approval regime, subject to which pharmaceutical companies can compete with each other.[24] It is essential that the current crisis should similarly bring about more effective and appropriate financial regulation while still encouraging beneficial competition and innovation.

An international regulatory system already existed pre-crisis with a view to setting minimum standards for banks and so to channeling competition into appropriate behavior. This took the form of the agreement known as Basel II, which has three "pillars:" minimum capital requirements, regulatory supervision, and risk disclosure to facilitate market discipline.[25] Clearly, the application of this framework has proved inadequate in the face of complex financial innovations and distorted incentives.

The following elements of regulation are additional to a necessary review of the standard components of Basel II.[26] First, incentives given to individuals within banks must not be one-sided (i.e. paying bonuses for short-term profit with no downside for long-term losses). Recent European debate has been side-tracked into crude proposals to limit the scale of bonuses, whereas it is their incentive effect that is crucial.

Recent European debate has been side-tracked into crude proposals to limit the scale of bonuses, whereas it is their incentive effect that is crucial.

Second, while credit default swaps and other elements of diversification and insurance must be allowed as prudent trading activities, they should not be traded by banks multiple times as bets on future prices or defaults.[27] Liquid markets also need to be created to get genuine prices for all supposedly safe assets.

Third, banks should be charged ex ante (i.e. before they get into a mess) for the explicit (and implicit) guarantees they receive from government, and the size of these charges should reflect the risk profile chosen by each particular bank, including the amount of debt financing relative to its equity base.[28]

Fourth, idiosyncratic assets, collateralized debt obligations ("CDOs"), and other complex or opaque financial innovations might be required to pass regulatory scrutiny and receive positive approval from a regulatory body, and not from a credit rating agency which is beholden to issuers for fees and supplementary services.[29] Credit ratings could be privatized at a later date once an appropriate regulatory regime is established.

Finally, and arguably most important, a credible bankruptcy regime must be established for banks so that contagion is contained. This is likely to require preemptive action by a monitoring central bank (and not the daily regulator which may be reluctant to admit that it has failed to keep the bank on track).

In conclusion, the banking system combines the two explosive characteristics of contagious failures and universal need by every other business. This combination means that major banks cannot be allowed to fail. The risk this entails and the recklessness it encourages mean that tough prudential regulation is essential. This is all the more important because recent bailouts only reinforce the moral hazard.

However, it is important to regulate appropriately so as not to stifle competition and innovation. This requires targeting regulation clearly at the problems (e.g. externalities, distorted incentives) and not a knee-jerk political response against the wrong target (e.g. competition, securities to diversify risks). With appropriate regulation and the standard tools of competition policy in place, competition among private banks can be left to work to the benefit of efficient businesses and consumers. The appropriate regulatory framework is necessary to align competition and welfare, bringing sustainably low prices for banking services and safe, innovative product development.

Finally, there is no reason why a government should not use their 'bailout' stakes in banks to restructure them into less contagion-prone (probably smaller) institutions. In Europe, the Commission is likely to use its state aid powers to require some degree of restructuring, but it remains to be seen whether this will be designed as an ad hoc punishment or a genuine attempt to redress properly identified problems.

III. Competition Versus Bailouts for the Rest of the Economy

The banking crisis stifled lending and the consequent credit crunch triggered a global recession. Minor banking crises do not always bite on the real economy, but history tells us that when a banking crisis does bite, it bites the economy's leg off. We are very much in the latter category today. A comprehensive IMF study of all systemically important banking crises for the period 1970 to 2007 covering 42 crises in 37 countries shows the average fiscal costs of crisis management to be 13 percent GDP, though they can be as high as 55 percent.[30] The consequent recessions are even more damaging with average cumulative losses equivalent to 20 percent GDP in the first four years, but ranging from 0-98percent GDP.[31]

It is from this perspective that we must view the massive fiscal stimuli that many governments put in place as an attempt to limit the decline and shorten the period of stagnation. The size of required fiscal stimulus could have been much less if bank finance was working properly. Even on an optimistic scenario, however, there will be a deep and protracted recession that is seeing numerous firms fighting for their survival. In these circumstances, should we abandon competition policy, particularly as it relates to state aid? I consider only aid to specif-

ic firms or industries, and not general fiscal or employment measures that are reasonably neutral in their impact on competition.[32]

Competitive markets certainly work to the benefit of consumers and efficient firms when financial markets are oiling them well. In good times, firms expand and enter new markets as they seek to attract customers and spending away from rivals. Profits are made by those who have invested well, produce efficiently, and make the most attractive product offers (i.e. those who provide what consumers want at a better price than offered by rivals). In bad times, firms contract and leave the market as they adjust to reduced customer spending. Losses are made by those who fail to provide what their customers want or who set prices that are too high (i.e. those who make unattractive offers).

Firms with the least attractive products or highest costs exit the market. Exit is as fundamental as entry in making markets work well. It is part of natural selection, leaving room for efficient firms to expand and new firms to enter. The same essential story applies to shops, restaurants, steel and cars. The role of competition policy is to ensure that firms do not conspire to evade this harsh but socially productive competitive discipline by fixing prices, excluding efficient rivals, merging with significant competitors, or receiving discriminatory state subsidies or protection.

In the absence of the special features discussed in section 2, subsidies undermine market outcomes and processes.[33] The problem most familiar to the European debate on State aid is that subsidies create international distortions to competition. Inefficient firms receiving subsidies take market share from more efficient foreign suppliers. This can result in retaliation and a mutually destructive subsidy war funded by taxpayers.

THE PROBLEM MOST FAMILIAR TO THE EUROPEAN DEBATE ON STATE AID IS THAT SUBSIDIES CREATE INTERNATIONAL DISTORTIONS TO COMPETITION.

However, the problems are not only international. Subsidies undermine the market mechanism because the prospect of a bailout leads to reckless behavior, as is so vividly illustrated by the banks. It also leads to "rent seeking" as the most successful CEOs become those who can best work the political system for subsidies, and not those who efficiently produce the best and most innovative products. There is abundant evidence of the failure of politicians or civil servants to pick winners. More insidiously, there is also a negative effect on efficient firms and entrants who are incentivized to hold back on investment and aggressive marketing if they know that inefficient rivals will hang on to segments of the market with inappropriate product offers and bloated capacity without fear of the consequences.

In structurally competitive industries (i.e. in the absence of sunk costs, state subsidies, or entry barriers), entry into and exit from a market can rapidly adjust to demand changes. Firms respond to expected prices relative to average costs to

trigger entry and exit. Incentives change in the presence of sunk (i.e. non-recoverable) costs; for example, not only will they want to stay in the market as long as variable (non-sunk) costs are covered, but they may want to hang on even if price falls below these costs as long as there is a prospect of the market recovering.[34] Thus, firms will be more cautious to enter and slower to exit. This provides a natural balance for such markets with less entry when demand is high and less exit in recession. Profits in good times balance losses in bad times and properly working financial markets will appreciate this and provide the necessary financial buffer.

Both economic theory and most of the empirical evidence suggest that an unhindered exit process is at least reasonably efficient.[35] The research shows that in the absence of intervention the market selects the best adapted firms to survive. The least efficient plants exit first, including those too small to achieve available economies of scale. If firms are equally efficient, then the largest downsize first. Once these adjustments have been made, if demand is insufficient relative to economies of scale and the toughness of competition, there may be a period of attrition with prices below cost until one of the remaining firms exits. This is a painful process for all in the industry and the transaction costs are substantial but it has the desirable attribute of leaving a sustainably efficient and competitive market structure.[36]

BOTH ECONOMIC THEORY AND MOST OF THE EMPIRICAL EVIDENCE SUGGEST THAT AN UNHINDERED EXIT PROCESS IS AT LEAST REASONABLY EFFICIENT.

How do things change when financial markets fail to provide lubrication and instead throw grit into the economic system? Problems can be caused at two levels. First, banks and other providers of finance play a vital role in appraising investment projects and the long-term viability of firms. It is possible that arbitrary financial constraints due to the banking crisis might force the exit of a firm that serves consumers better than a rival; yet the inefficient rival might survive because it happens to have a stronger line of credit.[37] Second, financial constraints on customers may depress demand for a whole sector if purchases are widely funded by borrowing (e.g. construction, cars, machinery), which might result in the scrapping of skills and assets that would be productive once the credit crunch clears.

These possibilities only serve to emphasize the need to get banks lending. As argued earlier, recapitalizations and loan guarantees have proved expensive yet insufficient to indirectly get the banks to lend. It would have been better had the governments taken active control of those banks they are subsidizing. These banks could have been run by independent trustees for the duration of the recession and with a policy of lending on "market investor" unsubsidized terms.[38] The idea is to correct the cause of the problem, the credit crunch, and to avoid giving politically determined subsidies to specific firms or industries. The resultant

loan book would then be attractive when the bank is privatized as soon as the market conditions allow.

There are two highly unattractive alternatives. Either no intervention, so competition is distorted and firms reliant on bank funding are affected asymmetrically, or finance determined by the "Department for Industry" where firms will be helped according to political impact and not according to previous reliance on bank funding.[39] The key lending skills lie within the banking sector whereas government departments find an easy route through grand gestures to big firms and big industries (even if the recipients were in long-term decline pre credit crunch).

With appropriate measures to get banks lending, are some "real sector" firms still "too big to fail" in a recession? "Too big" may be interpreted in several ways. The firm might be a monopoly provider, a large direct employer, or a firm supporting a large supply chain or distribution network. For a monopoly provider such as the owner of a rail network or a vital tunnel, the asset does not disappear if the owner gets in financial difficulty. If the assets have any positive value they can be bought out of administration and operated under new ownership. If the firm is not a monopoly but a large employer, then its viable assets could also be bought out of administration. It is inefficient to subsidize current shareholders and it would be harmful if it received preferential treatment over an efficient rival. The same argument applies to a long supply chain in, for example, the car industry.

WITH APPROPRIATE MEASURES TO GET BANKS LENDING, ARE SOME "REAL SECTOR" FIRMS STILL "TOO BIG TO FAIL" IN A RECESSION?

More subtly, it is possible that an efficient and an inefficient manufacturer may share key suppliers who benefit from economies of scale. The loss of a major customer may put such suppliers at risk and so potentially harm the efficient manufacturer's supply chain. However, an efficient supplier can respond by expanding into the market opportunities created by the exit of the inefficient firm and scaling down.[40] This is the way markets work to select efficient producers and subsidies interfere with this process. Subsidies to support a whole industry may appear less distortionary, but they inevitably divert demand and resources away from substitute products and so shift the pain. No other sector of the economy shares the pair of characteristics that set banks apart for state intervention in the current crisis.[41]

There is no doubt that restructuring is painful. However, the pain is less than the harm caused by industrial subsidies, as experienced by: efficient rivals who suffer reduced market share; customers who are offered costly and unattractive products; taxpayers whose real income falls; or the elderly, the sick, and school children who suffer from diverted public spending. It is important that those thrown out of work should receive strong support both financially and in retraining, but it is they who should receive the subsidies and not the shareholders and senior executives of failing firms. It is the latter who benefit most from bailouts.

IV. The Positive Role for Competition Policy During the Recession

Most of the analysis so far has related to State aid because this is the competition policy front line in a recession. History provides some worrying lessons also for other dimensions of competition policy. Anticompetitive agreements and mergers cause long-term harm which gets discounted heavily in a crisis. In international trade policy, there is a well known and strong correlation between recession and protection, with causation going both ways and feeding a negative spiral.[42] Effective enforcement of national competition policy in most of the world is relatively recent, so has yet to be challenged by recession. However, the United States has had the Sherman Act since 1890 and the last 120 years have seen numerous wars and slumps. Both types of crisis have dampened enforcement of the Act and the consequences have been particularly bad during recessions. Business cooperation can be bought (superficially cheaply) by politicians: "Antitrust laxity is often the government's first bargaining chip when it urgently needs something from industry."[43]

BUSINESS COOPERATION CAN BE BOUGHT (SUPERFICIALLY CHEAPLY) BY POLITICIANS: "ANTITRUST LAXITY IS OFTEN THE GOVERNMENT'S FIRST BARGAINING CHIP WHEN IT URGENTLY NEEDS SOMETHING FROM INDUSTRY."

Much has been made of the similarities between the current crisis and the Great Depression, especially the fiscal role of the New Deal. A closer look, however, does not settle one's nerves.[44] Franklin D. Roosevelt was persuaded by industrialists that it was necessary to suppress the enforcement of competition policy to gain cooperation and he agreed to this as an integral part of the deal. In twelve months from June 1935, the Interior Department received identical bids from steel firms on 257 different occasions, and these bids were 50 percent higher than foreign steel prices. It has been estimated that wholesale prices in 1935 were 24 percent higher than they should have been and even by 1939 they remained 14 percent higher. Cartel prices fed through to unrealistic wages and Cole & Ohanian estimate that unemployment was 25 percent higher than it would have been otherwise. They suggest that the depression may have lasted seven years longer than necessary.[45]

Fortunately, international institutions facilitating political and economic dialogue are now well established and genuinely global (e.g. WTO, G20), as has been made necessary by global economic integration. This has undoubtedly helped with the initial responses and rhetoric of policy intervention. However, there are dangerous signs. In the United Kingdom in October 2008, the Office of Fair Trading ("OFT") recommended that the Competition Commission should investigate the proposed merger of Lloyds-TSB and HBOS, but this advice was overridden by the Secretary of State.[46] This was the first case of such an intervention since the reforming Enterprise Act of 2002 was meant to take

mergers out of political decision making.[47] The merger has turned out to be a financial disaster and the interventions discussed in section 2 would undoubtedly have been better. As it stands, the United Kingdom (like the United States) now has a more concentrated banking structure which will be even more vulnerable to systemic failure unless prudential regulation is very much improved.

In spring 2009, politicians across the globe were thinking loudly about subsidizing specific firms, particularly in the car industry. The U.S. administration offered major subsidies to General Motors and Chrysler, though in the end not enough to stop them filing for bankruptcy protection. In France, President Sarkozy offered 6 billion EUROs in government support for Renault and Peugeot-Citroen subject to two conditions—no factories located in France would be closed and reassurance regarding jobs in France—before the European Commission intervened. Italy and Spain also produced major car subsidy plans. Intervention then switched to apparently more neutral car scrappage schemes to stimulate demand (though this is still biased towards the car sector and is a costly way of bringing forward the purchase of a durable good at the expense of lower demand next year).

More widely, the traditional instruments of trade protection are also visible. For example, tariffs were raised in India on some steel products, in Russia on cars, and in Ecuador on 940 different products. The EC re-introduced subsidies for the export of milk and milk products. Most of these at least work within WTO rules (e.g. raising tariffs within legal limits) but it remains likely that anti-dumping duties will return as a battleground: 2008 saw a 28 percent rise in applications over the previous year, the first rise since 2001.

National procurement has also been tied to fiscal policy. The February 2009 $800 billion fiscal stimulus bill of the new Obama administration included "Buy American" clauses (e.g. for steel to be used in state projects), though the original plan was modified in the face of potential retaliatory action by the EU. Paul Krugman has argued that, in the absence of an internationally coordinated fiscal stimulus, these clauses may not be protectionist in that they need not reduce trade below the viable alternative. As he puts it: "My fiscal stimulus helps your economy by increasing your exports — but you don't share in my addition to government debt."[48] He continues that if all countries were adopting a similar fiscal stance, "Buy American" would be unnecessary, but as they are not, it might be a second best way to get the economy moving. This is a coherent argument but it is politically impossible to limit the procurement bias to the appropriate level. The danger is that a sequence of "special cases" will result in a flood (which is why it is important to understand precisely why the precedent of the banks is so inappropriate for other sectors).

It is difficult to prevent discriminatory interventions even within the EU. Article 87 of the European Treaty prohibits state aid that may distort trade between Member States but permits non-distortionary forms of aid. For example,

the Commission requires that aid to banks should be: non-discriminatory, priced according to market investor principles,[49] and subject to behavioral restraints against aggressive growth at the expense of non-subsidized banks.[50] The last needs interpreting carefully in the context of banks failing to make sufficient loans (see section 2).

The EC has also invoked Article 87(3)(b) of the EC Treaty, which permits further, but strictly limited, aid intended to remedy a serious disturbance in the economy of a Member State. In December 2008, it adopted a "temporary framework" to allow Member States to tackle the effects of the credit crunch on the real economy in a minimally distortive way.[51] One aim was to restrict aid only to firms in difficulty due to the financial crisis and not allow aid for firms in long-term decline.[52] The EC rules are aimed at keeping the playing field level internationally within Europe. They are imperfectly adhered to, but they provide a helpful model for national rules in the current crisis.[53]

As a supranational organization, the EU is as tight and powerful as international cooperation comes, and it is backed by the legal force of a strong treaty, yet it still has difficulty keeping its members in line.

As a supranational organization, the EU is as tight and powerful as international cooperation comes, and it is backed by the legal force of a strong treaty, yet it still has difficulty keeping its members in line. The global institution charged with reducing impediments to international trade, the World Trade Organization ("WTO"), has far less control over its membership and has a very limited mandate.[54] Nevertheless, it can have a significant reporting role for changes in national trade policies, it can host talks to resolve disputes, and it can speak especially for those developing countries that have little retaliatory power in negotiations.[55] The lack of powers over sovereign states means that if diplomacy fails, the only credible bargaining chip is retaliatory tariffs or subsidies. Of course, actual trade wars are mutually destructive and the aim is that governments will realize this so the threat does not have to be implemented. By late summer 2009, it seems that, following the initial panic, the political urge for protectionist measures has moderated.

Pressure on the mainstream implementation of antitrust and merger policy comes both from short-term crisis management and, more insidiously, in an urban myth that "too much competition" may have contributed to the crisis. In earlier sections, I have given examples of crisis mergers and also rehearsed the long-term benefits of a competitive economy. As the discussion of bank regulation should make clear, the latter does not mean completely laissez-faire capitalism, but regulation targeted at allowing competition to thrive without creating negative externalities. Recession will create challenges for competition policy in each of the traditional areas:[56]

- Agreements between firms: "Crisis cartels" are liable to form when prices drop, and such coordination becomes addictive.[57] Seductive excuses may emerge along the lines of fixing prices in order to protect the number of post-recession suppliers. However, such cartels are more likely to delay recovery and fossilize an inefficient market structure. Firms in an industry may also try to get together to agree an "ordered reduction in capacity." Such cartels have occasionally been allowed in Europe under Article 81(3) in the past, but this would be misguided as collusion is unlikely to select the most efficient market structure (see section 3).[58] A potential problem relates to fines for prosecuted cartels because an otherwise appropriate fine might push a cartel member into bankruptcy during a financial crisis. If fines are not adjusted down, this may result in fewer firms in the market. However, this possibility should not be overplayed. Fines are generally set at a level that is insufficient for optimal deterrence because they often do not even cover the profits generated by cartels, let alone take proper account of the low probability of detection.[59] Cartels also allow inefficient firms to survive in the market. Overall, it is likely to be undesirably anticompetitive to adjust fines down in times of recession.

- Abuse by a single firm: There is a potential danger of a financially strong firm taking the opportunity to foreclose a smaller or more financially constrained rival. Recession, especially one induced by financial crisis, may prove fertile ground for unfair means to tip a rival over the edge. Competition authorities must be alert to such foreclosure though they should not simply protect inefficient rivals. Low demand growth may also facilitate entry deterrence strategies. For some foreclosure problems, the appropriate remedy may be to require access to a key facility or technology. The terms of such access are then crucial to making the remedy effective. Should the current recession result in deflation, that could create problems for previously agreed access terms.

- Mergers: The failing firm defense has been applied, at least implicitly, for bank takeovers in the last year, though there have been fewer such merger justifications in the real sector. If a firm is clearly going bankrupt, and if a particular merger is the least anticompetitive way to ensure the survival of efficient resources in the industry, then such mergers should be allowed.[60] But this is simply a statement of sensible policy in any circumstances and there is nothing special about the current recession in this respect.[61] It is only the frequency of this argument that may test the authorities. For mergers that do not involve a failing firm, divestiture remedies may be made more difficult if appropriate buyers cannot be found due to financial constraints. Should this arise, the fallback option has to be full prohibition at least pending the emergence of viable buyers.[62] There may also be a rise in opportunistic merger proposals with little economic logic but motivated by differential access to finance and an anticipated rise in the stock market. There is no particular reason why such mergers should raise com-

petition concerns. Finally, the economic justification of declining demand and low margins may be offered to justify the need for a more concentrated "equilibrium" market structure brought about by a horizontal merger, but merger control should focus on expected demand and not be transfixed by the last twelve months in appraising any competition concerns.

V. Conclusion

History suggests that competition policy will be increasingly under threat as the recession bites. Businesses under pressure will draw a plausible, though inappropriate, analogy between their own industry and banking bailouts. Those already in trouble before the crisis will grab at the opportunity to plead their case. Politicians seeking short-run popularity will think it is little sacrifice to cast aside the long-term benefits of competition to bribe businesses to support their pet schemes. And if the backlash against selfish, reckless bankers gets confused with the democratic benefits of competitive markets, it may even become tempting for politicians to knock competition policy directly as a populist gesture towards centralized industrial policy.[63]

Careful analysis of the sources of the crisis and a clear understanding of the unique double contagion in banking are crucial prerequisites for developing appropriate policy responses. Certainly, taxpayer money was needed to put the financial system on life support until it can pump sufficient finance on its own. Tighter prudential regulation of banks is self-evidently necessary. I have further argued that it would have been quicker, more direct, and less costly early in the crisis to nationalize troubled banks and instruct them to lend on commercial terms before contagion into the real economy got out of hand. However, no other sector of the economy justifies such exceptional treatment and it would be a great mistake to go backwards to replace competition policy with interventionist industrial policy. Similarly, it would be a mistake to impose regulation beyond that necessary to reduce the likelihood of a future financial crisis.

A STRONG AND ACTIVE COMPETITION POLICY, INCLUDING TIGHT CONTROL OF STATE AID, ENSURES THAT BUSINESS ENERGIES ARE NATURALLY GUIDED INTO SATISFYING CONSUMER NEEDS AND ARE NOT DIVERTED INTO COZYING UP TO BUSINESS RIVALS OR LOBBYING POLITICIANS.

A strong and active competition policy, including tight control of state aid, ensures that business energies are naturally guided into satisfying consumer needs and are not diverted into cozying up to business rivals or lobbying politicians. It has taken many years for enough politicians to appreciate this. In most countries outside the United States, competition policy of sufficient force has only begun to take root over the last decade. This makes it politically

fragile and the misleading "precedent" of bailing out the banks must not be allowed to make competition policy another casualty of the crisis.

---

1 The support of the Economic and Social Research Council is gratefully acknowledged. This paper has benefitted greatly from comments by Jayne Almond, Rob Anderson, Steve Davies, John Fingleton, Alan Gregory, Gerald Gregory, Shaun Hargreaves Heap, John Kay, John Kwoka, Phil Strahan, and from seminar and workshop participants at BERR (now BIS), CCP, UK Competition Commission, WTO, and the International Industrial Organization Society Conference in Boston. None of them can be held responsible for the views I express.

2 This is a revised version of CCP Working Paper 09-04 which was written in early March 2009. I still refer to the "current crisis" although there are signs in late summer 2009 that the worst of the "crisis" may be over. The consequent recession, unemployment, and government indebtedness will have repercussions for many years to come.

3 These flows were mainly from Japan and developing economies with trade surpluses (notably China and oil exporters) into the most financially developed countries (notably the United States and United Kingdom) seeking a safe haven for their savings and a place to hold reserves to counter possible future exchange rate crises. John Taylor makes the case that slack monetary policy, especially during 2003-05, and inappropriate policy responses to the evolving crisis should bear substantial blame; *see* John Taylor, *The financial crisis and the policy response: an empirical analysis of what went wrong*, (November 2008, mimeo).

4 Symptomatically, the banks were much more reluctant to cut their own bonuses unless required by governments to do so.

5 *See* the introductory chapter in Bruce Lyons (Ed.), Cases in European Competition Policy: the Economic Analysis, (2009), CUP.

6 I use the word "banks" as shorthand for all financial institutions that intermediate and insure transactions by firms and consumers.

7 *For example*, by February 16, 2009, the current crisis had seen the European Commission approve 43 separate financial sector aid schemes by Member States, and was investigating 11 more. These covered 19 Members including all 15 who joined pre-2004. *See* EC MEMO/09/67.

8 *See* John Vickers, *The financial crisis and competition policy: some economics*, GCP Magazine (December 2008), available online at www.globalcompetitionpolicy.org.

9 Confidence can also be important for firms whose purchasers do not receive the full benefit of the product at the time of purchase (e.g. insurance, airline tickets booked in advance, warranties, network products).

10 *For example, see*: Joseph Aharony & Itzhak Swary, *Additional evidence on the information-base contagion effects of bank failures*, J. Banking & Fin, 20, 57-69 (1996); Aigbe Akhigbe & Jeff Madura, *Why do contagion effects vary among bank failures?* J. Banking & Fin, 25, 657-80 (2001); Bong-Chan Kho, Dong Lee & Rene Stulz, *US banks, crises and bailouts: from Mexico to LTCM*, Am. Econ. R. P&P, 90.2, 28-31 (2000).

11 *See* Lawrence J. White, *Financial regulation and the current crisis: a guide for the antitrust community*, American Antitrust Institute working paper (2009) for an informed account of institutional problems in the U.S. mortgage system. His Table 5 shows that the 15 largest U.S. financial institutions each had equity (own capital) of less than 10 percent of assets. Such high leverage meant that an across-the-board fall in asset values of 10 percent would have moved each of them into negative net worth.

12 White (2009), *Id*, develops some differences between: a "small bank" which can relatively easily be saved by a central bank and its good assets sold to another bank; and a "large bank" that has more uninsured deposits and securities, much higher exposure to derivatives, and is a source of substantial counterparty risk.

13 If a supermarket goes bust, its rivals shed few tears as they bid to buy productive assets from the administrator and seek to supply the bankrupt chain's former customers. I return to the case of car manufacturers in section 3.

14 Empirical evidence for U.K. firms is provided by a series of surveys of SMEs conducted for the Department for Business, Innovation and Skills, published as *Business Barometer* (April 2009, URN 09/P75C). Although the successive samples of SMEs are not strictly comparable, the trend in responses is worrying because the percentage of SMEs unable to obtain finance from the first bank approached has increased sharply. In December 2008, it was 33 percent, up from 14 percent a year earlier. By April 2009, the figure had risen to 41 percent. The main reason given by refusing banks to SMEs was that their business sector was too risky. Needless to say, this is a self-fulfilling prophecy. Second most frequently mentioned was insufficient collateral (also self-fulfilling as property prices decline with the withdrawl of mortgage finance). For a related theoretical analysis of the advantage to larger firms, *see* Xavier Boutin, Giacinta Cestone, Chiara Fumagalli, Giovanni Pica & Nicolas Serrano-Velarde, *The deep pocket effect of internal capital markets*, CEPR Discussion Paper 7184, (2009).

15 Even in more normal times, the role of banks in mobilizing savings and allocating investment funds means that an appropriately competitive banking system is, in turn, crucial for developing the structure and competitiveness of other markets in the economy. For reviews of competition in banking, *see*: Allen Berger, Asli Demirguc-Kunt, Ross Levine & Joseph Haubrich, *Bank concentration and competition: an evolution in the making*, J. Money, Credit and Banking, Vol.36.3 (Part 2) pp.433-451 (2004); *Stijn Claessens Competition in the financial sector: overview of competition policies*, IMF Working Paper 09/45; Xavier Vives, *Competition in the changing world of banking*, Oxford R. Econ. Pol'y, No.17, pp.535-45, (2001).

16 In normal financial times, a firm whose product is not contaminated (or which can be swiftly made safe) will be able to obtain a loan to tide it over until the scare subsides. Banks cannot provide this service to themselves.

17 This is despite the huge costs. The principal bailout schemes in the EU totaled a nominal EURO 3 trillion, which is 24 percent EU GDP. However, three-quarters of this has been in guarantees, most of which will not be called in, and 10 percent was capital injections, some of which have increased in value. Recession due to contagion into the real sector will almost certainly be more costly in the long run.

18 Excessively risky activities included exposures to complex securitizations, trading of derivatives, and off-balance sheet activities that undermined capital adequacy. In particular, CDOs based on mortgages have been central to the failure of banks in the current crisis; and multiple trading of credit default swaps also created problems as the economic crisis deepened and firms became unable to repay loans. Problems were multiplied by ratings agencies wrongly attributing AAA status to these complex derivatives despite the absence of market prices (they had to rely on highly complex, fragile computer models). Furthermore, these CDOs often stayed within the banking system unsafely hidden off-balance sheet in structured investment vehicles ("SIVs"). The cavalier attitude to risk was not only found in U.S. and U.K. banks developing "innovative" financial products. For example, Austrian banks lent excessively to Eastern European consumers and firms and Japanese banks bought corporate equities. Weak corporate governance of banks played an important role in allowing this.

19 The latter is part of a monetary policy of "quantitative easing" (in the United Kingdom, United States, and several other countries) but it still supports the banks by providing liquidity.

20 Nationalization would also have permitted clearing out the senior executives of failed banks without undeserved compensation packages.

21 The EC does this routinely for state aid cases, and the OFT fulfilled this monitoring function with Northern Rock during its first year in public ownership. *See Office of Fair Trading, Northern Rock: the impact of public support on competition*, OFT1068, (March 2009).

22 This was a feature of the Japanese economy in the "lost years" of the 1990s.

23 Nationwide was paid £1.6 billion by the government to take over Dunfermline. Both were building societies.

24 It has to be acknowledged that the nature of pharmaceuticals customers, particularly national health authorities and price regulators, creates a tangle through which competition policy must operate in most countries; *see* Stephen Davies & Bruce Lyons *Mergers and Merger Remedies in the EU*, Edward Elgar, Ch. 8 and 9 (2007) for a discussion of competition and merger control in pharmaceuticals markets.

25 Basel II was agreed in 2004 and modified in 2005, so in principle it should have been up-to-date with modern banking. There are lessons to be learned about regulatory complexity and delegated responsibilities.

26 A core element of these standard components is Tier 1 asset requirements. These should be strengthened and made less pro-cyclical (the current fixed ratios mean that, in a recession, capital gets written off, which means loans must be reduced, which deepens recession). Also, the value of assets at risk needs to take account of apparently improbable severe crises (sometimes known as the "fat tails" problem in the distribution of returns). Consideration might also be given to limiting loan sizes relative to asset value, if this can be shown to contribute to asset price bubbles. For more macroeconomic suggestions, *see* Mathias Dewatripont, Xavier Freixas, & Richard Portes [eds.] *Macroeconomic Stability and Financial Regulation: Key Issues for the G20*, CEPR, (2009).

27 This distinction between diversifying risk and simply betting on markets is often confused. A related confusion is over investment banking which in recent years has been increasingly associated with trading activities (as distinct from project funding). There are good reasons to join retail and traditional investment banking and to trade securities *for the specific purpose* of diversifying risk. However, given the necessity of taxpayer bailouts of failing banks, there are very good reasons to separate huge trading (i.e. betting) activities which certainly do not justify being underwritten by the taxpayer but which seem to have grown to dominate "investment banking." This should be the context for the reintroduction of an appropriately modified Glass-Steagall Act.

28 Viral Acharya & Julian Franks, *Capital budgeting at banks: the role of government guarantees*, Oxera Agenda (February 2009) argue that government guarantees of bank survival have driven the cost of debt finance down to risk-free levels, which has encouraged excessive leverage.

29 Unfortunately, banks cannot be trusted to assess their own strategic risks. Paul Moore, former head of group regulatory risk at HBOS was dismissed (with a reputed £0.5m gagging payment) for pointing out in 2003 and 2004 that the bank was taking on too much risk in relation to excessive growth in lending (evidence to the House of Commons Treasury Committee; February 10, 2009). It is unlikely that this overruling of risk managers was unique to HBOS or to concern over lending growth. The Icelandic bank Kaupthing, Singer & Friedlander dismissed its heads of both risk and compliance when they complained about risky practices (Channel 4 News, February 24, 2009). In both the HBOS and Kaupthing cases, the concerns were also reported to the FSA (the U.K. financial regulator) but neither bank was reprimanded. In 2003, Ron den Braber warned his bosses at RBS that their models were underestimating risk (Financial Times, March 10, 2009). Other similar, sometimes anonymous, stories have been reported in newspapers in relation to excessive risks in the trading of complex derivatives (e.g. Sunday Times, February 22, 2009). The systemic problem is a failure to balance upside risk with the downside.

30 Luc Laeven & Fabian Valencia, *Systemic Banking Crises: A New Database*, IMF Working Paper WP/08/224, (2008).

31 John Boyd, Sungkyu Kwak, & Bruce Smith, *The real output losses associated with modern banking crises*, J. Money, Credit and Banking, 37.6, Dec., 977-999 (2004) (*see particularly* p. 978 and Table 4) estimate even larger output losses. A study of 23 such crises found only four countries that attained their pre-crisis trend level of output within 17 years. Typically, there was a drop in output, followed by a period of stagnation, until a return to the trend growth rate. The same study measures the accumulated loss of output this entails in several ways, and depending on which they take, the authors calculate the average capitalized loss as equivalent to between 7 months and 3 years of real GDP. One example of a crisis of this order of magnitude is the Norwegian banking crisis and recession in the early 1990s.

32 Competitively neutral macroeconomic stimulus is necessary for Keynesian reasons. Subsidies for retraining, regions, environmental protection, and fundamental R&D may rightly be given to correct a specific externality or for distributional reasons. However, it is sometimes difficult to make the sharp distinction between these so-called "horizontal" state aids and the more discriminatory and so distortionary sectoral- or firm-specific aid.

33 *See* the EAGCP advice on Rescue and Restructuring Aid which was written shortly before the current crisis: available at http://ec.europa.eu/dgs/competition/economist/eagcp.html.

34 This can be thought of as an option value of being in the industry should demand pick up. Similarly entry is delayed by the option value of not having committed to the sunk costs of entry. *See* Avinash Dixit, *Entry and exit decisions under uncertainty*, J. Political Econ.97.3, 620-38 (1989); *also* Robert S. Pindyck, *Sunk Costs and Risk-Based Barriers to Entry*, NBER Working Paper #14755, (2009).

35 *See* Marvin B. Lieberman, *Exit from Declining Industries: "Shakeout" or "Stakeout"?*'21 Rand 4, 538-554 (Winter, 1990) for empirical evidence and references to the theoretical foundations and other empirical findings. *See also*: Andrew B. Bernard & J. Bradford Jensen, *Firm Structure, Multinationals, and Manufacturing Plant Deaths*, LXXXIX.2 R. Econ. & Statistics, 193-204 (May 2007); and Mary E. Deily, *Exit Strategies and Plant-Closing Decisions: The Case of Steel*, 22 Rand 2, 250-263, (Summer, 1991).

36 This is not a claim that all free market structures are ideal in the theoretically abstract sense of what might be designed by a perfect planner with all the available information.

37 Highly leveraged or indebted firms are more likely to exit before their less leveraged rivals, at least in concentrated markets. *See* Dan Kovenock & Gordon M. Phillips, *Capital Structure and Product Market Behaviour: An Examination of Plant Exit and Investment Decision*, 10 R. Financial Studies 3, 767-803 (Autumn, 1997).

38 Lending should depend on ability to repay (outside the immediate crisis period), which does not mean a return to the precarious policies of the last decade. This form of state lending is accepted by the European Commission under what is known as the "market economy investor principle" and is relevant for both State aid and State owned firms.

39 Beyond political impact, it tends to be declining industries with concentrated market structures that have the greatest incentive to invest in lobbying activities because they face a smaller free-rider problem in reaping the rewards. *See* Frederick Robert-Nicoud and Richard Baldwin, *Entry and asymmetric lobbying: why governments pick losers*, 5.5 J. Eur. Econ. Assoc. (2007), 1064-93.

40 Arguments may also be made in relation to agglomeration economies by which a region develops a network of supply links and support services that benefit many independent firms. However, it is unlikely that even the current recession could overturn genuine long-term agglomeration economies. Detroit is sometimes used as an example from the car industry. However, it is notable that Japanese car manufacturers mostly chose to locate far from Detroit for their successful entry into the United States.

41 Nevertheless, specific sectors clearly have an incentive to obscure this fact and firms may collude in search of State aid. For example, GM and Chrysler approached Washington together, and Renault and Peugeot-Citroen approached Paris together.

42 For a review, *see* Kyle Bagwell & Robert W. Staiger, *Protection and the Business Cycle*, (January 2003), mimeo.

43 Daniel A. Crane, *Antitrust enforcement during national crises: an unhappy history*, GCP Magazine, 9 (December 2008). *Available* online at www.globalcompetitionpolicy.org.

44 The examples and estimates in this paragraph are taken from Harold L. Cole & Lee E. Ohanian, *New Deal Policies and the Persistence of the Great Depression: A General Equilibrium Analysis*, 112 J. Pol. Econ. 4, (2004).These findings have been challenged by Gauti Eggertsson, *Was the New Deal Contractionary?'* Federal Reserve Bank of New York Staff Report no. 264 (2006), but Cole & Ohanian are more persuasive. Note also the Robinson-Patman Act (1936) prohibiting price discrimination and various other "fair trade" laws that were arguably too interventionist, were also passed during the Great Depression, as was the Smoot-Hawley Tariff Act (1930).

45 There is also evidence that lack of competition unnecessarily prolonged the 1990s Japanese recession. *See* Michael E. Porter & Mariko Sakakibara, *Competition in Japan*, 18 J. Econ. Perspectives 1, 27-50, (2004).

46 In the state of panic at the time, the Secretary of State was supported by a powerful triumvirate of the Bank of England, Financial Services Agency, and Treasury on grounds of short-term financial stability. John Vickers argues that this aim might have been achieved in a less anticompetitive way, *see* John Vickers, *The financial crisis and competition policy: some economics*, GCP Magazine (December 2008). The merger creates a balanced duopoly in SME banking in Scotland, with the other duopolist being the crippled and near-nationalized RBS (see #158-9 of the OFT's *Anticipated acquisition by Lloyds TSB plc of HBOS plc: Report to the Secretary of State for Business Enterprise and Regulatory Reform*, (24 October 2008), available at http://www.oft.gov.uk/shared_oft/press_release_attachments/LLloydstsb.pdf). In the United States, emergency takeovers of Bear Stearns, Merrill Lynch, and Wachovia among others may have also been problematic.

47 The Act does allow for such a political decision on the grounds of public interest though this was intended to be interpreted narrowly, with national security as the only stated example plus a public interest provision to maintain media plurality (R. Whish, Competition Law, 898 (2001)). A new public interest "to ensure the stability of the UK financial system" had to be created in a formal Order to be passed urgently by both Houses of Parliament. Note that national security and media plurality are appropriately long-term considerations for a merger, whereas this merger's contribution to financial stability could only have been short-term at best. In fact, subsequent events have shown that HBOS was sitting on a loss of £10 billion in bad debts that Lloyds TSB failed to notice in its highly compressed and partial "due diligence." Consequently, two banks have been crippled instead of just one.

48 Paul Krugman, *Protectionism and stimulus*, on his blog dated 1 February 2009: http://krugman.blogs.nytimes.com/2009/02/01/protectionism-and-stimulus-wonkish/.

49 The Market Economy Investor Principle ("MEIP") allows a State injection of funds as long as this is on "terms which a private investor would find acceptable in providing funds to a comparable private undertaking when the private investor is operating under normal market-economy conditions;" OJ C307, 13.11.1993, #11.

50 For a succinct explanation of EC state aid rules as applied to banking, *see* Christopher Vajda, *The banking crisis and EC state aid rules*, Butterworths, 67-69, (2009).

51 *Temporary framework for State aid measures to support access to finance in the current financial and economic crisis*, Communication from the Commission, (26 November 2008). By the end of July

2009, 24 countries had taken advantage of the new rules and 55 non-bank aid schemes had been approved by the Commission under the Temporary Framework. *See* EC MEMO/09/67 and MEMO/09/380. Specific allowable measures are: up to EURO 0.5m cash grant per firm, provided the aid does not favor exports or domestic over imported products (which will be very hard to police); reductions of 15 percent (25 percent for SMEs) on loan guarantee premia for loans up to the size of the annual wage bill; relaxed rules on interest rate subsidies; 25 percent subsidies (50 percent for SMEs) for investment in green production; and provision of risk capital for SMEs. It is interesting to compare the incidence of these general schemes for the real sector with the bank rescues over the same period, where 18 Member States had 66 bank rescue or more general bank guarantee-type schemes approved by the EC.

52 More precisely, the relaxation is limited to SMEs plus firms that were not in difficulty before July 1, 2008.

53 *See* Dewatripont & Seabright, *Wasteful public spending and state aid control*, J. Eur. Econ. Ass'n 4.2/3, 513-22 (2006), on the commitment value of EU State aid rules. The United States has no equivalent to the EC for reviewing rescue and restructuring aid. One commentator suggests the United States needs a DoJ Deputy Assistant Attorney General for emergency restructuring to represent the interests of competition. *See* Albert Foer, *'Too big to fail?' The role of antitrust law in government-funded consolidation in the banking industry*, statement before the U.S. House of Representatives Judiciary Committee, sub-committee on courts and competition policy (March 17, 2009.

54 *For example*, the Doha round of trade liberalization was started in 2001 and is still struggling for agreement.

55 Other international institutions are also advocating an appropriate role for competition policy during the recession. For example, the Director-General of the OECD, Angel Gurría, has called for strong competition policy to speed recovery, OECD press release (February 19, 2009).

56 For further examples, *see* John Fingleton, Competition policy in troubled times, (speech dated January 20, 2009), *available at* http://www.oft.gov.uk/.

57 *See, for example*: Margaret Levenstein & Valerie Suslow, *What determines cartel success?* 44 J. Econ. Lit. 1, 43-95 (2006), pp. 43–95; Simon Evenett, Margaret Levenstein & Valerie Suslow, *International cartel enforcement: lessons from the 1990s*, World Econ. 24.9, 1221-45 (2001).

58 *See also* Andre Fiebig, *Crisis cartels and the triumph of industrial policy over competition policy in Europe*, Brooklyn J. Int'l L., XXV.3, 607-38; and Richard Whish, *Competition Law*, Butterworths, 577-8 (2003).

59 *See for example*: Gary Becker, *Crime and Punishment: An Economic Approach*, J. Pol. Econ 76.2, 169-217 (1968); Mitchell Polinsky & Steven Shavell, *The Economic Theory of Public Enforcement of the Law*, J. Econ. Lit. 38, 45-76 (2000); and M. Motta, Competition Policy: Theory and Practice, (2004).

60 The Lloyds TSB HBOS merger was not allowed on a classic failing firm defense, which is that if a firm is going to exit the market anyway, there will be no additional loss of competition due to the merger. As already described, the ministerial intervention in that case was on public interest grounds supposedly "to ensure the stability of the UK financial system."

61 The OFT appreciates this in its *Restatement of OFT's position regarding acquisitions of "failing firms"* December 2008, OFT1047.

62 Behavioral remedies may be feasible for some mergers, particularly if there is a natural way of enforcing them and if the competition problem is expected to be short-lived.

63 For example, Olivier Besancenot has achieved instant popularity in France by setting up the Nouveau Parti Anti-Capitaliste ("NPA").

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

The U.S. Industry Under Duress: Fit, or Finished?

John E. Kwoka, Jr.

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

The U.S. Auto Industry Under Duress: Fit, or Finished?

*John E. Kwoka, Jr.**

In the latter half of the 20th century, the U.S. auto industry truly lost its way. It squandered its competitive advantage, allowed itself to become vulnerable to forces beyond its control, lost its markets one by one to foreign rivals, and stared into the abyss of its complete demise. Only U.S. government intervention on a previously unimaginable scale prevented that outcome. A great debate emerged over the causes of the auto industry's collapse, the rationale for government intervention, and the effects of competition between government-owned and private auto companies. That debate was leapfrogged by events that forced decisions about intervention and ownership. But events have not obviated the need for examining these questions, since the U.S. government is now even more deeply involved in the U.S. auto industry. In addition, this experience may serve as a model or argument for other troubled sectors. This paper[1] seeks to cast light on some of the issues raised by government intervention.

*John Kwoka is the Neal F. Finnegan Distinguished Professor in the Department of Economics in Northeastern University.

I. Introduction

For more than a century the auto industry has been at the center of U.S. manufacturing. It has provided jobs to millions of people, wealth to tens of millions, and products by the hundreds of millions. The jobs it created were high quality jobs; jobs that taught skills, promoted mobility into the middle class, provided health care, and conferred retirement security on generations of Americans. Its products were useful and often exciting, capturing consumers' imagination, responding to their thirst for the open road, and permitting a lifestyle that came to be associated with the American Dream. The wealth it created enriched its owners and managers, its suppliers and communities, and its stockholders throughout the country. This is an extraordinary record unmatched by any other industry in this or any country.

BUT IN THE LATTER HALF OF THE 20TH CENTURY, THE U.S. AUTO INDUSTRY TRULY LOST ITS WAY.

But in the latter half of the 20th century, the U.S. auto industry truly lost its way. It squandered its competitive advantage, allowed itself to become vulnerable to forces beyond its control, lost its markets one by one to foreign rivals, and stared into the abyss of its complete demise. Only U.S. government intervention on a previously unimaginable scale prevented that outcome. A great debate emerged over the causes of the auto industry's collapse, the rationale for government intervention, and the effects of competition between government-owned and private auto companies. That debate was leapfrogged by events that forced decisions about intervention and ownership. But events have not obviated the need for examining these questions since the U.S. government is now even more deeply involved in the U.S. auto industry. In addition, this experience may serve as a model or argument for other troubled sectors.

This paper seeks to cast light on some of the issues raised by government intervention. We begin with a review of the root causes of the crisis in the U.S. auto industry and then discuss whether there is a principled basis for government intervention in the industry. For the latter question, we set out several rationales that have been offered for government intervention and analyze their possible economic foundations. We then go on to examine the nature of competition in an industry that now consists of both private companies and publicly-owned enterprises. We conclude with some observations about the role and effects of public policy in these circumstances.

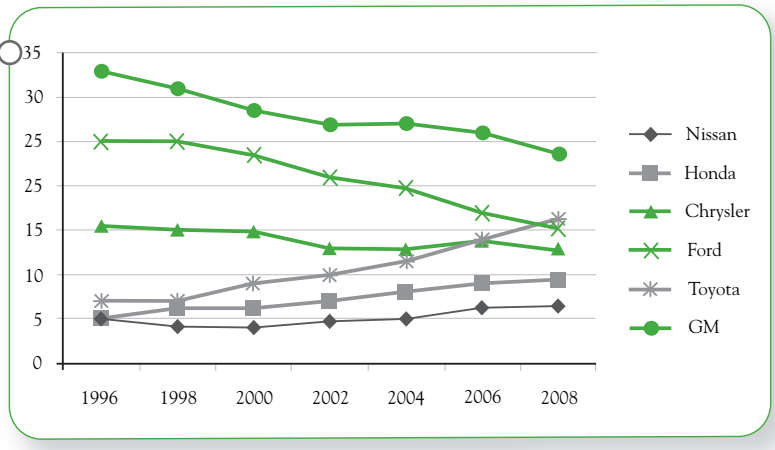
II. Autos: An Accident Waiting to Happen

Fifty years ago U.S. companies manufactured virtually all motor vehicles sold in this country. In 1965 General Motors ("GM") alone sold 50 percent, with Ford at 26 percent and Chrysler at 14 percent, leaving only 10 percent total for American Motors and a few small volume imports, primarily European. But as

Ford celebrated its centennial in 2003, and more especially as GM did so in 2008, both companies' market shares had been cut in half and were in steady decline. As shown in Figure 1, both of those companies were losing market share at a rate of about one percentage point per year, a trend that showed no signs of abating. As a consequence, GM's stock price had declined from its peak of $94 in 2000 to about $25, and continued its steady retreat. The company was on its way to reporting a loss of $30 billion in 2008—approximately $10,000 per vehicle sold. Ford's losses in 2008 totaled $15 billion while Chrysler lost $8 billion—an astonishing total of $53 billion.

Figure 1

Market share of U.S. vehicle sales



But if that seemed bad, it was about to get worse—much worse. U.S. demand for light vehicles (cars, SUVs, minivans, and pickups), which had been running at a rate of about 16-17 million units per year, started to weaken during the summer of 2008, but with the financial crisis and the macro recession in the fall, demand truly collapsed. By early 2009, the annual sales rate was less than ten million units—a rate last seen in 1982. In the first quarter of 2000, GM's sales plunged 45 percent, those of Ford and Chrysler by similar amounts. U.S. sales by Toyota, Honda, and other manufacturers initially held up as buyers shifted from larger U.S. vehicles to smaller products made by their Japanese and Korean rivals, but after that initial shift wore off, sales of foreign nameplate vehicles suffered the same precipitous decline.

The effects of such demand declines are readily understood from some simple economics. Automobile manufacture is a high fixed-cost business, so that sales declines result in revenue losses that substantially exceed cost reductions in the short- and medium runs. The upshot is large financial losses. Thus, GM lost more than nine billion dollars in the fourth quarter of 2008, and another $6 bil-

lion in the first quarter of 2009. Ford lost $1.4 billion and Chrysler $2 billion. GM's share price fell to $1.15 and its total market capitalization was less than $3 billion.

These financial effects prompted concern over the long-term viability of the three traditional Detroit-based companies. Not wanting the auto industry to collapse on its watch, the Bush administration stepped in with interim measures to ensure the survival of the U.S. companies. The harder questions were left to be more fully addressed by the incoming Obama administration.

THE HARDER QUESTIONS WERE LEFT TO BE MORE FULLY ADDRESSED BY THE INCOMING OBAMA ADMINISTRATION.

Before discussing those measures, it will be helpful to identify the various causes of the extreme sales and financial difficulties faced by the GM, Ford, and Chrysler—the "Detroit 3"—difficulties that for the most part exceeded those faced by foreign auto companies.[2] We divide these causes into short- and long-run problems, noting the interaction between the two along the way.

A. SHORT-RUN PROBLEMS

The obvious short-run problem that faced the U.S. auto sector starting in late 2008 was the collapse of demand. While past recessions had produced sales declines on the order of 15-25 percent, the magnitude of decline starting in the spring of 2008 was without precedent. This sales collapse had both macro- and micro-roots, as we shall now enumerate.

1. Macro Causes

- The Great Recession, which caused adverse changes in the major determinants of demand for autos—income, employment, and consumer confidence.
- The credit crunch, which affected auto suppliers' ability to finance operations, dealers' ability to finance inventory, and consumers' ability to purchases vehicles without full cash payment.

2. Micro or Industry-specific Causes

- Since auto purchases are postponable, demand has wider swings than for many other products. Most consumers can simply stay out of the market for a period of time, returning (often en masse) when consumer confidence and other conditions are more favorable. This has historically resulted in considerable volatility in auto sales.
- "Overselling" of cars in recent years, as the auto companies boosted short-term sales by substantial discounting, cheap credit, and off-mar-

ket sales to rental fleets. These strategies "pull" future sales to the present, but when that future arrives, some part of naturally arising demand has already been satisfied. That has left current demand even shorter than would normally be the case.

B. LONG RUN PROBLEMS

The other set of forces adversely affecting the Detroit 3 has been a number of longstanding, deep-seated, and largely unaddressed structural problems. These have left the U.S. auto companies vulnerable to various threats, including threats posed by the advent of Japanese cars, environmental and regulatory constraints, and periodic high gas prices. In each case the companies have been caught unprepared, denied responsibility, sought to avoid fundamental change, and permanently lost sales and employment. In the context of the present demand collapse, these problems have made matters far worse for the Detroit 3. These problems fall into four broad categories.

THE OTHER SET OF FORCES ADVERSELY AFFECTING THE DETROIT 3 HAS BEEN A NUMBER OF LONGSTANDING, DEEP-SEATED, AND LARGELY UNADDRESSED STRUCTURAL PROBLEMS.

1. The Product Itself

- Quality problems have long afflicted products coming out of Detroit. The initial inroads made by Japanese companies were due to offering cars with high quality at budget prices. Over the years, the Detroit 3 made enormous progress in closing the quality gap, but with the exception of a handful of vehicles, their defect rates remain significantly above the target established by ever-improving Japanese products and newly-emerging Korean competitors.[3]

- Even when Detroit has succeeded in manufacturing vehicles with defect levels comparable to its state-of-the-art competitors, domestic cars have often suffered from uninspired design, poor features, and the bad reputation of their car company. The current model of the Chevy Malibu illustrates this combination of high production quality but mediocre sales.[4]

- Above all, in recent years Detroit had become fixated on two high-volume and high-profit vehicles developed here; namely, minivans and sport utility vehicles. As shown in Figure 2, minivans first appeared in the early 1980s, their sales cresting at about 1.25 million units per year in the mid-1990s. As the minivan fad started to fade and foreign competitors moved into that segment, SUV sales took off, soon dwarfing the minivan boom. But SUVs went through a similar product cycle, causing Detroit to rather frantically search (unsuccessfully) for a new "hit" product.

Figure 2

U.S. vehicle sales (millions of units)

The importance of the "hit" vehicle strategy is threefold. First, it does not constitute a viable long-term business strategy, since it relies on the ability to come up with a never-ending series of new large-volume-and-profit products. Such persistent success is no more likely than with repeated betting—at some point all winning streaks end. Second, this strategy resulted in the Detroit 3 increasingly becoming truck companies rather than car or diversified vehicle companies, shifting resources and attention away from traditional passenger cars. In recent years GM has produced more light trucks than cars, Ford twice as many trucks, and Chrysler three times as many. Third, during this same time Toyota, Nissan, Honda, and others remained focused on cars, continued to improve those vehicles, and took an ever larger share of U.S. passenger car sales. While car sales had been declining as a fraction of total light vehicles, the Japanese (and now Korean) companies positioned themselves advantageously for the time when demand for passenger cars recovered.

2. Production Cost Problems

- Detroit has long suffered from an operating costs disadvantage relative to production in Japan (even after transportation costs) and then rela-

tive to production at Japanese transplant factories here in the United States. The evidence indicates that several Detroit assembly and manufacturing plants now have become cost competitive, although on average GM, Ford, and Chrysler plants remain somewhat less efficient than their competition.[5]

- Retirement costs and health benefits represent a substantial burden on the Detroit 3. The age and health status of their workforces is said to result in a per-vehicle cost differential relative to Japanese producers of perhaps $1100-1300.[6] Among the three Detroit companies, GM's predicament over time grew to be the most serious. It had about 4.6 retirees per active UAW member, compared to 2.1 and 1.6 for Ford and Chrysler, respectively (all this before the events of 2009). Notably, GM relieved itself of some of the long-term consequences by negotiating an arrangement with the UAW in 2006 that granted the union control of the health care fund in trade for smaller annual contributions by GM. While at the time this appeared to be a model for restructuring health care obligations, any long-term benefits have been overwhelmed by events.

3. Management Weaknesses and Failures

- The Detroit 3 have long suffered from management weaknesses—weaknesses of senior personnel and weaknesses in major decision-making. The Detroit culture has been stubbornly insular, focused on themselves, and in denial about outside threats. Until the present, no CEOs have been drawn from outside the industry. Examples of bold thinking—the GM-Toyota joint venture, Ford's green initiatives—have been few, far between, and without the kind of systemic impact necessary to transform the industry. By contrast, examples of short-sighted thinking abound: GM's 2006 decision to focus on trucks (including the Hummer brand); its progressive abandonment of Saturn over the past decade; Chrysler's near-complete devolution into a truck company; and all three companies' excessive number of divisions, products, and dealers are among many decisions that have sapped their competitive strength.

THE DETROIT CULTURE HAS BEEN STUBBORNLY INSULAR, FOCUSED ON THEMSELVES, AND IN DENIAL ABOUT OUTSIDE THREATS.

- Compounding these management failures at the Detroit 3 have been governance failures. Rather than a committed board of directors prodding and, if necessary, replacing weak management, there has been a tolerance of mediocrity. CEO after CEO has presided over vast losses of sales, share, and capitalization, all without penalty. Once having the largest market capitalization of any manufacturing company, GM was worth less than $2 billion by the end of year 2008. It has been calculated that between 1980 and 1990, GM and Ford destroyed $110 billion in capital and between 1997 and 2008 another $190 billion.[7]

Remarkably, despite this record, the Boards of Directors of the Detroit 3 have long exhibited nearly unwavering support for their CEOs. GM's board repeatedly expressed support for its recent CEO, even after decision after decision damaged the company.[8] The degree to which the boards were part of the problem was demonstrated by the failed efforts by Ross Perot in the 1980s and later by Kirk Kerkorian through his associate Jerome York to shake up GM's board. Both investors—savvy, well-financed, and experienced—essentially threw up their hands in exasperation and departed the Detroit scene.[9]

4. Public Policy

- The policy that perhaps has been most damaging to the long-term interests and health of the U.S. auto industry has been the country's commitment to cheap gas. Cheap gas has spurred the boom in sales of low-mileage vehicles favoring the Detroit 3, but it has also laid the foundation for these companies' vulnerability to gas price shocks and similar events. And these events have occurred with some regularity, with devastating effects on sales of vehicles by the Detroit 3 with their commitment to large low-mileage vehicles.

- A second government policy that has served the companies poorly has been the "alternative technology" fiction. Beginning with the Partnership for a New Generation of Vehicles in the 1990s and more especially with the Freedom Car of the Bush administration, the federal government has very publicly heralded programs apparently designed to assist the Detroit 3 in developing new, high-mileage, low-emissions power plants.[10] These have not yielded any such benefits, leading many to conclude that the purpose of these programs was more public relations than substance.

  In this respect, the contrast between U.S. and Japanese car companies could not be clearer. While official U.S. policy has been promoting fuel cell technology—a very long term and very difficult technology to implement—Toyota introduced the Prius in this country in 2001. That simple but sophisticated hybrid gas-electric vehicle instantly became a sales hit and badge of distinction to Toyota, eventually forcing Detroit to respond with its own hybrid vehicles.

C. COMPETITION: THE ROAD NOT TAKEN

Many of the U.S. auto industry's problems stem from the fact that the Detroit 3 have behaved as if they had no viable competitors, or at least none that mattered. This raises the question of the role—actual or potential—for competition policy with respect to the industry: Could competition policy have played a more constructive role in altering the structure or behavior of this tight-knit oligopoly?

Could competition policy have played a more constructive role in altering the structure or behavior of this tight-knit oligopoly?

There have been some notable efforts. In the 1960s the Justice Department Antitrust Division ("DOJ") conducted a preliminary investigation into competition in the U.S. auto industry. The Federal Trade Commission ("FTC") followed this up with an Omnibus Auto Industry Investigation in the late 1970s, examining various structural and behavioral issues that were contributing to the lack of competition. The FTC also weighed in on the competitive effects of increased regulatory stringency, the Chrysler bailout of 1980, import restraints, and various alliances and partial equity agreements among the U.S. and Japanese car companies that sprang up in the 1970s and 1980s. But growing competition from abroad rendered most of these concerns moot: Import restraints were relaxed and then eliminated, while Japanese companies set up factories in the United States.

The upshot was that weak competition among the Detroit 3 was overtaken by ever-stronger competition from new, foreign rivals. It might have been hoped that GM, Ford, and Chrysler would respond to that competition; but if they failed to do so, consumers now had alternatives available to them. Further failures of the Detroit 3 would be "their" problem—a private loss—rather than something that would adversely affect the public interest - automotive consumers.[11]

D. THE OUTCOME

Remarkably, even in the face of competitive threats of the first order, the Detroit 3 failed to take the necessary steps to preserve and strengthen their position. The result has been predictable but more extreme than might be expected. Sales of vehicles built by the Detroit 3 have fallen to their lowest level in decades, and will not be restored. Auto manufacturing has lost hundreds of thousands of jobs. UAW membership, once as high as 1.5 million, is now less than one-third that number, and falling. The auto companies' finances have jeopardized health benefits and retirement security for millions of workers and their dependents. The companies have been forced to close numerous plants and thousands of dealerships, creating financial distress for countless communities around the country.

Nowhere are these effects more acute than in Detroit—the Motor City—and surrounding communities. Fifty years ago Detroit had the highest per capita income of any city in the country, as well as the highest rate of home ownership. It now ranks at the bottom in terms of median income per household and has one of the highest poverty rates (34 percent) among large cities.[12] Nearby Flint Michigan, once home to 100,000 auto workers, now has 5,000 workers. Major parts of that city are abandoned and, in recognition of the permanence of this downsizing, the city is contemplating simply bulldozing some parts to the ground.[13]

Interestingly, however jarring and difficult this outcome, it had been tacitly accepted by all parties as a method of adjustment for the industry. The government, the unions, and the companies themselves no longer seemed resistant to the notion of a long, slow decline for the auto industry. The companies seem

resigned to progressively retreating to whatever vehicles remain profitable. The new industry equilibrium would involve fewer plants, workers, and products. In this context the role of policy would be limited to easing that decline in order to permit all parties more time for adjustment. Unemployment insurance, worker retraining, and community assistance would all slow and smooth out the decline, but not seek to fundamentally alter or prevent it.

> THE GOVERNMENT, THE UNIONS, AND THE COMPANIES THEMSELVES NO LONGER SEEMED RESISTANT TO THE NOTION OF A LONG, SLOW DECLINE FOR THE AUTO INDUSTRY.

This tacit understanding held up until the financial crisis and great recession starting in the fall of 2008. A sales decline of 40 percent took the companies off this glide path, jeopardizing their very existence. The result was a perceived role—indeed, need—for government intervention on an unprecedented scale. And that intervention was not simply intended to restore the companies to that glide path. Rather, as we shall see, it rethought the new equilibrium to which the companies were headed, requiring fundamental changes in management and products as conditions for assistance.

III. The Rationale For and Role of Government

Federal and state governments have long played an important role in the U.S. auto industry. This role routinely has involved tax and other financial benefits, as well as unemployment insurance and similar indirect assistance. With one exception, however, the government has not provided company-specific assistance where private capital markets declined to do so. That exception, of course, was the federal government bailout of Chrysler in 1980. Here we briefly discuss that experience, with particular attention to its stated rationales, and then address the various rationales that have been advanced for the more comprehensive assistance provided to the U.S. auto companies during the great sales collapse.

A. THE CHRYSLER BAILOUT

Almost exactly thirty years ago, Chrysler sought federal government backing for $1.2 billion in loans. This was the culmination of a long slide in Chrysler's sales and financial condition, itself the result of poor products, costly production, and management mistakes over the preceding decade. The debate over granting Chrysler assistance raised now-familiar issues: Many argued for letting the market work its will, asserting that Chrysler's problems were of its own making and so it should bear the consequences. Advocates of assistance alluded to a Congressional Budget Office study that estimated Chrysler's demise would cost 360,000 jobs. The government responded.

The Chrysler Corporation Loan Guarantee Act was passed in December, 1979. It provided for $1.2 billion in federally guaranteed loans—then an unprecedent-

ed amount—at a modest fee of 1 percent per year. In turn, Chrysler had to issue $50 million in new stock, sell off $300 million of assets to strengthen its cash position, secure $2 billion in concessions from workers, banks, local governments, dealers, and suppliers, and agree to a federal oversight board. Other branches of government offered additional breaks, including reduced fuel economy standards and relaxed emission standards for the next model year. In addition, import restrictions on Japanese cars were being negotiated and became binding in 1981.

Chrysler sought to do its part. It secured new loans from its banks (guaranteed by the federal government), concessions from the UAW (deferral of payment to the union pension fund plus wage cuts), and some assistance from its suppliers. It did not, however, fundamentally change its management, UAW contract, or health and pension obligations. And over the course of the following three years a recovery of the overall auto market permitted it to pay back its loans together with $350 million in interest.

As a practical matter, it seems beyond dispute that Chrysler would not have survived its crisis of the 1970s without government intervention. On the other hand, the company failed to do what was necessary to truly become more competitive in the long run, and so could be said simply to have postponed its day of reckoning. Moreover, the success of federal government intervention—by whatever criteria or for whatever length of time—in no way ended the debate over the merits of such intervention. All of those issues have returned in the present context.

B. CURRENT RATIONALES FOR INTERVENTION

In the recent debate over policy intervention for the auto industry, several rationales have been advanced. In this section we analyze four of the primary reasons—supplier effects, warranty issues, stranded assets, and spillovers. In each case we seek to bring some economics to bear on the merits of the argument.

1. Supplier Effects

The argument over supplier effects has been stated as follows: A huge sales decline followed by financial difficulties or bankruptcy for one of the Detroit 3 would result in a substantial decrease in orders and production at its suppliers, thereby jeopardizing the suppliers' financial viability. Then other auto companies dependent on the same suppliers would find their supplies of necessary parts at risk, creating further disruption at the auto manufacturing stage.[14] This degree of supply interdependence implies systemic effects from the demise of a single auto manufacturer. Its demise can bring cripple the operations of its horizontal competitors through supply chain disruptions.[15]

At first glance the economic basis for this argument is not altogether apparent. Suppose that one particular supplier of, say, auto seats is a major supplier to one particular auto manufacturer. It the latter's product sales collapse, the supplier

itself may well face financial jeopardy.[16] The concern is that the supplier's financial difficulty in turn threatens the supply of seats to another auto manufacturer. Several factors, however, are likely to mitigate this concern:

- The second manufacturer could shift its purchases toward the supplier in question in order to ensure adequacy of its business. Of course, the supplier might still not have enough business to remain in operation. If not, the second manufacturer still has options.
- It could increase purchases from other suppliers with whom it already deals. Ordinarily, this strategy might be limited by the other suppliers' ability to expand, but in times of generally weak product sales, this is unlikely to be a binding constraint.
- The manufacturer could enter into contractual arrangements with suppliers from which it had not previously purchased seats. This strategy is likely to take some additional time to implement, however, due to contract, product, and operational issues that would need to be resolved.
- Even if it went bankrupt, the supplier's assets would not likely disappear. Rather, a new enterprise would probably emerge from bankruptcy proceedings and operate as the successor supplier—although this again might take some time.
- Finally, the auto manufacturer in question might provide bridge funding or simply take over the supplier in order to ensure its continued operation. One complication might be that other auto manufacturers would likely cease doing business with the supplier now controlled by their rival, with adverse effects on the supplier's overall business.

These arguments help to explain both the merits of reliance upon markets and the limitations of that approach. Prevention of significant harm to the rival manufacturer would seem to depend on such things as the ability of that manufacturer to shift its purchases among suppliers quickly; its ability to negotiate alternative arrangements without undue delay; the ability of other suppliers to take on the additional orders; and/or the willingness of the manufacturer to intervene directly to support the supplier—none of which is a certainty.

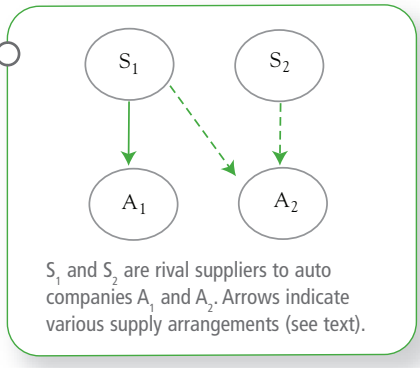
By imposing somewhat greater structure on the issue, we can illustrate a further competitive concern. Suppose there are only two auto companies, A_1 and A_2, each buying seats from two different suppliers, S_1 and S_2. Further, suppose seat manufacturing is subject to scale economies. And to make this example relevant, suppose finally that manufacturer A_1 goes bankrupt. Its sales fall precipitously and so do its purchases of seats.

Two possible cases follow: First, suppose that S_1 and S_2 are direct competitors at contract renewal time, with strong competition between them and price at the competitive level. In Figure 3, if A_1 buys primarily from S_1, collapse of A_1 may

then cause S_1 to collapse, leaving S_2 as the sole (i.e., monopoly) supplier to A2. The effect is to create a vertical monopoly with the usual double marginalization since now S_2 no longer faces competition in its supply to A_2. As a result, final product quantity falls and its price rises.

Figure 3

Supplier effects



S_1 and S_2 are rival suppliers to auto companies A_1 and A_2. Arrows indicate various supply arrangements (see text).

This scenario is analogous to the often-analyzed vertical foreclosure scenario, which presumes a two-by-two vertical arrangement, in which the merger of one supplier with one manufacturer leaves the other manufacturer subject to the now sole supplier of a necessary input.[17]

Secondly, we can relax some of the strong assumptions of the previous case. Suppose that both auto manufacturers engage in dual sourcing; that is, purchase of some supplies from both S_1 and S_2. In this case the supplier that is more dependent on the collapsed auto company A_1 suffers the greater demand decline and faces the greater financial jeopardy. Even if that supplier is more efficient, it could be forced into bankruptcy itself. Alternatively, it could seek a higher price from the manufacturer that is reliant upon it, but the manufacturer cannot grant that higher price without raising its own price in the final product market, disrupting its own operation and that of the final market.

These scenarios illustrate the manner in which competition may be harmed by the elimination of a supplier to the auto companies. Which of these scenarios apply, and to what degree the adverse effects emerge, depend on such things as the degree of product differentiation in the suppliers' products, the degree of substitutability against other products, and the periodicity of contracting for the supply product, as well as the considerations noted with respect to the more general arguments for concern over supplier demise.

2. Warranty Issues

A second argument for policy intervention into the auto industry has been the possible deterrent to consumer purchases from a company that faces some probability of bankruptcy. Clearly when most buyers purchase a vehicle, their deci-

sion involves some expectation about warranty coverage and service. But the warranty that they purchase may become worthless if the company in fact goes into bankruptcy. Recognizing this, some potential buyers may decide not to purchase from that company in the first place.

Moreover, if enough potential customers become concerned and postpone or divert their purchases, they may precipitate the very outcome they fear: Postponed sales may cause the firm to go bankrupt even if confidence in its warranties and collective continued purchase would have sustained it. This self-fulfilling prophecy is similar to the contagion effect in finance, whereby sufficiently widespread concern over an institution's viability can itself cause its non-viability.

From underlying economics, we might expect a potential customer to make his/her purchase decision with due consideration for expected product quality, warranty coverage, and the likely future "all-in" costs of the product. In this respect car warranties are similar to other aftermarket parts and services that are often bundled with the primary product being sold.[18] Examples of bundled future products and services include support for software, copiers and parts, and cameras and dedicated lenses.

RATIONAL BEHAVIOR WOULD IMPLY THAT CONSUMERS MAKE "LIFE-CYCLE" PURCHASE DECISIONS, PROPERLY ACCOUNTING FOR THE EXPECTED ON-GOING COSTS OF THE PRODUCT IN THEIR INITIAL PURCHASE DECISIONS.

Rational behavior would imply that consumers make "life-cycle" purchase decisions, properly accounting for the expected on-going costs of the product in their initial purchase decisions. A warranty that is bundled with the product is intended to provide insurance with respect to those possible future costs and hence strengthen current demand. If a company's viability is in question, that might significantly affect those expectations, devalue its warranty, and jeopardize its current product sales. This very argument had been made in the Chrysler bailout, and has recently been made again as a reason for policy intervention in the current U.S. auto market.

What is unclear in practice is whether consumers actually make decisions with such attention to future costs. Considerable evidence casts doubt on this proposition. Numerous studies indicate that, in deciding between more-efficient but higher purchase price appliances (e.g., refrigerators) vs. cheaper and less-efficient versions, consumers routinely opt for the latter at differentials that imply personal discount rates ranging from 25 percent to 130 percent, and sometimes as much as 300 percent.[19] This discounting of the future has been formalized into models of hyperbolic discounting, in contrast to the exponential discounting that is claimed to be rational and consistent.[20]

Somewhat oddly, if consumers in fact do not act "rationally" in accounting for future costs, warranties matter less in the purchase decision, and any adverse effects on an auto company in financial trouble may not be so great. There is little systematic evidence on this question.[21] Anecdotally, during the first six months of 2009 when both GM and Chrysler were in bankruptcy, Ford's sales were modestly better (or less bad) than those of its rivals, leading some observers to conclude that Ford was benefitting from customer concerns about its rivals' financial difficulties.

3. Stranded Assets

Economic models and policy prescriptions are commonly of a partial-equilibrium nature; that is, they address problems of individual markets rather than of a systemic nature. The implication of this approach is that other agents and markets are operating normally and stand ready to absorb and adjust to disruptions in the market in question. The failing firm defense to a merger, for example, has been created to prevent the disappearance of assets from production altogether.[22] Similarly, worker retraining programs, which are intended to facilitate worker adjustment to layoffs in a declining sector, presume other sectors have employment opportunities.

These presumptions may not hold, especially in an economic environment such as that in the U.S. in 2008-09 where overall market demand for autos has collapsed and employment opportunities have evaporated. Under these circumstances, physical capital that is no longer needed or layoffs of human resources may not have alternative uses or employment opportunities, at least not of a comparable nature or for a long time. In such cases the unused assets will be either lost or devalued. Indeed, this has long been a problem with worker retraining programs, which often result in jobs that do not fully use the skills of the unemployed workers or pay wages comparable to those from lost jobs.[23]

The high wages of U.S. auto workers reflect higher skills (i.e., greater human capital) relative to other manufacturing jobs, as well other considerations such as the monopoly wage effect, a union wage differential, and pure historical considerations. While these make it difficult to disentangle the skill differential, it is clear that alternative employment opportunities for laid off auto workers will involve considerable less skilled jobs and correspondingly lower wages. The skills embodied in those workers will simply be lost ("stranded").

A similar scenario applies to physical capital dedicated to auto production by the Detroit 3. Those machines and buildings, primarily located in the upper Midwest, often of considerable age and no longer state of the art, nonetheless, represent assets with some remaining productive capability. Yet these too will simply be lost resources to society, since there are almost surely no buyers interested in acquiring and operating such capital.

4. Spillover Effects

The final issue concerns spillover effects onto other sectors and institutions. We have already noted the serious effects that the collapse of auto sales have had on dealers, suppliers, and communities. To that might be added at least the following spillovers from the demise of an auto manufacturer:

- State and local finance. Michigan has the highest concentration of auto-related employment of any state, and the correspondingly lowest rate of job growth in the past decade. Bankruptcy would exacerbate the state's financial problems, at least in the short and medium runs.

- State unemployment insurance funds might be substantially depleted if any of the Detroit 3 were to suffer permanent closure. Even sustained by federal stimulus money, the potential loss of a hundred thousand jobs exceeds the ability of traditional unemployment plans. Health insurance for current and past employees would be jeopardized, with spillovers onto state and local health insurance plans or, worse yet, deterioration of the health of such individuals or major increases in uncompensated care at hospitals.

- The federal pension benefit guarantee fund would be overwhelmed by a huge numbers of individuals in the plans of the Detroit 3, if those plans collapsed and were taken over by the Pension Benefit Guaranty Corporation ("PBGC"). GM's plan covers 673,000 workers, Ford's 332,000, and Chrysler's 255,000.[24] Takeover of any of these plans by the PBGC would result in reduced benefits to these workers, reduced premium payments to the PBGC and, likely, a domino effect as the other companies also handed off their obligations.

- Banks and other creditors of the company, who would be left with little or nothing if any of the major Detroit companies ceased operations.

WHETHER AND TO WHAT EXTENT SUCH "EXTERNALITIES" DESERVE A RESPONSE FROM PUBLIC POLICY IS AN OPEN QUESTION.

Not all of these effects involve true economic externalities. At most, some—like pension guarantees—are really financial externalities; the kind of spillover costs that modern financial institutions by their very nature often create. Others, such as the effects on banks, may simply be effects that such agents should have foreseen, although their failure to do so will have wider repercussions, e.g., such banks will not be able to provide credit, thereby harming other businesses who are not parties to the problem at hand. Whether and to what extent such "externalities" deserve a response from public policy is an open question.

IV. Current Policy and Issues

Federal government assistance for the Detroit 3 began with some interim steps taken in December, 2008, and then far more substantial measures in the first half of 2009. Here we review these two phases of policy, and then discuss some of the on-going and future issues raised by such intervention.

A. THE AUTOMOTIVE INDUSTRY FINANCING AND RESTRUCTURING ACT (2008)

As GM, Ford, and Chrysler slid toward financial crisis during the last quarter of 2008, representatives of the companies testified before Congress that they needed a total of $34 billion in loans to survive the recession. The Bush administration sought to prevent the companies' immediate bankruptcy, but wished to avoid deep intervention that would put their mark on the industry. The result was the Auto Industry Financing and Restructuring Act, which was introduced in December 2008 and provided $14 billion in short-term bridge loans while the companies and the new administration worked out a more permanent plan. The interest rate of these loans would be 5 percent for the first five years and 9 percent thereafter. As a condition of participation, each firm would have to submit a plan for "viability" that required approval by March 31 in order to secure further aid.[25]

Both GM and Chrysler entered the program, obtaining the initial loans that ensured their short-term survival ($9.4 billion to GM, $4 billion for Chrysler). Although Ford was experiencing similar sales declines and losses on operations, it had made a series of financial moves in 2006 that fortuitously provided it with $25 billion in cash and lines of credit. This was far more than its rivals and sufficient to allow Ford to cover its own losses and remain fully independent. This choice would set up an important dichotomy in the U.S. auto sector, with two companies with deep government involvement competing against another that remained private throughout.

As events unfolded in early 2009, both GM's and Chrysler's sales declines and losses exceeded expectations. GM's January sales were 49 percent below the same month in 2008, while Chrysler sales fell 55 percent.[26] These results underscored the need for further interim financing, and indeed in February 2009 GM requested an additional $16.6 billion, Chrysler $5 billion. Both received substantial additional funds prior to the March 31 deadline as the more permanent policy took shape.

B. THE AUTO INDUSTRY FINANCING PROGRAM (2009)

Both GM and Chrysler submitted the necessary restructuring plans in February. The assessments of their adequacy by the new administration were made public on March 30. In each case it was determined that while the company had taken some necessary actions, it ultimately had "not satisfied the terms of the loan

agreement."[27] GM's plan was said to be "in its current form, ...not viable" and required "substantial restructuring." Chrysler's plan was "not likely to lead to viability on a standalone basis," so that the company "must seek a partner in order to achieve the scale and other important attributes it needs to be successful."[28] Each company was provided a modest amount of additional interim financing.

HAVING CROSSED THIS THRESHOLD OF INVOLVEMENT, THE GOVERNMENT PROCEEDED TO TAKE CHARGE OF THE FATES OF THESE TWO COMPANIES.

Having crossed this threshold of involvement, the government proceeded to take charge of the fates of these two companies. It established criteria for further assistance and warned that bankruptcy or acquisition might be required for viability. And it announced several new initiatives to provide assistance to the industry. These were targeted at suppliers, customer warranties, consumer credit, and communities. We discuss these in turn.

1. The Supplier Support Program

The Supplier Support Program ("Program") provides guaranteed payments to GM's and Chrysler's "Tier 1" suppliers, essentially their largest suppliers—companies such as TRW, Lear, Visteon, and American Axle. Under the Program suppliers can sell their eligible receivables to a special purpose facility funded by each automaker and operated by the Treasury Department. This facility ensures payment to suppliers, thereby protecting their financial health and ensuring continued delivery of crucial parts and supplies. The program is funded by fees charged to the car companies (up to 5 percent) and a supplier fee (2 or 3 percent, depending on when they want payment). Total government backing was set at $5 billion.

2. Warranty Commitment Program

Concern over the possible sales-deterrent effect of bankruptcy prompted the establishment of the Warranty Commitment Program. Under this program GM and Chrysler contributed cash and the Treasury Department provided a loan to a facility that would pay for warranty repairs on new vehicles sold during the company's restructuring period. If GM or Chrysler were to go into bankruptcy, a program administrator would identify a qualified provider of warranty services for covered vehicles, with the provider paid out of the fund. The auto companies would contribute 15 percent of the expected warranty cost on each vehicle, and the U.S. Treasury 110 percent. The Warranty Commitment Program was projected to cost the government $1.1 billion. It was terminated in June 2009 as the companies emerged from bankruptcy. It was never used by GM—whether Chrysler did so is unclear—and the entirety of government funds, together with interest, was repaid to the Treasury.

3. American Recovery and Reinvestment Act (2009)

The American Recovery and Reinvestment Act of 2009 ("ARRA") was a further policy response to the contraction of the U.S. auto industry, this focusing on the resulting job losses and community effects. A Director of Auto Recovery was designated as the point person under the Act, charged with coordinating government efforts, ensuring access to ARRA resources, deploying teams to communities facing plant closures, and attracting new industries to affected communities, among other things. These policy actions can be viewed as attempting to slow and ease the transition process as workers and communities, especially in the Midwest, were being forced to move away from their auto-oriented economic base.

C. MATCHING ISSUES TO POLICIES

There is a noteworthy correspondence between the four issues listed above as possibly justifying government intervention in the auto sector and the actual policies that have been implemented:

- Concern over supplier solvency has been translated into a government program to separate the fate of suppliers from the fate of the auto companies to whom they sell. This policy has gone far toward ensuring the viability of the supply sector and avoiding the kind of ripple effects some have been concerned about.

- The fear that consumers might avoid purchasing from a financially-troubled company due to uncertainty about its warranty was addressed by a program separating the security of the warranty from confidence in the manufacturer. While it was never clear how important this concern was in actual practice, nor how effective government backing for a warranty would be, the Warranty Commitment Program served to minimize, if not eliminate, this issue.

- The plight of auto workers and communities that might be stranded by economic dislocation was the focus of the ARRA. While the program may have offered some incremental assistance, the problems of workers and communities seem distinctly non-marginal and largely beyond the scope of this program.

- Enhanced unemployment insurance, extended COBRA provisions, stimulus money to the states, and direct assistance to banks and other lenders all represent policies that, while not specific to the auto sector, served to address various adverse spillovers associated with the sudden downturn in the industry.

With the exception of this last category, the costs of the various programs directed at the U.S. auto sector are summarized in Table 1.

Table 1

Cost of Assistance

Category of Assistance	Government Budgeted Costs
Loans to Auto Makers	$22.9 Billion
Assistance to Finance Companies	$7.4
Supplier Support Program	$5.0
Warranty Commitment Program	$1.1
Total	$36.4 Billion

V. Detroit—Plus Washington, Turin, Ottawa, and the UAW

At the outset of this essay, the question of the role of proper government in private industry was raised. We return to that question, now, but with a focus on the actual intervention by the federal government in Chrysler and, more especially, GM. In addition, we are interested in the important question of competition in an industry now consisting of a majority publicly-owned company in competition with privately-owned rivals. We address these in turn.

A. GOVERNMENT INTERVENTION IN CHRYSLER AND GM

In accepting federal bailout money, both Chrysler and GM subjected themselves to wide-ranging government oversight and influence. In the case of Chrysler the fundamental principle for reorganization was the view that it needed a merger partner to survive. In preparation, the government sought an agreement among Chrysler's various constituencies for a fundamentally changed company and for a division of the costs of such reorganization. When bondholders balked at the proposed terms, Chrysler was put into bankruptcy. Bankruptcy was a process many feared would take years and cause the destruction of the company anyway. In fact Chrysler was in bankruptcy court an astonishingly brief period—exactly 36 days.

What emerged was a transformed company—55 percent owned by the UAW (through its retirement fund, and nonvoting), 35 percent by Fiat, 8 percent by the U.S. government, and 2 percent by the Canadian government. Fiat ultimately could take a controlling stake in the new Chrysler. Substantial ownership by another auto company was crucial to this deal, as the U.S. government sought essentially to hand Chrysler over to a company that would manage its remaining valuable assets and preserve as many U.S. jobs as possible, while proceeding with closure of numerous dealerships and plants.

By making clear that it was prepared to force auto companies into bankruptcy, the government's hand was strengthened in dealing with GM. As was the case of Chrysler, GM entered Chapter 11 and also emerged remarkably quickly—in 40

days. But the plan for GM was quite different than that for Chrysler: the cutting of 34 percent of its already-reduced work force; closure of 37 percent of its dealerships and 13 out of 47 plants; dropping half of its brands; substantial wage concessions; revision of health and retirement benefits; replacement of senior management; and a substantially new board of directors. Perhaps most fundamentally and significantly, in trade for a considerable additional cash infusion, the U.S. government took a 60 percent ownership interest in the new GM. An additional 17.5 percent ownership stake is held by the UAW retirement fund (as in the case of Chrysler, nonvoting), with the remainder divided between the Canadian government and bondholders. The cast-off assets of GM were transferred to a new entity, Motors Liquidation, where they were indeed to be liquidated.

FOR GM, AS WELL AS FOR THE GOVERNMENT, THESE ARE ROLES THAT WOULD HAVE BEEN INCONCEIVABLE AT ANY TIME IN THE PAST.

This plan transforms GM—still the major player in the U.S. auto industry—into a majority publicly-owned company. For GM, as well as for the government, these are roles that would have been inconceivable at any time in the past. They raise, however, some issues with respect to public ownership that are familiar. We address these next.

B. PUBLIC OWNERSHIP AND COMPETITION

Economic theory stresses several possible rationales for public ownership. One line of reasoning notes the possible informational advantages of ownership over regulation as a method of social control.[29] A second argument points to the possible advantages of public ownership in providing services with important noncontractible attributes.[30] There is a significant body of empirical evidence that supports the proposition that public ownership may, under certain conditions, indeed result in superior performance.[31] On the other hand, conventional free-market economics stresses the potential for public ownership to be unresponsive to consumer preferences, inefficient in production, and technologically stagnant. There is considerable evidence in support of that view as well.[32]

Whatever the merits of this last argument, public ownership of GM is not explained by these considerations: It was very much the product of necessity rather than a government plan for deep involvement in the auto industry. Moreover, the government has stated its firm intention to relinquish its ownership stake as soon as practical, although the relevant time horizon would appear to be a few years. Interestingly, there appears to be no economic theory directly relevant to this case of an enterprise subject to public ownership under duress but which will be returned to the private sector. As a result, we must rely upon theory and insights from the case of more permanent and purposeful public ownership in order to frame our discussion here.

There would appear to be two distinct categories of issues that might be raised about public ownership of GM. The first concerns decisions made by and for the

company and, specifically, how they might differ because of its public owners. The second involves the unusual nature of competition between a publicly-owned auto company and other, privately-owned firms. We shall address each of these issues in turn.[33]

1. Decisions Under Public Ownership

As noted above, standard concerns about publicly-owned enterprises are that they are inefficient, out of touch with consumer preferences, and technologically stagnant. Whatever the evidence may be for these concerns in general, none obviously applies to the case of GM. That company's record with respect to costs, quality, consumer responsiveness, and technology are well-known, and that record is distinctly and indisputably inferior. Indeed, it was only government pressure in the context of GM's bankruptcy filing that accomplished in 40 days what that company had not been willing or able to do in the preceding thirty years; namely, to bring its costs substantially under control.

With respect to products and technology, GM's record in the private sector has already been similarly poor. Many of its core products have suffered from inferior design and have been slighted in favor of large low-mileage vehicles. Its reliance upon the latter left GM unprepared for periodic gas price spikes or economic downturns. Moreover, GM has pursued alternative technologies slowly and grudgingly—a strategy that has also left it well behind marketplace changes.

In all these respects the privately-owned GM deserves no praise or credit. Still, there are indications of what other pitfalls may await public ownership. One example concerns the oft-stated interest by the government in ensuring that GM (and indeed, the other companies) transition away from reliance on large vehicles to smaller, fuel-efficient ones. This may be desirable on many grounds, but this preference would seem inconsistent with another government policy—long-standing support for cheap gas. Moreover, it runs the risk of inducing GM and other companies into production of vehicles for which there is insufficient demand (at least after the recession passes).[34]

A second example also underscores concern about meddling with the business decisions of a publicly-owned enterprise. Legislation passed by the House of Representatives forbids termination of GM's and Chrysler's dealers in bankruptcy, requiring instead that any proposed terminations be handled via state dealer laws that, in practice, serve as nearly ironclad obstacles to termination.[35] To the extent that public ownership of GM results in the imposition of product preferences or the creation of impediments to change, such actions would represent a cautionary sign about such ownership. The administration seems cognizant of these hazards, and indeed has put a number of safeguards in place to minimize any interference.[36] Nonetheless, such dangers remain real.

2. Public-Private Competition

The economics literature addresses a further set of questions relevant to the GM case; namely, the nature of competition between a publicly-owned and a privately-owned company in the same market. A majority government-owned GM will be in direct competition with a fully private Ford, a largely private Chrysler, and various other auto companies, mostly privately owned.[37] A common concern in such a mixed setting is that since the publicly-owned firm does not have to maximize profit or even break even, it may set prices lower than those that would be decided by a privately-owned company. Such "unfair competition" seems unlikely here, however, since the institutions for GM's pricing decisions are at some distance from the apparatus of government ownership, and in any case the government has no obvious interest in having GM forgo profits by pricing low. Rather, its interests would seem to be to establish GM as a viable entity so that its stock sale yields revenue to the government.

THERE IS A LITTLE THEORETICAL WORK AND ONLY A MODEST AMOUNT OF EMPIRICAL EVIDENCE CONCERNING THE OUTCOME OF PUBLIC-PRIVATE COMPETITION.

There is a little theoretical work and only a modest amount of empirical evidence concerning the outcome of public-private competition.[38] A study of two Australian airlines, one public and the other private, concluded that the latter had superior operating efficiency,[39] while a study of Canadian railroads found no evidence of inferior performance by the publicly owned competitor.[40] This latter study concluded that competition forced comparable efficiency regardless of ownership—a conclusion shared by a number of subsequent studies.

VI. Conclusion

The difficulties faced by the U.S. auto industry in 2008 were in no small measure of its own making. Longstanding problems were exposed and exacerbated by the enormous decline in demand for autos of all sorts, but especially those produced by GM, Ford, and Chrysler. Thus, while the auto industries in other countries suffered as well, their sales declines were generally more modest and their recovery did not necessarily involve the massive restructuring required just for survival of the Detroit 3. That restructuring has now passed its first milestone as GM and Chrysler emerge from bankruptcy—each with a new majority owner, new managers, a new wage contract, new products, and a mission to succeed in a transformed auto market.

But what of the process by which this was achieved? Should the government have been so heavily involved, first in keeping two of the Detroit 3 afloat, and then by imposing its model on the companies? In fact, are these companies simply too big to fail, for all intents and purposes requiring government intervention when they are at risk of collapse? Or are there other principled reasons for

intervention? Some tentative conclusions about aspects of this question may be ventured.

First, it would seem evident that in modern industrialized economies there are companies whose decisions and operations are strongly affected with the public interest. Assisting such companies when under financial duress raises the widely cited concerns over moral hazard but those concerns do not refute the basic proposition.

Second, this concern arises for a firm (or industry) with some combination of size and interdependence with other sectors of the economy. Even a sizeable free-standing industry (if such can be imagined) might not bring calls for intervention , although a more modest one with an expansive network of suppliers and distributors and with dedicated production facilities and thus limited flexibility is a more likely candidate for intervention.

Third, there are valid economic arguments for government intervention in industries under duress. This review has covered some that are production-related, whereas others are more financial in nature. Particularly the latter deserve careful examination, since the financial institutions of a modern economy would seem to implicate a vast array of agents—more than perhaps intended.

THAT THE ROOT PROBLEMS WERE IN SO MANY CASES THE FAULT OF THE INDUSTRY ITSELF IS NOT IRRELEVANT BUT IT ALSO WOULD SEEM NOT APPROPRIATELY USED AS A TRUMP CARD AGAINST INTERVENTION.

As for the U.S. auto industry, its size and interdependence with other sectors would seem to establish a prima facie case for intervention under the extreme circumstances of the past year. That the root problems were in so many cases the fault of the industry itself is not irrelevant but it also would seem not appropriately used as a trump card against intervention. The challenge will be to ensure that the companies adopt different operating and management strategies so that such intervention is not again required for the foreseeable future. ◀

---

1 This paper is an outgrowth of presentations at the 2009 International Industrial Organization Conference, Boston, and the International Labor Organization Roundtable on Auto Sector Issues, Geneva. Helpful comments from Bruce Lyons and session participants, as well as excellent research assistance by Kathy Downey, are gratefully acknowledged.

2 *See Pain in the Auto Industry Extends Beyond Detroit*, WALL STREET J. (November 21, 2008). For further discussion of the longstanding problems of the U.S. auto industry, *see* Kwoka, *Automobiles: Overtaking an Oligopoly*, INDUSTRY STUDIES, 3rd ed.

3 *U.S. Auto Makers Gain a Bit on Japan's Quality*, WALL STREET J. (November 10, 2006).

4 *In the Chevy Malibu, GM's Pride and Its Challenge*, WASHINGTON POST (July 8, 2009). The importance of reputation effects was demonstrated much earlier but quite dramatically with the GM-Toyota joint venture, which began production of identical cars for the two companies in 1982. Toyota's version—

the Corolla—sold well, while GM's identical vehicle—badged as the Chevy Nova—languished on dealers' lots.

5 *Harbour Report Says Detroit 3 More Productive But Still Losing Money Per Vehicle*, The Auto Channel (June 5, 2008).

6 GM's total health care expenditures reached $5.6 billion in 2008. These totals are sometimes converted into an hourly wage differential, but the reasoning behind that characterization is flawed. *See The Tragedy of General Motors*, Fortune (February 8, 2006); *also*, *$73 an Hour: Adding It Up*, New York Times (December 10, 2008).

7 Between 1997 and 2008, GM invested $310 billion and Ford $155 billion. GM's total depreciation on physical plant was $128 billion, implying a net $182 billion in capital loss. Ford's was smaller—on the order of $8 billion. In each case the companies ended the ten-year period with negligible market value. The invested capital was essentially lost. David Yermack, *Just Say No to Detroit*, Wall Street J. (November 5, 2008).

8 There were apparently some limits to the board's tolerance. When its CEO refused to acknowledge the possibility of bankruptcy, GM's board publicly broke with that absurd view.

9 *Tragedy*, *supra*, note 6.

10 *See Hoping Not to Repeat The Mistakes of the Past*, New York Times (November 22, 2008).

11 As we shall see, this view does not address the worker, community, and social effects of its failures.

12 *Detroit Urged to "Stand Up" Against Poverty*, Michigan Chronicle (August 14, 2009).

13 *An Effort to Save Flint, by Shrinking It*, New York Times (April 21, 2009).

14 Underscoring the degree of acceptance of this argument, even Ford and Toyota—not in financial jeopardy—argued for assistance to GM in order to protect its own suppliers. *See*, *for example*, *The Ripple Effect of a Potential GM Bankruptcy*, Time (November 28, 2008).

15 Ford's CEO has stated that "should one of the other domestic companies declare bankruptcy, the effect on Ford's production operations would be felt within days—if not hours....Ford plants would not be able to produce vehicles." Alan Mulally, testimony before the Senate Banking Committee, November 18, 2008. A senior Toyota executive seconded this concern: "We share many of the same suppliers, so if one of our suppliers has difficulties with Chrysler, GM, or Ford, there's a good chance they are going to have difficulty for us....We don't want anyone going bankrupt." *Toyota: "We Really Don't Want Anyone to Go Bankrupt,"* The Truth About Cars (August 13, 2008).

16 This much is clearly correct. As Chrysler went into bankruptcy, for example, it owed more than $25 million to each of four different suppliers and lesser amounts to many more. It stated that "without a clear timeline for when [its bankruptcy] situation will end and production will resume, [there will be] massive suppliers bankruptcies that will stop Chrysler from resuming production." *Chrysler's Bankruptcy Staggers Affiliates*, Wall Street J. (May 2, 2009). An even clearer example of dependency between supplier and manufacturer is the case of Delphi, itself spun off from GM in 1999. While huge in its own right, Delphi's fate has remained inextricably linked with that of its former parent.

17 The difference with the present case, of course, is that the precipitating event is the collapse of a firm at each stage, diminishing competition and raising prices in the remaining supply chain. Some of the adverse effects may be moderated if the remaining supplier and manufacturer also merge, since that will eliminate the double marginalization but not the monopoly effect that derives from the single remaining (and integrated) firm.

18 Third-party warranty coverage is sometimes available, but it tends to be both difficult and more expensive to purchase.

19 Richard Thaler, The Winner's Curse, Ch. 8 (1992).

20 George Loewenstein et al, Time and Decision, (2003).

21 A *Consumer Reports* survey found that 78 percent of respondents said they were "unlikely" (64 percent said they were "very unlikely") to buy a new car from a bankrupt auto company. ConsumerReports.org. (March 30, 2009).

22 Kwoka & Warren-Boulton, *Efficiencies, Failing Firms, and Alternatives to Merger: A Policy Synthesis*, Antitrust Bull. (Summer 1986).

23 A recent study found that earnings of workers laid off in the 1982 recession remained 15-20 percent below their prior levels essentially indefinitely. Von Wachter, Song, & Manchester, *Long-term Earnings Loss Due to Job Separation Mass-Layoffs During the 1982 Recession*, NBER Working Paper (April 2009). Evidence regarding worker retraining has long been discouraging. *See*, *for example*, *U.S. Study Says Job Retraining Is Not Effective*, New York Times (October 15, 1993).

24 *Plight of Carmakers Could Upset All Pension Plans*, New York Times (April 24, 2009).

25 Other provisions of the Act involve restrictions on executive bonuses and golden parachutes, a requirement for bondholders to exchange bonds for stock, union wage concessions, and elimination of the "jobs bank" that provided full compensation for some laid-off workers, and oversight by a "car czar." At the Bush administration's insistence, nearly half of the loan amount was initially to come from a fund set aside for production of plug-in hybrids. The remainder was diverted from the Troubled Asset Relief Program ("TARP").

26 Ford's sales were 40 percent lower, while Japanese manufacturers posted slightly smaller declines.

27 "GM February 17 Plan: Viability Determination," U.S. Treasury Department, March 30, 2009.

28 "Chrysler February 17 Plan: Viability Determination," U.S. Treasury Department, March 30, 2009.

29 Shapiro & Willig, *Economic Rationales for the Scope of Privatization*, in The Political Economy of Public Sector Reform and Privatization, Suleiman and Waterbury, eds. (1990).

30 Hart, Schleifer, & Vishney, *The Proper Scope of Government: Theory and an Application to Prisons*, Q. J. Econ. (1997).

31 *See*, *for example*, Kwoka, *"The Comparative Advantage of Public Ownership: Evidence from U.S. Electric Utilities,"* Canadian J. Econ. (2005).

32 Megginson & Netter, *From State to Market: A Survey of Empirical Studies on Privatization*, J. Econ. Lit. (2001).

33 We do not address questions arising from worker ownership, since the UAW stake in GM (and in Chrysler) is non-voting. To the extent that enterprise decisions are nonetheless informally influenced by union presence, that would introduce a further complexity (as would the fact of union ownership, rather than direct employee ownership).

34 *See Industry's Big Hope for Small Car Fades*, Wall Street Journal (March 23, 2009).

35 *House Wants Dealerships Reinstated*, Wall Street Journal (July 17, 2009).

36 *Obama May Find It Tough Not to Meddle in GM Affairs*, Reuters (June 1, 2009).

37 Ford itself has issued a public statement asking for a level playing field against a government-owned GM. Ford Statement on GM Bankruptcy Filing (June 1, 2009).

38 Theoretical work by Cremer & Cremer, for example, focuses on competition between a private profit-making enterprise and an employee-owned competitor. *See* Cremer & Cremer *Duopoly with Employee-Controlled and Profit-Maximizing Firms: Bertrand vs. Cournot Competition*, J. of Comp. Econ. (1992).

39 Davies, *Property Rights and Economic Efficiency–The Australian Airlines Revisited*, J. L. Econ. (1977).

40 Caves and Christensen, *The Relative Efficiency of Public and Private Firms in a Competitive Environment: The Case of Canadian Railroads*, J. Pol. Econ. (1980).

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

The Approach to State Aid in the Restructuring of the Financial Sector

Lorenzo Coppi & Jenny Haydock

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

The Approach to State Aid in the Restructuring of the Financial Sector

*Dr. Lorenzo Coppi & Dr. Jenny Haydock**

The unprecedented nature of the financial crisis in autumn 2008 led the European Commission to approve a series of state support measures for the financial sector under Article 87(3)(b), which allows for aid to be considered compatible with the common market if it is "to promote the execution of an important project of common European interest or to remedy a serious disturbance in the economy of a Member State." There was a consensus that the very serious financial crisis constituted a "serious disturbance" to the European economy.

There are minimal precedents on the use of Article 87(3)(b), and therefore the Commission has advanced a framework for the analysis, especially in its communication, *The Return to Viability and the Assessment of Restructuring Measures in the Financial Sector in the Current Crisis under the State Aid Rules*. This paper discusses that communication and the appropriate framework for analyzing aid to the financial sector given under Article 87(3)(b).

*Dr. Lorenzo Coppi is a Vice President in the London office of Charles River Associates, with over twelve years of experience in advising clients in Europe and in the United States on the economics of antitrust in a full range of competition cases. Dr. Jenny Haydock is a Senior Associate in the same office, where she has worked on a variety of cases, including State Aid investigations, mergers, allegations of abuse of dominance, and cartel behavior.

I. Introduction

In this article, we comment on the European Commission's Communication: *The Return to Viability and the Assessment of Restructuring Measures in the Financial Sector in the Current Crisis under the State Aid Rules*, (the "Restructuring Communication"), published by the European Commission (the Commission) on its website on July 23, 2009.

The unprecedented nature of the financial crisis in autumn 2008 led the Commission to approve a series of state support measures for the financial sector under Article 87(3)(b), which allows for aid to be considered compatible with the common market if it is "to promote the execution of an important project of common European interest or to remedy a serious disturbance in the economy of a Member State." There was a consensus that the very serious financial crisis constituted a "serious disturbance" to the European economy. The effects of the crisis persist today and are likely to be felt for some time.

Given their urgency, the measures had to be approved without a full analysis of the State aid being granted, but the Commission set a timeframe of six months to review them and to determine whether they were compatible with State aid rules. The Commission is now in the process of carrying out this ex-post evaluation of the State aid provided to the financial sector.

There are minimal precedents on the use of Article 87(3)(b), and therefore the Commission has advanced a framework for the analysis in various communications. This paper discusses the appropriate framework for analyzing aid to the financial sector given under Article 87(3)(b) and concludes that:

- The recent financial crisis was an event of exceptional severity which had many characteristics of a market failure (major confidence crisis, evidence of panic, irrational behavior, bank runs, market breakdown).
- As a result of these market failures and of the significant risk to the economy that these represented, Member States provided aid to financial institutions which was approved under Art. 87(3)(b).
- Because, from the standpoint of economic efficiency, aid given to remedy a significant disturbance in the economy is much better justified than standard Rescue and Restructuring aid, aid given under Art. 87(3)(b) should be considered using a different approach from that of Rescue and Restructuring aid under Art. 87(3)(c).
- The appropriate approach to evaluating aid given under Art. 87(3)(b) is the Balancing Test indicated by the Commission in its *Common Principles for an Economic Assessment of the Compatibility of State Aid Under Article 87.3*.
- The Balancing Test requires the Commission to weigh the costs of the aid, in terms of market distortions, against the benefits in terms of

financial stability (the remedying of a "serious disturbance"), and not to require structural compensatory measures for their own sake.

- The Commission should distinguish between aid necessary to remedy the market failures of the financial crisis and aid given to banks to cover losses from flawed or risky business models. This article advances a framework for quantifying the proportionate and the additional tranches of aid.

THE COMMISSION SHOULD DISTINGUISH BETWEEN AID NECESSARY TO REMEDY THE MARKET FAILURES OF THE FINANCIAL CRISIS AND AID GIVEN TO BANKS TO COVER LOSSES FROM FLAWED OR RISKY BUSINESS MODELS.

- Banks that received only proportionate aid (or that can repay the additional aid in full) should be considered "structurally sound" (as defined in this article) and should not be required to present a restructuring plan. As illustrated in this article, this criterion is equivalent to one which considers as structurally sound only those banks that had enough capital to withstand the fair economic value of the losses that emerged during the financial crisis (that is, those banks which would have been solvent in the absence of the market failures which characterized the financial crisis).

- While both the proportionate aid and the additional aid should be considered compatible aid under Art. 87(3)(b), it is reasonable to require that they give rise to different levels of compensatory measures.

- Because proportionate aid only addresses an exceptional, systemic market failure, it is unlikely to result in appreciable moral hazard or distortion of competition in the relevant product markets, and thus—at most—only behavioral measures should be required for this type of aid.

- Additional aid, on the other hand, can be thought of as the aid necessary to bail-out financial institutions which were not structurally sound regardless of the crisis. It may be reasonable to impose some level of structural compensatory measures with regard to this type of aid, but these measures should be proportional to the tranche of additional aid granted, rather than to the full amount of aid.

- Even in the case of additional aid, because of the large number of institutions that received some form of aid and because of the specificities of the financial sector, moral hazard is often a more significant concern than distortions of competition in the relevant product markets, which are not likely to be significant.

- This suggests that—even in the case of additional aid—burden-sharing and behavioral measures (which target moral hazard directly) are often more appropriate than asset sales (which attempt to remedy distortions of competition in the relevant product markets).

- In addition to not being necessary, significant asset sales run the risk of endangering financial stability and slowing the return to fully-functional financial markets, thereby jeopardizing the very goal of the aid, and should therefore be considered very carefully.

The plan of the paper is as follows: In the next section we discuss the financial crisis and its market failure features. We then discuss the appropriate approach to evaluating state aid in the context of this exceptional financial crisis and we conclude that the standard Rescue and Restructuring ("R&R") framework developed in the context of Art. 87(3)(c) is wholly inadequate to assess aid given under Art. 87(3)(b). In section III, we propose a framework to analyze aid under Art. 87(3)(b) which is consistent with the Commission's own policy (the Balancing Test), and we devise a rigorous framework to identify the structurally unsound banks, as well as to separate the aid granted into a proportionate and an additional tranche. Section IV discusses the implications for the compensatory measures that the Commission is evaluating and Section V concludes.

II. The Inadequacy of the Standard Art 87(3)(c) R&R Framework to Analyze Aid under Art 87(3)(b)

For some months now, Europe and the rest of the world has been in the grip of a profound and pervasive financial crisis, more severe than any since the 1930s. As the European Commission stated in its April 2009 update of the State Aid Scoreboard, "The world economy is currently experiencing its severest financial and economic crisis in almost a century."[1]

As the Commission has highlighted, the crisis "equally affected financial institutions whose difficulties stemmed exclusively from the general market conditions which had severely restricted access to liquidity . . . the crisis hit also banks that could normally not be considered 'companies in difficulties'."[2] Several financial institutions that entered the crisis in good health saw their financial positions gradually deteriorate as a result of the worsening of the financial crisis and its effects on the real economy.

> IT IS INDEED THE CASE THAT THE FINANCIAL SECTOR IS UNIQUE, BOTH IN TERMS OF ITS ROLE IN THE ECONOMY AS A WHOLE AND, ALSO, IN TERMS OF THE INTERDEPENDENCE OF RIVAL FIRMS.

Faced with this unprecedented crisis, the European Commission provided guidance in the form of various communications to Member States as to the State aid rules applicable to State support for financial institutions during the crisis.[3] The Commission has recognized the exceptional nature of the crisis and the need for an unprecedented response, given the "systemic nature of the crisis" and to the "interconnectivity of the financial sector" which renders it unique.[4]

It is indeed the case that the financial sector is unique, both in terms of its role in the economy as a whole, and also in terms of the interdependence of rival firms.[5] As well as playing a crucial role in the economy as a whole, through the provision of loans and other banking services crucial to the running of any business, firms in the financial sector are interlinked and interdependent in a way that is not the case in other industries: first, because of interbank lending and other interactions; and second, because the reputation of and faith in the whole sector can be shaken by the removal of faith in just one institution.[6]

The Commission correctly stated that—given the exceptional nature of the crisis and the uniqueness of the financial sector—the R&R framework was not appropriate to analyze aid during the financial crisis and that the crisis required a fresh approach to State aid[7] and, for this very reason, the Commission resorted to Art. 87(3)(b).[8]

We agree with the Commission that the framework of R&R under Art. 87(3)(c) is not appropriate to analyze aid that has been given under Art. 87(3)(b). The rationale and context of Article 87(3)(b) is different from that of standard R&R aid. The role of ad-hoc R&R aid is to rescue and restructure firms that would have failed under normal market conditions. As indicated in the R&R guidelines, R&R aid is an extreme measure which usually is not consistent with the efficient functioning of a competitive market (see paragraphs 4 and 8 of the R&R guidelines).

We agree with the Commission that the framework of R&R under Art. 87(3)(c) is not appropriate to analyze aid that has been given under Art. 87(3)(b).

There is broad economic consensus that ad-hoc R&R aid has tenuous justifications from the standpoint of economic efficiency and can, in fact, result in serious market distortions. By simply keeping an otherwise failed firm in business, or making it stronger in the market than it otherwise would be, the aid can create inefficiencies, in which less-efficient firms serve customers who could be more efficiently served by other firms. Furthermore, as with all aid, it can distort future incentives for firms as they anticipate future State aid and create moral hazard. Because of its tenuous justification from the perspective of market efficiency, ad-hoc R&R aid is subject to fairly strict (almost punitive) "compensatory" measures in order to ensure that the normal functioning of competitive markets is not hindered by State intervention.

This is not the case with aid awarded under Article 87(3)(b). This aid has a clear justification: "To promote the execution of an important project of common European interest or to remedy a serious disturbance in the economy of a Member State." A serious disturbance is likely to involve significant market failures; the correction of which is quite justified from an economic perspective since, properly performed, it renders the market more efficient, rather than less so.

Thus, while R&R aid is not consistent with the efficient functioning of markets, aid under Article 87(3)(b) is compatible with economic efficiency and, indeed, this aid attempts to return markets which have been hit by a serious disturbance to a normal and efficient situation. It is important that these observations are always carried forward in order to fully consider their implications in terms of the nature and motives of the aid and, thus, the appropriate action to be taken at the time.

The aims of restoring viability to the financial sector and protecting future financial stability, are sensible and well-justified goals and should constitute the guiding principle of any State aid analysis under Art. 87(3)(b). Of course, there may be costs involved in such interventions, but these may well be outweighed by the positive effects. And it is here that the analysis of such State aid must start, in line with the Commission's own "Balancing Test," as outlined in the Common Principles as an overarching methodology to assess State aid under Art. 87(3): "The assessment of the compatibility of an aid is fundamentally about balancing its negative effects on trade and competition in the common market with its positive effects in terms of a contribution to the achievement of well-defined objectives of common interest."

The application of the Balancing Test serves to ensure that three key objectives stated in the Restructuring Communication—stabilizing the financial system, ensuring that aid is kept to the minimum, and minimizing distortions of competition—do not clash against each other. In the next section, we apply the Balancing Test to the State aid given in the context of the financial crisis.

III. Applying the Balancing Test to Art. 87(3)(b)

The Balancing Test proposed by the Commission to analyze Aid under Art. 87(3) has three pillars, which can be formulated as three key sets of questions:

- Is the aid aimed at a well-defined objective of common interest? Why is the State aid needed? Why can the private sector not deliver the objective?
- Is the aid well designed to deliver the objective of common interest? Is aid appropriate? Is there a positive incentive effect? Is the aid proportionate to the problem tackled?
- Are distortions of competition and trade limited?

In the rest of the section we provide answers to these questions.

A. AID AIMED AT A WELL-DEFINED OBJECTIVE: THE FINANCIAL CRISIS AS A MARKET FAILURE

The Common Principles correctly state that the main economic rationale of State aid is to remedy a market failure (and/or to improve equity and social cohe-

sion). In this case, it is useful to distinguish between two types of market failures: those that lead to a breakdown of financial markets; and the negative externality that a failed financial institution imposes on other financial institutions and, ultimately, on the economy.

As to the market failures that led to the financial crisis, although the specific reasons for and the concatenation of events that led to the financial crisis in autumn 2008 are still being debated, it appears clear that risk mispricing, unrealistic expectations and short-termism, excessive leverage, asset price bubbles, and moral hazard followed by panic have been important elements of the financial crisis.

All these factors can be characterized as "market failures" in economic terms. As a result of these market failures, credit markets and, in particular, money markets had completely broken down in autumn 2008, endangering even structurally sound banks. Banks found their balance sheets interspersed with impaired assets for which there was no market, despite most of these assets having positive value outside of a crisis situation.

As a result of these market failures, credit markets and, in particular, money markets had completely broken down in autumn 2008, endangering even structurally sound banks.

It is a well-recognized problem in economics that information asymmetry may lead to market failure and, possibly, market breakdown. If a proportion of the goods available are known to be of low value, but these goods cannot be differentiated by buyers from goods of higher value, the price that buyers are willing to pay for the goods will fall to the point that no seller in possession of a higher-value product will offer it for sale, and so the market for the goods will shrink and may collapse. This is known as the "lemons" problem, after Nobel laureate George Akerlof's seminal contribution.[10] Thus, even those banks with certain valuable assets found it difficult or impossible to gain an acceptable price for them, because buyers could not distinguish them from so-called "toxic" assets which were of little or no value.

This "lemons" problem quickly spread from affecting single assets to affecting entire financial institutions. As investors became uncertain as to the quality of financial institutions' assets, concerns started to surface about the health and viability of financial institutions. As FED Chairman's Ben Bernanke stated: "At the root of the problem is a loss of confidence by investors and the public in the strength of key financial institutions and markets."[11] Once this crisis of confidence started to develop, investors became unwilling to lend to financial institutions, and financial institutions became unwilling to lend to each other, effectively precipitating the whole financial system into a potentially fatal liquidity crisis. At this point, the "lemons" problem had become a confidence crisis caused by uncertainty over whether the banks would survive.

While the "lemons" problem had, in part, caused the breakdown in financial markets, a second source of market failure exacerbated it and escalated it into a severe crisis. Given the interconnectivity of the financial system, the failure of one bank imposes a negative externality on another bank. In economics, a negative externality means that there is likely to be overprovision of a good if the market failure is not corrected, meaning in this case that banks would fail even when the failure was not efficient, from society's perspective. This second source of market failure is indeed what can potentially turn a financial crisis into a "serious disturbance of the economy," which requires State aid in order to be remedied.[12] However, this externality is a basic feature of the financial markets, and thus not exceptional (what was exceptional was the number of banks that could have failed and thus the potential severity of this negative externality).

Compare these market failures to the typical situation of ad-hoc R&R aid under Art. 87(3)(c), in which markets continue to function normally; there are no specific market failures such as externalities, asymmetric information, coordination failures, or incomplete markets. In that situation, it is typically a single firm—the recipient—that has failed to compete in normal market circumstances. It is clear that the justification for and the wider benefit of aid under those circumstances are significantly less compelling.

In the next section we propose a methodology for identifying aid to remedy the first type of distortion (the "lemons" problem)—which can be considered capital support provided by the state that is short-term in nature and a prudent bolstering of banks' capital positions—from "bail-out aid" to banks that did not have a viable business model (independently of the crisis) and thus should restructure.

B. AID AS A WELL-DESIGNED INSTRUMENT: EFFICACY AND PROPORTIONALITY

The second leg of the Balancing Test assesses whether the aid is a well-designed instrument. This requires aid to satisfy three principles: (i) that aid is an appropriate tool to tackle the market failure(s) identified; (ii) that aid can bring about a solution to the market failure problem; and—most importantly for the purpose of assessing aid under Art. 87(3)(b)—(iii) that aid is proportionate to the problem tackled.

1. Aid Was an Effective Tool to Tackle the Financial Crisis and Achieved its Goals

State aid has been provided to address the market failures discussed above and to avoid banks failing: the guarantees have been granted to avoid bank runs and to allow interbank markets to become more liquid; the impaired asset schemes have been introduced to allow a "fair value" pricing of illiquid assets; and the recapitalizations provisions have been necessary to allow banks to make risk provisions

to cover remaining impaired (or risky) assets in their balance sheets and because banks are now forced to put back more equity capital to cover their loans.

Aid was necessary in order to remedy the market failure and/or to avoid the risk of systemic financial failure. In particular, one of the most effective ways to solve the "lemons problem" is by means of an asset guarantee, as it removes the unobservable risk of default that is the root cause of the breakdown in financial markets.

AID WAS NECESSARY IN ORDER TO REMEDY THE MARKET FAILURE AND/OR TO AVOID THE RISK OF SYSTEMIC FINANCIAL FAILURE.

There is a consensus that these instruments were appropriate, and that the State support measures have been successful to bringing some degree of stabilization to the financial markets (even though financial markets are not yet completely back to normal). The first two criteria of the "appropriateness" leg of the Balancing Tests are therefore satisfied.

2. The Use of the Proportionality Principle to Differentiate Between Different Types of Aid

The proportionality principle requires that aid is kept to the minimum necessary to achieve its benefits. The reasoning behind this is perfectly legitimate. It is important to make sure that State support is not used to bolster a bank's financial position beyond what is required by the current market circumstances, and that it is not used to pay shareholders. In other words, even a serious crisis does not permit the writing of a "blank check."

In this context it is important to distinguish the aid that was proportionate to solve the crisis-specific market failures, from the additional aid that was necessary to rescue banks that would have been unviable in any case. While both types of aid can be provided "to remedy a serious disturbance in the economy of a Member State," the aid proportionate to solving the breakdown in financial markets (the "lemons problem" and the ensuing crisis of confidence) should be sufficient to allow structurally sound banks to return to viability, will result in only limited distortions of competition, and thus should not be subject to compensatory measures.

Any additional aid would be rendered necessary by the excessively risky actions of banks, rather than by the failure of the financial markets. Even if this additional aid was necessary to avert a potential catastrophe, it must be considered that the reason why this second type of aid was necessary was that some banks had overstretched themselves, and thus had contributed to creating the breakdown in financial markets in the first place. Thus this aid may warrant some compensatory measures, to avoid significant distortions of competition.

3. A Practical Approach to Applying the Proportionality Principle

A practical way to apply this proportionality principle, to distinguish between different tranches of aid, would be the following:

- Start from the bank's balance sheet before the financial crisis.
- The Committee of European Banking Supervisors (CEBS) and the European Central Bank (ECB) should establish the parameters of a "no financial crisis" market scenario.
- Value the bank's assets on the basis of commonly agreed "fair economic value" methodologies—this should result in a certain amount of required write-downs from the pre-crisis market value of the bank's asset base (the "fair economic value losses").
- For that part of the asset base which affects regulatory capital, calculate the difference between the value of the asset base during the crisis and the pre-crisis fair economic value of the asset base (the "market failure losses").[13]
- Calculate the amount of capital needed to bring the bank's capital from the previous regulatory minimum to the new level required by the market as the result of the crisis of confidence (the "crisis of confidence capital increase").[14]

The sum of the "market failure loss" and the "crisis of confidence capital increase" is the amount of aid proportionate to remedy the breakdown in financial markets. Any aid additional to this amount should be considered additional aid which the bank can either repay, or—if it cannot be repaid—it should be the basis on which structural compensatory measures should be calculated. As a corollary, if the bank can repay this additional aid, then it should be considered a structurally-sound bank that does not need to restructure. This is equivalent to saying that a bank is structurally sound if it had enough surplus capital (in addition to the regulatory minimum) to cover the fair economic value of its losses.

THE SUM OF THE "MARKET FAILURE LOSS" AND THE "CRISIS OF CONFIDENCE CAPITAL INCREASE" IS THE AMOUNT OF AID PROPORTIONATE TO REMEDY THE BREAKDOWN IN FINANCIAL MARKETS.

A stylized example may help clarify this point. Imagine that before the crisis a bank had 100 in assets which had a market value of 100, and 100 in liabilities, of which 17 was in capital, well above the regulatory minimum of 5. Assume that after September 2008, the assets' market value fell to 70.[15] The reasons for this fall in value were two-fold. First, the assets were likely overpriced to some degree before the crisis; that is, they exceeded their fair economic value. We might therefore imagine that the fair economic value of the assets was, in fact, 90. However, the remainder of the fall in the value of the assets was a reflection of the market failure identified above; namely, that a "lemons" problem meant that potential purchasers were

unable to gauge the true value of the assets—the market had broken down to some degree—and assets were undervalued by 20 (this can be considered the "market failure loss"). In addition, as the result of the systemic crisis of confidence discussed above, the market was uncertain whether the bank would survive and thus required it to hold at least 10 in capital (in addition to any capital necessary to absorb the likely losses), instead of the regulatory minimum of 5.[16]

Assume that, in order to tackle the crisis, the State provided a recapitalization of 30. The proposed test would indicate that the amount of aid proportionate to remedy the breakdown in financial markets would be 25; that is: (i) the difference between the fair economic value of the assets pre-crisis, 90, and the market value of the assets during the crisis, 70, plus (ii) the 5 needed to bring the bank's capital from the previous regulatory minimum to the new level required by the market, 10, as the result of the crisis of confidence. In addition, there would be an amount of additional aid equal to 5 (the 30 of aid granted minus the proportionate tranche of 25).

In this specific example, the bank still has 17 of capital (as the 30 in losses were entirely covered by the State aid), which includes 12 of capital above the regulatory minimum ("surplus capital"). This amount of surplus capital is enough to cover the fair economic value losses of 10 (that is, the difference between the pre-crisis market value of the assets, 100, and their fair economic value, 90). This implies that the bank is structurally sound and it would have been able to cover its losses and survived in the absence of the market failures discussed above. Another implication of the bank having enough surplus capital, and thus being structurally sound, is that it would be able to repay the 5 of additional aid while maintaining the new minimum capital requirement of 10 (in fact—with 12 in capital after repaying the additional aid—the bank would even be above the new regulatory minimum). Thus a structurally sound bank only needed aid proportionate to remedy the temporary breakdown in financial markets.

THIS APPROACH DOES NOT NECESSARILY MEAN THAT A STRUCTURALLY SOUND BANK WOULD BE ABLE TO REPAY THE ENTIRE AID (INCLUDING THE PROPORTIONATE TRANCHE) ONCE THE CRISIS IS OVER AND THE MARKET VALUE OF THE ASSETS RISES TOWARDS THEIR "FAIR VALUE."

In contrast, we can imagine a different example, in which a bank had only 10 of capital before the crisis—that is, 5 of surplus capital—and thus would not be able to cover all the fair economic value losses with its surplus capital, or to repay any of the additional aid. This second bank should not be considered structurally sound; it should be required to present a restructuring plan, and compensatory measures should be considered, although limited to the tranche of additional aid (i.e. 5).

This approach does not necessarily mean that a structurally sound bank would be able to repay the entire aid (including the proportionate tranche) once the

crisis is over and the market value of the assets rises towards their "fair value." In fact, the financial crisis, by spilling over to the real economy and by throwing it into one of the most severe recessions of the last fifty years, has changed the economic outlook and, thus, the fair economic value of the assets. This effect should also be ascribed to the market breakdown and—as we explain in the next section—should not result in a requirement to consider compensatory measures. To the extent that the market failures addressed by the aid are temporary in nature, it would be reasonable to expect that the State should be able to claw-back a certain amount of aid as the market failures diminish, allowing the value of the assets to climb towards their fair economic value. As the crisis of confidence eases, banks can go back to more efficient capital adequacy ratios.

Note that, for the sake of simplicity, we have assumed that the 30 in aid was a pure grant, without any form of remuneration or equity participation for the state. To the extent that some remuneration was received by the state (including equity participation), the value of this remuneration should be netted out of the aid, as financing should be considered aid only to the extent that it exceeded the value of the remuneration or equity participation.

AID RENDERED NECESSARY BY THE CRISIS ITSELF IS UNPROBLEMATIC; WHILE AID RENDERED NECESSARY BY THE RECKLESS ACTIVITIES OF CERTAIN BANKS MUST BE SUBJECT TO FURTHER SCRUTINY.

The upshot of this analysis is much the same as the Commission's position: Aid rendered necessary by the crisis itself is unproblematic; while aid rendered necessary by the reckless activities of certain banks must be subject to further scrutiny. As a result, prudent banks will face fewer, if any, restructuring or compensatory measures than more reckless banks. But this analysis makes it clear that the relevant issue is the cause of the aid (to remedy a market failure or to cover real economic losses), rather than the form or the simple amount of the aid received—although the higher the amount of aid, the more likely it is that there is at least some additional aid. A more reckless bank will have received more additional aid than a less reckless bank, and some more prudent banks will have received no additional aid.

4. Using the Proportionality Principle to Distinguish Between "Structurally Sound" and "Structurally Unsound" Banks

The Commission's communications clearly identify the need to distinguish between banks that are "fundamentally sound" and whose difficulties stem exclusively from the general market conditions and those banks whose structural solvency problems are linked to their particular business models or investment strategies. It should be remembered that the purpose of such a distinction is to assess how best to respond to State aid measures, rather than to simply identify profitable and unprofitable banks under current circumstances (although such an analysis need not be irrelevant). The key is the soundness or otherwise of banks

at the point at which State aid was granted, since this sheds light on the motives of the aid.

From a policy perspective, it is important that those banks which have engaged in overly risky investment strategies or that had unsound business models are not allowed to receive aid without any compensatory measures being imposed on them. If no compensatory measures were imposed, the aid would create a significant moral hazard problem and would risk fostering another crisis like the one we are experiencing. Thus, this approach of differentiating between banks is—in principle—reasonable. However, the manner in which the Commission has made this distinction is too simplistic, and does not go to the heart of the purpose of such an analysis.

FROM A POLICY PERSPECTIVE, IT IS IMPORTANT THAT THOSE BANKS WHICH HAVE ENGAGED IN OVERLY RISKY INVESTMENT STRATEGIES OR THAT HAD UNSOUND BUSINESS MODELS ARE NOT ALLOWED TO RECEIVE AID WITHOUT ANY COMPENSATORY MEASURES BEING IMPOSED ON THEM.

First, it is too simplistic to consider all banks that have received certain types of aid, a certain amount of aid, or have received aid in different tranches, as necessarily "unsound," and all others as "sound."[17] While these simple screening devices may be useful to identify cases for more in-depth review, they should not constitute a presumption that the recipient is structurally unsound and would not have survived, even in the absence of the widespread market failures which characterized the financial crisis.[18] It is important to look more closely at the motives behind the aid and the reasons for its provision and, in particular, at whether it was justified on the basis of the market failures that led to the liquidity crisis and the general breakdown in confidence.

Second, and more importantly, even if a bank is not "structurally sound," it does not follow that all the aid given to the bank in question must be considered distortionary. Rather, some of the aid may very well be justified on the basis of the general market conditions and of the liquidity crisis, even if that amount would not have sufficed to maintain the solvency of the bank. Only the aid above and beyond what was justified on the basis of the breakdown of financial markets should be considered aid given for the purpose of sustaining a "structurally unsound" bank, and thus be subject to closer scrutiny.

The test we propose in this article provides a more analytical approach than that currently applied by the Commission, and it is very simple:

"A bank should be considered structurally sound if, at the time of the crisis, it had sufficient "surplus capital" (i.e. above the regulatory minimum) to

> absorb all future "fair economic value losses" (calculated adopting a scenario in which the financial crisis has not considerably affected the real economy and thus the economic outlook). If this test is passed, it must mean that a bank would have survived absent the market failures which characterized the recent financial crisis, and should therefore be considered structurally sound."

Only banks which did not have enough capital above the regulatory minimum to cover the fair economic value losses should be considered structurally unsound. These banks would have received some amount of additional aid and would not be able to repay it. Therefore, once the additional aid has been identified, the question must then be asked whether the bank can, or will be able to, repay that additional aid. If it can, we might consider that aid "erroneously" given, quite understandably, with a desire to ensure the bank's survival, but to an extent that actually overestimated that bank's exposure. The bank is still sound, it would have remained sound were it not for the "serious disturbance," and it should therefore not be expected to restructure or pay compensatory measures. There is also no reason not to allow the bank to choose the method of repayment: from its own resources, from those of debt holders, or from an asset sale on the open market.

On the other hand, if a bank did not have enough surplus capital to cover the fair economic value losses, it would not be in a position to soon repay the additional aid while remaining viable, and we must therefore consider that the additional aid was necessary because that bank had invested heavily in risky, overpriced assets. It should be given the possibility to repay—through its own choice of method—as much of the additional aid as possible while remaining viable. Then, as under the R&R guidelines, such a bank should be susceptible to demands to restructure and/or provide compensatory measures, but only to reflect the size of the remaining additional aid.

The reason for the different treatment of these two tranches of aid (the proportionate aid and the additional aid) lies in the different amount of distortions that these types of aid are likely to have, as we explain in the next section.

C. DISTORTIONS OF COMPETITION ARE LIMITED

It cannot be denied that State aid has potentially distortionary effects on the market, nor that economic efficiency demands that such distortions be minimized. The Communication points to various potential market distortions arising from State aid given to the banks: it may reinforce the market position of the aid recipient relative to that of its unaided competitors; it may help perpetuate failed business models; it may reduce the incentive to compete; and it may create moral hazard by encouraging excessive risk taking.

These distortions of competition may result in various types of market inefficiencies: allocative and productive inefficiencies (as non-efficient banks are shielded from competition); dynamic inefficiencies (as incentives to compete are reduced); and risks to financial stability (as the result of moral hazard).

It is quite correct for the Commission to be concerned about potential market distortions caused by State aid; indeed this is the key reason why, in economic terms, State aid can be problematic. In the Commission's own Common Principles, distortions of competition (along with effects on trade) are given as the negative effects of aid which must be weighed against the positive effects in the Balancing Test. While, in this case, the positive effects of avoiding an economic catastrophe must be considered to exceed any potential costs, it is not unreasonable to consider the distortions involved. Furthermore, when considering the appropriate measures to be taken in response to a finding of additional aid, the need to minimize distortions should be considered.

It is useful to distinguish between two main types of distortions of competition: those arising from moral hazard; and those arising from potential distortions of competitions in the product market(s).

1. Moral Hazard

In the context of financial markets, moral hazard is often identified as the most significant distortion that may be generated by the State aid. As several commentators have put it, there is the possibility that aid could "sow the seeds of the next crisis."[19] It is clear that an implicit promise of any aid in future may affect firms' incentives going forward. In particular, this promise may result in moral hazard, which arises when a firm or individual is protected from the "downside" of its risks, incentivizing inefficiently risky behavior. We believe that such possible moral hazard distortions are the most significant potential market distortion arising from the aid in this case.

IN THE CONTEXT OF FINANCIAL MARKETS, MORAL HAZARD IS OFTEN IDENTIFIED AS THE MOST SIGNIFICANT DISTORTION THAT MAY BE GENERATED BY THE STATE AID.

The tranche of proportionate aid should not give rise to very significant moral hazard. This is because that aid is exceptional in nature, as it is only justified on the basis of a very unusual complete breakdown in financial markets. Thus, proportionate aid will only affect banks' expectations that, should another complete breakdown in financial markets arise in the future, aid of a similar magnitude could be granted to banks again. But banks should expect that, in normal circumstances, only the standard approach to bank restructuring will be applied. It is not clear why this set of expectations should reduce the incentives for dynamic competition or increase moral hazard. It is recognized that the severity of this crisis was exceptional, and that the government response to the crisis was exceptional as well. Proportionate aid should not substantially affect the way

banks behave in normal market circumstances, since it would not be seen as a precedent for intervention in such normal circumstances.

Banks may still expect that—should they fail in the absence of a complete breakdown in financial markets—additional aid would be provided to them in order to avoid the negative externality arising from the interconnected nature of financial institutions, and this may fuel moral hazard. In fact, the banks already had these expectations, and it was the realization that such expectations might be unfounded that made the bankruptcy of Lehman Brothers such a traumatic event for the financial system. Nonetheless, we discuss in section IV how carefully-considered and appropriate compensatory measures can be used to minimize the risk of moral hazard from additional aid.

2. Limited Distortions of Productive Efficiency in the Relevant Product Markets

The second type of distortion of competition is the productive inefficiency arising from allowing inefficient players to survive and to maintain their market share. In the case of systemic aid under Art. 87(3)(b), this source of distortion is less important than moral hazard. The aid received by banks, and especially proportionate aid, does not automatically result in a distortion of competition. The need for proportionate aid arises from the market failure that affected all banks; by definition, it is symmetric in nature.

IN THE CASE OF SYSTEMIC AID UNDER ART. 87(3)(B), THIS SOURCE OF DISTORTION IS LESS IMPORTANT THAN MORAL HAZARD.

Unlike normal ad-hoc R&R aid, the entire sector has benefited from government intervention and aid was generally available to every bank that demanded it, and thus a level-playing field was largely preserved. This was especially the case when government intervention took the form of guarantee, asset purchase, or recapitalization schemes open to all banks operating in a Member State. In many cases banks were even encouraged to participate in aid schemes, regardless of their true need to receive it. Schemes open to all banks are, by nature, likely to be considerably less distortionary than aid reserved to a sub-set of institutions, who may then be unfairly advantaged in the market.[20]

Even when recapitalization has been carried out on an ad-hoc basis, it does not necessarily confer an advantage over competitors, given the strings attached to the State aid. Banks are often wary of accepting public money if they can avoid it, as they fear it will open the door to more public scrutiny of their policies and strategies. Most European banks that were in a position where they could opt out of the recapitalization and asset purchase schemes chose to do so. Ten U.S. banks have repaid the aid that they have received under the U.S. Troubled Asset Relief Program, ("TARP"). This illustrates that those schemes need not confer a com-

petitive advantage (otherwise all major banks would choose to participate in the schemes if possible).

In addition, to the extent that the aid actually remedied a systemic problem the banks faced—rather than simply being a hand-out—the aid arguably limited any negative impact on efficiency, in the sense of keeping inefficient firms alive. By remedying the problems caused by write-downs and a lack of faith in the financial sector, the aid actually removed some of the inefficiencies which had rendered it necessary in the first place. In this sense it was not the same as aid used, for example, to keep alive an inefficient manufacturer, which would lead to potentially significant productive inefficiencies.

This is indeed the very nature of aid under Article 87(3)(b): by remedying a true market failure, rather than "papering over the cracks" of a firm's failings, it may not create an efficiency imbalance in the market. Thus it is reasonable to conclude that there are no significant distortions in the relevant product market(s) associated with proportionate aid under Article 87(3)(b). It is interesting to note that much of the economics commentary on the banking crisis has focused on the need to minimize the cost to taxpayers, and the need to avoid moral hazard and thus a repeat of the crisis, with much less consideration given to the potential for productive inefficiencies or distortions of competition between market participants.[21]

This may be different in the case of additional aid, as—by its nature—it is aid given to rescue a structurally unsound bank. However, it is clear that, by preventing the collapse of large interconnected banks, the aid has avoided a disaster for the European banking sector. Unlike in the usual case of R&R aid ex Art. 87(3)(c), aid given under 87(3)(b) has an immediate and tangible benefit on competitors—in this case by preserving the stability of the financial system and avoiding the domino effect of bank failures.

THE COMMISSION CANNOT THEREFORE SIMPLY ASSUME THAT THE SHEER FACT THAT SOME BANKS NEEDED AID WHILE OTHERS DID NOT INDICATES THAT THERE WAS A DISTORTION OF COMPETITION IN THE RELEVANT PRODUCT MARKETS.

The Commission cannot therefore simply assume that the sheer fact that some banks needed aid while others did not indicates that there was a distortion of competition in the relevant product markets: a careful and thorough analysis of the actual distortions of competition needs to be carried out in each case. In the case of additional aid, the assessment of whether the aid reinforces a recipient's market power needs to be made by reference to the competitive conditions in the particular markets in which the recipient is active, and requires a detailed analysis of: market definition, the recipients' market positioning and that of their rivals, barriers to entry and expansion, the presence of any friction in the market, and the degree of rivalry between market participants. In other words, in light of the externality that

interconnected banks impose on each other, a proper analysis needs to be carried out and distortions of competition in the product market(s) cannot simply be assumed.

In conclusion, proportionate aid is unlikely to result in a significant distortion of competition, but it is possible that some additional aid may create moral hazard and some distortions of competition in the relevant product market(s). Given the systemic nature of the crisis and several particular features of the financial systems, such distortions of competition are likely to be limited, and primarily related to moral hazard.

D. CONCLUSIONS ON THE BALANCING TEST

There can be little doubt that the positive effects of the aid outweigh the negative, distortionary effects, given the importance of avoiding a serious economic catastrophe. Furthermore, a proportion of the aid must be considered to pass the Balancing Test by being well-designed and proportionate to remedying a breakdown in financial markets. According to the Commission's own Common Principles, this implies that that aid should be compatible under Art. 87(3). There are two corollaries of this conclusion: Compatible aid does not require compensatory measures (see paragraph 73 of the Common Principles); and Compatible aid does not need to be repaid, or at least not immediately, as discussed in more detail in the next section.

We do however acknowledge—in line with the Commission's thinking on this issue—that additional aid used to rescue structurally unsound banks should be treated differently from proportionate aid given to remedy the breakdown in financial markets. Since additional aid was rendered necessary by the risky activities of the recipient bank, as opposed to the market failure which prompted the use of Article 87(3)(b), and to the extent that it is not repaid, it must be considered that it has more distortionary effects than proportionate aid. These distortionary effects may be mitigated by certain compensatory measures.

IV. The Implications for Compensatory Measures

The differences between ad-hoc R&R aid under Art. 87(3)(c) and stabilization aid under Art. 87(3)(b), as highlighted in the Communication, have important implications for determining the appropriate compensatory measures.

A. STRUCTURAL COMPENSATORY MEASURES ARE NOT APPROPRIATE IN THE CASE OF PROPORTIONATE AID

Structural compensatory measures (such as divestments and reductions in capacity) might have a place in ad-hoc R&R aid ex Art. 87(3)(c) as the rescued firm should have exited the market as a result of the normal exercise of market forces and, thus, competitors should be "compensated" for a rival remaining in the mar-

ket. Further, all firms must be diverted from the moral hazard associated with anticipating that they will be "saved." We note—however—that the Commission's Economic Advisory Group on Competition Policy ("EAGCP") has recently commented on R&R aid, noting that compensatory measures should serve to minimize distortions (moral hazard and "competitive externalities"), rather than being aimed *per se* at "compensating" competitors.[22] We agree entirely with this position.[23]

Nevertheless, structural compensatory measures are not justified in the case of proportionate aid under Art. 87(3)(b) where banks would not have failed had normal market forces continued to operate. This is implicitly recognized by the Commission in its assessment in the Restructuring Communication that only certain banks need to engage in "more substantial restructuring," and that such a measure is designed to "restore viability." There are at least three reasons why compensatory measures are not justified for proportionate aid under Art. 87(3)(b).

THERE ARE AT LEAST THREE REASONS WHY COMPENSATORY MEASURES ARE NOT JUSTIFIED FOR PROPORTIONATE AID UNDER ART. 87(3)(B).

First, to the extent the Balancing Test has shown that the aid is compatible with Art. 87(3)(b), the Commission has no justification or power to demand compensatory measures. Second, even if the Commission had the power to impose them, compensatory measures might be conceivable only when a bank has benefited from the aid in a manner which is disproportionate with respect to benefit and support for the financial sector as a whole, whereas—in this case—the proportionate aid is common to most banks. Third, even if they were justified, it is not clear that structural compensatory measures are necessarily consistent with achieving the goals of the aid under Art. 87(3)(b), as stated at paragraph 2 of the Communication: (i) attain financial stability and maintenance of credit flows; (ii) limit distortion of competition and effects on trade; and/or (iii) limit moral hazard and maintain banks' competitiveness.

If concerns remain about distortions of competition—primarily driven by moral hazard issues—the fact that proportionate aid was rendered necessary by a sector-wide market failure leading to a sector-wide crisis means that regulation and behavioral compensatory measures, rather than mandated asset sales or other structural compensatory measures, would be most appropriate.

B. COMPENSATORY MEASURES MAY BE JUSTIFIED IN THE CASE OF ADDITIONAL AID BUT MUST BE DETERMINED CAREFULLY

We have explained that part of the aid granted may constitute additional aid and, as such, it may have more distortionary effects than proportionate aid. It may therefore be reasonable to try to minimize the distortionary effects by imposing some compensatory measures on banks which have received substantial addi-

tional aid, provided that these measures do not endanger the goal of achieving financial stability by returning banks to viability.

We emphasize that any such measures should apply only to the additional aid; a finding of additional aid should not mean that all the aid granted to an institution becomes susceptible to the same compensatory measures. It would be unreasonable to treat banks that received very small amounts of additional aid as harshly as banks that received large amounts of additional aid, even using the excuse that any level of additional aid means that the financial institution was kept alive by the aid, and that the market should be brought back to the "no aid counterfactual" in which the bank would have been liquidated. Consistent with the EAGCP recommendation, we believe that compensatory measures should only be undertaken to remedy as much as possible the loss in efficiency that the aid generated, and thus what is important is not simply the "no aid counterfactual" but the difference—in terms of departure from economic efficiency—between the situation generated by the aid and the no aid counterfactual. It is clear that keeping alive a very inefficient player (which requires large amounts of additional aid) creates a significantly larger departure from economic efficiency than keeping alive a marginal player (which requires very small amounts of additional aid), and thus the latter should be subject to significantly fewer compensatory measures.

C. ASSET SALES ARE UNLIKELY TO BE THE MOST EFFECTIVE COMPENSATORY MEASURE

While other burden-sharing measures may, to some extent, address moral hazard (which, we argue, is the most significant potential distortion of competition), it is difficult to see how compensatory measures involving asset sales can efficiently achieve this goal.

WHILE OTHER BURDEN-SHARING MEASURES MAY, TO SOME EXTENT, ADDRESS MORAL HAZARD (WHICH, WE ARGUE, IS THE MOST SIGNIFICANT POTENTIAL DISTORTION OF COMPETITION), IT IS DIFFICULT TO SEE HOW COMPENSATORY MEASURES INVOLVING ASSET SALES CAN EFFICIENTLY ACHIEVE THIS GOAL.

Asset sales tend to affect most directly the current shareholders of the bank. It is reasonable that shareholders bear the brunt of the losses incurred by banks. However, of all the stakeholders, this group is the one which is likely to be the least subject to moral hazard, for at least three reasons. First, asset sales target the current shareholders of a bank, which need not be the owners who were in place before and during the crisis. Numerous banks throughout Europe have changed hands in recent months, some now being partially or entirely state-owned. It is not clear that measures which are felt by those who did not own the banks while the risky behavior at issue took place will have a strong effect on moral hazard going forward.[24]

Second, shareholders have suffered significantly as a result of the crisis,[25] and while aid may have salvaged some shareholder value, the cost of such aid has been significant, so it is difficult to see how shareholders would be prone to significant levels of moral hazard.

Third, and perhaps more importantly according to many commentators, the most significant source of moral hazard has not come from distorted incentives for shareholders, but rather from distorted executive incentives and failings in bank's governance, which encouraged the pursuit of short-term profits and risk taking and which existed—and continue to exist—independent of any State aid.[26] Behavioral compensatory measures that align executives' incentives to the long-term profitability and viability of banks may be the best solution to the moral hazard problem, but these need to apply to all banks—sound and unsound, and regardless of whether they received aid—and thus should be imposed through sectoral regulation rather than on an ad-hoc basis using State aid law.[27]

Therefore, asset sales are unlikely to be the best way to tackle moral hazard while maintaining financial stability. One might take a somewhat different view of burden sharing which targets debt holders (particularly subordinated debt holders). Burden sharing may address moral hazard on the part of subordinated bondholders if the restructuring forces them to convert their bonds into stocks. Since, in many cases, the aid meant these debt holders kept all of their investment and continued receiving interest, it is important to consider the moral hazard they face. To the extent that these debt holders have a direct influence on banks' behavior, burden sharing to minimize the moral hazard they face going forward may be justified and effective. As to structural compensatory measures, the sale of assets would not directly impact debt holders or bank executive compensation.

As well as being ineffective in tackling the most significant source of potential inefficiencies, asset sales may also be damaging to the Commission's overall goal of financial stability.

As well as being ineffective in tackling the most significant source of potential inefficiencies, asset sales may also be damaging to the Commission's overall goal of financial stability. This is for several reasons. First, mandatory asset disposals may actually worsen a bank's solvency or future solvency if there is not a corresponding reduction in liabilities, assets are sold below book value, or the sales price is materially below the value of foregone earnings. Achieving the right balance between a combined disposal of assets and liabilities and ensuring the bank's solvency and viability is very difficult. In the current market circumstances, banks would most likely have to divest their most profitable assets, which would reduce the bank's ability to be viable and improve its solvency by retaining earnings. The result of assets sales would thus likely run counter to the aid's objectives.

A particular problem from the point of view of the bank sector is that—unlike any other sector—competitors can take on the divested assets only if they can

raise a corresponding amount of capital to maintain their capital adequacy ratios at a prudent level (which at the moment is above the minimum regulatory level). This tends to reduce the ability of competitors to take on divested assets, and thus to be "compensated." This is particularly important given that many banks suffer from re-ratings of their Risk Weighted Assets due to more prudent risk management policies, deteriorating asset prices, and the wave of downgrades of bonds by rating agencies. An added complication of compensatory measures during a systemic crisis is that there are many sellers and few buyers, so it may be difficult to sell a significant portion of assets without depressing their prices to a point which might create another financial crisis.

Another distortion of competition typical of the usual R&R aid ex Art. 87(3)(c) is that State aid may sustain the recipient's output and this may displace ("crowd out") the output that would have been provided by the recipient's competitors. For aid under Art. 87(3)(b) to have a "crowding out" effect, it must be the case that the aid recipient's rivals have the capacity and the willingness to increase lending. These conditions are not met in much of the European financial sector, as the credit contraction has limited banks' ability to lend. Wholesale funding markets still do not allow refinancing of long term wholesale funding and banks therefore need to rely heavily on the European Central Bank ("ECB") for their liquidity. This is likely to make asset sales even more difficult and closer to fire sales.

PERHAPS MORE IMPORTANTLY, GIVEN THE MARKET CONSTRAINTS ON THE ABSORPTION OF DIVESTED ASSETS, IT IS VERY LIKELY THAT COMPENSATORY MEASURES WILL RESULT IN A REDUCTION IN THE LEVEL OF THE ASSETS AVAILABLE IN THE MARKET OVERALL.

Perhaps more importantly, given the market constraints on the absorption of divested assets, it is very likely that compensatory measures will result in a reduction in the level of the assets available in the market overall. As assets constitute, for the most part, short and long-term loans provided by the banking sector to the economy, this would have exacerbated the monetary contraction which is already very serious, potentially damaging the opportunity of recovery in the real economy.

V. Conclusions

In conclusion, while we think that some behavioral compensatory measures can be efficiently imposed on banks that received aid under Art. 87(3)(b), structural compensatory measures should only be considered with regard to the tranche of the additional aid; that is, that aid that was above and beyond what was necessary to remedy the effects of the market failures that lead to a breakdown of financial markets ("the proportionate aid").

Even in this case, other burden-sharing measures and measures focusing on governance and executive pay may be more efficient than asset sales in addressing the main distortionary effect of the aid: moral hazard. Finally, asset sales risk undermining the key goal of the aid granted under Art. 87(3)(b)—returning banks to viability and stabilizing the financial system—so they should only be imposed only when there is compelling evidence of distortions of competition in the product market(s).

1 *See* page 3 of the Scoreboard.

2 *See* page 8 of the Scoreboard.

3 The "Banking Communication" (October 2008), the "Recapitalisation Communication" (December 2008), the "Impaired Assets Communication" (February 2009), and the "Restructuring Communication" (July 2009).

4 As referred to in the Restructuring Communication.

5 For a discussion of the uniqueness of the financial sector and its implications for State aid policy, *see* Bruce Lyons, *Competition Policy, Bailouts and the Economic Crisis*, CCP Working Paper 09-4, 2009.

6 Another source of uniqueness of the financial sector is that banks have inherently unstable balances sheets with long-term assets (loans and investments) and short-term liabilities (deposits), which make them particularly vulnerable to crises of confidence.

7 "The general erosion of confidence within the banking sector in October 2008 led to serious difficulties in accessing liquidity. The crisis had become systemic and equally affected financial institutions whose difficulties stemmed exclusively from general market conditions severely restricting access to liquidity. It thus became doubtful whether the R&R Guidelines were still providing an appropriate framework to tackle the crisis, as the crisis also hit banks that could normally not be considered "companies in difficulties." Furthermore, urgent structural action became necessary in many cases." Spring 2009 update of the State Aid Scoreboard, page 8.

8 This framework is rarely applied; it is significant that there is no body of case law defining the relevant criteria to be applied under Art. 87(3)(b).

9 A market failure is a situation in which the market alone fails to provide the optimal level of a good or service (that is, a Pareto efficient solution). The concept of market failure is linked to the first fundamental theorem of welfare economics. The market fails to deliver an efficient outcome whenever there are incomplete markets, asymmetric information, coordination failures, consumers and producers do not behave competitively, and/or no equilibrium exists. In these circumstances all economic actors can, in principle, be made better off by removing the market failure, possibly through the use of state aid. For a definition of market failure, *see* John O. Ledyard, *Market failure*, The New Palgrave Dictionary of Economics, 2nd Ed., Steven N. Durlauf & Lawrence E. Blume eds., (2008).

10 George A. Akerlof, *The Market for 'Lemons': Quality Uncertainty and the Market Mechanism*, 84 Q. J. Econ. 3, pp. 488–500, (1970).

11 Bernanke's remarks at the President's Working Group Market Stability Initiative Announcement, Washington, D.C.,on Oct 14, 2008 (*available at* http://www.federalreserve.gov/newsevents/speech/bernanke20081014a.htm).

12 Charles Goodhart & Dirk Schoenmaker, *Fiscal Burden Sharing in Cross-Border Banking Crises*, 5 Int'l J. Central Banking 1, pp. 141-165, (2009).

13 We understand that financial institutions need to perform credit analyses of credit related assets which are subsequently audited. These audited figures can be used to determine the true economic value of these assets.

14 Note that this amount should only consider the amount of extra capital that the market requires financial institutions to have as a result of the crisis of confidence discussed above, and should therefore exclude: (i) the additional capital necessary to cover the likely write-downs that can be expected as a result of the analysis in the previous points (both the fair economic value and the market failure losses); and (ii) any non-transitory increase in capital adequacy requirements in recognition of the fact that previous regulatory requirements were inappropriate, as any increase in capital required for these reasons should not be considered as having been caused by a market failure.

15 Note that we are assuming that the full amount of this revaluation directly affects regulatory capital.

16 Note that in this stylized example the increase in the minimum capital from 5 to 10 is the net of two effects: (i) the increase in the minimum capital adequacy ratio as a result of the crisis of confidence, and (ii) the reduction in the Risk Weighted Asset base as the result of the 30 losses, which tends to reduce the amount of necessary capital. Note also that—in this stylized example—we abstract from the fact that the book value of the capital may be different from its market value. We also assume that the "normal" regulatory minimum remains at 5: i.e. none of the increase from 5 to 10 can be considered a non-transitory increase in regulatory requirements as a result of a permanent change in regulatory policy.

17 The Commission explicitly mentions that institutions which have received a certain amount of aid, and institutions which have received asset relief in additional to some other aid, will be susceptible to restructuring demands. For example, *see* footnote 4 of the Restructuring Communication: "The criteria and specific circumstances which trigger the obligation to present a restructuring plan have been explained in the Banking Communication, the Recapitalisation Communication and the Impaired Assets Communication. They refer in particular, but not exclusively, to situations where a distressed bank has been recapitalised by the State, or when the bank benefiting from asset relief has already received State aid in whatever form that contributes to coverage or avoidance of losses (except participation in a guarantee scheme) which altogether exceeds 2% of the total bank's risk weighted assets. The degree of restructuring will depend on the seriousness of problems of each bank." *Also note* paragraph 55 of the Impaired Assets Communication: "In-depth restructuring would also be required where the bank has already received State aid in whatever form that either contributes to coverage or avoidance of losses, or altogether exceeds 2% of the total bank's risk weighted assets, while taking the specific features of the situation of each beneficiary in due consideration."

18 The Commission does not seem to take into consideration even relatively simple qualitative indicators of whether a bank was structurally sound, such as: absence of interventions or warnings by the financial regulators; absence of history of aid measures in the past; absence of indications from analysts and rating agencies that there would be anything wrong or particularly risky in a bank's strategy; and/or share price or traded debt movements indicating an early loss of confidence by investors. Although these qualitative indicators are imperfect and an analysis based on them would not be as rigorous as the approach outlined in this article, they would certainly provide a better measure of an institution's viability than the simplistic approach which the Commission seems intent on applying based on the form and amount of aid received.

19 For example, *see* Luigi Zingales, *Yes We Can, Secretary Geithner*, Economists' Voice, (February 2009). *Also see* Thomas F. Cooley, *Moral Hazard on Steroids*, Forbes, (March 2009).

20 Or, as John Vickers put it when writing about the October 8 U.K. scheme: "Given that the crisis is systemic and one of inadequate capital, not just insufficient liquidity, schemes on the lines of the U.K.

plan announced of October 8 make good economic sense. While state bailouts of arguably insolvent institutions are deeply unattractive, the realistic alternatives were still worse. The scheme is broadly competitively-neutral among U.K. institutions, and positive for other countries, many of whom have emulated the package. So while it is surely state aid, it is not seriously competition-distorting aid." John Vickers, *The financial crisis and competition policy: some economics*, GCP Magazine, (Dec-08), *available at* www.globalcompetitionpolicy.org.

21 *See*, for instance, Jeremy Bulow & Paul Klemperer, *Reorganising the Banks: Focus on the Liabilities, Not the Assets*, Economists' Voice (March 2009); and Zingales, *supra* note 19. *Also see* Douglas Diamond, Steve Kaplan, Anil Kashyap, Raghuram Rajan, & Richard Thailer, *Fixing the Paulson Plan*, Wall Street J, (September 26, 2008); and June 2009 comments by former Bank of England Deputy Governor John Gieve, reported at http://www.bloomberg.com/apps/news?pid=20601085&sid=a.sawnjO6kws.

22 *See* pp. 9 & 10 of the EAGCP Commentary on European Community Rescue and Restructuring Aid Guidelines, (February 2008).

23 This position also seems to be supported within the Commission. Georges Siotis, of the Chief Economist Team, noted that:

- "For non-financial institutions, compensatory measures typically consist of asset disposals and/or capacity reductions that "compensate" competitors for the survival of the distressed firm.
- For financial institutions, the disappearance or downsizing of a bank may actually hurt competitors"

See slide 18 of Georges Siotis, *The current financial crisis and EU Competition Policies*, at the ECRI/DIW/CEPS Conference, June 10, 2009 (*available at* http://www.ecri.eu/new/system/files/Siotis_2009-06-10.pdf).

24 In the case of banks in which the State is now a significant shareholder, compensatory measures may result in a "double-whammy" for tax payers: they had to bail-out banks and now they have to face a drop in the value of their "investment" as the result of asset sales.

25 For example, shares in RBS lost around 80 percent of their value over the 12 months to July 2009. AIG shares lost around 98 percent of their value in the same period.

26 *See*, for instance, Marco Becht, who highlights how the current crisis has "brought to light classic examples of board failure on strategy and oversight, misaligned or perverse incentives, empire building, conflicts of interest, weaknesses in internal controls, incompetence, and fraud." Marco Becht, *Corporate Governance and the Credit Crisis*, Macroeconomic Stability and Financial Regulation: Key Issues for the G20, Mathias Dewatripont, Xavier Freixas, &Richard Portes eds. CEPR, (2009).

27 It should be considered that a bank cannot unilaterally change its executive pay structure without incurring heavy costs in terms of lost talent. While it would be to the advantage of all bank shareholders to do so, there may be a coordination failure preventing imposing different incentive structures on the management. Regulation would be necessary to impose this more efficient incentive structure.

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Merger Review of Firms in Financial Distress

Ken Heyer & Sheldon Kimmel

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Merger Review of Firms in Financial Distress

*Ken Heyer & Sheldon Kimmel**

In recessions, we expect to see an increase in both the number and share of mergers where at least one of the parties is having difficulty independently staying afloat. This raises the importance of adopting a sound framework for analyzing merging firms in some form of financial distress.

This paper[1] concludes that, while it can be hard to evaluate a failing firm defense under the Merger Guidelines, the principles underlying the test are generally sound, even when the overall economy is going through very difficult times. The recent severe downturn may lead to more proposed mergers between financially distressed firms, but it does not imply that looser standards ought to be applied when evaluating them.

*The authors are economists at the Antitrust Division of the U.S. Department of Justice. The Antitrust Division encourages independent research by its economists. The views expressed herein are entirely their own and are not purported to reflect those of the Department.

I. Introduction

The current global economic recession raises serious challenges, not only for those devising and implementing macroeconomic policies, but also for those working in the field of competition policy.[2] While it is hard to predict our economy's short-run future, and while recent trends provide encouragement that we are beginning to emerge from the sharp recession of 2008-2009, history strongly suggests that recessions are not a thing of the past. The current and any future recessions provide, and will continue to provide, challenges for policymakers.

At times when increasing numbers of firms are in financial distress, we shouldn't be surprised to see more mergers where at least one of the parties is having difficulty staying afloat. This raises the importance of the appropriate standards to apply to such mergers.

A QUESTION ALSO ARISES, OR UNDOUBTEDLY WILL ARISE SHORTLY, AS TO WHETHER OUR ECONOMY'S RECESSION MEANS THAT MERGER ANALYSIS SHOULD EMPLOY A MORE FORGIVING SET OF REQUIREMENTS FOR MERGERS PROPOSED BY FIRMS IN SOME SIGNIFICANT FINANCIAL DISTRESS.

The relatively demanding conditions under which the federal competition authorities permit an otherwise anticompetitive merger are based on what is widely referred to as the "failing firm defense" and are relatively clear. As stated in § 5.1 of the U.S. Department of Justice and Federal Trade Commission Horizontal Merger Guidelines, they are as follows:

> "A merger is not likely to create or enhance market power or facilitate its exercise if the following circumstances are met: 1) the allegedly failing firm would be unable to meet its financial obligations in the near future; 2) it would not be able to reorganize successfully under Chapter 11 of the Bankruptcy Act; [FN. Citing the relevant statute omitted] 3) it has made unsuccessful good-faith efforts to elicit reasonable alternative offers of acquisition of the assets of the failing firm that would both keep its tangible and intangible assets in the relevant market and pose a less severe danger to competition than does the proposed merger; and 4) absent the acquisition, the assets of the failing firm would exit the relevant market."

While the language of the Guidelines itself is clear, the underlying rationale is not so widely understood and appreciated. Particularly when competition authorities will be faced with a disproportionately large number of proposed mergers for which some version of a failing firm defense may be offered, it is

important to remind ourselves of the principles underlying that defense and, more broadly, the appropriate framework for analyzing merging firms in some form of financial distress.

A question also arises, or undoubtedly will arise shortly, as to whether our economy's recession means that merger analysis should employ a more forgiving set of requirements for mergers proposed by firms in some significant financial distress. Our view is that, properly understood and applied, the Merger Guidelines' failing firm requirements are appropriate even in these difficult economic times. Although a weak economy may mean that more transactions will pass muster under this standard, those that do not should be blocked in troubled economic times for the same reasons they should be blocked in more "normal" times. The alternative would be a reduction in competition and harm to consumers and the economy as a whole.

II. Some Basic Merger Economics

At the outset, it is worth reviewing some basic economics relevant to merger policy generally. This will help establish familiar principles relevant to our subsequent discussion of firms that are in financial distress and are flailing, but may not be failing.

A. THE BENEFITS OF COMPETITION

Competition, or more accurately the benefits generated by the process of competition, provides the central underlying rationale for antitrust law and competition policy. Competition, properly defined to include competition to obtain monopoly power by best satisfying the demands of consumers, tends to allocate society's scarce resources most efficiently. This, in turn, maximizes the value that society can squeeze out of its resources.

Eliminating competition clearly helps improve the profitability of firms seeking to eliminate competition. This is, after all, why firms often seek protection from rivalry. This enhanced profitability, however, comes at the expense of consumers. And even more importantly, it comes at the expense of the economy as a whole. As Adam Smith[3] noted back in 1776,

> "The interest of the dealers, however, in any particular branch of trade or manufacturers is always in some respects different from, and even opposite to, that of the public. To widen the market and to narrow the competition, is always in the interest of the dealers. To widen the market may frequently be agreeable enough to the interest of the public; but to narrow the competition must always be against it, and can serve only to enable the dealers, by raising

> their profits above what they would naturally be, to levy, for their own benefit, an absurd tax upon the rest of their fellow-citizens. The proposal of any new law or regulation of commerce which comes from this order ought always to be listened to with great precaution, and ought never to be adopted till after having been long and carefully examined, not only with the most scrupulous, but with the most suspicious attention. It comes from an order of men, whose interest is never exactly the same with the public, who have generally an interest to deceive and even to oppress the public, and who accordingly have, on many occasions both deceived and oppressed it."[4]

Although some might try to defend the wealth transfer from customers to producers that a reduction in competition causes, a reduction in competition also leads to completely indefensible economic distortions that impair the functioning of the economy. In the short run, there is the well-known "deadweight loss" generated by a monopolist's profit incentive to restrict output below the competitive level. And in the longer run, eliminating competition weakens the incentive of firms to beat out rivals for the patronage of consumers by, for example, lowering costs (i.e., leaving more of society's scarce resources for other purposes), reducing prices, and producing more desirable products.

This provides the economic basis not only for blocking mergers whose primary effect may be to substantially eliminate competitive constraints, but also for antitrust laws prohibiting cartels, and for public policy permitting firms to enter markets or expand sales in competition with one another.

B. MERGERS AND EFFICIENCY

Of course, none of this means that mergers between rivals should always be prohibited. Federal antitrust authorities explicitly acknowledge that such mergers can be economically beneficial. Indeed, in evaluating the net consequences of proposed mergers—even ones between significant rivals—it is common practice for the agencies to analyze the extent to which the proposed merger may produce efficiencies—cut costs, improve quality, promote innovation. And, to the extent that these efficiencies are what the agencies refer to as "cognizable," the agencies will perform an integrated analysis of the merger's likely net economic effect.[5]

COGNIZABLE EFFICIENCIES COME IN MANY FORMS, AND CAN BE ESPECIALLY DIFFICULT FOR COMPETITION AUTHORITIES TO DISCERN AND EVALUATE, PARTICULARLY EX ANTE.

Cognizable efficiencies come in many forms, and can be especially difficult for competition authorities to discern and evaluate, particularly ex ante. Not only are the efficiencies themselves often difficult to evaluate, but it can be even more difficult to determine the extent to which they are truly specific to the merger—i.e., are

unlikely to be achieved in the absence of the proposed merger or some other means having comparable anticompetitive effects.

In some circumstances, merger-specific efficiencies may be so large that the merger will generate net economic benefits (for example, lower prices) even after accounting for possibly greater market power by the merged firm or anticompetitive coordination by the merged firm and its remaining rivals. And in situations where the totality of the evidence indicates there is no significant risk of anticompetitive effects, mergers are generally cleared without any requirement that the firms demonstrate merger-specific efficiencies.

III. Mergers Involving Firms in Financial Distress

Traditional antitrust review applies to mergers between financially viable long-term competitors whose pre-merger independence appears to limit the exercise of market power.

This review includes, inter alia, an evaluation of any cognizable efficiencies, the likelihood of sufficient and timely entry, and, of course, competitive effects analysis.

Where one of the merging parties is in a financially weakened state, special considerations may apply. It is frequently suggested that such mergers should be treated more leniently under the antitrust laws; however, the arguments for and against such a policy are often left insufficiently explored and inadequately defended. In some circumstances treating such mergers with greater leniency may be appropriate. In other circumstances, however, it is not.

For the moment, assume we are being asked to evaluate a proposed merger between two significant competitors and that, absent the struggling financial position one happens to be in at the moment, the merger would appear to raise very serious competitive concerns. Also assume we are evaluating a merger of two of only three or four firms that have been competing in a relevant market, that the firms proposing to merge both have high market shares, and that sufficient entry is unlikely to be timely. In such circumstances, the competition authority may confront the following scenarios.

A. THE FIRM WILL NOT BE A CONSTRAINING COMPETITOR IN THE FUTURE BECAUSE ITS PRODUCTIVITY IS DECLINING AND/OR BECAUSE THE SUPPLY OF KEY INPUTS IT OWNS IS BEING EXHAUSTED

Particularly (though not only) where one of the merging firms is in a financially distressed condition, one needs to consider whether historical evidence of the firm's role in the market is a reasonable proxy for the role it is likely to play going forward, but for the merger. If the firm is not likely to be an effective competitor absent the merger, then the merger is unlikely to produce an adverse effect on competition.

Antitrust analysis is forward, rather than backward looking, and competition authorities should rely on historical evidence only to the extent that this helps inform us regarding the future.[6] Where it does not, we need to look elsewhere. For example, a firm may own some assets that will continue to be productive for many years and others whose supply will be exhausted shortly. In some cases it may be clear how little life key productive assets have left in them (e.g., a once-rich coal seam may be almost completely mined, a valuable patent may be nearing expiration, or a factory may reside in a building that has been condemned and cannot be economically brought up to code). The actual or prospective loss of a firm's key assets, including the exhaustion of valuable scarce inputs, is relevant to its future competitiveness as a standalone firm and thus to the competitive implications of a merger—even one between rivals with historically high market shares.

Other circumstances where a firm's future competitiveness cannot be assumed to be similar to the past might include situations where the firm is on the verge of bankruptcy, in particular:

B. THE FINANCIALLY DISTRESSED FIRM IS UNABLE TO MEET ITS FINANCIAL OBLIGATIONS BECAUSE IT IS POORLY RUN AND/OR BECAUSE THERE IS A SIGNIFICANT EXOGENOUS DECLINE IN DEMAND FOR ITS PRODUCT

To the extent that a to-be-acquired firm is unable to meet its financial obligations, a key question for competition authorities is whether, absent the merger, the firm would be able to reorganize effectively. The fact that the firm's creditors may be forced to take a one-time financial loss is not terribly relevant to competition analysis unless the firm's assets will be liquidated and the firm, even under the ownership of a third party, will be unable to continue competing effectively. Even if the firm is unable to reorganize successfully under the bankruptcy laws, competition authorities properly consider what will become of the firm's assets in the event that the proposed merger does not take place. If these assets would likely be purchased by a firm that presents no (or fewer) competitive problems and would continue being employed as an independent competitive force in the market, then the mere fact of current financial distress does not imply that the proposed merger is necessarily benign.

To the extent that a to-be-acquired firm is unable to meet its financial obligations, a key question for competition authorities is whether, absent the merger, the firm would be able to reorganize effectively.

For this reason, the failing firm defense in the Merger Guidelines requires that the relevant assets of the failing firm be shopped before competition authorities will approve a potentially anticompetitive merger. It is also why it could be more accurate to refer to the failing firm defense as an "exiting assets defense." If the

assets would likely remain in the market–even if in the hands of some other player–then permitting the merger may well be anticompetitive.[7]

For example, consider a manufacturer claiming to be a less effective competitor going forward because it owes billions of dollars due to an unanticipated fall in the demand for its product. The firm may claim that inability to service this debt, perhaps due to frozen credit markets, leaves it incapable of financing those investments necessary to remain productive. "How," it might ask, "can competition authorities reasonably object to our being acquired even by a major rival?"

IT IS ALSO WHY IT COULD BE MORE ACCURATE TO REFER TO THE FAILING FIRM DEFENSE AS AN "EXITING ASSETS DEFENSE."

In examining this situation more closely, we shall see that when financial distress does justify permitting such a merger it is because the acquisition likely generates efficiencies. We shall discuss efficiencies and their relationship to the failing firm defense in somewhat greater detail in section IV below.

For our flailing manufacturer to be unable to meet its financial obligations and still be a desirable acquisition target by a major rival, the rival must believe that the firm's troubles are only temporary. Otherwise, it would not want to spend good money buying a failed entity. The acquirer may hold to this belief for a number of reasons:

1. The struggling firm will turn around even without an injection of capital from this particular acquirer. Perhaps the firm is basically sound but lost a lot of assets (e.g., due to a fire or a flood or embezzlement) and those short-term losses won't be repeated. Or, perhaps other sources would be willing to supply the firm with needed short-term credit. In such situations, the "but for" scenario is one where the financially distressed company remains a viable competitive force in the absence of the competitively suspect merger; hence financial distress does not justify permitting the merger.

2. The firm never will return to profitability (even as a division of the acquiring firm), but the acquirer does not know this and is making a bad bet. While this possibility certainly cannot be ruled out, likely mistakes by people spending their own money seem a weak justification for departing from sound principles of competition policy. Merger here should be permitted–assuming no less anticompetitive buyer appears with an offer to purchase and continue operating the firm or use its assets to compete in the market.

3. Only the acquiring firm can nurse the target firm back to health. The rival may have superior expertise in running this type of firm, be highly knowledgeable about the future prospects of the market, and/or be able to capture synergies by combining its skills with the struggling firm's assets. These factors may provide an economic justification for

the acquirer to wish to inject needed capital into the acquired firm's enterprise.

Under such circumstances, financial distress works as a failing firm defense because there is an underlying efficiency defense; the acquirer understands and can, perhaps uniquely, release the potential of the currently flailing firm.

IV. Efficiency Analysis and the Failing Firm Defense

While §5.1 of the Guidelines does not explicitly mention efficiencies, the failing firm defense implicitly relies upon the merger generating efficiencies. No firm would want to buy a competitor that meets the conditions given in §5.1 unless it believes that it can change what had been a failing firm (pre-merger) into what will be a profitable division (post-merger). Potentially that might be achieved in a number of ways, including cutting the failing firm's costs (an efficiency), or raising its revenues. One way of increasing revenues is by enhancing product quality (also an efficiency).

Revenues can, of course, also be enhanced by raising a firm's price without raising its quality or providing benefits to consumers. This sounds a lot like an anti-competitive effect. Recall, however, that for the failing firm defense to be satisfied, it must be determined that the firm's assets would be exiting the market "but for" the merger. If that takes place, then these assets would be providing no competitive constraint in the market at all. Thus, if the conditions for the failing firm defense are satisfied, competition authorities would have no reason to object to the merger even if it were known with certainty that price was going to increase.

SUCH A CLAIM IS NOT CREDIBLE UNLESS IT IS ACCOMPANIED BY AN EFFICIENCIES DEFENSE.

A failing firm defense commonly begins with the merging parties claiming that they can survive as a merged firm, but that one of them will not survive without the merger. The claim is that, one way or the other, there will be one less firm in the industry and so the merger itself will not affect the number of firms in the industry. Therefore, the parties allege, the merger is harmless.

Such a claim is not credible unless it is accompanied by an efficiencies defense. One can usefully divide the mass of failing-firm stories into two types: one where the acquiring firm wants to buy the failing firm's assets because it believes it may be able to improve the performance of those assets so much that they will be worth maintaining, and one where the acquiring firm has no such hopes. Consider those two types of stories in detail.

In the case of the efficiency story, an improvement in performance would tend to increase output above what it otherwise would have been, which is good both

for the merging parties and for society as a whole. If the improvement in performance is large enough, that pro-competitive effect could justify the merger as beneficial to consumers.

On the other hand, one can imagine a firm wanting to acquire the assets of a failing competitor even though there is no chance that this will improve the productivity of those assets.[8] While such transactions have their defenders, there are sound reasons for believing that such transactions will lead to an expected increase in price.[9]

To understand, think about what happens when the owner of a failing firm becomes the manager of these same assets as a division of the acquiring firm (for simplicity, assume that he isn't given any additional responsibilities). By the failing firm definition, there was nothing that this manager could have done to make those assets profitable pre-merger. Post-merger, assuming that the merger doesn't promise efficiencies, there is nothing he can do to make his costs any lower.

Prospects for revenue enhancement aren't good either, unless price rises (a point we'll come back to). Indeed, his efforts to enhance his division's revenues are potentially going to be undermined by the managers of the firm's other divisions (who might well prefer for him to go out of business so that they could sell at higher prices).

In such circumstances, a merger turns what had been a failing firm into what is now a failing division, unless prices rise post-merger. This implies that the most profitable way for its new owner to raise prices post-merger is to shut the failing division down. Of course, this is something the acquiring firm presumably knew all along, and so its motivation for buying the plant must be a fear that the acquired firm wasn't really going to shut down on its own (i.e., wasn't failing). In that case, the merger transforms a firm that wasn't doing well, but wasn't failing, into a division that is failing until shut down. The merger actually exacerbates the "failure" that it is supposed to solve.[10]

MORE ON THE REQUIREMENT OF A SHOP FOR THE ALLEGEDLY FAILING FIRM'S ASSETS

As discussed above, one key requirement of the failing firm defense is that the relevant assets be shopped to see if they would continue operating in the market in the hands of a less anticompetitive acquirer. If the financially distressed firm receives a bid from another firm, however, it may not be the case that this acquirer will employ the assets in the market of concern. Assets are often fungible and have alternative uses to which they can be put. Perhaps a competing bidder will liquidate them entirely. Given such uncertainty, should the competition authority be troubled by the possibility that the alternative purchaser might not continue employing the assets in its market of concern?

It should not. If the initial bidder were seeking to obtain the financially distressed firm's assets simply to exercise greater market power, it should be happy to see those assets exit the market via purchase by someone else. Therefore, if under these circumstances the initial bidder insists on paying more than others for the relevant assets, then it is a safe bet that there are efficiencies underlying the purchase and it ought to be permitted.

A more difficult scenario is where the shop turns up an alternative bidder who bids less but does seem likely to keep the relevant assets operating in the relevant market. In this case the initial bidder may be seeking to acquire the assets for purposes of exercising greater market power or achieving efficiencies (or both).

Determining the net effect of permitting the troubled assets to go to the highest bidder is, in principle, similar to asking—in the non-failing-firm context—whether a merger threatening competitive harm ought to be permitted because there are sufficiently large and cognizable efficiencies. The major difference between the two cases is that in the failing firm context the "but for" scenario is that the relevant assets will be operated by some new purchaser, rather than by the current owner. The need for competition authorities to gauge the future competitiveness of the relevant assets after they are in the hands of a new owner may make prediction even more difficult and uncertain.

Another issue that may arise when shopping assets–particularly during a severe economic downturn–is that divestitures to be mandated under a Consent Decree might not find any buyer willing to pay greater than liquidation value. In this context, does the failure to find any willing buyer necessarily demonstrate that the merger itself satisfies the conditions for a successful failing firm defense?

In this context, does the failure to find any willing buyer necessarily demonstrate that the merger itself satisfies the conditions for a successful failing firm defense?

The answer is no. There may be other reasons why no third party is willing to purchase the assets. Perhaps the package of assets being shopped is the "wrong" collection of assets and cannot be used by anyone to compete profitably in the market of concern. Alternatively, it may be that although no third party would be willing to purchase and operate the divested assets (perhaps because of an inability to secure credit), both of the two merging firms would continue independently in operation. In such circumstances, the "but for" scenario would be continued rivalry between the merging firms, and a merger raising serious competitive concerns should be permitted only where cognizable efficiencies outweigh these feared harms.

V. *International Shoe v. FTC*[11]: An Early Application of Failing Firm Analysis

The failing firm defense has taken many different forms, but is hardly new. It has been around since at least 1930, when the Supreme Court issued its opinion in the *International Shoe* case and spelled out, for the first time, an acceptable failing firm defense.

The motivation for the merger between International Shoe and McElwain followed from a severe downturn in the orders that the failing firm (McElwain) had been receiving. As the Court noted:

> "Beginning in 1920 there was a marked falling off in prices and sales of shoes, as there was in other commodities; and, because of excessive commitments which the McElwain Company had made for the purchase of hides as well as the possession of large stocks of shoes and an inability to meet its indebtedness [the company's officers] concluded that the company was faced with financial ruin, and that the only alternatives presented were liquidation through a receiver or an outright sale. New orders were not coming in."[12]

International Shoe, on the other hand, "had so conducted its affairs" that its problem was an inability to fill the demand for its shoes.

Generations of scholars have apparently misread *International Shoe* [Indeed, in 2001 the DC Court of Appeals said in *FTC v. Heinz* and *Milnot* that "the Supreme Court has not sanctioned the use of the efficiencies defense in a section 7 case"[13]] Although it has been consistently overlooked, an "efficiency defense" for the parties' merger is very clearly recognized in *International Shoe*:

> "During the early months of 1921, [International Shoes'] orders exceeded the ability of the company to produce, so that approximately one-third of [301] them were necessarily canceled.… It is perfectly plain from all the evidence that the controlling purpose of the International in [buying the McElwain shoe company] was to secure additional factories, which it could not itself build with sufficient speed to meet the pressing requirements of its business."[14]

Thus, the Court found that International was buying McElwain so that, post-merger, it could fill the demand for the brands that it owned with the relatively good but under-utilized plants of McElwain. This would combine their complementary strengths to generate efficiencies. The Court held specifically that the merger was legal

> "[i]n the light of the case thus disclosed … the purchase of its capital stock by a competitor (there being no other prospective purchaser) … to facilitate the accumulated business of the purchaser … does not substantially [303] lessen competition or restrain commerce within the intent of the Clayton Act."

The key fact in *International Shoe* was that a downturn in one firm's business matched an upturn in another firm's business, allowing a merger to combine the different strengths of each firm, making the merged firm stronger than either had been on their own. Moreover, by holding that the only alternative to the merger was liquidation, the Court also appears to have found that no less anticompetitive purchaser would be willing to purchase the assets and keep them operating in the market of concern–i.e., that the merger satisfied an "exiting assets" requirement.

Although the Court clearly laid out an efficiency defense, it did not actually use that precise phrase. This is hardly surprising, however, since it wasn't until 1967 that the term "efficiency defense" appeared in any Supreme Court decision.[15]

VI. Failing Firm Analysis during Tough Economic Times

Although not necessarily easy to apply, the logic of the existing failing firm defense, and the conditions required to be met before otherwise anticompetitive mergers are approved, seems sound. Are there reasons why these conditions ought to be loosened during tough economic times (such as those we are experiencing today)?

ARE THERE REASONS WHY THESE CONDITIONS OUGHT TO BE LOOSENED DURING TOUGH ECONOMIC TIMES (SUCH AS THOSE WE ARE EXPERIENCING TODAY)?

Historically, economic downturns have often led to attempts to get new regulations or laws that restrict "unfair" or "excessive" competition (i.e., a downturn can be seen as evidence that free markets have failed, supporting moves away from free markets). One of the largest examples is the National Industrial Recovery Act, which may have been an attempt to fix what the Great Depression was thought to have shown to be broken. Regardless

of its motivation, the NIRA allowed hundreds of industries to legally meet collusively in smoke-filled rooms to limit competition as spelled out in their collusive agreements (i.e. codes of "fair competition"). Available evidence suggests that its effect on the economy was very harmful.[16]

In thinking about whether or why a departure from traditional failing firm principles might be warranted, one might consider that those applying the failing firm principles are not all-knowing and invariably make mistakes from time to time. Available information is imperfect and costly to obtain, and the future is uncertain. Talented and hard working though they may be, the staffs of competition agencies will not always get it right. And if, during tough economic times, the costs of wrongly finding a firm to be viable exceed the costs of wrongly finding a firm to be failing, then arguably competition authorities should bear greater risk and more readily permit acquisitions of firms in financial distress.

At least as a theoretical matter, this possibility cannot be rejected. Absent supporting empirical evidence, however, the reverse seems as likely to be true; during times of financial distress the costs of getting it wrong might argue in favor of adopting an even tougher stance. Perhaps policy should tilt against insulating financially distressed firms from the forces of competition.

Without persuasive evidence one way or the other, an agnostic approach seems prudent. Indeed, although it may be impolitic to say so, a feature of all "bailouts"–including relaxing antitrust standards to permit anticompetitive mergers–is that they help allow inefficient management and labor agreements to stay in place instead of permitting the marketplace to force the painful changes that should be made.

Clearly there is always harm from blocking a merger that would have cut the costs of the failing firm. One might also argue that such a cost is relatively high during an economic downturn because it is easier to redeploy assets in booming times than in downturns (e.g., you'll have better luck finding a job if you're the only unemployed person in a strong economy than if you joined all the millions looking for a job in a downturn). On the other hand, one might argue that the cost of allowing a merger to create market power is greater during a downturn (since entry may be likelier during a boom). As usual, the competition authorities have to weigh all these possibilities, hoping to strike the right balance.

Another argument offered for permitting financially distressed firms to be acquired in what might be anticompetitive deals is that it is good for the economy to permit flailing firms to get the highest possible value for their assets when they are going under—even if it leads to greater market power and short-run harm to consumers. If competition authorities refuse to provide such a "safety net" there is likely to be less entry in the first place. Therefore, one might be tempted to argue in favor of relaxing the relatively demanding conditions required for a successful failing firm defense.

Such arguments are not appealing, either in good economic times or in bad ones. The amount of entry is based on potential entrants' expected profits. Truncating the amount that a firm stands to lose in tough times (by allowing anticompetitive mergers) provides an incentive for more than the optimal amount of entry–and perhaps also to entry being skewed towards markets where investors believe that a failing firm defense would be applied most leniently. Profits aren't capped when a firm does especially well (nor, we would argue, should they be). For similar reasons, a floor shouldn't be placed under a firm's losses by providing it with an antitrust free pass when it seeks to exit via an anticompetitive merger.

FOR SIMILAR REASONS, A FLOOR SHOULDN'T BE PLACED UNDER A FIRM'S LOSSES BY PROVIDING IT WITH AN ANTITRUST FREE PASS WHEN IT SEEKS TO EXIT VIA AN ANTICOMPETITIVE MERGER.

VII. Conclusion

Evaluating whether a proposed merger satisfies the Merger Guidelines' failing firm defense is often difficult. The principles underlying the test are, however, generally sound. Moreover, these principles remain appropriate even when the overall economy is going through very difficult times. Severe economic downturns may lead to more proposed mergers between financially distressed firms, but it does not imply that looser standards ought to be applied when evaluating them.

VIII. Appendix[17]

Avoidable costs (costs that a firm can avoid by going out of business) have to be at the heart of any model used for considering the failing firm issue. To illustrate the way failing firm analysis works, this Appendix considers what may be the simplest case (the same forces apply also in more complicated cases, but a simple case is easier to understand). Consider a duopoly where each firm has an avoidable cost and each one's marginal cost is constant out to its capacity, although one firm is said to be failing because its marginal cost is relatively high. For the moment, consider the possibility (that will lead us to a contradiction, so that possibility will ultimately be rejected, but for the moment suppose) that both the failing firm's plant and its acquirer's plant will operate post-merger, and the merger creates no efficiencies.

Since the failing firm's marginal cost is relatively high, to the degree it is possible to shift any output from the failing firm's plant to the acquirer's plant, that necessarily increases total profit. So if both plants still operate post-merger, the acquirer's plant must be operating at full capacity (i.e., unable to accept any more orders being shifted to it). Therefore, the merger cannot improve the environment facing the failing firm's plant (i.e., post-merger its competitor will produce at full capacity which is the worst environment the acquired plant can

face). Therefore the merger doesn't improve the failing firm's profitability, which is as bad (or worse) post-merger as it was pre-merger. Therefore, it will be shut down post-merger, contradicting the earlier assumption that both firms operate post-merger.

However, it makes no sense to go through the expense of a court battle to be allowed to pay good money to buy and then shut a plant that was failing on its own (i.e., if a plant is going to be shut whether it's acquired in a merger or not, it makes no sense to pay to be the one that gets to shut it). Thus, we also reject the initial assumption: If the buyer has no efficiencies to add to the acquisition, then the "failing firm" will fail only if its competitor is allowed to buy it.

---

1 An earlier version of this paper appeared as Economic Analysis Group Discussion Paper EAG 09-1 March 2009.

2 Over the period from January 2008 to December 2008 monthly bankruptcy filings rose from 72,179 to 97,682. Monthly bankruptcies peaked in March 2009 at 134,578 before dropping somewhat to 127,699 in June 2009–the last month for which we have data, *see*: http://www.uscourts.gov/Press_Releases/2009/bankrupt_f2filmn_jun2009.xls

3 Adam Smith, The Wealth of Nations, Vol. 1, p.278.

4 Smith's recommendation about regulation is worth highlighting even though this paper focuses more narrowly on the issue of failing firms. Distilling the above quote down to just its regulatory recommendation, Smith concludes that since "The interest of the dealers … in any particular branch of trade or manufacturers is always in some respects different from … that of the public.… The proposal of any new law or regulation of commerce which comes from this order … ought never to be adopted till after having been long and carefully examined, not only with the most scrupulous, but with the most suspicious attention."

5 Cognizable efficiencies are "merger-specific efficiencies that have been verified and do not arise from anticompetitive reductions in output or service. Cognizable efficiencies are assessed net of costs produced by the merger or incurred in achieving those efficiencies." Section 4, Horizontal Merger Guidelines.

6 U.S. v. General Dynamics, 415 U.S. 486, 501.

7 In the context of mergers that raise little risk of substantially harming competition, normal market forces can be relied upon to move assets to their most efficient uses. For this reason the Merger Guidelines properly refrain from imposing a requirement that the assets be transferred to whoever the government may feel is the most efficient alternative purchaser. Such a requirement is necessary and imposed only when there are substantial competitive risks.

8 The "productivity" (or, more generally, the "efficient use") of the assets discussed in the text is not restricted simply to how well they can be used physically to produce, but also how well they can be used to enhance economic value. We are here considering proposed acquisitions that have no prospect of permitting either to take place.

9 Farrell & Shapiro demonstrate formally that "If a merger generates no synergies, then it causes price to rise." *See* the proof of their Proposition 2 in the Appendix to Joseph Farrell & Carl Shapiro, *Horizontal Mergers: An Equilibrium Analysis*, The Am. Econ. Rev. (March 1990).

10 The Appendix provides, in the context of a particular example, a more formal demonstration of these points.

11 International Shoe v. FTC, 280 U.S. 291 (1930).

12 *Id.* at 299.

13 246 F.3d 708, 721).

14 *Supra* note 11, at 300-301

15 Indeed, Westlaw finds the phrase "efficiency defense" being used in only one Court decision, ever: FTC v. Proctor & Gamble, 386 U.S. 568.

16 One recent study, Cole & Ohanian, *New Deal Policies and the Persistence of the Great Depression: A General Equilibrium Analysis*, J. Pol. Econ., (2004) found that "New Deal cartelization policies are a key factor behind the weak recovery [during 1934-9], accounting for about 60 percent of the difference between actual output and trend output." The authors point not simply to the National Industrial Recovery Act (NIRA)–which was struck down in 1935 as unconstitutional—but to government failure to enforce the antitrust laws even after 1935. They write that "the government openly ignored collusive arrangements in industries that paid high wages" until the 1938 appointment of Assistant Attorney General Thurman Arnold. Cole & Ohanian note further that "The number of new cases brought by the DOJ rose from just 57 between 1935 and 1939 to 223 between 1940 and 1944."

17 *See also* Kimmel, *The Supreme Court's Efficiency Defense*, Supreme Court Econ. R., (2004).

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Review of Reverse-Payment Agreements: The Agencies, the Courts, Congress, and the European Commission

William H. Rooney & Elai Katz

Amy R. Fitzpatrick, Michelle Leutzinger, & Peter J. Scoolidge

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Review of Reverse-Payment Agreements: The Agencies, the Courts, Congress, and the European Commission

*William H. Rooney and Elai Katz**

with Amy R. Fitzpatrick, Michelle Leutzinger, and Peter J. Scoolidge

Two bills seeking to ban reverse-payment agreements are currently pending in Congress, and the European Commission has declared that such agreements, depending on the circumstances, may violate European competition laws. Meanwhile, several U.S. Courts of Appeals have upheld reverse-payment settlements as lawful if the restrictions in the settlement are within the scope of the patent. This article provides an overview of the treatment of reverse-payment agreements by the agencies, the appellate courts, Congress, and the European Commission, without advocating a view on the legality of such agreements or the merits of court decisions, proposed legislation, or investigations relating to them.

*William H. Rooney, Esq. is a partner in the law firm of Willkie Farr & Gallagher LLP. Elai Katz, Esq. is a partner in the law firm of Cahill Gordon & Reindel LLP. Their practices focus on civil and criminal antitrust matters and both have had significant experience at the intersection of antitrust and intellectual property law. Amy R. Fitzpatrick, Esq., is an associate in the Washington, DC office of Willkie Farr & Gallagher LLP with significant experience in complex litigation and antitrust matters; Michelle Leutzinger, Esq., is an associate in the New York office of Willkie Farr & Gallagher LLP; and Peter J. Scoolidge, Esq., is an associate in the New York office of Cahill Gordon & Reindel LLP.

I. Introduction

Over the last decade, branded and generic pharmaceutical companies, the federal antitrust agencies, antitrust practitioners, federal courts, legislators, and the European Commission have grappled with the legality of patent settlements and other agreements that involve "reverse payments." Reverse payments are so-termed because, in contrast to circumstances in which the alleged infringer pays the patent holder for a license to enter the market, the patent holder pays the alleged infringer supposedly not to enter the market during some or all of the term of the allegedly infringed patent. Such agreements have been challenged as antitrust violations by the Federal Trade Commission ("FTC") and the plaintiffs' bar, but have been upheld in the settlement context by most appellate courts. Two bills seeking to ban reverse-payment agreements are currently pending in Congress, and the European Commission has declared that such agreements, depending on the circumstances, may violate European competition laws.

This article provides an overview of the treatment of reverse-payment agreements by the agencies, the appellate courts, Congress, and the European Commission without advocating a view on the legality of such agreements or the merits of court decisions, proposed legislation, or investigations relating to them. We begin by briefly describing the Hatch-Waxman statutory framework within which reverse-payment agreements have arisen.

II. Statutory Framework for Reverse-Payment Agreements: The Hatch-Waxman Act

Reverse-payment agreements originated in response to patent infringement litigation that arose out of the Drug Price Competition and Patent Term Restoration Act of 1984, commonly referred to as the Hatch-Waxman Act.[1] The Hatch-Waxman Act was designed to increase competition and lower prices for consumers by accelerating the entry of generic drugs while, at the same time, maintaining the incentives to develop new drugs. The Hatch-Waxman Act permits companies to file with the Food and Drug Administration ("FDA") an abbreviated new drug application ("ANDA") for generic products that are shown to be bioequivalent to FDA-approved branded products. The ANDA procedure permits generic manufacturers to bypass the costly and lengthy new drug application ("NDA") process and to receive faster FDA approval to market the generic products.[2]

Every ANDA filing must include one of four certifications addressing the potential of the generic product to infringe a patent covering the reference-branded drug as to which the generic drug is bioequivalent. The certifications claim that:

(I) no patent was filed for the reference drug;

(II) the patent has expired;

(III) the patent expires before the ANDA filer will begin marketing the product; or

(IV) the patent is invalid or would not be infringed by the generic product.[3]

The last is referred to as a "Paragraph IV" certification. Paragraph IV filings are a means by which generic companies police brand-company assertions of patent protection and may expedite the entry of generic competition before the asserted patent expires.

For the purpose of describing the context in which reverse-payment agreements arise, a brief summary of the relevant rules surrounding Paragraph IV certifications and patent infringement litigation follows.[4] The Hatch-Waxman Act encourages Paragraph IV filings by rewarding the first generic manufacturer to file a Paragraph IV certification on a given drug with a 180-day exclusivity period during which the first-filer can market the drug without competition from other ANDA-approved generic drugs.[5] Should the patent holder initiate patent infringement litigation, however, the first-filer cannot enter the market for 30 months after the date that the patent holder receives notice of the Paragraph IV certification, a provision commonly referred to as the "30-month stay."[6]

THE LITIGATION THAT FOLLOWS PARAGRAPH IV CERTIFICATIONS HAS PROVIDED THE CONTEXT IN WHICH REVERSE-PAYMENT AGREEMENTS HAVE EVOLVED.

The ANDA filer must notify a patent-holder within 20 days of making such a certification, and the patent-holder then has 45 days to initiate suit.[7] Brand companies frequently initiate patent-infringement litigation on the basis of the Paragraph IV certification—that is, the brand company disputes the generic company's statement that the brand company's allegedly applicable patent is invalid or will not be infringed by the imminent generic entrant. The litigation that follows Paragraph IV certifications has provided the context in which reverse-payment agreements have evolved.

III. The FTC's Initial Response to Reverse-Payment Agreements

A. THE EARLY CONSENT AGREEMENTS—HYTRIN AND CARDIZEM CD

In the mid- and late-1990s, Paragraph IV certifications increased significantly, as did the FTC's focus on competition in the healthcare sector. The FTC began to investigate reverse-payment agreements in the Paragraph IV patent-litigation context and expressed skepticism as to the legality of the practice. FTC officials described the practice as "gaming" the Hatch-Waxman Act—claiming that such agreements were designed to eliminate competition and share the resulting monopoly profits.[8] The antitrust bar watched the progress of the reverse-payment

investigations with interest, as the agreements presented challenging antitrust issues in the increasingly important pharmaceutical context.

In 2000, the FTC announced a settlement with Abbott Laboratories and Geneva Pharmaceuticals with respect to their "interim agreement" pending the conclusion of the then-current infringement litigation over Abbott's blood-pressure drug, Hytrin.[9] During the course of the Hytrin infringement litigation, Abbott had agreed to pay $4.5 million per month in exchange for first-filer Geneva's promise not to release its generic Hytrin until the earlier of the resolution of the parties' patent litigation or the entry of another generic competitor. Geneva had also agreed not to transfer or relinquish its 180-day right of exclusivity. The Geneva-Abbott agreement was entered three days after Geneva was granted FDA approval of its generic drug.[10]

Under the terms of the settlement with the FTC, Abbott and Geneva agreed not to enter into future agreements involving restrictions on relinquishing exclusivity or involving restrictions on entering the market with a non-infringing product.[11] They also agreed to submit for court approval, along with notice to the FTC, any future interim agreement involving payments to generic companies to stay off the market.[12] The Hytrin agreement reflected the FTC's skepticism of reverse payments in the Hatch-Waxman litigation context.

Following the Hytrin and Cardizem consent decrees, some concluded that reverse payments in the Hatch-Waxman context were risky.

Within a year, the FTC also announced a settlement with Hoechst Marion Roussel, Inc. ("HMR") and Andrx Corporation regarding their agreement in the context of patent-infringement litigation over HMR's angina drug, Cardizem CD.[13] That settlement followed the FTC's challenge of HMR and Andrx's interim agreement in which Andrx had agreed that, while the patent litigation remained unresolved, Andrx would neither market its generic Cardizem CD following FDA approval nor relinquish its 180-day right of exclusivity.[14] In return, HMR would give Andrx quarterly payments of $10 million with payment to begin following FDA approval. At the time the parties entered the agreement, HMR's 30-month stay on Andrx's entry was scheduled to expire within a year. The agreement further stipulated that HMR would make an additional payment to Andrx if Andrx eventually prevailed in the patent litigation.[15] The restrictions imposed on HMR and Andrx by the settlement with the FTC were largely the same as those contained in the FTC's settlement with Abbot and Geneva.[16]

Following the Hytrin and Cardizem consent decrees, some concluded that reverse payments in the Hatch-Waxman context were risky. The story, however, was just beginning to unfold.

B. ADMINISTRATIVE CHALLENGES—IN THE MATTER OF SCHERING-PLOUGH CORPORATION

Furthering the enforcement gains obtained in the Hytrin and Cardizem matters, the FTC pursued an investigation of an allegedly disguised reverse payment in connection with infringement litigation over Schering-Plough's prescription potassium deficiency drug, K-Dur 20. Schering-Plough, Upsher-Smith Laboratories, and American Home Products Corporation had entered into settlements resolving Paragraph IV patent litigation instead of interim agreements during the pendency of the infringement litigation that were used in the Hytrin and Cardizem matters. In the Schering-Plough settlements, Schering-Plough paid cash amounts to the generic companies, but did so in return for licenses to certain intellectual property that the generic companies had developed or were in the process of developing. The FTC questioned the *bona fides* of the payments, suspecting that the payments were, in fact, reverse payments.

American Home Products settled with the FTC in a consent decree with relief similar to that obtained in the Hytrin and Cardizem decrees.[17] Schering-Plough and Upsher, however, chose to litigate the case with the FTC in an action before an Administrative Law Judge.[18]

The Administrative Law Judge found that the challenged license agreements were *bona fide* and lawful and dismissed the complaint.[19] The FTC staff appealed that decision to the full Commission, which reversed in a lengthy opinion that described the Commission's view on the lawfulness of reverse payments under the antitrust laws. As an initial matter, the Commission found that the Schering-Plough payment was disproportionate to the value of the Upsher licenses and that the payment was, in fact, tantamount to a "reverse payment."[20] The FTC found that the "quid pro quo for the payment was an agreement by the generic to defer entry beyond the date that represents an otherwise reasonable litigation compromise."[21]

An appeal to the Eleventh Circuit Court of Appeals followed, the result of which is discussed below. Reverse-payment cases brought by private plaintiffs were also making their way through the federal courts during the same time period.

IV. The Federal Courts' Treatment of Reverse-Payment Agreements

A. SIXTH CIRCUIT: IN RE CARDIZEM CD ANTITRUST LITIGATION (2003)

The first appellate court to address reverse payments was the Sixth Circuit in a private action that arose from the Cardizem interim agreement between Andrx and HMR that was the subject of the FTC-Cardizem consent decree.[22] As noted above, the Cardizem agreement did not settle the underlying patent litigation but provided that the generic manufacturer would neither enter before a speci-

fied period nor relinquish its 180-day exclusivity period, thus precluding entry of other generic competitors under then-applicable Hatch-Waxman rules. The district court further observed that the agreement also prohibited Andrx from marketing "non-infringing or potentially non-infringing" drugs.[23]

The Sixth Circuit in *Cardizem*, in an opinion by Judge Oberdorfer,[24] sitting as an appellate judge by designation, treated the interim agreement as a *per se* unlawful horizontal market allocation. The Sixth Circuit noted that the agreement did not settle the litigation, contained a clause that precluded Andrx from "relinquish[ing] or otherwise compromis[ing]" its 180-day period of exclusivity, and restrained Andrx from marketing "noninfringing and/or potentially noninfringing" drugs.[25] *Per se* treatment is typically reserved for limited categories of restraints of trade so familiar to the courts that a conclusive presumption of illegality is appropriate. The opinion stated that:

> "[T]he Agreement … [is] a classic example of a per se illegal restraint of trade.… [I]t is one thing to take advantage of a monopoly that naturally arises from a patent, but another thing altogether to bolster the patent's effectiveness in inhibiting competitors by paying the only potential competitor $40 million per year to stay out of the market.… [T]he fact that this is a 'novel' area of law [does not] preclude per se treatment"[26]

DIFFERENT JUDICIAL PERSPECTIVES ON REVERSE PAYMENTS WERE ABOUT TO EMERGE—PARTICULARLY WITH RESPECT TO SETTLEMENT AGREEMENTS THAT DO NOT LIMIT EXCLUSIVITY RELINQUISHMENT AND ARE WITHIN THE SCOPE OF THE PATENT.

Thus, the court found the agreement akin to classic examples of restraints that the Supreme Court has subjected to the *per se* rule, including "naked, horizontal restraints pertaining to prices or territories."[27] Different judicial perspectives on reverse payments were about to emerge—particularly with respect to settlement agreements that do not limit exclusivity relinquishment and are within the scope of the patent.

B. ELEVENTH CIRCUIT: VALLEY DRUG CO. V. GENEVA PHARMACEUTICALS, INC. (2003) AND SCHERING-PLOUGH V. FTC (2005)

In *Valley Drug Co. v. Geneva Pharmaceuticals, Inc.*,[28] the Eleventh Circuit reviewed a district court decision holding the interim agreement between Abbott and Geneva over Hytrin (the same interim agreement that was challenged by the FTC and the subject of the 2000 consent agreement with the FTC), as well as a final settlement between Abbott and Zenith Goldline Pharmaceuticals, to be *per se* unlawful.[29] The Eleventh Circuit's reversal of the

district court's decision was issued while the Schering-Plough matter was under consideration by the FTC Commissioners and led some to re-examine the FTC's theories on reverse payments.

In addressing the issue of reverse payments, the Eleventh Circuit in *Valley Drug* started with the observation that a patent was at issue and that patents grant a lawful right of exclusion. As such, the court held that, "[b]ecause the district court failed to consider the exclusionary power of Abbott's patent in its antitrust analysis, its rationale was flawed."[30] It further held that an agreement that involves restrictions on competition no greater than "the exclusionary potential of the patent" does not violate the Sherman Act.[31] The Eleventh Circuit referred to patent-immunity law,[32] which the Federal Circuit would later address in *In re Ciprofloxacin Hydrochloride* Antitrust Litigation.[33]

The Eleventh Circuit thus started its analysis from patent law, not antitrust law, and outlined a test that it later summarized in *Schering-Plough* as follows: "the proper analysis of antitrust liability requires an examination of: (1) the scope of the exclusionary potential of the patent; (2) the extent to which the agreements exceed that scope; and (3) the resulting anticompetitive effects."[34] The court also explained that, on the record before it, the presence or size of a reverse payment from the patent holder to the alleged infringer did "not alone demonstrate that the Agreements had obvious anticompetitive tendencies above and beyond Abott's potential exclusionary rights under the [relevant] patent."[35] The Eleventh Circuit later clarified in *Schering-Plough* that the patent infringement action may be susceptible to an antitrust suit "[i]f the challenged activity simply serves as a device to circumvent antitrust law."[36]

Although the decision in *Valley Drug* preceded the FTC's decision in Schering-Plough, the FTC did not follow *Valley Drug* or devote considerable resources to discussing the opinion, except to acknowledge *Valley Drug*'s rejection of the *per se* standard.[37] The analytical perspective of the FTC was significantly different from that of the Eleventh Circuit, as the FTC focused its assessment with the antitrust laws first in mind. From the FTC's antitrust perspective, the reverse payment was centrally relevant as it appeared to be the consideration (or the sharing of monopoly rents) for an anticompetitive agreement that was facilitated by the claimed misuse of the provisions of the Hatch-Waxman Act.[38] Although the FTC did not declare reverse-payment settlements *per se* unlawful, the circumstances in which the reverse payment would not be anticompetitive were narrowly confined.

ALTHOUGH THE FTC DID NOT DECLARE REVERSE-PAYMENT SETTLEMENTS *PER SE* UNLAWFUL, THE CIRCUMSTANCES IN WHICH THE REVERSE PAYMENT WOULD NOT BE ANTICOMPETITIVE WERE NARROWLY CONFINED.

Not surprisingly, Schering-Plough appealed the FTC's decision in its case to the Eleventh Circuit, as was its right under the FTC Act.[39] Schering-Plough thus

pitted the FTC's view that reverse payments are fundamentally anticompetitive against the different and more patent-oriented view presented in *Valley Drug*. Although the FTC tried to reconcile the two, the Eleventh Circuit's decision in *Schering-Plough* confirmed that, in the Eleventh Circuit, the *Valley Drug* patent-oriented framework prevailed. The Eleventh Circuit thus reversed the FTC decision in *Schering-Plough* and held that the K-Dur settlement was lawful under the *Valley Drug* analytical framework.[40]

The FTC sought certiorari, which prompted the Solicitor General (with the Antitrust Division of the Department of Justice ("DOJ")) to argue that the issues had not been sufficiently developed in the lower courts and to suggest that certiorari not be granted.[41] The Supreme Court denied certiorari,[42] thereby ending the first FTC-litigated reverse-payment matter with a victory for the pharmaceutical companies. Meanwhile, other cases involving reverse payments were making their way to other appellate courts.

C. SECOND CIRCUIT: IN RE TAMOXIFEN CITRATE ANTITRUST LITIGATION (2005, AMENDED 2006)

The Second Circuit in *In re Tamoxifen Citrate* Antitrust Litigation affirmed the dismissal under Federal Rule of Civil Procedure 12(b)(6) of a reverse-payment challenge involving a metastatic breast-cancer drug, tamoxifen citrate.[43] The Second Circuit held that a reverse payment to settle an appeal from a judgment of patent invalidity did not violate antitrust law where the exclusionary effects of the settlement did not exceed the scope of the patent grant.[44] The court joined the Eleventh Circuit in rejecting a "categorical[ ] condemn[ation of] reverse payments," [45] and declined to base the lawfulness of a settlement following a judgment of patent invalidity upon predictions of an appellate court's future assessment of the patent's validity.[46]

Plaintiffs in Tamoxifen, rather than arguing for *per se* unlawfulness, instead claimed that the reverse payment was unlawful because "[t]he value of the consideration provided to keep [the generic manufacturer's] product off the market . . . greatly exceeded the value [the generic manufacturer] could have realized by . . . entering the market with its own competitive generic product."[47] The court rejected that approach as failing to consider sufficiently the incentives of a patent holder, even one that is relatively confident of the validity of its patent.[48] Instead, the court opted for the *Schering-Plough* and *Valley Drug* analysis that considers "whether the 'exclusionary effects of the agreement' exceed the 'scope of the patent's protection.'"[49] The Second Circuit noted in its discussion that plaintiffs did not allege that the underlying patent was obtained through fraud or that the underlying infringement lawsuit was "objectively baseless."[50]

D. FEDERAL CIRCUIT: IN RE CIPROFLOXACIN HYDROCHLORIDE ANTITRUST LITIGATION (2008)

In re Ciprofloxacin Hydrochloride Antitrust Litigation involved the settlement of Paragraph IV litigation between Bayer and generic manufacturer Barr Laboratories, Inc. over Barr's 1991 ANDA filing for generic ciprofloxacin hydrochloride (ciprofloxacin), a synthetic antibiotic.[51] Under the terms of the settlement, Bayer agreed to pay Barr $49.1 million and either to supply Barr with ciprofloxacin for resale or to make quarterly payments through December 31, 2003. Barr also agreed to convert its Paragraph IV certification to Paragraph III and not to market generic ciprofloxacin until after Bayer's patent expired. In addition, Barr agreed to affirm the validity and enforceability of the patent and admit infringement.[52] Advocacy groups and direct and indirect purchasers of ciprofloxacin filed a complaint against Bayer and Barr, alleging that the settlement agreement was an illegal market allocation.[53]

The Eastern District of New York granted summary judgment for defendants and plaintiffs appealed.[54] Prior to the Second Circuit's approval of the reverse-payment settlement in *Tamoxifen*, defendants in *Cipro* sought to transfer the appeal from the Second Circuit to the Federal Circuit. Because the *Cipro* indirect-purchaser plaintiffs included in their complaint a state-law claim similar to a federal Walker-Process claim that involved a substantial question of patent law, the Second Circuit found that the Federal Circuit had jurisdiction over the indirect-purchaser appeal. The Second Circuit, however, denied the motion to transfer with respect to claims by the direct-purchaser plaintiffs.[55]

THE FEDERAL CIRCUIT FOUND THAT THE DISTRICT COURT HAD PROPERLY APPLIED A RULE OF REASON ANALYSIS BY PLACING THE INITIAL BURDEN ON THE PLAINTIFF TO SHOW THAT THE SETTLEMENT HAD AN ADVERSE EFFECT ON COMPETITION IN THE RELEVANT MARKET.

The Federal Circuit affirmed the district court's grant of Bayer's and Barr's motion for summary judgment against the indirect-purchaser plaintiffs. The district court reasoned that all anticompetitive effects caused by the settlement agreement were within the exclusionary zone of the patent and thus could not be redressed by antitrust law.[56] The Federal Circuit found that the district court had properly applied a rule of reason analysis by placing the initial burden on the plaintiff to show that the settlement had an adverse effect on competition in the relevant market, in this case the market for ciprofloxacin.[57]

In addition, the Federal Circuit held that, in the absence of fraud in procuring the patent or sham litigation, a court "need not consider the validity of the patent in the antitrust analysis of a settlement agreement involving a reverse payment."[58] That is, under the Federal Circuit's holding in *Cipro*, a *bona fide* litigation as to a patent's validity or application can be settled within the scope of the exclusionary zone of the patent.[59]

The Federal Circuit observed that the same result is reached by starting from the doctrine of patent immunity.[60] The court cited authorities indicating that, where the patent holder does not extend the exclusionary power obtained from the patent beyond the scope of the patent, the patent holder is generally immune from the application of the antitrust laws.[61] The Federal Circuit indicated that, while the district court conducted its analysis under the antitrust laws, it was implicitly respecting and affirming the traditional doctrine of patent immunity, which displaces the antitrust laws within the exclusionary zone of a patent:

> "[T]he [district] court simply recognized that any adverse anti-competitive effects within the scope of the . . . patent could not be redressed by antitrust law. This is because a patent by its very nature is anticompetitive; it is a grant to the inventor of the right to exclude others from making, using, offering for sale, or selling the invention. . . . Thus, a patent is an exception to the general rule against monopolies and to the right of access to a free and open market. The district court appreciated this underlying tension between the antitrust laws and the patent laws when it compared the anti-competitive effects of the Agreements with the zone of exclusion provided by the claims of the patent.
>
> * * * * *
>
> [T]he essence of the Agreements was to exclude the defendants from profiting from the patented invention. This is well within Bayer's rights as the patentee."[62]

The Federal Circuit also observed that the Eleventh Circuit in *Valley Drug* "did not advocate application of [an antitrust] analysis, finding such an analysis to be inappropriate given that the anticompetitive effects of the exclusionary zone of a patent are not subject to debate."[63] The Federal Circuit pointed to the Second Circuit's analysis in *Tamoxifen* in which the Second Circuit had concluded that the presence or size of a reverse payment "is not enough to render an agreement violative of the antitrust laws unless the anticompetitive effects of the agreement exceed the scope of the patent's protection."[64]

In summary, the Federal Circuit concluded that the outcome of the case was the same under both antitrust law and patent law:

> "[I]n cases such as this, wherein all anticompetitive effects of the settlement agreement are within the exclusionary power of the patent, the outcome is

> the same whether the court begins its analysis under antitrust law by applying a rule of reason approach to evaluate the anti-competitive effects, or under patent law by analyzing the right to exclude afforded by the patent. The essence of the inquiry is whether the agreements restrict competition beyond the exclusionary zone of the patent. This analysis has been adopted by the Second and Eleventh Circuits and by the district court below and we find it to be completely consistent with Supreme Court precedent."[65]

Plaintiffs' petition for certiorari in the United States Supreme Court seeking review of the Federal Circuit's decision was denied on June 22, 2009.[66]

V. Recent Agency Positions—*FTC v. Cephalon, Inc.*, *FTC v. Watson Pharmaceuticals, Inc.*, and the DOJ's Amicus Brief in *Arkansas Carpenters Health and Welfare Fund v. Bayer AG*

The FTC is apparently seeking to produce a split in the circuit courts on the lawfulness of reverse payments to encourage Supreme Court review.[67] To that end, in February 2008, the FTC filed a complaint in the United States District Court for the District of Columbia against Cephalon, Inc.[68] The FTC alleged that Cephalon willfully maintained its monopoly power with respect to its branded prescription narcolepsy drug, Provigil (modafinil), through a course of allegedly anticompetitive conduct that included entering into settlement agreements with potential generic competitors that, the FTC claims, included reverse payments.[69]

THE FTC IS APPARENTLY SEEKING TO PRODUCE A SPLIT IN THE CIRCUIT COURTS ON THE LAWFULNESS OF REVERSE PAYMENTS TO ENCOURAGE SUPREME COURT REVIEW.

The FTC filed suit in federal court rather than pursuing the conduct through the FTC's administrative process (as was done in Schering-Plough) perhaps to avoid an appeal to a circuit in which the law on reverse payments appears to be largely settled (e.g., the Eleventh or Second Circuit). The FTC is seeking a permanent injunction barring Cephalon from enforcing the terms of the agreements with the four generic companies that prevent those companies from marketing generic versions of Provigil before 2012.[70] The *Cephalon* case was transferred to the Eastern District of Pennsylvania in the Third Circuit. Motions to dismiss in the cases are pending. The FTC action is accompanied by private actions also challenging the *Cephalon* settlements.[71]

More recently, the FTC challenged payments by Solvay Pharmaceuticals, Inc. to generic manufacturers of its testosterone-replacement drug AndroGel—Watson Pharmaceuticals, Inc. and Par Pharmaceutical Companies, Inc—in connection with a co-marketing arrangement and a patent-infringement settlement agreement that defers generic entry until 2015. The FTC filed a complaint in the United States District Court for the Central District of California, alleging violations of the FTC Act, Sherman Act, and California unfair competition laws.[72]

According to the FTC, while Solvay's patent for AndroGel expires in 2020, ANDA first-filer Watson received FDA approval to market its generic AndroGel in 2006.[73] As alleged in the Commission's complaint, Solvay had estimated that a generic launch in mid-2006 would result in a loss of 90 percent of its sales within the year and in a decline in annual profits by about $125 million.[74] The FTC claims that Solvay agreed to pay Watson $19 million for the first year and an estimated $30 million annually for the next five years,[75] and also agreed to pay Par $12 million annually for six years, purportedly in connection with co-marketing or back-up manufacturing arrangements.[76]

The Commission relied on arguments by Watson and Par in their Paragraph IV litigation with Solvay to allege that Solvay's patent was unlikely to exclude generic competition and that the settlement agreement was an anticompetitive agreement to share monopoly profits.[77] In a statement released with the filing of the Solvay complaint, then-Commissioner Leibowitz indicated that the Commission will continue to challenge such patent settlements as anticompetitive.[78] The district court in the Central District of California in April granted defendants' motion to transfer the case to the Northern District of Georgia in the Eleventh Circuit, where the underlying patent-infringement suit was litigated.[79]

WHILE STILL A "RULE OF REASON" ANALYSIS, THIS "PRESUMPTIVELY UNLAWFUL" APPROACH PLACES A HEAVIER BURDEN ON THE DEFENDANTS THAN THE DOJ HAD PREVIOUSLY ADVOCATED.

Finally, in July 2009, in response to an invitation from the Second Circuit to address the challenge to the *Cipro* settlement in *Arkansas Carpenters Health and Welfare Fund v. Bayer AG*, the DOJ advocated that reverse-payment settlements be treated as "presumptively unlawful."[80] The DOJ argued that, if the settlement allows no generic competition until patent expiration, defendants generally will be unable to rebut the presumption with a reasonable explanation for the payment. Even if both parties believe the patentee is likely to win the validity litigation, the DOJ would view the settlement as anticompetitive because "it eliminates the possibility of competition from the generic" before the patent's expiration.[81]

While still a "rule of reason" analysis, this "presumptively unlawful" approach places a heavier burden on the defendants than the DOJ had previously advocated. In 2008, arguing against a *per se* approach, the DOJ expressed caution in impeding Hatch Waxman settlements:

> "In [the context of Hatch-Waxman settlements], per se illegality could increase investment risk and litigation costs to all parties. These factors run the risk of deterring generic challenges to patents, delaying entry of competition from generic drugs, and undermining incentives to create new and better drug treatments or studying additional uses for existing drugs."[82]

Then, the DOJ also emphasized the government's strong policy of encouraging the settlement of litigation to explain its reservations with a *per se* illegality rule.[83] The DOJ, through the Solicitor General, even confronted the FTC position by submitting an amicus brief to the Supreme Court in *Schering-Plough* that recommended that the Court deny the FTC's petition for certiorari.[84] In its brief, the DOJ highlighted competing policy considerations between patent rights and antitrust laws and asserted that "the mere presence of a reverse payment in the Hatch-Waxman context is not sufficient to establish that the settlement is unlawful."[85]

In contrast, in its more recent amicus brief to the Second Circuit in *Arkansas Carpenters*, the DOJ argued that the *Tamoxifen* standard "inappropriately permits patent holders to contract their way out of the statutorily imposed risk that patent litigation could lead to invalidation of the patent while claiming antitrust immunity for that private contract."[86] The DOJ also cautioned against embedding a patent trial within an antitrust trial, acknowledging that its current views are in tension with its previous call for an examination of the patent infringement claim's merits.[87] The DOJ argued that it is "neither necessary nor appropriate to determine whether the patent holder would likely have prevailed in the patent infringement litigation."[88] Instead, the DOJ advocated that the court base liability "on whether, in avoiding the prospect of invalidation that accompanies infringement litigation, the parties have by contract obtained more exclusion than warranted in light of that prospect."[89]

VI. Pending Legislation Seeks to Prohibit Reverse-Payment Agreements

Some in Congress do not believe that the appellate courts have been properly analyzing reverse-payment agreements and have proposed legislation to limit or prohibit such agreements. For example, Senator Herbert Kohl (D-WI) and Representative Bobby Rush (D-IL) introduced in the Senate and House, respectively, legislation that would specify the legal treatment of reverse-payment agreements. The Kohl bill (S. 369), which is entitled the "Preserve Access to Affordable Generics Act,"[90] was initially drafted to ban reverse-payment agreements and has since been modified to treat reverse-payment agreements as presumptively unlawful.[91]

The Kohl bill would amend the FTC Act to declare presumptively unlawful any agreement "resolving or settling, on a final or interim basis, a patent infringement claim" in which a generic drug company (1) "receives anything of value" from the brand company, and (2) "agrees to limit or forego research, development, manufacturing, marketing, or sales of the [generic] product for any period of time."[92] The Kohl bill would allow the presumption of unlawfulness to be overcome by "clear and convincing evidence that the procompetitive benefits of the agreement outweigh the anticompetitive effects."[93]

Excluded from prohibition under the Kohl bill are agreements in which (1) the value that the generic company receives is no more than the right to market its product prior to the expiration of the allegedly infringed patent or other statutory exclusivity; (2) the payment is for reasonable litigation expenses not exceeding $7.5 million; or (3) the brand company covenants not to sue for patent infringement by the generic product.[94] The Kohl bill would also authorize the FTC to exempt, by rule, certain agreements that it finds will further competition and benefit consumers.[95] The Senate Judiciary Committee on October 15, 2009, voted to place the Kohl bill on the legislative calendar for consideration by the full Senate.[96]

The Rush bill in the House (H.R. 1706 entitled the "Protecting Consumer Access to Generic Drugs Act") would treat violations as an unfair method of competition under section 5 of the FTC Act.[97] The Rush bill would prohibit agreements in which an ANDA filer "receives anything of value" and "agrees not to research, develop, manufacture, market, or sell, for any period of time, the [generic] drug."[98] An exception is made for generic companies receiving no more than the right to market the drug and a waiver of the patent holder's claim for damages based on prior marketing of the drug. The Rush bill also authorizes the FTC to exempt, by rule, certain agreements that it finds will further competition and benefit consumers.[99]

As anticipated, the Obama Administration seems to support the legislative restriction of reverse payments.

As anticipated, the Obama Administration seems to support the legislative restriction of reverse payments. The FTC has been a vocal advocate for legislation addressing the reverse payments issue for some time. FTC Chairman Leibowitz has indicated that he views the elimination of reverse payments a top priority in antitrust enforcement under the new administration:[100] "The new administration does seem to recognize that [pay-for-delay settlements are] a real problem for consumers, [and] fixing it … would actually help pay for healthcare reform."[101] Indeed, then-Senator Obama (along with nine other Democratic senators) co-sponsored a previous version of the Kohl bill in 2007.[102]

The fact and form of any legislative response to reverse payments, however, remain the subject of debate.

VII. The European Commission Examines Settlement Practices in the Pharmaceutical Industry

In an inquiry into competitive practices in the European pharmaceutical sector, the European Commission ("EC") investigated over 200 brand and generic companies for the period from 2000 through 2008. On July 8, 2009, the EC issued a final report of its inquiry[103] that examined (among a variety of other subjects) settlements and other agreements between patent holders and generic companies, their effect on generic entry, and the cost of pharmaceutical products. The EC found that just under half of the 207 total settlement agreements concluded between patent holders and generic companies during the time studied imposed a restriction on the generic company's ability to market its medicine.[104] Of those restrictive settlements, 45 percent included a value transfer from the patent holder to the generic company in the form of a direct payment, license, or distribution agreement.[105]

Twenty-three settlement agreements, or approximately 10 percent of all settlements and 23 percent of settlements that restricted entry, included cash payments totaling over 200 million euros.[106] In six of the 23 agreements, the generic company agreed not to enter the market until a court judgment on patent infringement had been decided. In the remaining 17 cases, the generic company agreed either to exit or not enter the market until after the brand company's patent expired.[107] The report also provided a brief overview of the U.S. assessments of such settlement agreements, discussing the FTC enforcement measures in the *Cephalon* and *Solvay* cases and the Eleventh Circuit's decision in *Schering-Plough*.[108]

While the report encourages EU member states to pass legislation to create a unified patent and litigation system, no EU legislation to ban reverse-payment settlements has been proposed.

The EC report identified for further scrutiny "[s]ettlement agreements that limit generic entry and include a value transfer from an originator company to one or more generic companies [as] potentially anticompetitive agreements."[109] In a statement issued with the release of the report, the European Commissioner for Competition, Neelie Kroes, said that "[t]he first antitrust investigations are already underway, and regulatory adjustments are expected to follow dealing with a range of problems in the sector."[110]

While the report encourages EU member states to pass legislation to create a unified patent and litigation system,[111] no EU legislation to ban reverse-payment settlements has been proposed.

VIII. Conclusion

The Federal Circuit decision in *Cipro*, the most recent appellate judicial analysis of reverse-payment settlements, has synthesized the approaches in the Second and Eleventh Circuits in finding that reverse payments within the exclusionary scope of the patent do not violate the antitrust laws. The Federal Circuit employed both rule of reason and patent-immunity principles in reaching that conclusion. The FTC continues to challenge reverse-payment settlements, with the apparent goal of producing a circuit split and attracting Supreme Court review.

Congress continues to consider various responses to reverse-payment agreements. The EC is beginning to review the settlement of patent litigation in the context of its competition laws, and its pharmaceutical sector report has shown the EC's interest in the treatment of such settlement agreements by the U.S. courts and enforcement agencies.

---

1 Pub. L. No. 98-417, 98 Stat. 1585 (1984) (codified as amended at 21 U.S.C. § 355 (2003)). The Hatch-Waxman Act was originally passed in 1984 and sponsored by Senator Orrin Hatch (R-UT) and Representative Henry Waxman (D-CA).

2 21 U.S.C. § 355(j) (2003).

3 *Id*. § 355(j)(2)(A)(vii)(I)-(IV).

4 For a thorough discussion of the numerous and complex rules relating to Paragraph IV certification *see, e.g.*, In re Tamoxifen Antitrust Litig., 466 F.3d 187, 190-93 (2d Cir. 2006).

5 21 U.S.C. § 355(j)(5)(B)(iv).

6 *See Id*. § 355(j)(5)(B)(iii). A court may also shorten or lengthen the thirty-month period pursuant to this section.

7 *Id*. §§ 355(j)(2)(B)(ii)(I), (j)(5)(B)(iii); 21 C.F.R. § 314.95(f) (2007).

8 *See, e.g.*, Jon Leibowitz, Commissioner, FTC, Prepared Remarks at the Antitrust in HealthCare Conference: Health Care and the FTC: The Agency as Prosecutor and Policy Wonk (May 12, 2005), 6, available at http://www.ftc.gov/speeches/leibowitz/050512healthcare.pdf; Deborah Platt Majoras, Chairman, FTC, Prepared remarks at the World Congress Leadership Summit, New York: The Federal Trade Commission: Fostering a Competitive Health Care Environment that Benefits Patients (Feb. 28, 2005), 10, *available at* http://www.ftc.gov/speeches/majoras/050301healthcare.pdf; Timothy J. Muris, Chairman, FTC, Prepared Statement of the FTC before the Committee on Energy and Commerce Subcommittee on Health (Oct. 9, 2002), 2, *available at* http://ftc.gov/os/testimony/107hearings.shtm.

9 In re Abbott Labs., No. C-3945, 2000 WL 681848 (F.T.C. May 22, 2000). "Interim agreements" do not purport to resolve the underlying patent litigation, but rather typically have specified the generic company's competitive conduct during the pendency of the patent litigation.

10 *Id*. at 4-5.

11 *Id*. at 7.

12 *Id*. at 8.

13 In re Hoechst Marion Roussel, Inc, No. 9293 (decision and order) (May 8, 2001), *available at* http://www.ftc.gov/os/caselist/d9293.shtm.

14 The FTC alleged in its complaint that accompanied its consent order that the HMR-Andrx agreement also had the purpose and intended effect of deterring Andrx from selling "non-infringing or potentially non-infringing" drugs. In the Matter of Hoecht Marion Roussel, Inc., No. 9293 (Complaint), 6, (Mar. 16, 2000) *available at* http://www.ftc.gov/os/2000/03/hoechstandrxcomplaint.htm.

15 *Id.* at 4-5.

16 *See* In re Hoechst Marion Roussel, Inc, No. 9293 (decision and order) at 5.

17 *See* In re Schering-Plough Corp. (Consent Order as to American Home Products Corporation) (Apr. 2, 2002), *available at* http://www.ftc.gov/os/caselist/d9297.shtm.

18 *See* In re Schering-Plough Corp., No. 9297 (Initial Decision) (July 2, 2002), *available at* http://www.ftc.gov/os/adjpro/d9297/index.shtm; In re Schering-Plough Corp., 136 F.T.C. 96, No. 9297 (Opinion of the Commission) (Dec. 18, 2003), *available at* http://www.ftc.gov/os/adjpro/d9297/index.shtm, rev'd sub nom. Schering Plough Corp. v. FTC, 402 F.3d 1056 (11th Cir. 2005), cert denied, 548 U.S. 919 (2006).

19 In re Schering-Plough Corp., No. 9297 (Initial Decision).

20 In re Schering-Plough Corp., 136 F.T.C. 956, No. 9297 (Opinion of the Commission) at 1052.

21 *Id*. at 988.

22 In re Cardizem CD Antitrust Litig., 332 F.3d 896 (6th Cir. 2003), cert. denied sub nom, Andrx Pharmaceuticals, Inc. v. Kroger Co., 543 U.S. 939 (2004).

23 In re Cardizem CD Antitrust Litig., 105 F. Supp. 2d 682, 699 (E.D. Mich. 2000); *see also* Cardizem, 332 F.3d at 908 n.13.

24 The Honorable Louis F. Oberdorfer is a Senior Judge for the U.S. District Court for the District of Columbia.

25 Cardizem, 332 F.3d at 902-03, 907-08 & n. 13 (citation and internal quotations omitted).

26 *Id.* at 908.

27 *Id*. at 907-8.

28 344 F.3d 1294 (11th Cir. 2003), cert. denied, 548 U.S. 919 (2006).

29 *See* In re Terazosin Hydrochloride Antitrust Litig., 164 F. Supp. 2d 1340, 1348-54 (S.D. Fla. 2000), rev'd, 344 F.3d 1294 (11th Cir. 2003).

30 Valley Drug, 344 F.3d at 1306.

31 *Id*. at 1311.

32 *Id*. at 1306-9.

33 544 F.3d 1323 (Fed. Cir. 2008).

34 Schering-Plough, 402 F.3d 1056, 1066 (11th Cir. 2005) (citing Valley Drug, 344 F.3d at 1312), cert. denied, 548 U.S. 919 (2006).

35 Valley Drug, 344 F.3d at 1310-11.

36 Schering-Plough, 402 F.3d at 1067-68 (citing Asahi Glass Co. v. Pentech Pharmaceuticals, Inc., 289 F. Supp. 2d 986, 991 (N.D. Ill. 2003)).

37 In re Schering-Plough Corp., No. 9297 (Opinion of the Commission) at 971-72.

38 *See Id*. at 987-88.

39 15 U.S.C. § 45(c) (2007) (permitting an appeal of FTC decisions to any circuit where the respondent resides or where the challenged conduct was used).

40 Schering-Plough, 402 F.3d 1056, 1075-76 (11th Cir. 2005), cert. denied, 548 U.S. 919 (2006).

41 *See* Brief for the United States as Amicus Curiae in FTC v. Schering-Plough Corp., 548 U.S. 919 (2006) (No. 05-273), at *16-20, available at http://www.usdoj.gov/atr/cases/f216300/216358.pdf.

42 FTC v. Schering-Plough Corp., 548 U.S. 919 (2006) (No. 05-273).

43 In re Tamoxifen Citrate Antitrust Litig., 466 F.3d 187, 190 (2d Cir. 2006).

44 *Id*. at 213-16.

45 *Id*. at 207.

46 *Id*. at 203-04.

47 *Id*. at 208.

48 *Id*. at 210 (citing Valley Drug, 344 F.3d at 1310 ("Given the asymmetries of risk and large profits at stake, even a patentee confident in the validity of its patent might pay a potential infringer a substantial sum in settlement.")).

49 *Id*. at 212 (citing Schering-Plough, 402 F.3d at 1076).

50 *See Id*. at 208-09, 212-13, 217; *see also* Schering-Plough, 402 F.3d at 1066-68 (citing Asahi, 289 F. Supp. 2d at 991). The appellate courts have not stated that liability would attach to generic defendants for settling a matter where the branded company had filed baseless litigation or obtained the patent through fraud.

51 544 F.3d 1323, 1327-28 (2008).

52 *Id*. at 1328-29.

53 *Id*. at 1329.

54 In re Ciprofloxacin Hydrochloride Antitrust Litigation, 363 F. Supp. 2d 514 (E.D.N.Y. 2005); Cipro, 544 F.3d at 1330.

55 In re Ciprofloxacin Hydrochloride Antitrust Litigation, No. 05-2863-CV (2d Cir. Nov. 7, 2007) (unpublished order).

56 *Id*. at 1330 (citing Cipro, 363 F. Supp. 2d at 523-40).

57 *Id*. at 1332.

58 *Id*. at 1336.

59 *Id*. at 1337.

60 *Id*. at 1336 ("[T]he outcome is the same whether the court begins its analysis under antitrust law by applying a rule of reason approach . . . or under patent law by analyzing the right to exclude afforded by the patent").

61 *Id*. at 1333 (citing United States v. Gen. Elec. Co., 272 U.S. 476, 485 (1926), which states: "It is only when [the patent owner] . . . steps out of the scope of his patent rights . . . that he comes within the operation of the Anti-Trust Act"; E. Bement & Sons v. Nat'l Harrow Co., 186 U.S. 70, 91 (1902), which states: "The very object of [the patent] laws is monopoly, and the rule is, with few exceptions, that any conditions [imposed by the patentee] which are not in their very nature illegal . . . will be upheld by the courts"; In re Tamoxifen, 466 F.3d at 201-02; Valley Drug, 344 F.3d at 1312; United States v. Studiengesellschaft Kohle, m.b.H, 670 F.2d 1122, 1127 (D.C. Cir. 1981)).

62 Cipro, 554 F.3d at 1333 (internal quotations and citations omitted).

63 *Id*. at 1335 (citing Valley Drug, 344 F.3d at 1312 n.27).

64 *Id*. at 1336 (citing Tamoxifen, 466 F.3d at 212-13).

65 *Id*. (citing Walker Process Equip., Inc. v. Food Mach. & Chem. Corp., 382 U.S. 172, 175-77 (1965) for the proposition that "there may be a violation of the Sherman Act when a patent is procured by fraud, but [that otherwise] a patent is an exception to the general rule against monopolies").

66 Ark. Carpenters Health and Welfare Fund v. Bayer AG, __ U.S. __, 104 S.Ct. 3587 (2009).

67 *See* Statement of Commissioner Jon Leibowitz before the Subcommittee on Commerce, Trade, and Consumer Protection (May 2, 2007), *available at* http://www.ftc.gov/speeches/leibowitz/070502reversepayments.pdf (stating that "[i]t's public knowledge that [the FTC is] looking to bring a case that will create a clearer split in the circuits").

68 FTC v. Cephalon, Inc., No. 08-0244 (D.D.C. Feb. 13, 2008) (Complaint), available at http://www.ftc.gov/os/caselist/0610182/080213complaint.pdf.

69 *Id*. at 25.

70 *Id*. at 27.

71 *See*, *e.g.*, In re Modafinil Antitrust Litigation, 06-01797 (E.D. Pa.); Apotex, Inc. v. Cephalon, Inc., 06 CV 02 768 (E.D. Pa.).

72 *See* Complaint, FTC v. Watson Pharmaceuticals, Inc., 09-00598 (C.D.Cal., Feb. 12, 2009).

73 *Id*. at ¶¶ 2, 44.

74 *Id*. at ¶ 50.

75 *Id*. at ¶ 66.

76 *Id*. at ¶ 73.

77 *Id*. at ¶¶ 58-92.

78 Jon Leibowitz, Concurring Statement of Commissioner Jon Leibowitz, Feb. 2, 2009, available at http://www.ftc.gov/os/caselist/0710060/index.shtm ("I strongly support our two-pronged approach to eliminating these unconscionable deals. First, we will continue to challenge patent settlements that are anticompetitive and force consumers to pay more for much needed drugs. Second, we will advocate for legislation along the lines of the bipartisan measure (introduced last Congress by Senators Kohl, Obama, Grassley, Durbin, and Schumer as well as Representatives Waxman, Dingell, and Rush), which would offer a simple, effective and straightforward solution to the problem by banning payments from the brand to the generic while permitting legitimate settlements.").

79 *See* Order Transferring Cases, FTC v. Watson Pharmaceuticals, Inc, No. 09-00955 (C.D.Cal. Apr. 8, 2009). Defendants' motions to dismiss are pending as of September 2009, with discovery scheduled to close in January 2010.

80 *See* Brief for the United States in Response to the Court's Invitation in Arkansas Carpenters Health and Welfare Fund v. Bayer AG, (2d Cir. 2009) (No. 05-2852), at 21-32, *available at* http://www.usdoj.gov/atr/cases/f247700/247708.pdf.

81 *Id*. at 29-30.

82 Cong. Rec. S1195 (Feb. 26, 2008).

83 *Id*.

84 Brief for the United States as Amicus Curiae in FTC v. Schering-Plough Corporation (No. 05-273) (2006).

85 *Id*. at 11.

86 *Id*. at 14.

87 *Id*. at 26, n.9.

88 *Id*. at 24.

89 *Id*. at 25.

90 S. 369, 111th Cong. § 3 (2009). The bill was formerly S. 316, 110th Cong. § 3 (2007), which expired in committee. The former Kohl bill was co-sponsored by Senator Leahy and eight other Democratic Senators. Senator Leahy is no longer a co-sponsor. Current co-sponsors include Senators Grassley (R-IA), Brown (D-OH), Feingold (D-WI), Durbin (D-IL), Collins (R-ME), Franken (D-MN), and Klobuchar (D-MN).

91 Jessica Dye, "Senate Panel Plans Looser Rules On Pay-For-Delay," Law360 (September 24, 2009).

92 S. 369 § 3 (2009). The bill reflects changes adopted unanimously by the Senate Judiciary Committee on September 24, 2009. As originally introduced, the Kohl bill would have amended the Clayton Act to

declare such agreements *per se* unlawful, excluding agreements in which the value that the generic company receives is no more than the right to market its product prior to the expiration of the allegedly infringed patent. S. 369, 111th Cong. § 3 (Feb. 3, 2009).

93 S. 369 § 3.

94 *Id*.

95 S. 369 § 3.

96 S. 369—111th Congress: Preserve Access to Affordable Generics Act. (2009). In *GovTrack.us (database of federal legislation)*. Retrieved Oct 27, 2009, from http://www.govtrack.us/congress/bill.xpd?bill=s111-369 (Oct. 26, 2009).

97 H.R. 1706 § 2(c).

98 H.R. 1706, 111th Cong. § 2 (2009). The Rush bill was formerly H.R. 1902, 110th Cong. (2007), which expired upon the conclusion of the 110th Congress.

99 H.R. 1706, 111th Cong. § 3 (2009).

100 Jon Leibowitz, Concurring Statement of Commissioner Jon Leibowitz, Feb. 2, 2009, *available at* http://www.ftc.gov/os/caselist/0710060/index.shtm. ("Eliminating these pay-for-delay settlements is one of the most important objectives for antitrust enforcement in America today.").

101 Anna Edney, Congress Daily, "FTC Eyes Aggressive Action On Generic Drugs," Feb. 19, 2009, *available at* http://lostintransition.nationaljournal.com/2009/02/ftc-eyes-aggressive.php.

102 Cong. Rec. S11505 (Sept. 12, 2007); *see also* Barack Obama, A New Era of Responsibility: Renewing America's Promise, Feb. 26, 2009, at p. 28, *available at* http://www.whitehouse.gov/omb/assets/fy2010_new_era/a_new_era_of_responsibility2.pdf.

103 European Commission, Pharmaceutical Sector Inquiry: Final Report, Competition DG (July 8, 2009) ("EC report").

104 *Id*. at 270 ¶ 743.

105 *Id*. at 270 Figure 106.

106 *Id*. at 279 ¶ 768.

107 *Id.* at 278 ¶ 767.

108 *Id*. at 287-89.

109 *Id*. at 524 ¶¶ 1573.

110 Press Release, Antitrust: shortcomings in pharmaceutical sector require further action, July 8, 2009, *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/09/1098&format=HTML&aged=0&language=EN&guiLanguage=en.

111 EC report at 525 ¶ 1578.

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Whistling Past the Graveyard: The Problem with *Per Se* Legality Treatment of Pay-for-Delay Settlements

Michael Kades

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Whistling Past the Graveyard: The Problem with the *Per Se* Legality Treatment of Pay-for-Delay Settlements

*Michael Kades**

Arguably, the most important debate in antitrust jurisprudence involves pay-for-delay patent settlements in which the brand company pays the generic to stay out of the market. As a matter of economics, it will generally be most profitable if the brand and the generic firm avoid the possibility of competition and share the resulting monopoly profits; however, such settlements will reduce competition and increase the costs of drugs. If pay-for-delay settlements are legal, parties will enter them to the detriment of consumers. Current cases, in particular the *Tamoxifen* and *Ciprofloxacin* decisions, however, have gone a long way towards adopting just such a standard, a standard that is already having an effect. By adopting an approach without regard to its implications or erroneously suggesting that pay for delay settlements are an ineffective way to delay competition, courts are essentially whistling past the graveyard. In addition, economics and empirical evidences explain why eliminating the 180-day exclusivity will not solve this problem. Unless changed the legal rule articulated in the *Tamoxifen* and *Ciprofloxacin* decisions will costs consumers billions of dollars.

*The author is an attorney advisor to Chairman Jon Leibowitz of the Federal Trade Commission. Both in that role and as a staff attorney in the Health Care Division of the Bureau of Competition he has worked on numerous matters advocating the illegality of certain patent settlements between brand and generic pharmaceutical cases, including *In re Schering Plough*, where he was on the trial team and argued the appeal on behalf of complaint counsel. He was also involved in the Commission amicus briefs in *Tamoxifen* and *Ciprofloxacin* and the investigation of Cephalon's Provigil settlements among others. The author wishes to express deep gratitude to Brad Albert, Mary Giovagnoli, Scott Hemphill, Elizabeth Hilder, Elizabeth Schneirov, Joel Schrag, and Dave Schmidt. Their thoughtful comments improved this article immensely. Kathryn Vajs provided invaluable help in editing and overcoming technological snafus.

I. Introduction

The phrase "Whistling Past the Graveyard" has many related meanings and appears in sources as varied as Robert Blair's *The Grave* and a Don Henley rock song. In Blair's poem, a school boy whistles "aloud to bear his courage up" as he passes by the graveyard, a scary and dangerous place. Rather than avoiding the danger (using a different route) or finding protection (walking with the group), the boy whistles, which, although it might make him feel better, does nothing to eliminate any real danger of the moment. In Don Henley's song *If Dirt Were Dollars*, the lines are an indictment of those who blithely ignore the problems surrounding them:

> "Gods finest little creatures
> Looking brave and strong
> Whistling past the graveyard
> Nothing can go wrong
> Quoting from the scriptures
> With patriotic tears…
> These days the buck stops nowhere
> No one takes the blame"[1]

Ignoring the implications of their antitrust analysis is precisely how courts have approached the issue of pay-for-delay patent settlements, also known as reverse-payment settlements or exclusion-payment settlements. Although these agreements occur only within the narrow range of pharmaceutical patent litigation, their growing prevalence makes them the subject of arguably the most important debate in antitrust jurisprudence.

The danger, or the graveyard, is that recent decisions have gone a long way towards adopting a rule of *per se* legality with respect to these settlements which, in turn, will dramatically increase prescription drug prices. Under the developing rule, a patent settlement is almost always legal if it allows the alleged infringer to enter the market at patent expiration or earlier. Such a settlement violates the antitrust laws only if the patent was obtained by fraud, the litigation is a sham, or the settlement blocks entry of a totally unrelated product. In other words, in virtually all pharmaceutical patent litigation, the brand may pay the generic to stay off the market until the expiration of the last patent. Under such a rule, payments to delay entry should occur in virtually every case because such settlements will be profitable for both parties. These settlements will delay generic entry, forcing consumers to pay substantially higher prices for prescription drugs. Already these deals are having an impact. A recent analysis by

Federal Trade Commission ("FTC") economists estimates that these types of deals cost consumers $3.5 billion per year.[2] That number will only increase if the courts settle on a rule of *per se* legality.

Courts are whistling past this disastrous result; brusquely dismissing it, or offering solutions that are ineffective. Decisions enunciating broad principles of legality generally ignore the implications of this rule. When they do consider it, they comfort themselves with the hypothesis that branded drug manufacturers cannot afford to pay off enough generic competitors to truly delay competition. Although there are many reasons to see this as little better than whistling, the most obvious is the specific regulatory framework under the Hatch-Waxman Act, the law controlling pharmaceutical approvals. In particular, the first company to file for generic approval generally receives 180 days of market exclusivity, meaning that the FDA will not approve a second generic filer during that time. Delaying the first-filer's entry, then, delays everyone else's. Although there are ways for subsequent filers to eliminate the first-filer's exclusivity, the first filer's exclusivity still creates a heightened barrier. Further, the parties to the settlement can structure it in such a way as to limit the subsequent filer's incentives to pursue a patent challenge.

COURTS ARE WHISTLING PAST THIS DISASTROUS RESULT; BRUSQUELY DISMISSING IT, OR OFFERING SOLUTIONS THAT ARE INEFFECTIVE.

In a variation on this theme, others have suggested that the problem is not the reverse payments, but is instead the 180-day exclusivity. In their view, we can avoid the ill effects of delayed entry if we eliminate the 180-day exclusivity for first filers who settle. Although there may be policy reasons for eliminating the 180-day exclusivity or creating incentives for subsequent challengers to eliminate that exclusivity, it will neither end the practice nor solve the problems created by pay-for-delay settlements. Even if there was no 180-day exclusivity, pay-for-delay settlements would still be profitable and delay generic entry for three reasons. First, the number of potential generic competitors may be low. Second, the transaction costs of the payments will be low. Third, it may be less expensive for a branded firm to pay off multiple generics rather than just one.

A fuller understanding of the pay-for-delay problem begins with a background of the regulatory and legal context of brand-generic patent settlements. Recent cases articulate a rule of *per se* legality for settlements in which entry occurs no later than patent expiration. This article does not explore the legal flaws in that proposed rule of *per se* legality; rather, the point is that no one should be fooled about the cost of such a rule. Both the brand and the generic will always earn more if the brand pays the generic to stay out of the market until patent expiration; therefore, these types of deals will only become more prevalent and delay generic entry longer. Finally, economics and empirical evidences explain why eliminating the 180-day exclusivity will not solve this problem.

II. Background on Patent Pharmaceutical Litigation

There are two basic ways to obtain approval for a prescription medication. First, a company can file a New Drug Application ("NDA") in which it demonstrates the safety and efficacy of its product, among other things.[3] Satisfying these standards requires costly clinical trials. It also provides a list of patents which cover the product, which the FDA makes public.[4] Most branded drugs receive approval through the NDA process.

Companies seeking to sell a generic version of a drug can follow a second path, called an Abbreviated New Drug Application ("ANDA"). Instead of repeating the expensive safety and efficacy testing of the brand product, they can establish that their product is bioequivalent to the brand product.[5] In addition, when the generic files its application, it must provide a certification as to each patent the brand has listed as covering its product. It can certify that there are no patents listed for the branded drug; that the listed patents have expired; that the generic will not sell the product until the listed products expire; or that the listed patents are invalid or not infringed by the generic's product.[6]

If the generic makes the last certification (that the patent is not infringed or is invalid), known as a paragraph IV certification, it must give notice to the brand company. In turn, if the brand company sues the generic within 45 days, the Food and Drug Administration may not approve the generic company's ANDA for thirty months.[7] In effect, just by filing suit the brand gets the equivalent of a thirty-month preliminary injunction[8]—without any showing on the merits.

In effect, just by filing suit the brand gets the equivalent of a thirty-month preliminary injunction—without any showing on the merits.

At the same time, the Hatch-Waxman Act gives the first generic applicant to make a paragraph IV certification (the "first filer") a valuable advantage. The FDA may not approve a second generic filer until 180 days after the first filer begins marketing its product.[9] Under current law, there are various ways a first filer can forfeit its exclusivity, but it generally requires a second filer either to prevail in a patent infringement action brought by the brand company or to win a declaratory judgment action that the patent is invalid or not infringed.[10] The first filer does not lose its exclusivity by settling the patent litigation. In other words, if the first filer agrees, as part of a patent settlement, not to enter until a year before patent expiration, the 180-day exclusivity period would prevent the FDA from approving any other generic until six months before patent expiration.

III. The Economics of Patent Settlements

Brand companies frequently sue generic companies for patent infringement. Generally hundreds of millions and, not infrequently, billions of dollars are at stake for the brand company. If the generic company successfully defends against the infringement claim, competition occurs. The generic will quickly take as much as 80 percent of the brand's prescriptions in a matter of months.[11] Although initially priced at roughly a twenty percent discount of the brand price, the generic price can fall to as little as twenty percent or less of the brand price when multiple generics enter.[12] In turn, consumers save billions of dollars from generic entry. Conversely, if the brand triumphs, it preserves hundreds of millions or even billions of dollars of additional revenue.

Weeding out weak patents or designing around narrow ones has helped control prescription drug costs. Many patents protecting brand products are weak or sufficiently narrow that they do not block generic entry. In those cases, successful generic challenges ensure that the patent does not deprive consumers of the benefits of generic competition. In the period 1992 to 2000, of those cases that went to trial, brand companies successfully protected their monopoly from all competition less than 30 percent of the time.[13] Another study showed that in the pharmaceutical industry, the alleged infringer won roughly two-thirds of the decisions in the Federal Circuit.[14]

ON FOUR BLOCKBUSTERS ALONE, CONSUMERS ARE EXPECTED TO SAVE OVER 16 BILLION DOLLARS BECAUSE OF GENERIC ENTRY PRIOR TO PATENT EXPIRATION.

The savings to consumers are substantial. On four blockbusters alone, consumers are expected to save over 16 billion dollars because of generic entry prior to patent expiration. See Table 1.[15] In each of those cases, the patent was no bar to competition. The generics' victory in the patent litigation ensured that consumers received the benefits of competition earlier rather than later.

Table 1

Examples of Generic Entry Prior to Patent Expiration from Successful Patent Challenges

Drug	Generic Entry Date	Years Prior to Patent Expiration	Brand Sales Prior to Generic Entry	Consumers Savings
Zantac	1997	5	$1.6 billion	$2.45 Billion
Taxol	2000	11	$1.6 billion	$3.5 Billion
Prozac	2001	2.5	$2.5 billion	$2.5 Billion
Buspar	2001	17	$600 million	$8.8 Billion

Of course, not every case goes to trial; many cases settle. As a matter of basic negotiation theory, parties should settle cases when the value of the settlement is greater than the perceived or expected value of pursuing the case to judgment.[16] When combined with the underlying economics of the pharmaceutical industry, this is a recipe for anticompetitive settlements.

Because generic entry causes total profits to fall, it is profitable for the parties to avoid entry if the branded company can compensate the generic company. The generic competes at a lower price, so the profits it makes by competing are lower than the amount the brand loses by facing competition. A simple pie chart, depicted in Figure 1, helps illustrate this point. In the top pie ("Monopoly"), the brand has all of the profit without generic competition. Once generic entry occurs (in the "Competition" pie), the brand loses substantial profits. Although the generic company earns profits, depicted by the dark green-shaded segment, consumers save money by buying the lower-priced generic, represented by the gray-shaded slice. In technical terms, the joint profits of the duopoly are less than the monopolist's profits.[17]

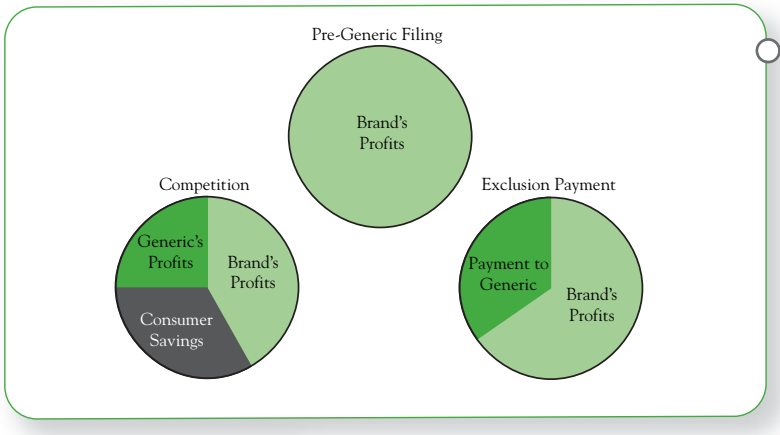


Figure 1

Incentives to Pay for Delay

Left to their own devices, both the incumbent monopolist and the entrant are better off if they eliminate competition and share the monopoly profits. As shown in the "Exclusion Payment" pie, the generic earns more from taking a payment not to compete than by entering the market. The dark green-shaded slice in the Exclusion Payment pie is larger than the green slice in the Competition pie; the brand is sharing a portion of its monopoly profits with the generic. Similarly, the light green-shaded slice of the Exclusion Payment pie, representing the branded company's profits, is larger than the light green-shaded slice of the Competition pie because the brand—despite paying the generic—earns more than if it faces competition.

During a patent litigation, there is obviously uncertainty about whether there will be a monopoly or competition after the court's decision. While that uncer-

tainty affects what the parties expect to earn from pursuing litigation, it still does not change the fact that, for the vast majority of situations, the brand and the generic can each earn more if the generic agrees not to compete and the brand pays the generic more than the generic values the litigation. The weaker the patent, the more willing the brand is to pay the generic. Graphically, the pie charts of Figure 1 need only a slight modification to account for the uncertainty of the litigation. As Figures 2 and 3 show, we start with the same pre-generic-filing pie, with the brand earning its monopoly profits.

The "Expected Competition" pie replaces the Competition pie. Compared to the Competition pie in figure 1, the slices of the generic's profits and the consumer savings are all discounted by the chance that the patent will block competition while the brand's expected profits are larger to account for the possibility that it may win the patent suit. In Figure 2, the patent is weak, so the expected profits of the brand and the generic are similar to what would happen if there is competition. In contrast, in Figure 3, the patent is strong; both the generic's expected profits and the expected consumer savings are small, but the brand's expected profits are much larger than what it would earn if there is competition.

Figure 2

Incentives to Pay for Delay (Weak Patent)

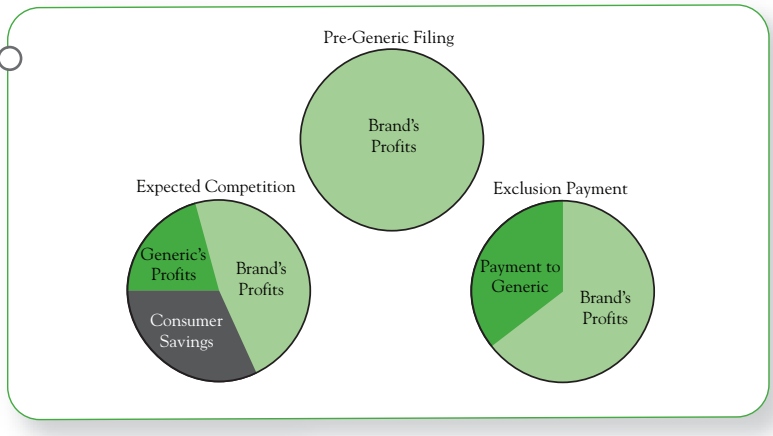
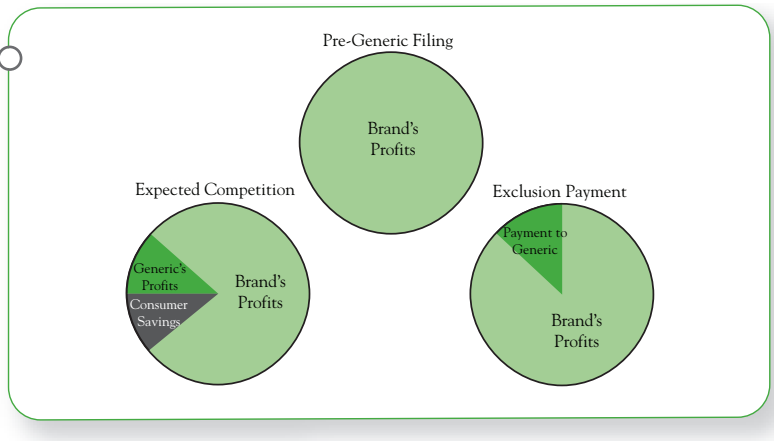


Figure 3

Incentives to Pay for Delay (Strong Patent)



The important point is that, whether competition is certain (Figure 1), the patent is weak (Figure 2), or the patent is strong (Figure 3), the brand and the generic are better off preserving the monopoly by having the branded firm pay the generic company not to enter. Left to their own devices—in a world where they face no antitrust constraints—most brand-generic patent cases should settle with the brand paying the generic not to compete during some portion of the remaining patent term, making what have become known as reverse payments, exclusion payments, or pay-for-delay settlements. The strength of the patent—how likely it is to block competition—would not determine when there is competition; rather, the profits the branded firm earns by eliminating the threat of generic competition and the brand's willingness to share those profits determines when competition would occur.

Not surprisingly, payments are more likely to protect weak patents than strong ones. The brand has more to lose from litigation and is willing to pay more to avoid that possibility than if the patent is strong. Because the payment involves dividing up the gray slice—the avoided consumer savings—and because the value of eliminating those consumers' savings is greatest when the patent is weak, payment becomes an especially attractive method for holders of weak patents to prevent competition.

At the same time, and this is the point of Figure 3, allowing branded firms with strong patents to pay off their generic competitors still harms competition. As Hovenkamp, Janis, & Lemley have explained, even where the patentee has a 75 percent chance of winning, the reason it "is willing to make this payment is precisely because there is a 25 percent chance that the patent would be held invalid or not infringed and the market would become competitive."[18]

The proposition that brands and generics can earn more from the brand paying the generic not to enter than they can from litigating or settling without a payment is uncontroversial. In the 1980s, Michael Meurer identified the general result that allowing a monopolist patent holder to make a lump sum payment to the alleged infringer would eliminate litigation and preserve monopoly profits.[19] More recently, Carl Shapiro explained this result specifically in the context of brand-generic patent settlements: "For this reason, the FTC has a sound basis for its skepticism about 'reverse cash payments' from the patent holder to the challenger."[20] Even those who try to justify these agreements grudgingly acknowledge that brands and generics can earn more by eliminating potential competition and sharing the resulting profits.[21]

IV. Current Case Law

Herein is the danger, or the graveyard, so to speak. Allowing brands to settle patent litigation by paying generics will eliminate competition and cost consumers billions of dollars a year, but that is precisely the legal rule the courts are moving toward.

The *Tamoxifen* decision in the Second Circuit and the *Ciprofloxacin* decision in the Federal Circuit were not subtle in their analysis and approach. Both set out a fairly straightforward rule: Unless the patent was obtained by fraud or the litigation is a sham, a settlement in which the brand pays the generic not to enter the market with an allegedly infringing product until patent expiration is legal.

HEREIN IS THE DANGER, OR THE GRAVEYARD, SO TO SPEAK. ALLOWING BRANDS TO SETTLE PATENT LITIGATION BY PAYING GENERICS WILL ELIMINATE COMPETITION AND COST CONSUMERS BILLIONS OF DOLLARS A YEAR, BUT THAT IS PRECISELY THE LEGAL RULE THE COURTS ARE MOVING TOWARD.

The *Tamoxifen* decision was the first clear articulation of this broad rule of legality, granting a motion to dismiss a pay-for-delay challenge. Tamoxifen is an anticancer drug.[22] The brand, AstraZeneca, sued the first generic, Barr, but Barr won before the trial court.[23] While the case was on appeal, the parties settled. Barr agreed to stay out of the market until six months before patent expiration, and AstraZeneca paid Barr 21 million dollars.[24] After settling with Barr, AstraZeneca won three patent cases against successive generic filers.[25] Private plaintiffs challenged the Barr agreement as an unreasonable restraint of trade.[26] The trial court in the antitrust case dismissed the action, holding that the plaintiffs failed to state a claim upon which relief could be granted.[27] In affirming the lower court, the Second Circuit explained:

> "Unless and until the patent is shown to have been procured by fraud, or a suit for its enforcement is shown to be objectively baseless, there is no injury to the market cognizable under existing antitrust law, as long as competition is restrained only within the scope of the patent."[28]

In the Second Circuit's view, absent those exceptions, a settlement could restrain competition beyond the scope of the patent only if the generic agreed to stay out of the market beyond the patent's expiration or the agreement covered unrelated products.[29] Because the patent *might* block competition, the brand could pay to ensure that result:

> "So long as the patent litigation is neither a sham nor otherwise baseless, the patent holder is seeking to arrive at a settlement in order to protect that to which it is presumably entitled: a lawful monopoly over the manufacture and distribution of the patented product."[30]

Therefore, the Second Circuit concluded that "We do not think that the fact that the patent holder is paying to protect its patent monopoly, without more, establishes a Sherman Act violation.[31]

In *Ciprofloxacin*, the Federal Circuit adopted the Second Circuit's holding that there could be no antitrust liability for any exclusion unless it was greater than what the patent might prevent.[32] Ciprofloxacin is an antibiotic. The brand, Bayer, sued Barr, the first generic.[33] Before reaching trial, the parties settled. Bayer paid Barr close to $400 million, and Barr agreed to stay off the market until six months before patent expiration.[34] The Federal Circuit upheld the lower court's decision to grant summary judgment in favor of the defendants.[35] Because the patent might block entry, the Federal Circuit reasoned that the means the brand chooses to achieve that result, whether litigation or payment, made no difference:

> "[T]here is no legal basis for restricting the right of a patentee to choose its preferred means of enforcement and no support for the notion that the Hatch-Waxman Act was intended to thwart settlements."[36]

One could ask why the legal right to exclude embodied in the patent includes the right to use one's economic power— the sharing of monopoly profits—to eliminate competition. Or, put another way, if the patent's ability to block competition is uncertain, why should the patent-holder be able to guarantee its monopoly by paying its potential competitor to stay out of the market? In essence, the Second Circuit and the Federal Circuit believe that the patent-holder can buy off its potential competitor so long as the patent infringement claim is not a sham.

V. Implications of the Developing Rule

What happens if the rules of the Second and Federal Circuit become the law of the land? If it is legal to pay for delay, it will certainly be in most parties' interest. The effect of such a rule is obvious—it will reduce pharmaceutical competition and increase the cost of prescription drugs.

As to the first point, it is a matter of simple logic easily illustrated by probability charts. Suppose that there are six branded products. Suppose further that they are distinct. They contain different active ingredients; employ different methods of actions; are sold in different markets; and are produced by different companies. In each case, a generic has filed an application to sell its product before the expiration of the respective brand firm's patent, and each brand has a one in three

chance of winning. If all six cases go to judgment, as Figure 4 shows, on average two of the brand companies will win their suits, and there will be a monopoly until patent expiration (represented by green on the timeline). At the same time, on average, four generic firms will win, and there will be competition (represented by gray on the timeline). Of course, not all cases go to trial; some settle. While many factors affect the terms of the settlement, one would expect—if there are no payments to the generic—that the results would roughly correlate with the probable outcome of the patent litigation. So, if three cases settled without payments, as Figure 5 shows, the settlements might roughly prevent entry for one-third of the remaining patent life and allow competition for the remaining two-thirds of the patent life.

Figure 4

Results of Litigation

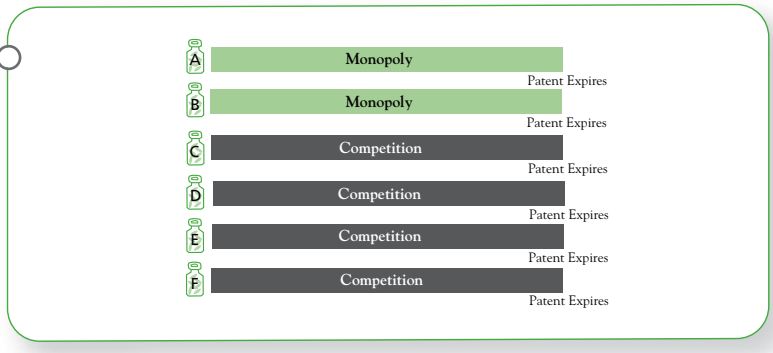
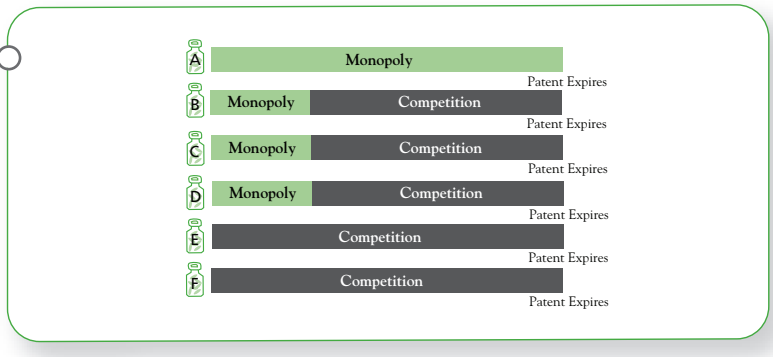


Figure 5

Settlements Without Payments



If, however, the brand may pay the generic, then, as depicted in Figure 6 below, the brand should simply pay the generic to stay out of the market until patent expiration.

Even if the brand has a relatively strong chance of winning, the payment eliminates "the incremental chance that the market would be competitive."[37] For example, if each brand had a two-thirds chance of winning its patent litigation,

then the amount of green and gray would roughly reverse in Figures 4 and 5. If they litigate, the brand on average wins four cases, and the generic wins two. A settlement with just a date would roughly protect two-thirds of the patent life. But, if payments are legal, one would still expect that the vast majority of settlements would reflect Figure 6. Overall, allowing payments will eliminate competition that would otherwise occur.

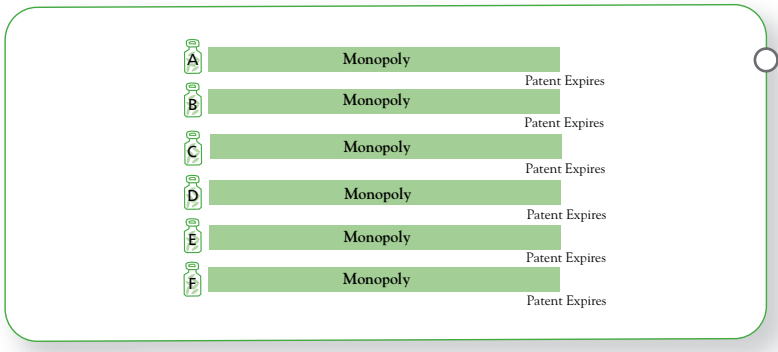


Figure 6

Impact of *Per Se* Legality

As a practical matter, the *Tamoxifen* and *Ciprofloxacin* decisions are already having a substantial effect. Based on the analysis of economists at the Federal Trade Commission, settlements with payments delay entry 17 months longer than settlements without payments.[38] Virtually all of the settlements with a payment to the generic occurred after the *Tamoxifen* decision.

Going forward, the impact could be staggering. Economists at the Federal Trade Commission estimate that, if nothing changes, settlements with payments for delay will cost consumers $3.5 billion dollars per year.[39] At the end of 2008, brands were attempting to block generic entry on products with roughly $90 billion in sales.[40] This is the current universe that the *Tamoxifen* and *Ciprofloxacin* rules could affect. On average, 15 percent of cases settle each year, and 24 percent of those settlements involve a payment. In turn, the delay of 17 months means consumers will lose 17 months of mature generic competition in the life of the product.[41] For a mature market, the average consumer savings is roughly 77 percent.[42] One can take the consumer savings, the length of delay, likelihood of settlement, and the pool of drugs for which settlements can be reached and calculate the harm to consumers.[43] Applying a different methodology and looking at the cost already incurred by consumers, Professor Hemphill estimated that settlements involving payments have already cost consumers 12 billion dollars.[44]

ECONOMISTS AT THE FEDERAL TRADE COMMISSION ESTIMATE THAT, IF NOTHING CHANGES, SETTLEMENTS WITH PAYMENTS FOR DELAY WILL COST CONSUMERS $3.5 BILLION DOLLARS PER YEAR.

While the description of any methodology may be dry, the implication is not. Under the *Tamoxifen* and *Ciprofloxacin* rule, uninsured patients will pay more for their drugs, insured patients will face higher copayments and premiums, and employers and the government will see their prescription drug costs rise.

Moreover, the FTC economists' estimate is almost certainly low if the *per se* rule of legality becomes the law of the land. Despite its setbacks, the FTC currently remains active in investigating and challenging patent settlements it considers to be anticompetitive, having attacked settlements on two separate products in the last two years. The FTC's enforcement may lead companies to be more cautious. For example, it is much easier to defend a particular settlement if, despite compensation to the generic, it allows for entry before patent expiration.[45] If the law definitively legalized these settlements, one would expect the companies to pursue more profitable settlements. There would be more pay-for-delay settlements, and in each settlement, the generic—in exchange for a larger payment—would agree to a longer delay.

Such concerns are far more like Cassandra's warnings than Chicken Little's predictions because brand and generic companies have already shown their willingness, in the absence of legal constraints, to push settlements out to patent expiration. In the 1990s, there was a period when drug companies entered pay-for-delay deals before the FTC, state attorneys general, or private plaintiffs were aware of the practice.[46] Based on the most comprehensive available statistics, between 1992 and 1999, there were eight final settlements in which the brand paid the generic, and the generic agree to stay off the market for some period of time.[47] In six of those eight settlements, the generic agreed to stay out of the market until patent expiration.[48] Those numbers suggest that if pharmaceutical companies could enter such deals with abandon, consumers would have to pay for higher-priced branded drugs for a much longer period of time.

SUCH CONCERNS ARE FAR MORE LIKE CASSANDRA'S WARNINGS THAN CHICKEN LITTLE'S PREDICTIONS BECAUSE BRAND AND GENERIC COMPANIES HAVE ALREADY SHOWN THEIR WILLINGNESS, IN ABSENCE OF LEGAL CONSTRAINT, TO PUSH SETTLEMENTS OUT TO PATENT EXPIRATION.

An example sheds light on the implications of a *per se* legality rule. In 1999, in the middle of the Prozac litigation between Lilly (the brand firm) and Barr (the generic company), Barr offered to settle and walk away from the litigation if Lilly would just pay Barr $200 million.[49] Lilly refused because it believed such payments were illegal.[50] Barr won the case, launched two-and-a-half years before the patent expired, and consumers saved roughly $2.5 billion.[51] If *Tamoxifen* and *Ciprofloxacin* were law, Lilly would never have taken the risk of losing the profit stream for its blockbuster.

VI. Whistling Toward *Per Se* Legality

The courts have adopted this rule with no concern for its impact—indeed, with barely any recognition of it. At best, they have offered dismissive explanations that are no better than whistling by the graveyard. The *Ciprofloxacin* court was largely silent on the economic implications of its rule. The *Tamoxifen* court assured its readers that payments were not a way to prevent competition because there were simply too many generics to pay off. That assurance, however, should provide little comfort because the economic and regulatory structure of the pharmaceutical industry makes pay-for-delay settlements quite an effective strategy.

The *Tamoxifen* court understood the incentives. The court recognized that the brand earns more profit when it pays for delay than the brand and generic earn if there is competition.[52] The *Tamoxifen* court further recognized the brand's incentive to pay to protect those profits.[53] And, it acknowledged the natural result: "it seems to make obvious economic sense for the generic manufacturer to accept such a payment if it is offered."[54] The court acknowledged "a troubling dynamic": "the less sound the patent or less clear the infringement and, therefore, the less justified the monopoly enjoyed by the patent holder, the more a rule permitting settlement is likely to benefit the patent holder by allowing it to retain the patent."[55]

This "troubling dynamic" gave the court little pause because it believed that a patent holder cannot realistically pay off every generic competitor:

> "But the answer to this concern lies in the fact that, while the strategy of paying off a generic company to drop its patent challenge would work to exclude that particular competitor from the market, it would have no effect on other challengers of the patent, whose incentive to mount a challenge would also grow commensurately with the chance that the patent would be held invalid."[56]

In the *Tamoxifen* court's view, then, there could be a problem only if the brand paid off all the generics, an event the court discounted: "We doubt, however, that this scenario is realistic."[57]

As a matter of economic theory, this tune has some attraction: Paying entrants to stay out of the market will be ineffective if there are too many and the costs of finding and negotiating with them are too high. This perspective ignores the reality of the pharmaceutical market. Even if the brand could only pay one generic and not block anyone else, the delay before a second entered (depending

on the circumstance, the second generic may have to develop the product, get approval, and win its law suit) would still harm consumers significantly.

A. THE 180-DAY EXCLUSIVITY MAKES PAYING THE FIRST FILER AN EFFECTIVE STRATEGY

More specifically, the *Tamoxifen* court ignored the biggest obstacle to taking comfort in the market. Under the Hatch-Waxman Act, the first ANDA filer has an exclusivity that makes it much more difficult for subsequent filers to enter, so the brand may not need to pay multiple filers. Moreover, later filers typically have less incentive to vigorously contest the brand-name firm's patents, because they do not receive the bounty of the 180-day exclusivity.

The 180-day exclusivity is the most obvious reason why the *Tamoxifen* court's assessment is misplaced. The first generic filer to certify that the patent covering the brand product is invalid or not infringed by the generic's product is the first applicant and has 180 days of market exclusivity, meaning the FDA will not approve a second generic until 180 days after the first filer markets its product.[58] If the first filer accepts a settlement in which it agrees to delay entry for five years, then the FDA may not approve a second filer for 5.5 years. This problem is often referred to as the bottleneck problem because the first filer's delay blocks additional entry.

In principle, the law provides a solution. If the second filer wins its patent case in a non-appealable decision and if the first filer does not launch within 75 days of that decision, the first filer loses it exclusivity.[59] The FDA may then approve the second or any other filer.

IN PRACTICE, HOWEVER, BRANDS HAVE FOUND WAYS TO LIMIT THE INCENTIVES OF THE SUBSEQUENT GENERIC TO PURSUE ITS CHALLENGE THROUGH A GENERIC ACCELERATION CLAUSE.

In practice, however, brands have found ways to limit the incentives of the subsequent generic to pursue its challenge through a generic acceleration clause. The first filer settles with an agreed entry date. The settlement also provides that the first filer may enter earlier if another generic wins (the entry date is accelerated to the date of the subsequent filer's victory).[60] As long as the first filer launches within 75 days of the subsequent filer's victory, the first filer does not forfeit its exclusivity, and the second filer must still wait another 180 days before receiving approval. Even if it wins, the subsequent filer still must likely wait for the expiration of the 180-day exclusivity period.

A forerunner of this dynamic occurred in the *Altace* litigation. The first filer (Cobalt) settled. Then, the second filer (Lupin) litigated and won. Under the legal scheme in place, a final appellate court decision triggered the exclusivity. Lupin, however, still had to wait 180 days before it could receive final approval.[61] Lupin would have been in a worse position under the current law because it might have had to wait as long as 255 days after its victory for final approval

because only commercial marketing will trigger the 180 day exclusivity. And, as long as the first filer launches within 75 days of the subsequent generic's victory, the first filer will not lose its exclusivity.

The subsequent filer faces reduced incentives to pursue its litigation. The brand pays the first filer and agrees that the first filer can accelerate its entry if another generic wins its patent litigation. Then, the brand offers the subsequent filer a settlement without a payment and entry 180 days after the first-filer enters. Even if it wins the litigation, the subsequent filer will still enter 180 days after the first filer. The value of winning then is in expediting the process, but this may not justify the cost of going forward. In many cases, a payment to the first filer and an acceleration clause may eliminate the incentive to subsequent generics to vigorously pursue their cases.

B. EVEN WITHOUT THE 180-DAY EXCLUSIVITY, PAYMENTS ARE LIKELY TO BE A SUCCESSFUL STRATEGY

It might be tempting to think that eliminating the 180 day exclusivity for settling first filers will fix the problem of payments,[62] but it will still be relatively easy for brands to pay generics. Without a broader prohibition on compensation, eliminating the exclusivity is likely to lead to bigger payments to the first filer because the branded company would have to compensate the first generic for the loss of exclusivity, but the payments will still occur. First, there will never be hundreds of companies waiting to enter the generic market. In most cases, the number of entrants—especially those willing to fight patent litigation—will likely be in the single digits.[63] Further, if the settlement allowed the first filer to enter as soon as anyone else, that clause would continue to dampen the subsequent filer's incentive to pursue its patent challenge.[64]

Second, the transaction costs in reaching such a deal are relatively low. Because any generic seeking approval to sell its product has to give notice to the brand, the brand always knows who is trying to enter. Because the parties are in litigation, the transaction costs of negotiating the payment are less than the alternative of litigating.

Third, paying-off multiple generics may not substantially increase the overall cost of the strategy; it may cost a brand less to pay two generics to delay entry than one, and it may cost even less to pay three. Additional generics, all of whom are essentially producing an identical branded product, will drive down the price.[65] At the same time, additional generic entry does not increase generic output.[66] If prices fall and output is constant, then overall revenues will fall. In turn, with multiple generics, each company expects to earn less. Because the first generic takes such a large portion of the branded company's sales, subsequent generic entry has little additional effect on the branded product's sales. So, the brand receives roughly the same benefit from paying off multiple generics as it does with one. Looking back to Figure 1, the dark green slice for generic profits

shrinks with more generics, but the light green slice (the branded company's expected profits) remains the same.[67] Therefore, the brand may need to pay less to eliminate the potential competition than if there were only one generic.

Assume that the brand product has yearly sales of one billion dollars. A single generic, assuming it takes 80 percent of the brand's sales and prices at a 30 percent discount, will earn roughly 560 million dollars in revenue. In contrast, if five generics enter, they drive the price down to 33 percent of the brand price. The total generic revenue will fall to 267 million dollars. In other words, if the brand has to pay the full revenue of the generics, it would actually cost more than twice as much to buy off one generic than five generics. Although these numbers may overstate the disparity between revenue in a sole generic market and a multiple generic market,[68] they clearly illustrate that eliminating the 180-day exclusivity outright may have the unintended consequence of making pay-for-delay settlements more common. On a blockbuster product with 11 or more filers, it may cost very little to pay all the subsequent filers.

VII. Conclusion

This then is what we can conclude about the *per se* legality rule for pay-for-delay settlements. If these settlements are legal, they will occur more frequently, reduce generic entry, and raise the costs of prescription drugs by billions of dollars a year. Furthermore, we should take no comfort in the idea that there are too many generics to pay or that a simple change to the 180 day exclusivity will solve the problem.

ARGUABLY, COURTS DECIDE CASES, NOT PUBLIC POLICY; THEREFORE, THEY NEED NOT CONSIDER THE IMPLICATIONS OF THE RULES THEY DEVELOP. THAT IS AN ESPECIALLY ODD CLAIM IN ANTITRUST LAW.

Arguably, courts decide cases, not public policy; therefore, they need not consider the implications of the rules they develop. That is an especially odd claim in antitrust law, which over the last thirty years has imported ever more sophisticated economic reasoning and policy into its legal rules. If one believes *Tamoxifen* and *Ciprofloxacin* correctly state the law, it is hard to imagine how the results of the per se rule of legality represent sound public policy: the rule protects weak and narrow patents, prevents competition, and raises drug costs.

If the courts are whistling towards *per se* legality, Congress has shown a much deeper and more sophisticated appreciation for the problem. In the House, the Energy and Commerce Committee voted out a bill to eliminate the practice and incorporated it into health care reform. The Senate Judiciary Committee favorably reported a bill that would apply a much more stringent standard. The danger of pay-for-delay settlements is as real as its cost to consumers; the cost of whistling towards *per se* legality is far greater than whistling by the graveyard.

1 Don Henley, *If Dirt Were Dollars*, on The End of Innocence (Geffen Records 1989).

2 *See*, *infra* notes 28-37

3 21 U.S.C. § 355(d).

4 21 U.S.C. § 355(b)(1).

5 21 U.S.C. § 355(j).

6 21 U.S.C. § 355(j)(2)(A)(vii)(IV).

7 21 U.S.C. § 355(j)(5)(B)(iii).

8 The stay will also expire if the district court rules in favor of the generic firm in the rare event that this happens sooner.

9 *Id*. § 355(j)(5)(B)(iv).

10 *Id*. § 355(j)(5)(D). Under current law, this generally requires a victory at the court of appeals.

11 Studies relying on data from the 1980s and early 1990s found generics taking 38-50 percent of the brand's sales within the first year. *How increased Competition from Generic Drugs has Affected Prices and returns in the Pharmaceutical Industry*, A CBO Report (July 1998) (collecting studies), *available at* http://www.cbo.gov/ftpdocs/6xx/doc655/pharm.pdf (last visited September 23, 2009). More recently, generic substitution rates have accelerated. Within the first week of generic Prozac's launch, Merk-Medco achieved an 80 percent substitution rate for mail order customers. *See Merck-Medco Achieves 80 Percent Generic Substitution Rate in First Week*, BW HealthWire (August 20, 2001), available at http://www.allbusiness.com/health-care/health-care-regulation-policy-health/6112721-1.html (last visited September 24, 2009).

12 David Reiffen & Michael R. Ward, *Generic Drug industry Dynamics*, 87 Rev. Econ. & Stats. 37, 38 (2007); Seven N. Wiggins & Robert Maness, *Price Competition in Pharmaceuticals: The Case of Anti-infectives*, 42 Econ. Inquiry 247, 247.

13 Federal Trade Commission, *Generic Drug Entry Prior to Patent Expiration: An FTC Study*, 19-20 (July 2002), *available at* http://www.ftc.gov/os/2002/07/genericdrugstudy.pdf (last visited September 19, 2009).

14 Paul Janicke & Lilan Ren, *Who Wins Patent Infringement Cases?* 34 AIPLA Q.J. 1, 20 (2006), *also* John R. Allison & Mark A. Lemley, *Empirical Evidence on the Validity of Litigated Patents*, 26 AIPLA Q.J. 185, 205-06 (1998) (study of all patent validity litigation from 1989-1996 found 46 percent of all patents litigated to judgment held invalid).

15 Written Statement of GPHA (Kathleen Jaeger), *Closing the Gaps in Hatch-Waxman: Assuring Greater Access to Affordable Pharmaceuticals*, Senate Committee on Health, Education, Labor and Pensions, at 60-61, 107th Congress, May 8 2002, *available at* http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_senate_hearings&docid=f:79636.pdf; *FTC Charges Bristol-Myers Squibb with Pattern of Abusing Government Processes to Stifle Generic Drug Competition, available at* http://www.ftc.gov/opal2003/03/bms.shtm. Some of these products have a checkered past. For example, in both BuSpar and Zantac, there were payments to generic competitors to stay out of the market until an earlier patent expired. *See* Cmpl. ¶¶ 25-31, In re Bristol-Meyers, Docket C-4076, March 7, 2003, *available at* http://www.ftc.gov/os/2003/03/bristolmyerscmp.pdf (last visited September 19, 2009) (BuSpar) and C. Scott Hempill, *An Aggregate Approach to Antitrust: Using New Data and*

Rulemaking to Preserve Drug Competition, 109 Colum. L. Rev. 629, 649 (2009). It is unclear whether these figures are discounted for inflation.

16 *See generally*, Robert D. Cooter & Daniel Rubinfeld, *Economic Analysis of Legal Disputes and Their Resolution*, 26 J. Econ. Lit., 1067 (1989).

17 *See e.g.*, Denis Carlton & Jeffrey Perloff, Modern Industrial Organization, 4th edition, 8, (2005).

18 Herbert Hovenkamp, Mark Janis, & Mark A. Lemley, *Anticompetitive Settlements of Intellectual Property Disputes*, 87 Minn. L. Rev, 1719, 1759 (2003). For an explanation of how payments in cases with relatively strong patents harm competition as a matter of policy, *see* discussion *infra* at note 31.

19 Michael J. Meurer, *The Settlement of Patent Litigation*, 20 Rand J. Econ. 77, 78 (1989).

20 Carl Shapiro, *Antitrust Limits to Patent Settlements*, 34 Rand J. Econ. 391, 408 (2003).

21 Robert D. Willig & John P. Bigelow, *Antitrust Policy Towards Agreements that Settle Litigation*, Antitrust Bull. 655, 659 (2004) "Second, in contrast to the socially beneficial settlements with intermediate dates of entry, the agreements that are the most profitable for the firms (without consideration of legality) might entail the payment of a substantial sum by the patent holder to the entrant in exchange for the promise that the entrant will not actually come into the market, or at least not until the patent has lost its impact and financial value."

22 In re Tamoxifen, 466 F.3d 187, 193 (2nd Cir. 2005).

23 *Id*.

24 *Id*.

25 *Id*. at 195.

26 *Id*. at 196.

27 *Id*. at 197.

28 *Id*. at 213.

29 "Thus the stated terms of the Settlement Agreement include nothing that would place it beyond the legitimate exclusionary scope of Zeneca's patent: The Settlement Agreement did not have an impact on the marketing of non-infringing or unrelated products. *Id* at 214.

30 *Id*. at 208-209.

31 *Id* at 206.

32 In re Ciprofloxacin, 544 f.2d 1323, 1336. "The essence of the inquiry is whether the agreements restrict competition beyond the exclusionary zone of the patent. This analysis has been adopted by the Second and the Eleventh Circuits and by the district court below, and we find it to be completely consistent with Supreme Court precedent."

33 *Id* at 1328.

34 *Id*. at 1328-39 and *supra* note 5.

35 *Id*. at 1327.

36 *Id* at 1323, 1337.

37 Hovenkamp *et. al.*, *supra* note 18, at 1759 n. 176.

38 This estimate does not assume that every settlement with a payment would settle without a payment: "This does not mean that we are assuming that all settlements with payments would 'become' settlements without payments if the former were banned. Some would: others might involve litigation of the patent. But since settlements without payments will tend to reflect patent strength, they can provide a benchmark for the consumer impact of either alternative." Jon Leibowitz, *Pay-for-Delay Settlements in the Pharmaceutical Industry: How Congress Can Stop Anticompetitive Conduct, Protect Consumers Wallets, and Help Pay for Health Care Reform (The $35 Billion Solution)*, Speech to the Center for American Progress, Appendix at 14, *available at* http://www.ftc.gov.speeches/leibowitz/090623payfordelayspeech.pdf (last visited September 13, 2009).

39 *Id*.

40 *Id*. Appendix at 13.

41 *Id*. at 12, note 7.

42 *Id*. at 12.

43 *Id*.

44 Hemphill *supra* note 15 at 650. Professor Hemphill assumed deals on average delayed entry one year, and looked only at deals where there was sufficient public information available.

45 *See e.g.* in re Tamoxifen, 466 F.3d at 216.

46 In 2000, the Commission brought the first of its actions in this area. *See* Cmpl. In the Matter of Hoescht-Marion-Roussel, Carderm Capital, and Andrx Corp., FTC Docket No. 9293 (March 16, 2000).

47 Generic Entry Prior to Patent Expiration, An FTC Study, p. 31 (2002), *available at* http://www.ftc.gov/os/2002/07/genericdrugstudy.pdf (last visited September 19, 2009).

48 Generic Entry Prior to Patent Expiration, An FTC Study, at Table 3-3, p. 32 (2002), *available at* http://www.ftc.gov/os/2002/07/genericdrugstudy.pdf (last visited September 19, 2009).

49 David J. Morrow, *Trial is Getting Underway in Prozac Lawsuit*, N.Y. Times, (January 25, 1999), *available at* http://www.nytimes.com/1999/01/25/business/trial-is-getting-under-way-today-in-prozac-patent-lawsuit.html

50 *Id*.

51 *See* Table 1, *supra* note 17.

52 In re Tamoxifen, 466 F.3d at 209 (quoting In re Schering-Plough Corp., slip op. at 27, 2003 WL 22989651 (Fed. Trade Comm'n Dec. 8,2003), 2003 FTC LEXIS 187, vacated, 402 F.3d 1056 (11th Cir. 2005)).

53 *Id*.

54 *Id*.

55 *Id*.

56 *Id*. 466 F3d. at 211 (quoting In re Ciprofloxacin Hydrochloride Antitrust Litig., 363 F.Supp.2d at 514, 534 (E.D.N.Y. 2005).

57 *Id*. at 212.

58 21 U.S.C 355(j)(2)(B)(iv).

59 21 U.S.C. 355(j)(5)(D)(i)(I)(bb)(BB).

60 *E.g.* First Amended Complaint at 15-16, Federal Trade Commission v. Cephalon, Inc., No. 2:08-cv-02141-MSG (filed August 12, 2009); Protecting Consumer Access to Generic Drugs Act of 2009: Hearing Before the Subcomm. on Commerce, Trade and Consumer Protection, 111th Congress (statement of Barry Sherman, CEO of Apotex Inc., at 3-4).

61 *See generally*, Letter from Gary Buehler to Carmen M. Shepherd and Kate C. Beardsley, No. 2007N-0382 (January 29, 2009) Denial of Citizen Petition, *Rampiril Capsules and the 180-day exclusivity*.

62 For this view, *see* Testimony of Barry Sherman, *supra* note 59. Others have argued that a settlement with a delayed entry date and retain its exclusivity may function just like a payment. *See* Hemphill, *supra* note 15, at 651-653. Obviously eliminating the 180-day exclusivity for such settling first filers would eliminate that form of compensation.

63 *See* Reiffen & Ward, *supra* note 1124, at 48, Table 4. Of the 31 products they examined, 20 had less than 10 approved ANDAs. Even for the former blockbuster Zocor, only 11 companies have received ANDAs. *See* Approved Drug Products with Therapeutic Equivalence Evaluations, http://www.accessdata.fda.gov/scripts/cder/ob/docs/temptn.cfm.

64 *See supra* notes 59-63.

65 Richard G. Frank & David Salkever, *Generic Entry and the Pricing of Pharmaceuticals*, 6 J. Econ. and Mgmt. Str.75, 89 (1997).

66 *Id*.

67 Revenues are not the same as profits, but here they work as a reasonable estimate as to the overall impact of entry on generic profits. *See* Authorized Generics: An Interim Report, An FTC Study at 11 n. 14 (2009).

68 The numbers are illustrative only. The price discounts are consistent with Frank & Salkever's findings but not Reiffen's. *Compare* Frank and Salkever, *supra* note 64, at 84 (Figure 2) with Reiffen & Ward, *supra* note 12, at 44. In most cases, even a first filer will only face competition 180 days after it launches.

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Reversing the Trend? The Possibility that Rule Changes May Lead to Fewer Reverse Payments in Pharma Settlements

Anne Layne-Farrar

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Reversing the Trend? The Possibility that Rule Changes May Lead to Fewer Reverse Payments in Pharma Settlements

*Anne Layne-Farrar**

The article begins by laying out a simple framework that makes obvious the incentives at play in generic drug entry, brand challenges, and settlements between the two. Once this common understanding has been established, several rule changes that have taken place are summarized—one in the form of an amendment to Hatch-Waxman and another in a recent decision by the Court of Appeals for the Federal Circuit. These institutional changes may have the consequence of reducing the prevalence of reverse payments. This possibility suggests a different policy tact might be called for, one that shifts emphasis from determining whether or not reverse payments should be *per se* illegal to working with the incentives that firms already face and exploiting those incentives to reduce firms' inclinations to enter into anticompetitive reverse-payment settlements.

*Dr. Anne Layne-Farrar is a Director at LECG. She specializes in antitrust matters where the core issues are at the intersection of intellectual property economics and competition policy.

I. Introduction

The debate over so-called "reverse payments"—where a patent-holding brand name pharmaceutical firm makes a settlement payment to a generic competitor to prevent or delay the generic from entering the branded drug market[1]—reached a fevered pitch this year with the introduction of a legislative proposal aimed squarely at settlements involving such reverse payments. Specifically, Senator Herb Kohl (D.-Wisconsin) introduced the "Preserve Access to Affordable Generics Act"[2] that would make it "unlawful" for parties involved in pharmaceutical patent litigation to sign a settlement agreement in which the generic company:

(1) receives "anything of value;" and

(2) "agrees not to research, develop, manufacture, market, or sell the [generic product] for any period of time" but that did not

(3) "demonstrate by clear and convincing evidence that the pro-competitive benefits of the agreement outweigh the anticompetitive effects of the agreement."

That reverse settlements are deemed important enough for their own legislation is indicative of the heat the debate has generated. On one side of the reverse-payment debate is a camp, which includes the Federal Trade Commission Chairman and the Department of Justice's Chief Economist, that is calling for reverse settlements to be made *per se* illegal. The view maintained by this group is that *all* reverse payments are anticompetitive: The sole reason for a brand firm with patents for a commercially successful drug to make a reverse payment is to delay the generic firm's entry.[3] As Carl Shapiro wrote in 2003, prior to his appointment as the DOJ's Antitrust Division Deputy Assistant Attorney General for Economic Analysis, "Presumably the patent holder would not pay more than avoided litigation costs unless it believed that it was buying later entry than it expects to face through the litigation alternative."[4] Similarly, FTC Chairman Liebowitz points to reverse payments as

> "…yet another example of pharmaceutical companies turning competition on its head. Congress enacted the landmark 1984 Hatch-Waxman Act to encourage early generic entry and save consumers money, but these anticompetitive deals threaten to destroy that benefit and make crucial portions of the Hatch-Waxman Act extinct in all but name."[5]

Indeed, it is the Hatch-Waxman Act that makes reverse payments possible in the first place. Specifically, the 1984 Act enables generic firms to file an "abbre-

viated new drug application," ("ANDA"), with the Food and Drug Administration. To file an ANDA, the generic firm needs only to show that its generic version of the drug works the same as a previously approved pioneer drug. Within the filing, the generic firm must specify whether the pioneer drug's patent will still be in force at the time of the generic's entry. The option listed under Paragraph IV of the ANDA states that the pioneer drug's patent will *not* have expired at the time of generic entry. Paragraph IV ANDA filings trigger a 45 day window for the maker of the pioneer drug to respond by challenging the generic firm's entry as infringing on its patent. This challenge is a form of ex ante infringement case in which the infringement has not actually occurred but is expected to occur when the generic firm actually begins marketing its version of the drug.

THE COMPLICATION IS THAT WITH THE INCREASED EX ANTE GENERIC CHALLENGES HAVE COME SETTLEMENTS INVOLVING REVERSE PAYMENTS.

Because Hatch-Waxman enables generic drug firms to challenge a brand name drug without actually entering the market and without making any allegations of patent invalidity, the Act lowers the risk of and thus encourages patent challenges. As noted by Chairman Liebowitz above, this was one of the goals of Hatch-Waxman and by all accounts it has been fulfilled.[6] The complication is that with the increased ex ante generic challenges have come settlements involving reverse payments.

Not all economists and lawyers see reverse payments as inherently anticompetitive, however. On the opposite side of the debate are those who recognize that brand name drug makers can have legitimate, efficiency-based reasons for offering potential generic competitors reverse payments. First, in other contexts it is well recognized that settling a suit rather than litigating to conclusion saves resources and can be pro-competitive.[7] In recognition of the generally beneficial nature of settlements, Judge Richard Posner has observed,

> "*Any* settlement agreement can be characterized as involving "compensation" to the defendant, who would not settle unless he had something to show for the settlement. If any settlement agreement is thus to be classified as involving a forbidden "reverse payment," we shall have no more patent settlements."[8]

Additionally, some argue that "important economic realities … can make reverse payments pro-competitive."[9] For instance, Dickey et. al. list such factors as: the brand firm's risk aversion; information asymmetries between the brand and generic regarding the validity of the patents at issue; differing expectations

regarding likely litigation outcomes; and different discount rates as potential legitimate reasons for reverse-payment settlements. The argument here is not that all reverse payments are pro-competitive, but rather that some may be and thus all such settlements should not be banned as *per se* illegal.

MANY IN THE "NOT-ALL-BAD" CAMP POSIT THAT THE ISSUE AT THE HEART OF THE MATTER IS NOT AN ANTITRUST QUESTION, BUT RATHER WHETHER THE BRANDED DRUG PATENT(S) ARE STRONG.

Many in the "not-all-bad" camp posit that the issue at the heart of the matter is not an antitrust question, but rather whether the branded drug patent(s) are strong.[10] If so, the brand firm will most likely win the challenge so that settlement, even involving a reverse payment and some delay of generic entry, is welfare-enhancing because it eliminates litigation costs, does not deprive consumers of any reasonably expected period of lower drug costs, and still manages to make the two firms involved better off than if they had continued to litigate.

Related to the patent validity point is the issue of drug research. As is well documented by now, pharmaceutical research and development ("R&D") is extremely costly and time consuming. Studies have estimated that the average new drug takes somewhere between 10 to 15 years to go from lab to pharmacy, and that the journey can cost upwards of $1.3 billion (counting both direct and opportunity costs).[11] Moreover, the odds that any one drug tested will eventually be approved are quite small—some estimate on the order of 1 out of every 5,000.[12] With such large and risky upfront outlays necessary for innovation, patent protection plays a key role in ensuring the proper incentives for investments in new drugs. While not all patents will be valid, these industry dynamics suggest some caution in dealing with anything that can cut a patent term short, including pre-expiry generic challenge.

I will admit to falling into this latter more cautious not-all-bad camp. Once we admit the possibility that at least some settlements can be pro-competitive (or at least not harmful), we must move away from *per se* illegality and consider how best to achieve the desired policy objectives. Namely, we want to strike the right balance between upholding valid intellectual property rights and their pivotal (albeit long-term) role in spurring pharmaceutical innovation and the more immediate drug pricing benefits that early generic entry can provide consumers. Thus, if we assume that at least some reverse-payment settlements do not harm consumer welfare, then we need to explore policy options that have the potential to reduce harmful settlements without eliminating settlements altogether.

In the context of that assumption, the analysis presented here considers the various factors that affect a brand firm's decision to offer a reverse-payment settlement and questions whether and how those factors might be exploited to limit the occurrence of reverse payments in the first instance. If we are able to employ firms' natural incentives as a means to reduce the prevalence of reverse pay-

ments, we will have a smaller set of cases over which to debate the competition effects and welfare implications.

The article begins by laying out a simple framework that makes obvious the incentives at play in generic drug entry, brand challenges, and settlements between the two. Once this common understanding has been established, several rule changes that have taken place are summarized—one in the form of an amendment to Hatch-Waxman and another in a recent decision by the Court of Appeals for the Federal Circuit. These institutional changes may have the consequence of reducing the prevalence of reverse payments. This possibility suggests a different policy tact might be called for, one that shifts emphasis from determining whether or not reverse payments should be *per se* illegal to working with the incentives that firms already face and exploiting those incentives to reduce firms' inclinations to enter into anticompetitive reverse-payment settlements.

II. The Framework

To make clear the various forces at work in a generic firm's challenge of a patented brand name drug, it is helpful to layout a simple framework. Consider a market with at least two firms, a brand and a generic, referenced B and G respectively. G is considering filing a paragraph IV ANDA in regards to a drug that B is currently supplying. To avoid a complication that does not add insight to our discussion, assume that both firms face the same marginal costs of production, *mc*: while B clearly incurs research and development costs that G does not, once the drug is approved by the FDA assume that the cost of making and distributing it would be identical for both firms should G enter the market. If G files an ANDA and B responds by challenging G with patent infringement, both firms incur legal costs (L), although those costs may differ across the firms. With these basic assumptions in mind, we can turn to the various scenarios possible under this setup.

A. CASE 1: NO GENERIC ENTRY

G may decide, for whatever reason, not to enter the market in competition with B. In this baseline case, G earns profits from some outside option, π_o, say from pursuing a different generic drug. It is against this outside profit that G will evaluate the alternative of entering the market in competition with B. G will only file an ANDA for B's drug if it expects to earn more in this competition than it can otherwise earn through its alternative options.

B. CASE 2: GENERIC ENTRY WITH NO BRAND CHALLENGE

If G does decide to file a paragraph IV ANDA, several outcomes are possible. First, B may decide, for certain reasons,[13] not to challenge G's entry into the market. In this case, G would enter the market uncontested and compete with B for sales of the drug. As a result, the price of the drug would fall to the duopoly level.

The duopoly price (P_d) exceeds the competitive price (P_c),[14] which would prevail should several generic firms enter the market in competition with B, but it is lower than the monopoly price (P_m) charged by B prior to G's entry.

Under this scenario, both B and G would earn the duopoly return: $\pi_d = (P_d - mc)(\sigma X_d)$, where X_d is the aggregate quantity of the drug sold in the market given two suppliers only (B and G), a quantity that exceeds the monopoly quantity sold when B faced no competition ($X_d > X_m$), and σ is B's share of the market. If $\sigma = 0.5$, the two firms split the market evenly, but other divisions are certainly possible. Thus, in this case, consumer prices would fall, the aggregate quantity sold would increase, and consumers would be better off (in the short term at least) than if G had not entered. The brand firm, however, is typically worse off. Even though the aggregate quantity sold increases, it is generally the case that the brand firm's price falls by enough that the brand firm earns less than before, when it held a monopoly.[15]

C. CASE 3: GENERIC ENTRY WITH BRAND CHALLENGE

Entry with brand firm challenge is slightly more complicated in that there are two potential outcomes. First, B could win the litigation, in which case B would remain a monopolist for the residual term of its patent while G would not be able to enter until after patent expiry. Note that both firms' earnings would be reduced by the litigation expenses they incurred.

ENTRY WITH BRAND FIRM CHALLENGE IS SLIGHTLY MORE COMPLICATED IN THAT THERE ARE TWO POTENTIAL OUTCOMES.

On the other hand, B could lose the infringement challenge, in which case G would be free to enter the market immediately, before the patent expires, and compete with B. But in this case, B's patent would have been invalidated. If any other generic firms (say, firms G_2 through G_n) were interested and capable of entering the market, they would be free to do so without risking an infringement challenge by B. Hence, if B loses its challenge of G's entry, the resulting market could be more competitive; rather than a duopoly, the several firms in the market would earn a competitive return. Making the reasonable assumption that prices fall by more than quantities sold increase, we have $\pi_c < \pi_d < \pi_m$.[16] Again, B and G would also incur litigation expenses along the way.

D. CASE 4: GENERIC ENTRY WITH SETTLEMENT

The final possibility is that B settles with G. In this case, B could offer G a payment not to enter the market for some specified time, perhaps until the patent expires. With settlement, both firms still incur some litigation expense, but less

than they would have if the trial had run its course to a court decision (e.g., litigation expenses L are reduced by some fraction $0 < \lambda < 1$).

Make the realistic assumption that the settlement amount is a multiple of the earnings that G could have made in the duopoly market had B not challenged its entry: $S = \delta(P_d - mc)\sigma_G X_d$, where δ is the multiple and σ_G is G's share of market. If $\delta \leq 0$ then G pays B a licensing fee to enter the market. In this case, the payment is not "reverse" but rather flows in the typical direction found in patent infringement cases. If $\delta > 0$, however, the settlement involves a reverse payment from B to G. If $\delta > 1$ then the reverse payment amounts to more than the generic could have ever possibly earned by entering the market.

III. Important Decision Parameters

The above discussion points to a number of key factors in generic and brand firms' strategic decisions regarding early entry and competition. First, the difference in the brand firm's expected earnings as a monopolist and as a duopolist (competing with G), less the expected cost of litigation, is pivotal in the brand firm's decision to challenge the generic firm's entry. Assume for the moment that B knows with certainty it can win its challenge against G if it spends L_B on the litigation. Then, as long as $\pi_m - \pi_d - L_B > 0$ B's monopoly earnings are sufficiently above those it could earn in competition with G, even accounting for litigation costs, then it will want to challenge the generic's entry.

FIRST, THE DIFFERENCE IN THE BRAND FIRM'S EXPECTED EARNINGS AS A MONOPOLIST AND AS A DUOPOLIST, LESS THE EXPECTED COST OF LITIGATION, IS PIVOTAL IN THE BRAND FIRM'S DECISION TO CHALLENGE THE GENERIC FIRM'S ENTRY.

Of course, firms are never assured of winning a lawsuit, regardless of the money spent making their case and regardless of their views on patent validity. Thus B's assessment of the chances of winning the lawsuit will play a role as well, and will affect the expected cost of litigating (L_B). If, on the other hand, $\pi_m - \pi_d - L_B < 0$ then having a "monopoly" on the drug does not translate into supra-competitive earnings that warrant the expense of litigation, even if the brand firm was assured of winning the challenge. This latter scenario could hold, for example, if the brand drug faces competition from a number of close (albeit chemically distinct) substitutes.

The litigation challenge condition above is likely to hold in many instances. For example, when Merck's Zocor drug was alone on the Simvastatin market, it commanded in excess of $5 billion in annual revenues (profits are unavailable).[17] When the first generic entered, Merck's annual revenue fell to just below $3 billion (See Figure 1 below). After the second generic entered, Merck's annual revenue fell to below $1 billion.

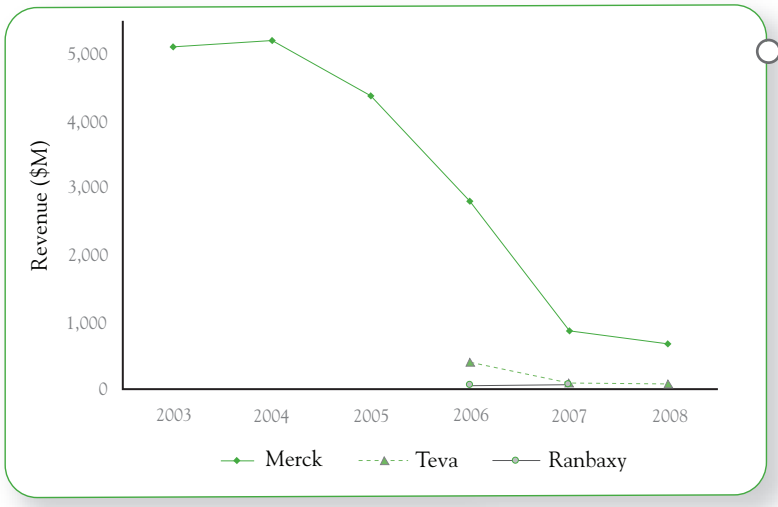


Figure 1

Litigation costs would surely not erode a $2 billion difference, meaning that Merck had very strong incentives indeed to challenge the first generic's entry into the production of Simvastatin. Available evidence suggests that such precipitous drops in earnings are not uncommon in the face of generic entry.[18]

At the same time that substantial profits are on the line for brand firms, the stakes are considerably smaller for generic firms. The chart below shows the ratio of generic to brand name drug prices as the number of generic firms on the market increases. Note that the first bar, which shows one generic firm earns on average 94 percent of the brand firm's price, is overstated because it includes so-called "branded generics" offered by the brand firm itself (or by a third party sponsored by the brand firm). In general, industry statistics suggest that if a brand earns $1 billion as a monopolist, the first generic will earn around $80 million.[19]

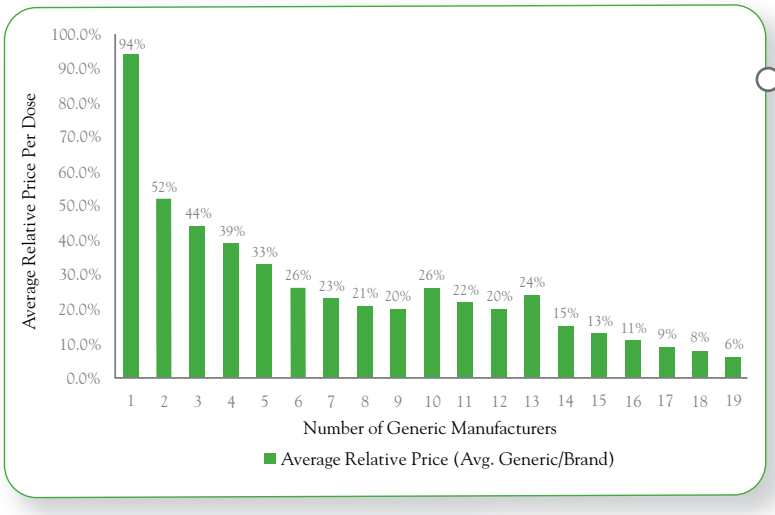


Figure 2

With these various options in mind, the generic will decide on whether to enter the market by evaluating its expected profits under each scenario. To do this, the generic will assign probabilities to possible outcomes.

Given that the condition for a brand firm to challenge a generic firm's paragraph IV ANDA filing ($\pi_m - \pi_d - L_B > 0$) is likely to be met with some frequency, the relevant question is whether litigation will proceed to conclusion or whether the parties will settle. Of course, with such large discrepancies in potential earnings, if the brand finds it worthwhile to challenge the entry the brand is also likely to be able to make a settlement offer that the generic firm cannot refuse. For instance, if the brand was earning $1 billion a year as a monopolist but its earnings would fall to $600 million should a generic firm enter the market, the brand would have up to $400 million (less anticipated litigation fees) with which to pay a generic to stay out of the market and the brand would still be better off than if it abstained from challenging the entry. Given that the generic firms generally expect to earn far less than the brand firm's lost sales, say on the order of $150 million as assumed above, there is clearly quite a bit of latitude for a settlement that can make both parties better off, given the generic filed the ANDA. The question, of course, is whether consumers are worse off, but recall that we have assumed that at least some fraction of these settlements are not harmful.

OBSERVE THAT WITH ANY SETTLEMENT, THE COURT MAKES NO RULING ON WHETHER THE BRAND FIRM'S PATENT IS VALID OR NOT. THIS HAS IMPORTANT CONSEQUENCES FOR OTHER GENERIC FIRMS THAT MAY BE CONSIDERING WHETHER TO ENTER THE MARKET.

Observe that with any settlement, the court makes no ruling on whether the brand firm's patent is valid or not. This has important consequences for other generic firms that may be considering whether to enter the market. These firms would still run the risk of patent infringement challenges from the brand firm, as the brand firm has made no concessions regarding the patent's validity. If the settlement involved a licensing fee paid by the first generic, then later generic entrants will likely wait for patent expiry given the generally slim margins they can earn in the marketplace. If the settlement involves a modest reverse payment, later generics will still be likely to wait for patent expiry as they would expect a lower settlement payment than the first entrant obtained because the brand firm has less to lose with a second entrant as compared to the first, and the generic would still have to incur litigation expenses to obtain that settlement. Only if the first settlement involved a lucrative reverse payment would later generics have incentives to attempt early entry, taking the first payment as a signal that the brand firm has ample supra-competitive earnings to allow for multiple reverse payments.

If the brand firm does not offer a settlement to the first generic filer, however, it faces the risk that the court will find its patent is invalid and/or not infringed. In that case, the brand is far more likely to face not one generic (the firm filing),

but potentially many generics since the primary risk associated with generic entry will have been removed. The privately beneficial nature of settlements is in fact one of the bones of contention in the debate over reverse payments; advocates of *per se* illegality point to the two parties' mutual benefit and argue that consumers must be worse off as a result.[20]

IV. How Settlement Decisions Are Made

Returning to the simple framework above, what condition must be met for the brand firm to attempt a reverse-payment settlement versus litigating to conclusion?[21] Assuming B challenges G, if the parties settle and the generic delays entry the brand firm would continue to earn its monopoly profit, but would have to pay out of that amount the settlement S and the litigation costs L_B, which is a fraction of the total litigation costs that would have resulted if the trial had run to completion: $\pi_m - S - \lambda L_B$. If the brand firm does not offer a settlement, then with some probability ρ it will win the case and continue to earn its monopoly profit, less the litigation expenses: $\rho(\pi_m - L_B)$. With probability $1 - \rho$ the brand firm will lose the case, multiple generic firms will enter the market once the patent is invalidated,[22] and the brand firm will earn a competitive profit: $(1 - \rho)(\pi_c - L_B)$. Combining these last two potential profits for the brand firm, we obtain the expected profit from completing the trial: $\rho\pi_m + (1 - \rho)\pi_c - L_B$. The relevant comparison is then whether the brand firm's expected earnings are higher when it settles or when it takes its chances with a trial.

Working through the algebra, we find that the brand firm will offer a settlement S when the following condition holds:

$$\pi_m - \pi_c > \frac{S - (1 - \lambda)L_B}{1 - \rho}. \quad (1)$$

In words, the condition implies that when the difference between the best and worst possible profit outcomes from the brand firm's perspective (monopoly earnings versus competitive earnings) is larger than the reverse-payment settlement amount required less the additional litigation costs needed to complete the trial, all weighted by the odds of losing the trial, then the brand firm will prefer to pay to settle the case.

Considering a few straightforward comparative statics helps to clarify the intuition behind the reverse-payment settlement condition (1). First, and most obviously, the higher monopoly profits are over the competitive level of profits, the more likely the brand firm will want to settle with the generic firm in order to avoid the risk of an invalidated patent—and with it the entry of multiple generic firms and the lower competitive profits that come with that entry.

Second, and not surprisingly, the smaller the litigation cost savings from settling as opposed to taking the trial to its conclusion, $(1 - \lambda)L_B$, the less likely a

reverse-payment settlement is (again holding everything else constant). This is a traditional component of any settlement decision and is not particular to pharmaceutical reverse-payment deals.

Third, the probability of losing the case also plays a key role in the brand firm's decision to offer a reverse-payment settlement or not. As ρ increases, the fraction on the right hand side of the settlement condition (1), $\frac{S - (1 - \lambda)L_B}{1 - \rho}$, will rise as well, making it less likely that the profits protected, $\pi_m - \pi_c$, are large enough to justify the settlement. In other words, as the probability of the brand firm's winning the infringement case increases, the chance of a settlement decreases. The converse is, of course, true as well. The mechanism at work here is the risk that an invalidated patent spurs multiple generic firms to enter, and reduces the brand firm's profits below the duopoly level possible with just one generic challenger.

V. The Potential for Multiple Generic Entrants

Thus far the simple framework above has done little more than make some well known incentives explicit. With this common ground understanding in hand, however, let us turn to some less obvious aspects of brand and generic firm competition: the possibility that multiple generic entrants might be able to reduce the prevalence of reverse payments.

LET US TURN TO SOME LESS OBVIOUS ASPECTS OF BRAND AND GENERIC FIRM COMPETITION: THE POSSIBILITY THAT MULTIPLE GENERIC ENTRANTS MIGHT BE ABLE TO REDUCE THE PREVALENCE OF REVERSE PAYMENTS.

Not too long ago, the rules were such that one and only one generic firm at a time had any incentive to file a paragraph IV ANDA. Under Hatch-Waxman, the first generic filer received 180 days of "exclusivity," during which the FDA provided no other generic company approval to market the same drug. The 180 days started to count down as soon as the first filer began selling its generic product or, in the case of a challenge from the brand firm, when the court ruled that the generic did not infringe the patent and could start selling its product. Of course, with the stakes so often high in these pharmaceutical cases, the parties quickly identified the loopholes in this system: a settlement does not involve a court decision and if the generic does not begin marketing its drug before patent expiry, the 180 day clock would not start ticking until then.

Thus, in 2002, an FTC report expressed concern over so-called generic entry "parking," whereby the first generic filer would "park" its exclusivity period, not competing with the brand firm before patent expiry but preventing other generic firms from entering the market before then.[23] Brand firms would issue unilateral covenants not to sue generic firms over the drug's key patent (but not nec-

essarily for all patents needed to manufacture the drug), but settlement with the first generic filer ensured that it would not enter the market before the brand's patent expired. Such deals essentially amounted to privately beneficial collusion: the brand firm maintained its monopoly while the generic firm obtained a beneficial deal from the brand firm while still maintaining its exclusivity period once it eventually did enter the market.

The courts inadvertently facilitated this practice by holding that later generic filers did not have standing to file for declaratory judgment on the brand firm patent's validity or infringement. Thus, the brand firm's covenant not to sue was seen as removing the threat of infringement suit while the first generic filer's failure to actually market the drug meant its exclusivity was not yet expired.

These obstacles have since been removed, however, with the last piece falling into place in 2008. Importantly, the Hatch-Waxman Act was amended in 2003 so that a first generic filer can now lose its exclusivity period.[24] Among the forfeiture events are: 1) failure to market the drug promptly; 2) failure to obtain FDA approval to market the drug in a timely fashion; and 3) the expiry of all the relevant patents. Specifically, the first filer will lose its exclusivity if it has not marketed its drug as of 75 days after receiving FDA approval to do so, 30 months after submitting its application, or immediately upon winning a court challenge from the brand firm. If the first generic filer loses its exclusivity period for one of these reasons, then no generic firm benefits from 180 days of exclusivity. The amendments also added the ability for generic firms to file a counterclaim to delist the brand firm's patent, giving generics an additional weapon in an entry bid.

In regards to the court's role in fostering generic drug parking, the Court of Appeals for the Federal Circuit's ruling in *Caraco Pharm. Labs., Ltd. v. Forest Labs., Ltd.* held in June 2008 that a brand firm's unilateral covenant not to sue does not moot later generic entrants' ability to challenge the brand firm's patent and that filing a paragraph IV ANDA is enough to satisfy standing.[25] As a result of these various changes, a second generic filer can now trigger the first filer's 180 day exclusivity period, or the first filer can lose its exclusivity altogether. In combination, the changes thus effectively remove the threat of parking.

THE QUESTION NOW BECOMES WHETHER THIS NEW FREEDOM FOR SECOND AND LATER GENERIC FILERS AFFECTS THE INCENTIVES FOR THE BRAND FIRM AND THE FIRST GENERIC FILER TO SETTLE WITH A REVERSE PAYMENT.

The question now becomes whether this new freedom for second and later generic filers affects the incentives for the brand firm and the first generic filer to settle with a reverse payment. To answer this question, return to our discussion of incentives. First, we would expect a lucrative reverse payment to act as a lure to other generic firms capable of entering the market. Seeing a relatively large payment signals to other generic firms that the brand firm does indeed have considerable monopoly profits at stake (e.g., that the left hand side

of condition (1) is much greater than the right hand side) and is likely capable of making additional payments to other rivals. Whereas before a brand firm could dismiss any such second, third, or later generic filers secure in the knowledge that the settlement terms struck with the first filer in combination with the 180 day blockade would effectively keep these competitors out until patent expiry (or close to it), under the new regime second filers can enter far sooner, either after 180 days or, if the FDA revokes exclusivity from the first filer for a failure to act, as soon as the trial is concluded (assuming a settlement with the second filer has not been reached).

Consider, for example, a patent with 5 years left on its term after the settlement with the first filer is reached. The first generic filer has agreed not to enter before the brand firm's patent term is up, but the second generic firm can file as soon as it learns of the settlement. The second filer will either be successful, in which case it can enter immediately, or the brand will settle with it too. Thus, for the second generic filer's part, as long as its threat of entry is credible, it has an incentive to file: it can either win the challenge and have some period of first mover advantage (de facto exclusivity before other generics can enter),[26] or it can reach a settlement and get paid by the brand firm. Weighing against these possible benefits, the second generic firm's downside is limited: with no actual sales on the market yet, it stands to lose its litigation expenses but would have little if anything to pay in damages to the brand firm (this is, indeed, the point of Hatch-Waxman, to lower the risks and hence provide incentives to challenge brand name drug makers). Once again the brand firm is faced with the litigate-settle decision, but this time it has already paid one reverse-payment settlement to the first filer, and has lost the first trial's litigation expenses as well.

Returning to the simple framework above, if the brand firm anticipates the possible chain of events at the time of its negotiations with the first filer, its decision is now based upon the following condition:

$$(2) \qquad \pi_m - \pi_c > \frac{S^1 - (1-\lambda)L_B^1}{1-\rho^1} + \frac{S^2 - L_B^2}{1-\rho^2}.$$

This follows because now the first reverse-payment settlement carries with it the knowledge that a second filer will surely come knocking.[27] It may be reasonable to assume that the first filer is the strongest generic challenger (that is, it would take the most sales away from the brand firm should it enter the market), but even if the second filer is less capable, two generic competitors instead of one removes the duopoly possibility from the list of market outcomes and typically implies lower earnings for the brand firm.[28] Thus, even if the settlement payment required for the second generic filer, S^2, is smaller than the first settlement, S^1, the aggregate settlement amount is nevertheless higher. The difference between the brand firm's best case scenario (monopoly profits after winning the first generic challenge) and its worse case scenario (losing patent validity and facing

multiple generic firm entry) must be higher than it was before in order to justify the first reverse-payment settlement amount.

If there are more than two generic firms that could credibly file paragraph IV ANDAs, then the brand firm would have to consider several additional settlements as it contemplated whether to settle with the first filer. Condition (2) could then expand with three or four settlement terms on the right hand side, each one making that first settlement less likely.

VI. The Implications of Multiple Filers

Two important points arise from the above line of reasoning. First, now that the path has been cleared for multiple generic filers, we may indeed see fewer reverse payments. As long as more than one generic firm can offer a credible entry strategy, the brand firm will consider the signal that its reverse-payment settlement with the first generic firm sends to other generic firms. If the first filer is paid amply for delayed entry, the second (third, etc.) generic rivals will see an opportunity to either acquire a lucrative settlement themselves or else to gain a first move advantage from the first filer given its failure to act, hence precipitating their earlier entry. Brand firms with marginal products, that is those with drugs that were just barely making the settlement hurdle before as their "monopoly" profits were not high enough to warrant a settlement payment under condition (1) will likely find that condition (2) is not satisfied at all. Depending on how many products fall into this category, we could see fewer reverse payments than we otherwise would have absent the amendments to Hatch-Waxman and the ruling in *Caraco*.

FIRST, NOW THAT THE PATH HAS BEEN CLEARED FOR MULTIPLE GENERIC FILERS, WE MAY INDEED SEE FEWER REVERSE PAYMENTS.

That is the positive side of the new regime. The second implication is not so positive. Namely, there is likely to be a second order negative effect on the first filer's incentives to file. If we think of the entry/litigate decisions in a game theory setting, the incentives to file first have weakened with the latest changes to Hatch-Waxman and the court standing requirements. Since it is now possible to lose the 180 day exclusivity period, the first filer has more at risk. If it accepts a lucrative reverse-payment settlement offer from the brand firm and agrees to delay its entry, the FDA can revoke its exclusive status. Thus, when it does eventually enter the market, it may find itself in second or even third place. Moreover, the size of any reverse payment from the brand firm is likely to be smaller now as well. Since the brand firm must consider all potential entrants as soon as the first filer emerges, it is likely to negotiate harder with the first generic filer for a lower settlement amount, thus avoiding a strong signal to other generic firms to enter early.

VII. Adding Product Differentiation

Thus far we have considered drugs in the abstract. However, the above arguments are likely to fit some drug categories better than others. For instance, drugs targeting relatively high risk conditions, such as heart disease or cancer, might offer more price resilience for brand firms.[29] If either patients or doctors have a high level of concern over the efficacy and reliability of a generic version, they may be more insistent that a brand name drug be prescribed. In that case, the doctor would indicate "no generic substitutes" on the prescription. Without such explicit physician instructions, however, pharmacies and insurance companies are likely to make generic substitutions as a matter of course in order to keep costs down.[30]

More broadly, any perception of higher quality or greater reliability on the part of doctors or patients will tend to offer the brand firm some relief from generic competition. It would be surprising if such perceptions enabled the brand firm to fully maintain its monopoly share or price, but any cushion against competition will tend to reduce the difference between the brand firm's monopoly earnings and its earnings under either duopoly or competition.

Looking back at the litigation condition, this time weighted by the odds of winning the suit, $\rho(\pi_m - \pi_d - L_B) > 0$, it is easy to see that any factor that softens competition (brand recognition, quality perceptions, etc.) will also soften the brand firm's incentives to challenge generic entrants. As π_d increases, the left hand side of the challenge condition decreases, meaning that it is less likely to be greater than zero. Likewise, the factors that soften price erosion for the brand firm will also reduce the firm's incentives to offer generic entrants reverse-payment settlements in the case of litigation. We can see this by considering settlement condition (1) $\pi_m - \pi_c > \frac{S - (1-\lambda)L_B}{1-\rho}$. Again, the left hand side of the equation falls as π_c rises, making it harder for the brand firm to clear the inequality and offer the generic the needed reverse-payment settlement of S.

VIII. Policy Considerations

The potential for multiple generic firms to file ANDAs within a relatively short time frame, as discussed above, suggests an important policy question. If there is any possibility that at least some fraction of reverse payments are not harmful to consumers, then making such settlements *per se* illegal is not good policy. Instead, policymakers could consider how to better align incentives to encourage more generic firms to file paragraph IV ANDAs promptly. In other words, unless we are certain that every single reverse payment lowers consumer welfare, the logic presented above offers an alternative route to reducing potentially anticompetitive reverse payments—one that does not require inflexible legislation that

might eliminate some beneficial settlements and might erode important incentives to invest in pioneer drugs.

Little attention appears to have been paid to working with the incentives already in place for brand and generic firms, as compared to debating the rules regarding what settlements should and should not be allowed. I would therefore like to close this article with a suggestion: that scholars and policymakers spend time brainstorming on ways to further amend Hatch-Waxman to encourage multiple generic filers to come forth earlier in the process. They might also consider whether new incentives should be put in place for subsequent filers, after the first generic has paved the way, as a means of restricting first settlements. With additional thought devoted to these paths, we might find that restrictive *per se* legislation is not needed.

IF THERE IS ANY POSSIBILITY THAT AT LEAST SOME FRACTION OF REVERSE PAYMENTS ARE NOT HARMFUL TO CONSUMERS, THEN MAKING SUCH SETTLEMENTS PER SE ILLEGAL IS NOT GOOD POLICY.

1 These settlement payments are referred to as "reverse" because the funds flow from the patent holder (the brand firm) to the would-be licensee (the generic firm), in reverse of the normal course of patent infringement suits where licensees pay patent holders to license valid patents.

2 Preserve Access to Affordable Generics Act, S. 369, 111th Cong. (2009).

3 Note that pharmaceuticals are unique in this regard. Pre-patent expiry challenges are relatively rare in other industries (outside of charging patent invalidity in response to infringement allegations). As explained below, they arise in pharmaceuticals as a result of the Hatch-Waxman Act, which in large part was aimed at increasing patent challenges in pharmaceutical markets. For a discussion of the details of Hatch-Waxman, *see* Thomas Cotter, *Refining the 'Presumptive Illegality' Approach to Settlements of patent Disputes Involving Reverse Payments*, 87 MINN. L. REV. 1789 (2003).

4 Carl Shapiro, *Antitrust Limits to Patent Settlements*, 34 RAND J. ECON. 391, 407-408 (2003); *see also* C. Scott Hemphill, *Paying for Delay: Pharmaceutical Patent Settlement as a Regulatory Design Problem*, 81 N.Y.U. L. REV. 1553 (2006) for a likeminded argument against reverse payments.

5 *See* Concurring Statement of Commissioner Jon Leibowitz, FTC v. Watson Pharmaceuticals et al. (Feb. 2, 2009), available at http://ftc.gov/speeches/leibowitz/090202watsonpharm.pdf.

6 FEDERAL TRADE COMM'N, GENERIC DRUG ENTRY PRIOR TO PATENT EXPIRATION: AN FTC STUDY (2002), *available at* http://www.ftc.gov/os/2002/07/genericdrugstudy.pdf. The study notes that "only 2 percent of generic applications sought [pre-patent expiry entry], but from 1998 to 2000, approximately 20 percent of the generic applications sought entry prior to patent expiration." at ii.

7 *See, e.g.*, Richard Posner, *An Economic Approach to Legal Procedure and Judicial Administration*, 2 J. LEGAL STUD. 399 (1973); Robert Cooter & Daniel Rubinfeld, *Economic Analysis of Legal Disputes and Their Resolution*, 27 J. ECON. LIT. 1067 (1989); Steven Shavell, *The Level of Litigation: Private Versus Social Optimality of Suit and of Settlement*, 19 INT'L REV. L. ECON. 99 (1999).

8 Asahi Glass Co. v. Pentech Pharmaceuticals, Inc., 289 F. Supp. 2d 986 (2003). Emphasis in original.

9 Bret Dickey, Jonathan Orszag, & Laura Tyson, *An Economic Assessments of Patent Settlements in the Pharmaceutical Industry*, (2008) *available at* http://www.compasslexecon.com/highlights/Documents/

Economic_Assessment_of_Patent_Settlements_Dickey_Orszag_and_Tyson.pdf. *See also*, Marc G. Schildkraut, *Patent-Splitting Settlements and the Reverse Payment Fallacy*, 71 Antitrust L. J. 1033 (2004); and Daniel A. Crane, *Exit Payments in Settlement of Patent Infringement Lawsuits: Rules and Economic Implications*, 54 Fla. L. Rev. 747 (2002).

10 *See, e.g.*, Ken Letzler & Sonia Pfaffenroth, *Patent Settlement Legislation: Good Medicine or Wrong Prescription?*, 23 Antitrust 81 (2009).

11 *See* Congressional Budget Office, Research and Development in the Pharmaceutical Industry, (CBO 2006) *available at* http://www.cbo.gov/ftpdocs/76xx/doc7615/10-02-DrugR-D.pdf; Joseph A. DiMasi & Henry G. Grabowski, *The Cost of Biopharmaceutical R&D: Is Biotech Different?*, 28 Managerial Decision Econ. 469 (2007); and Joseph A. DiMasi, Ronald W. Hansen, & Henry G. Grabowski, *The Price of Innovation: New Estimates of Drug Development Costs*, 22 J. Health Econ. 151 (2003).

12 Tufts Center for the Study of Drug Development, *Backgrounder: How New Drugs Move Throughout the Development and Approval Process*, Nov. 1, 2001 *available at* http://csdd.tufts.edu/newsevents/recentnews.asp?newsid=4 (last visited September 10, 2009).

13 The brand firm may be concerned that its patent will be deemed invalid, which would then invite additional generic entry; or it may view its monopoly earnings on the pioneer drug as insufficient to warrant expensive patent enforcement litigation. We explore the drivers of this decision below.

14 I am assuming here only that $P_c < P_d$, thus the "competitive" market may in fact be an oligopoly.

15 *See, e.g.*, Robert Cohen, *It's hard to beat generic—As drugs come off patent, major firms feel pinch*, The Star Ledger, Mar. 9, 2008 (noting how major drug makers such as GlaxoSmithKline, Merck, Bristol Myers Squibb, Pfizer, and Sanofi-Aventis faced revenue losses from generic competition for block-buster drugs); Val Brickates Kennedy, *Pfizer posts 18% drop in first-quarter profit; Pharma giant affirms 2008 financial outlook*, MarketWatch 10, Apr. 17, 2008 (noting an 18 percent drop in Pfizer's profits reflecting increased generic competition for top selling products such as Zyrtec and Lipitor); Peter Loftus, *Battle Over Merck Bone Drug Shows High Stakes Of Generics*, Dow Jones News Wires, Jan. 23, 2009 (noting that Merck stands to lose billions in revenue from generic competition for its drug *Fosamax*).

16 I am abstracting here from any brand name effects. In actuality, the pioneer firm may be able to maintain some price premium over the many generic competitors. Nevertheless, the fact remains that the brand firm's price will be significantly lower than the price it can command when it is the sole supplier of the drug, or even when it is a duopoly supplier with just one generic rival.

17 Chriss Schott, Jessica Fye, & Yuriy Prilutskiy, Merck & Co., Inc.: Updating Model Post Guidance, 4 (J.P. Morgan North America Equity Research, Dec. 5, 2008); Ken Cacclatore et al., Teva Pharmaceutical: Teva Able to Manage Through – Reiterate Outperform, 3 (Cowen & Company, Nov. 6, 2008); Ranbaxy, Annual Report 2006,10; Ranbaxy, Annual Report 2007, 16.

18 *See supra* note 15.

19 Based on market data for 40 branded drugs that faced generic competition between 1992 and 1998 in Atanu Saha et al., *Generic Competition in the U.S. Pharmaceutical Industry*, 13 Int'l J. Econ. Bus. 15 (2006).

20 *See* Jon Liebowitz, Chairman, Federal Trade Comm'n, "Pay-for-Delay" Settlements in the Pharmaceutical Industry: How Congress Can Stop Anticompetitive Conduct, Protect Consumers' Wallets, and Help Pay for Health Care Reform (The $35 Billion Solution), Address at the Center for American Progress (June 23, 2009), at 3-4, *available at* http://www.ftc.gov/speeches/leibowitz/090623payfordelayspeech.pdf. As noted above, this view has been challenged by those arguing that brand firms with strong patents are likely to win their infringement case and thus the settlement

saves litigation expenses and court resources without any negative impact on consumers. *See supra* note 10.

21 The possibility of a traditional settlement with license fees flowing from the generic firm to the brand firm is ignored here.

22 Even if the finding is that the patent is not infringed, the odds of additional generic entry should increase, as they can tailor their entry to not infringe as well.

23 *See* Federal Trade Comm'n, *supra* note 6.

24 Medicare Prescription Drug, Improvement, and Modernization Act of 2003, 42 U.S.C. §§101-1203 (2003).

25 Caraco Pharm. Labs., Ltd. v. Forest Labs., Ltd, 527 F.3d 1278, (2008).

26 Or the second filer may get a (generic) first mover advantage, which can be important in negotiating insurance provider deals. *See* Benjamin G. Druss et al., *Listening to Generic Prozac: Winners, Losers, and Sideliners*, 23 Health Affairs 210 (2004) *available at* http://content.healthaffairs.org/cgi/content/full/23/5/210 (noting how Barr Laboratories used its exclusivity period to contract with pharmacy benefits managers and large purchasers for its generic version of Prozac).

27 If there is some uncertainty as to whether a second generic will file, the second term on the right hand side can be weighted by a probability parameter. The general point remains, however, that a new term will be added to the settlement condition.

28 Observe that the probability that B wins its challenge against G^2 is written at ρ^2 in condition (2). It is likely, however, that the generic firms will not differ significantly in their planned production and sales of a generic version of the drug. If that is the case, the odds of winning the challenge hinge more on the strength of the patent and less on any particulars of the generic challenger. If the odds of winning are the same in each case, we can combine the two fractions so that:

$$\pi_m - \pi_c > \frac{S^1 - (1-\lambda)L_B^1 + S^2 - L_B^2}{1-\rho^1}.$$

Presumably, the litigation costs for the second case would be lower, being able to leverage work done for the first. However, these costs are for the full second trial as that case has not yet begun.

29 F. M. Scherer, *Pricing, Profits, and Technological Progress in the Pharmaceutical Industry*, 7 J. Econ. Perspectives, 97, 101 (1993); Jeffrey A. Dubin, *Empirical Studies in Applied Economics* 144 (Springer 2001).

30 *See* Federal Trade Comm'n, Consumer Protection – Facts for Consumers, http://www.ftc.gov/bcp/edu/pubs/consumer/health/hea06.shtm.

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Patent Settlements and Reverse Payments Under EU Law

Marc van der Woude

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Patent Settlements and Reverse Payments Under EU Law

*Marc van der Woude**

The purpose of this contribution is to explore the status types that settlements and reverse payments could have under Article 81 EC. It seeks to identify the elements of the legal tests which could possibly be applied to assess the legality of such settlements and, in particular, those providing for a value transfer from the originator to the generic firm. This will be done as follows: Section 2 summarizes the main findings of the Final Report on settlement agreements; Section 3 makes an inventory of relatively old case law that dealt with comparable issues or those related to patent settlement agreements; and Section 4 makes an attempt to distill a legal test from the two previous sections for the assessment of patent settlement agreements between originator and generic firms under EC competition law.

*Lawyer at the Brussels bar, Professor in Competition Law at the Erasmus University Rotterdam. The author would like to thank Delphine Gillet for her assistance.

I. Introduction

On July 8, 2009, the Directorate General for Competition of the European Commission ("the Commission") officially presented its Final Report on the pharmaceutical sector inquiry ("the Report").[1] This 500 page report essentially deals with two issues in which patent protection plays a central role, namely the delay in generic entry and a decline in innovation. Pharmaceutical companies not only rely on a wide range of patents (patent clusters and divisional patents) to oppose generic entry, but they also use their patents as defensive tools to prevent other originator companies from carrying out Research and Development ("R&D") activities. The report conveys the impression of a deficient European patent system combining a semi-unified patent delivery system and 27 different modes of patent protection. This system offers many possibilities to use patent laws for other purposes than stimulating innovation. It is therefore not surprising that the Final Report recommends the creation of a real community patent system supported by a unified judiciary. Nor is it surprising that this recommendation has the full support of the generic and innovating industry.

HOWEVER, PATENT AND REGULATORY ISSUES RELATING TO MARKET AUTHORIZATIONS AND REIMBURSEMENT RULES ARE NOT THE ONLY CAUSES OF THE DELAY IN GENERIC ENTRY AND DECLINING INNOVATION.

However, patent and regulatory issues relating to market authorizations and reimbursement rules are not the only causes of the delay in generic entry and declining innovation. The Report also refers to various commercial practices that could fall foul of antitrust rules. Concerning delayed generic entry, the Report distinguishes between two types of practices. The first category is unilateral in nature. It refers to smart and excessive use of patents, market authorizations, and reimbursement rules created by the originators concerned by the drop in prices and profits that normally occurs as a result of generic market entry. According to the Report, practically all originator companies have developed a tool-box of measures destined to delay such entry. The Commission's decision fining *Astra Zeneca* for having misled regulatory authorities offers an example of the unilateral use of some of these tools.[2]

The second category of measures is bilateral in nature. These measures involve both originators and generic companies. This category concerns settlement agreements, including settlement agreements providing for a value transfer from the originator to the generic firm, either in the form of a direct (reverse) payment, a license, or a distribution right. The Final Report notes that this type of agreement has attracted the attention of the U.S. antitrust authorities and cites various examples of the American case law, including the recent *Cephalon* and *Solvay* cases.

The Commission seems keen to explore whether these precedents can also be followed in the European context, so as to speed up generic entry. On the day of

the presentation of its Final Report, the Commission announced that it had initiated formal proceedings against Les Laboratoires Servier. The Commission is investigating whether the settlement agreements which this originator concluded with several generic companies concerning the marketing of the generic version of perindopril infringe Article 81 EC. In the absence of precedents, this procedure will break new legal ground.[3]

The purpose of this contribution is to explore the status types that settlements and reverse payments could have under Article 81 EC. It seeks to identify the elements of the legal tests which could possibly be applied to assess the legality of such settlements and, in particular, those providing for a value transfer from the originator to the generic firm. This will be done as follows: Section 2 summarizes the main findings of the Final Report on settlement agreements; Section 3 makes an inventory of relatively old case law that dealt with comparable issues or those related to patent settlement agreements; and Section 4 makes an attempt to distill a legal test from the two previous sections for the assessment of patent settlement agreements between originator and generic firms under EC competition law.

II. Settlement Agreements and the Sector inquiry

The Report characterizes a settlement agreement as a commercial agreement pursuant to which parties settle their patent related disputes, opposition procedures, and litigation. Settlement agreements can give rise to competition concerns where they lead to the delay of generic entry in return for a payment from the originator to the generic company. It should be noted, however, that the agreements with such features represent a relatively small minority. During the sector inquiry, the Commission examined 207 agreements concluded between 2000 and 2008. Most of them (52 percent) did not restrict generic market entry. As regards the other 48 percent, entry was restricted in various ways: an absolute ban on entry, postponed access, or access under a license from the originator.

In addition, of these 48 percent, most agreements did not provide for value transfers. Only 45 percent of the restrictive agreements provided for value transfers in the form of lump sum payments, the grant of distribution rights, or compensation for legal costs and/or the purchase of assets, such as stocks of products in the possession of the generic company. Moreover, these payments occurred in both directions: payments flowing from the originator to the generic firm and payments from the generic to the originator. This being said, the amounts of money transferred from the originator to the generic (200 million EUROs) are significantly higher than amounts flowing in the opposite direction (7 million EUROs).

In total, the Report gives a relatively dispersed picture of settlement agreements, which can be summarized as follows: 108 agreements without entry restrictions (of which 69 percent were without value transfer and 31 percent with such transfer) and 99 agreements with entry restrictions (of which 55 percent were without value transfer and 45 percent with value transfer). This picture does not justify the finding that payments from the originator to the generic firm are necessarily linked to entry restrictions. Value transfers and restrictions on generic entry are two different concepts that may or may not coincide, especially since value transfers can also take place from the generic to the originator. As the Final Report observes, patent settlements are fact-specific and are difficult to categorize in general terms.

VALUE TRANSFERS AND RESTRICTIONS ON GENERIC ENTRY ARE TWO DIFFERENT CONCEPTS THAT MAY OR MAY NOT COINCIDE, ESPECIALLY SINCE VALUE TRANSFERS CAN ALSO TAKE PLACE FROM THE GENERIC TO THE ORIGINATOR.

Even so, the Commission also sought to identify the reasons why pharmaceutical companies entered into settlement agreements. These considerations vary from originator to generic companies. For originators, the Final Report lists two main reasons: the relative strength of the patent rights at stake and the revenues generated by the patented products. In assessing the strength of their patent rights, originators particularly focus on the ability to obtain interim injunctions against generic entry. For generic firms, the relative strength of the patent rights also plays an important role, but less so than the litigation costs, suggesting that generics seem to prefer a settlement agreement over a legal war of attrition.

Interestingly, both the originator and generic firms attach much importance to the position of other generic entrants. If there are more generic firms likely to enter the market, the incentive for the originator to enter into a settlement agreement increases, because the agreement keeps his patent rights in place and, hence, their deterrent effect *vis-à-vis* other generic contestants. For the generic firm, it is important to secure a position as the first generic on the market. In general terms, prices of pharmaceutical products rapidly erode once several generic entrants have penetrated the market. Entering into a settlement agreement might mean that the generic contracting party is the only generic on the market.

Finally, the Report notes on several occasions that the description of the agreements and the U.S. enforcement practice against such agreements do not provide any guidance on whether certain types of agreements could be deemed compatible or incompatible with EC Competition law. The Report indeed states that "such an assessment would require an in-depth analysis of the individual agreement, taking into account the factual, economic and legal background."

This leads us to the next section, which deals with the question where such guidance can be found.

III. Guidance From Old Precedents

A. TRADEMARK DELIMITATION AGREEMENTS

As mentioned above, there are no precedents under EU competition law dealing with settlement agreements concluded between originator and generic pharmaceutical firms. However, in the earlier phases of EU competition law, the Commission and the Court of Justice had the chance to assess comparable issues when dealing with the compatibility of trademark delimitation agreements with Article 81 EC. These cases concerned the settlement of conflicts between owners of trademarks, which could be considered as giving rise to confusion. Where these settlements involved companies from different Member States, the settlement could lead to the allocation of national markets, and hence to splitting up the common market.

THERE ARE NO PRECEDENTS UNDER EU COMPETITION LAW DEALING WITH SETTLEMENT AGREEMENTS CONCLUDED BETWEEN ORIGINATOR AND GENERIC PHARMACEUTICAL FIRMS.

Obviously, this allocation of territories raised questions as to its compatibility with the market integration objective which, at that time, was still listed high on the Commission's priority list for antitrust enforcement. The Commission sought to reconcile this tension by assessing, or second guessing, what the outcome of the trademark conflict would have been. In the presence of a genuine trademark conflict, the Commission considered that trademark delimitation conflicts could not be regarded as restrictive in nature.

The *Sirdar/Phildar* case of 1975 shows, however, that this approach did not correspond to the Commission's initial position.[4] The case concerned a trademark settlement, pursuant to which Sirdar was allowed to use this trademark for the supply of knitting yarn in its home state, the United Kingdom, and its French counterpart, the Phildar trademark in France. Elsewhere, the trademarks would coexist. The Commission bluntly found that the agreement had as its object to restrict competition, since it restricted the possibility for both companies to sell in each other's territories.

Two years later, the Commission followed a more nuanced approach when assessing the trademark delimitation agreement concluded between two textile companies, namely J.C. Penney Co. from the United States and the Anglo-Irish ABF Group, which sold its products under the Penney's trademark.[5] The Commission considered that the agreement offered the least restrictive alternative to solve the dispute.[6] It noted that the application of national trademark law would have allowed each party to oppose imports by the other party in each other's territories. In addition, the exports affected by the agreement represented relatively small quantities.

The last official decision dealing with trademark delimitation issues dates from 1982.[7] It concerned a dispute between two producers of tobacco products, namely Segers and BAT. Their agreement sought to put an end to the alleged confusion between the Toltecs and Dorcet trademarks in Germany. In this case, the Commission also analyzed whether the agreement, which prevented Segers from importing certain products under the Toltecs trademark from the Netherlands into Germany, led to a more restrictive result than the result to which the unilateral assertion of trademarks right would have led. When applying this test, the Commission found that there could be no serious ground for phonetic or visual confusion between the Toltecs and Dorcet trademarks. It also noted that Segers had not availed itself of the possibility to have the Dorcet trademark removed from the German trademark register, despite the fact that this trademark was not effectively used. The Commission, therefore, qualified the settlement agreement as restrictive in nature, especially since BAT, the owner of the Dorcet trademark, had entered into a series of similar agreements.

BAT challenged the 50,000 EUROs fining decision before the European Court of Justice.[8] It held that the Commission was not competent to assess whether there was a real risk of confusion or not. This was, according to BAT, a matter of German trademark law and not of Community (competition) law. The Court rejected this argument. It acknowledged that trademark delimitation agreements are "lawful and useful if they serve to delimit, in the mutual interests of the parties, the spheres within which their respective trademarks may be used, and are intended to avoid confusion or conflict between them."

However, such agreements may be caught by the cartel prohibition of Article 81(1) EC, "if they also have the aim of dividing up the market or restricting competition in other ways." The Commission is therefore competent to intervene against such agreements. The Court specified in this respect that the "Community competition system does not allow the improper use of rights under any national trademark law in order to frustrate the Community's laws on cartels" (ground 33). As regards the facts of the case, the Court shared the Commission's analysis that the settlement agreement basically imposed undue restrictions on Segers' ability to import tobacco products in Germany. The agreement did not clearly specify to which tobacco products the conflict related. Nor did it contain any explanation why Segers waived its right to claim priority rights for its trademark. It also contained a restriction on advertising that did not bear "even the semblance of a connection with the question of the use of the trademark as such."

It follows from this overview that the Commission, as well as the Court, consider that trademark settlement agreements are not caught by Article 81 if they genuinely seek to avoid a real dispute between the parties, and that antitrust authorities are competent to make their own assessment of the risk of confusion and therefore of the authenticity of the dispute. The cartel prohibition applies, however, if the dispute is sham and if the settlement agreement just covers up a

market-sharing agreement. The prohibition also affects restrictive provisions that go beyond what is required to solve the dispute. In other words, the cartel prohibition does not apply to trademark delimitation agreements that are necessary and proportionate in view of solving a trademark conflict.

THE CARTEL PROHIBITION DOES NOT APPLY TO TRADEMARK DELIMITATION AGREEMENTS THAT ARE NECESSARY AND PROPORTIONATE IN VIEW OF SOLVING A TRADEMARK CONFLICT.

B. PATENT NO-CHALLENGE CLAUSES

There are no precedents under EU competition law explicitly dealing with patent settlements, but there are various decisions and judgments concerning no-challenge clauses: *i.e.* contractual provisions, which often appear in distribution, licensing, or joint venture agreements, and which prohibit the licensee from contesting the validity of the patents covering the licensed products. This case law may be relevant for assessing the legality of patent settlements under EU competition law, because no-challenge clauses are an integral part of most, if not all, of these settlements. Such clauses often embody the outcome of the settlement by specifying the respective patent rights of the parties and their commitment to respect these rights.

In the early stages of European competition law, no-challenge clauses were treated with suspicion. In the old *AOIP/Beyrard* case, the Commission held that a contractual restriction on the licensee's ability to contest the validity of the patent was contrary to the public interest:

> "Even if it is the licensee who is best placed to attack the patent on the basis of the information given to him by the licensor, the public interest in the revocation of patents which ought not to have been granted requires that the licensee should not be deprived of this possibility."[9]

This statement reflects a certain distrust in patents. They are seen as obstacles to commercial freedom.

This negative approach also influenced the Commission's legislative policy. The Commission indeed systematically excluded the benefit of its block exemption regulations for agreements containing no-challenge clauses: "Article 81(1) shall not apply to agreements including certain obligations, provided that these obligations are without prejudice to the (…) right to challenge the validity of the (…) patent."[10] This position changed with the adoption of the block exemption currently in force. Article 5(1) sub(c) of Regulation 772/2004 excludes patent no-challenge clauses from the scope of the block-exemption for technol-

ogy transfer agreements; but, unlike the preceding regulations, the presence of such a clause in a license agreement no longer implies that the entire agreement loses the benefit of the block-exemption.[11]

The Court's approach to patent no-challenge clauses also evolved over time. In the *Windsurfing* case, it followed the same rigorous approach as the Commission did in its early decisional practice.[12] The case concerned the legality of several contractual restrictions imposed by Windsurfing on its licensees. The Commission held that Windsurfing only held a patent on the rig and that the controversial licensing provisions were seeking to extend the scope of the patent protection to the board. In this context, the Commission objected to a clause that prevented Windsurfing's licensees from challenging the patents. Windsurfing did not accept this reasoning and appealed against the prohibition decision before the European Court of Justice.[13]

Windsurfing argued in the first place that the Commission was not entitled to assess the scope of its patents. This was, in its view, a matter of national law. Relying on similar grounds as those put forward in the *BAT* case referred to above, the Court acknowledged that the Commission is not competent to determine the scope of a patent, but accepted that the Commission can assess a patent's scope where this is relevant to determine whether or not Community competition rules have been infringed. This assessment is carried out only in the context of competition law procedures and does not bind national courts when they have to rule on the validity or scope of the patent under national law.

After having thus clarified the Commission's competence in patent-related matters, the Court examined the appeal against the Commission's prohibition of the patent no-challenge clause. The Court ruled that such a clause was not covered by the patent right itself and that it was contrary to the public interest:

> "such a clause clearly does not fall within the specific subject matter of the patent, which cannot be interpreted as also affording protection against actions brought in order to challenge the patent's validity, in view of the fact that it is in the public interest to eliminate any obstacle to economic activity which may arise where a patent is granted in error."

Without any further reasoning, the Court qualified the no-challenge clause as an unlawful restriction of competition.

In 1988, however, the Court of Justice took a more liberal stance regarding patent no-challenge clauses.[14] The case concerned a patent settlement between

Bayer and Mr. Süllhöfer who each held patents for construction panels. Under that agreement, Süllhöfer granted Bayer a non-exclusive, royalty-free license with the right to sublicense its patents in Germany, and a non-exclusive license subject to royalties in other Member States. From its side, Bayer granted Süllhöfer a royalty-bearing, non-exclusive license. Bayer also undertook not to challenge the validity of Süllhöfer's patents. The truce which this agreement was supposed to bring about was of short duration. Soon after its conclusion, the parties started to argue about its interpretation. In this context, the German courts stayed proceedings and requested the Court of Justice to rule on the validity of a patent no-challenge clause under Article 81 EC.

During the proceedings before the Court, the Commission argued that a non-challenge clause could not be considered as restrictive, when it is included:

> "in an agreement whose purpose it is to put an end to proceedings pending before a court, provided that the existence of the industrial property right which is the subject-matter of the dispute is genuinely in doubt, that the agreement includes no other clauses restricting competition, and that the no-challenge clause relates to the right in issue."

The Commission thus took the same position as the one adopted for the assessment of trademark delimitation agreements.

Bearing in mind that it had followed a similar approach in the *BAT* case, the Court's reaction to the Commission's argument can be qualified as surprising. The Court discards the suggestion that the legality of the no-challenge clause should be assessed in conjunction with the settlement agreement which it is supposed to support. The Court isolates the clause from the context of the settlement and analyzes it directly. It holds in the first place that where the license is granted for free, there can be no restriction of competition, because "the licensee does not suffer from the competitive disadvantage involved in the payment of royalties" (ground 17). Moreover, even where a license had been subject to payment, a no-challenge clause is not restrictive, "if the license relates to a technically outdated process which the licensee undertaking did not use" (ground 18). Finally, the Court pointed out that:

> "if the national court were to consider that the no-challenge clause contained in the license granted subject to payment of royalties does involve a

limitation of the licensee's freedom of action, it would still have to verify whether, given the positions held by the undertakings concerned on the market for the products in question, the clause is of such a nature as to restrict competition to an appreciable extent (ground 19)."

C. PRELIMINARY CONCLUSIONS

The case law examined above is relatively old; one should therefore be cautious in drawing conclusions. Even so, one can be relatively confident that the Commission and Courts will still apply the "least restrictive alternative test" as developed in the case law on trademark delimitation agreements. If the outcome of settlement is less restrictive than what the outcome of (protracted) litigation would have been, the settlement agreement can hardly be considered as restrictive. There is one proviso to this test: The agreement should not only be necessary, but also proportionate to solve the conflict. Restrictions that have no bearing with the underlying dispute will not benefit from the presumption that they are not restrictive in nature.

EVEN SO, ONE CAN BE RELATIVELY CONFIDENT THAT THE COMMISSION AND COURTS WILL STILL APPLY THE "LEAST RESTRICTIVE ALTERNATIVE TEST" AS DEVELOPED IN THE CASE LAW ON TRADEMARK DELIMITATION AGREEMENTS.

The Court accepts that the application of this test implies some form of second guessing of the relative strength of the patent rights at stake by the competent antitrust authority. It should be noted that this assessment is only made for the purposes of applying EC competition rules, and that it does not bind national courts when they are requested to determine the validity of patent rights under national patent law.[15]

It follows from the Commission's position in *Bayer v. Süllhöfer* that no-challenge clauses are, in its view, an integral part of settlement agreements, and that their legality should be assessed in conjunction with those agreements. The Court of Justice, however, seems to consider that the question as to whether or not a no-challenge clause restricts competition must be assessed in isolation.

However, it is also possible to interpret the *Bayer v. Süllhöfer* precedent in another way. It may indeed be considered that the facts of the case did not justify a complex assessment of the underlying patent dispute. Since the no-challenge clause related to a technology which Bayer did not use in any event, and for which it did not have to pay, applying EC competition rules can be regarded as a relatively hypothetical issue which did not merit much judicial attention.

The *Bayer v. Süllhöfer* case contains one important proviso; namely, the fact that a contractual provision which restricts the freedom of action of one of the parties does not suffice to trigger the prohibition of Article 81. Any agreement

must be assessed in its economic and legal context and will only be caught by this prohibition if it appreciably restricts competition. This applies to all agreements, including settlement agreements that cannot be justified by the underlying patent dispute. The least restrictive alternative test discussed above simply means that settlements meeting this test are generally not caught by Article 81 EC, but it does not inevitably mean that agreements failing this test are necessarily prohibited. Sham agreements are not necessarily restrictive agreements; they will therefore only be caught by Article 81(1) EC if they appreciably restrict competition in their economic and legal context.

IV. Assessing Patent Settlement Agreements Under Article 81

The requirement that all agreements must be assessed in their legal and economic context implies that there is, unlike U.S. competition law, no *per se* rule which could possibly apply to settlement agreements. It should be noted that this approach also applies to so-called hard-core restrictions such as price-fixing, market-sharing, or output restrictions. The fact that these restrictions cannot benefit from the presumption of legality conferred by the Notice of minor importance does not dispense the Commission or Courts from assessing whether they can, by their object or effect, restrict competition in a given legal and economic context.[16] In any event, one cannot reasonably argue that settlement agreements are akin to hard-core or naked restrictions which can be presumed to be anticompetitive. As shown by the Final Report on the sector inquiry, there is a large variety of settlement agreements and only a minority of these agreements is likely to give rise to competition concerns. Settlement agreements must therefore be assessed on a case-by-case basis.

The first step of this analysis concerns the question whether the settlement agreements have, as their object or effect, restricting or delaying generic market access. If they do not, they are unlikely to be caught by Article 81 EC. Some settlement agreements may even be pro-competitive. This is the case, for example, with settlement agreements which allow the generic firm to launch its product or which allow it to create generic market presence.

The issue of value transfers from originator to the generic is not relevant when assessing the restrictive nature of a settlement agreement. As shown by the Final Report, payments may take place in all sorts of manners and under all sorts of settlement agreements, including those that do not restrict competition. Reverse payments by themselves are and cannot be restrictive.

The second step of the analysis only applies to settlements that delay or restrict generic market access. Applying the trademark case law discussed under Section 2 by analogy, one could argue that Article 81 does not apply to such agreements where they are less restrictive than the outcome of patent litigation between the

originator and the generic. Indeed, if the originator fully succeeds in enforcing its patents, there will be no generic entry whatsoever. In other words, Article 81 does not apply to settlement agreements which produce restrictive effects which are less or equal to those resulting from the judgment on the merits of the originator's patents. The application of this "least restrictive alternative" test implies that the authority must make its own assessment of the relative strength of the patents at stake. This judge or authority must, in a certain sense, second guess what a specialized patent court or authority would have decided if the parties to the agreement had fought their dispute until the bitter end.

Here again, the presence or absence of payments does not seem relevant for carrying out this assessment. As a rule, the relative strength of a patent is a technical issue and not a financial one. Even so, a significant value transfer to the generic firm in a scenario where the originator's patent is *prima facie* weak, may constitute an indication that the originator was paying the generic firm to not enter the market, in particular when the parties to the agreement do not have any plausible explanation for the disproportionate nature of the payment. In other words, reverse payments may, in certain scenarios, offer circumstantial evidence for finding that the settlement agreement does not constitute the least restrictive alternative.

This brings us to third step of the analysis. The fact that a settlement restricts generic entry and that this effect cannot be justified by the patents invoked does not suffice to trigger Article 81(1) EC. This fact simply implies that the agreement restricts competition between the contracting parties, but does not imply that it significantly restricts it in the Common Market, as required by Article 81(1) EC. This last condition implies, as stated above, that the settlement agreement in question must be assessed in its legal and economic context. There are various situations in which a restriction of the competition between the parties does not necessarily lead to a restriction of competition in that wider context.

If the parties concerned only have a small market presence, the agreement is unlikely to have such an effect. The Notice on agreements of minor importance lays down the presumption that agreements involving parties whose market share does not exceed 10 percent do not appreciably restrict competition. So, if the market share of the parties to the settlement agreement remains below this threshold, the agreement is unlikely to lead to an appreciable restriction of competition.

However, the application of market share thresholds obviously requires the definition of a relevant market. Under its decisions to date, the Commission has defined relevant markets in the pharmaceutical sector on the basis of therapeutic indications: All drugs which can be prescribed for the same therapeutic indication are considered to be part of one and the same product market.[17] One may wonder, however, whether this traditional market definition method is always adequate to assess settlement agreements between originators and generics. As

illustrated by the Report on the sector inquiry, price levels in markets where no generic entry has taken place are significantly higher than the price levels prevailing in markets which have already turned generic.[18] A settlement agreement that delays generic entry may effectively keep price levels high and thus significantly restrict competition, if it is concluded between an originator and the first potential generic entrant. Such an agreement would prevent the market from turning generic and hence protect the higher price levels. By contrast, a settlement agreement concluded between firms that already operate in a market with generic market presence is unlikely to produce such effects.

One may wonder, however, whether this traditional market definition method is always adequate to assess settlement agreements between originators and generics.

Seen from this angle, it is also possible to refer to what the Court meant to say in *Bayer v. Süllhöfer* case. Agreements restricting the use of products that are not going to be used regardless, are unlikely to have a significant market impact triggering Article 81(1). If the pharmaceutical products covered by the settlement agreement are unlikely to be used or sold, the settlement agreement does not merit much attention from the antitrust enforcers.

This last comment leads to the more general question concerning the expediency of antitrust enforcement against settlement agreements. Obviously, fighting agreements which delay market entry and which create unnecessary costs for social security schemes is a good cause. It is less obvious that settlement agreements contribute significantly to this delay. The Final Report does not quantify the societal costs that could possibly be allocated to settlement agreements that delay market entry. It rather conveys a picture of a wide variety of agreements. The majority of these settlements do not restrict generic market entry.

Moreover, distinguishing restrictive settlement agreements from neutral or even pro-competitive settlement agreements is a complex task. A radical and harsh condemnation of settlement agreements and reverse payments is hard to reconcile with this complexity and may even have a counterproductive effect. If generic firms lose the option of concluding settlement agreements when they enter the market at the risk of being sued for patent infringements, they may decide not to enter the market at all. Finding the right dosage also applies to antitrust enforcement. ◀

1 Pharmaceutical Sector Inquiry, Final Report, DG Competition, Staff Working Paper, 08.07.2009, *available at*: http://ec.europa.eu/competition/sectors/pharmaceuticals/inquiry/staff_working_paper_part1.pdf.

2 2006/857/EC: Commission Decision of 15 June 2005, (Case COMP/A.37.507/F3 — AstraZeneca), *O.J.*, L 332, 30/11/2006, p. 24 – 25.

3 Please note that this article does not deal with settlement agreements concluded between originator firms.

4 Commission Decision of 5 March 1975, (IV/27.879 - Sirdar-Phildar), *O.J.*, L 125, 16/05/1975, p. 27 – 30.

5 Commission Decision of 23 December 1977, (IV/29.246 - Penneys), *O.J.*, L 060, 02/03/1978 p. 19 – 27.

6 The settlement agreement also provided that Penney America would pay ABF a certain sum in installments. The Commission's analysis did not deal with the legitimacy of these payments.

7 Commission Decision of 15 December 1982, (IV/C-30.128 Toltecs-Dorcet), *O.J.*, L 379, 31/12/1982, p. 19 – 29.

8 Judgment of the Court of 30 January 1985, BAT Cigaretten-Fabriken GmbH v Commission of the European Communities, Case 35/83, *E.C. Reports* 1985, p.363.

9 Commission Decision of 2 December 1975, (IV/26.949 - AOIP/Beyrard), *O.J.*, L 006, 13/01/1976 p. 8 – 15.

10 Article 2 of the Commission Regulation (EC) No 2349/1984 of 23 July 1984 on the application of Article 85 (3) of the Treaty to certain categories of patent licensing agreements, *O.J.*, L 113, 26/04/1985, p. 35; *See also* Commission Regulation (EC) No 556/1989 of 30 November 1988 on the application of Article 85 (3) of the Treaty to certain categories of know-how licensing agreements, *O.J.* L 61, 04/03/1989, p. 1.

11 Commission Regulation (EC) No 772/2004 of 27 April 2004 on the application of Article 81(3) of the Treaty to categories of technology transfer agreements, *O.J.*, L 123 , 27/04/2004 p. 11 – 17.

12 Commission Decision of 11 July 1983, (IV/29.395 - Windsurfing International), *O.J.*, L 229, 20/08/1983, p. 1 – 21.

13 Judgment of the Court (Fourth Chamber) of 25 February 1986, Windsurfing International Inc. v Commission of the European Communities, Case 193/83, *E.C. Reports* 1986, p. 611.

14 Judgment of the Court of 27 September 1988, Bayer AG and Maschinenfabrik Hennecke GmbH v Heinz Süllhöfer, Case 65/86, *E.C. Reports* 1988, p. 5249.

15 This is an important nuance to the primacy rule laid down in Article 16 of Regulation 1/.2003 according to which Commission decisions are binding for national courts and authorities having to rule in the same case.

16 Commission Notice on agreements of minor importance which do not appreciably restrict competition under Article 81(1) of the Treaty establishing the European Community (de minimis), *O.J.*, C 368, 22/12/2001, p. 13 – 15 ; Judgment of the Court of 30 June 1966, Société Technique Minière (L.T.M.) v Maschinenbau Ulm GmbH (M.B.U.), Case 56-65, *E.C. Reports* 1966, p. 235 ; Judgment of the Court of First Instance of 2 May 2006, O2 Germany v Commission, *O.J.* C 154, 01/07/2006, p. 15.

17 *AstraZeneca*, *supra* note 2., §358 and f.

18 Pharmaceutical Sector Inquiry, Final Report, *supra* note 1., p. 77 and f.

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

No Single Monopoly Profit, No Single Policy Prescription?

A Comment on *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit* by Einer Elhauge

Harry First

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

No Single Monopoly Profit, No Single Policy Prescription?

*Harry First**

I. Introduction

Professor Einer Elhauge's most recent article, *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit Theory*,[1] begins with a critique of the "thrall" in which the single monopoly profit theory has held tying law and ends with an affirmation of the current state of the law: "The [current] quasi-per se rule thus correctly condemns ties based on tying market power absent offsetting efficiencies, even without substantial tied foreclosure." I like the beginning and I like the destination. It's the journey that is not without some problems for me.

I divide this essay into two parts. First I want to talk about the goals of antitrust. Second I offer some comments on Professor Elhauge's approach to tying and the importance of the one monopoly profit theory.

II. Antitrust's Goals

The debate over the proper goals of antitrust policy is a long-standing one. Its last major iteration was in the late 1970s through the 1980s when the argument was over: a) whether economics was the sole source of wisdom for antitrust and economic efficiency the sole metric for desirable policy, or b) whether other disciplines and other values—roughly, democratic or social values—should also be considered. Economics and economic efficiency won out, in part on the argument that a single approach and a single value would provide surer (and better)

*Charles L. Denison Professor of Law, New York University School of Law. I thank Eleanor Fox, Dan Rubinfeld, and Oren Bar-Gill for their comments on an earlier draft and their helpful conversations about the themes of this essay.

outcomes than multiple approaches and goals which might not only be in conflict, but also hard to measure against each other.

Professor Elhauge's article is, in a sense, mute acknowledgement of the triumph of economic methodology and economic goals in antitrust. Its methodology is to attempt to solve all the dilemmas of tying and bundled discounts through economic arguments based on hypothetical supply and demand curves and predictions of consumer and producer behavior given certain initial (and restrictive) assumptions about price and demand ("Suppose, for example…"). But as the paper itself explicitly acknowledges, this economic methodology does not always lead to a sure outcome. These are arguments, after all, and Professor Elhauge is engaged in an effort to convince the reader that his economic arguments are superior to the economic arguments that other commentators have made. None of this is surprising, although it is a reminder that economics has not necessarily produced more certainty in antitrust decision-making.

PERHAPS MORE IMPORTANTLY, THOUGH, PROFESSOR ELHAUGE'S ARTICLE SHOWS THAT ECONOMICS DOES NOT NECESSARILY SETTLE THE QUESTION OF THE PROPER GOAL OF ANTITRUST.

Perhaps more importantly, though, Professor Elhauge's article shows that economics does not necessarily settle the question of the proper goal of antitrust. Professor Elhauge makes his view clear from the beginning of the article that "consumer welfare," rather than "total welfare," is the "governing antitrust standard." In juxtaposing "consumer welfare" against "total welfare," Professor Elhauge comes down firmly on one side of an important three-sided debate over antitrust's goals. I say "firmly" rather than "explicitly" because it is more in the telling, as Professor Elhauge works through the hypothetical gains and losses from tying, that it becomes clear that by "consumer welfare" he means the "consumer surplus," and that it is the consumer surplus whose diminution antitrust is intended to prevent. Indeed, critical to many of Professor Elhauge's arguments is his relentless focus on consumer surplus as the sole measure of antitrust policy (and a measurable measure at that).

If "consumer welfare" is to be the goal of antitrust, who could be against it? The answer is no one, which is why consumer welfare is such an attractive rhetorical label.The real issues come when one tries to get behind the label to see what its user has in mind and how easy, or hard, it is going to be to prove its reduction. Professor Elhauge points to Judge Bork's well-known rhetorical capture of the term, equating consumer welfare with the net effect on total welfare (consumer and producer), otherwise known as the deadweight welfare loss of allocative inefficiency. Elhauge captures the flag differently, focusing just on the effect on consumers. To put the dispute more graphically, Bork wanted to focus antitrust on a potentially small triangle "created" when monopolists reduce output to the profit-maximizing monopoly level. In this article Elhauge wants to focus antitrust on some larger triangles that reflect the consumer surpluses in tying and tied products at monopoly and competitive levels respectively, then examine how those

triangles could change with price-discriminating ties and, finally, see whether those changes indicate that producers are now able to take ("extract") some of the surpluses for themselves, thereby, presumably, making consumers worse off and creating antitrust liability—without regard to how output is affected.[2]

There is a lot more behind these two different views of "consumer welfare" than geometry, of course. The total welfare standard rests on the theoretical structure of welfare economics, focusing on the total wealth of society and seeking an allocation of productive resources in a way that best satisfies all consumers' numerous preferences (whatever these preferences may be and however they got them). But a total welfare standard also rests on a policy argument that we should be indifferent to the redistribution of consumer surplus to producers, either because producers are also consumers in an ultimate sense or because we have no good reason to prefer consumers over producers (even if the income of one group is distributed to the other).

Elhauge rests his argument for a consumer surplus standard on a reading of the legislative history of the Sherman Act (which shows that Congress had no concern for allocative efficiency and great concern for the ability of powerful firms to raise prices to buyers), as well as his argument that the Supreme Court has "never embraced a total welfare standard" but has, instead, viewed the Sherman Act as a "consumer welfare prescription." He also differs on the redistribution point, arguing that redistribution from consumers to producers is likely to be "undesirable because shareholders of monopoly firms generally have higher income than consumers."

But, as I said earlier, this is a three-way fight. In addition to battling Bork in the text, Elhauge battles Greg Werden in the footnotes. Specifically, Elhauge takes on what he says is Werden's view that "antitrust law protects not consumer welfare, but 'the competitive process.'" Putting aside the question whether Werden really sees no role for antitrust in protecting consumer welfare, Elhauge correctly challenges the "competitive process" goal for antitrust as being poorly defined. What exactly do we mean by it? More competitors? More competitive behavior? Can't be, because we allow mergers and we permit firms to collaborate. No, courts might say they are protecting the competitive process, but they only do so when consumer welfare—presumably meaning consumer surplus—is harmed.

If we can't show an effect on end-user buyers from, say, a buyer's cartel or a merger of manufacturing input suppliers, should antitrust then not apply?

Werden's riposte (although not made directly to Elhauge's article) is that "consumer welfare" is often a poorly-defined term. More to the point, with which consumers are we concerned? Textbook economic theory posits consumers who are people, thus pointing to the end-user buyer as "the consumer," but real-life markets and antitrust problems often involve intermediate buyers and sellers that are not people. If we can't show an effect on end-user

buyers from, say, a buyer's cartel or a merger of manufacturing input suppliers, should antitrust then not apply?[3]

The truth is that when we look at any of the proposed goals for antitrust, we can find something missing. I agree that "competitive process" is a fuzzy term, but we need something to get beyond the static account of neoclassical price theory. We need to explain how firms move from time 1 to time 2, to understand the mechanics of what incentives need to be maintained to push firms to lower price or to innovate, and to see what exclusionary practices can dampen those incentives. Preserving the "competitive process" acknowledges that we can't predict with precision how "consumer surplus" might be affected in the future, but we can examine the processes that are likely to achieve the results that consumer surplus tries to measure. At the same time, although figuring out the consumer surplus may not allow us to decide every case, it does help us understand how buyers can be hurt in some cases (even intermediate buyers) and we need not work through the complex economics of passing-on to know that effects in intermediate markets can affect capital flows or innovation or pricing in ways that are hard to trace in a complicated economy. As for a total welfare standard, it is true that such a standard might ignore immediate harm to buyers; but, still, we can't be completely indifferent to what happens to producer surplus. How else to understand antitrust's continuing concern for efficiencies?

But even this three-way fight leaves some important economic effects out of the calculus. What about consumer choice? Consumers value it, the courts have mentioned it.[4] Might this not be something worth paying attention to? What about innovation efficiency? There is now a danger that courts will pursue a naïve Schumpeterian view of the need for monopoly as an incentive to innovation. Should not antitrust pay more attention to conduct that suppresses the competitive incentives for innovation, independent of other measures of efficiency or consumer surplus? Indeed, innovation efficiencies may very well be more important to a progressive economy than either the static measures of allocative efficiency or consumer surplus.[5] What about a new (but, in a way, old) idea on the economic policy front, "too big to fail"? Does antitrust have to ignore this economic concern unless a plaintiff can prove some effect on consumer surplus or total welfare? Might attention to this economic problem be quite consistent with antitrust's traditional concern for large-firm mergers and concentration?

There is now a danger that courts will pursue a naïve Schumpeterian view of the need for monopoly as an incentive to innovation.

And then there is the lurking challenge of behavioral economics. "Consumers" and "producers" are the stick-figures of antitrust analysis. Antitrust economics has little to say about who these consumers and producers are and how they actually behave. Behavioral economics has a lot to say about how con-

sumers behave and how their preferences can be shaped by manipulating the systematic biases that they (we) exhibit when making decisions under conditions of uncertainty. How good a guide for policy is "consumer surplus," then, if all it measures is the sum of such fluid and manipulable preferences?

Producers are similarly under-described. What biases do firm managers exhibit under conditions of uncertainty, when deciding, for example, whether to enter a market? And what about non-manufacturing producers? Our hypotheticals may have moved from widgets to printer manufacturers (the one Professor Elhauge uses in his article), but what about retail distributors or service providers? How do they behave?

Finally, there are distributive concerns. It is possible to use distributive concerns to support some general preference for consumers as a class, as does Professor Elhauge, although the empirics behind the generalization may be unclear today in an economy where many people of modest means own stock and the wealthy are consumers of large amounts of luxury goods. But it is also possible to think of distributive concerns in more specific cases where business practices may have uncertain effects on the welfare of infra-marginal customers but substantial effects on customers who are priced out of the market, customers whom we might call "supra-marginal," or, better yet, "marginalized." For example, allowing resale price maintenance may permit a seller to project and protect an image of exclusivity, but it may also keep the goods away from the discounters that made those goods available to poorer people. Is that a just result? Why must we ignore the welfare of those marginalized customers? Perhaps we could even pay more attention to the marginalizing effect of monopoly pricing, as the following excerpt from *International Technologies Consultants v. Pilkington* indicates:[6]

> "Alistair Pilkington invented an ingenious new method of making high quality flat glass at high speed, much less expensively than by grinding and polishing it, in the 1950's. He thereby made a great contribution to cheap, good plate glass for everyone. . . . The patent enabled the Pilkington company to take exclusive benefit of the idea for a limited period of time, even though numerous other people necessarily knew the method almost immediately. * * * We do not know whether [the defendants] have conspired to prevent others from using the ideas in Pilkington's expired patents, in violation of the antitrust laws, by means of unjustified [trade secrets] litigation and threats of litigation. But if they have, as the complaint alleges, then the world is being deprived of the economic value of Alistair Pilkington's great invention. Indeed, in poorer areas of the world, doubtless people lack windows to let in the sun and keep out the rain, wind, cold, and insects, because of improper exploitation of monopoly pricing."

I think that the lack of consensus on the "ultimate metric" in antitrust (to use Professor Elhauge's words) not only reflects gaps in each argument, it reflects a weakness in the initial argument that there is an ultimate metric. Or, to return to the earlier debate over antitrust's goals, the lack of consensus casts doubt on whether there is a single goal against which antitrust law can be measured, as opposed to a complex set of goals against which competitive practices must be judged. To put it another way, there is no single policy prescription.

III. Tying and the One Monopoly Profit Theory

The one monopoly profit theory has certainly had an important impact on how we think about tying agreements. I'm not sure that commentators and courts have been held in thrall to it, or that its limits are not understood, but it certainly is a worthwhile scholarly endeavor to deal with the second step of the theory; that is, the argument that ties are imposed not to increase monopoly profits but, often, to price discriminate and that such price discrimination can expand output, which is welfare-enhancing. Professor Elhauge deals at length with ties that effect price discrimination (in various ways) and shows that monopoly profits (or, perhaps, price raising) might really be possible in the tied product market and that consumers will be hurt because they will have less consumer surplus between the tying and tied product, whatever might happen to output. Professor Elhauge's conclusions seem right to me.

WHAT STRIKES ME AS A LITTLE UNUSUAL IN PROFESSOR ELHAUGE'S TREATMENT OF THE SINGLE MONOPOLY PROFIT THEORY, THOUGH, IS THAT DESPITE THE ANNOUNCED TITLE OF THE ARTICLE, AND UNLIKE A GOOD MURDER NOVEL, THE VICTIM DOESN'T DIE IN THE END.

What strikes me as a little unusual in Professor Elhauge's treatment of the single monopoly profit theory, though, is that despite the announced title of the article, and unlike a good murder novel, the victim doesn't die in the end. There is no "death of the single monopoly profit theory." Rather the article ends this way:

> "The [current] quasi-per se rule thus correctly condemns ties based on tying market power absent offsetting efficiencies, even without substantial tied foreclosure. However, this so-called quasi-per se rule should not apply to products that have a fixed ratio and lack separate utility because those conditions generally negate anticompetitive effects absent substantial tied foreclosure."

And the article reaches this conclusion because "[t]ying cannot extract individual consumer surplus … if the products are used or tied in fixed ratios,

because then buyers would experience any tied product price increase as an increase in the marginal price of buying the tying product." In other words, the single monopoly profit theory is correct and, where it holds, current law is wrong.

Why is current law wrong, though? That a monopolist imposing a tying and tied product in fixed proportions can't earn additional monopoly profit doesn't make the tie presumptively lawful. Consumers are still denied a choice they might prefer in the tied product market and, in some cases, innovation in the tied product market might be dampened or suppressed. (What incentives will there be to innovate in complements if the monopolist can just tie the innovation out of existence?) There might also be other reasons why such a monopolist would impose such a tie—for example, to impede or deter entrants in the tied product market that might grow to challenge its monopoly position in the tying product market (a possibility that Professor Elhauge does recognize in the article). Why not stick with the presumption of illegality and shift the burden to the defendant to show an efficiency justification for refusing to sell the products unbundled? Why give in to the one monopoly profit theory?

Whether Professor Elhauge's life support for the single monopoly profit theory matters much to the actual case law, though, is questionable.[7] Take the three controversial tying cases that he discusses, *Kodak*, *Microsoft*, and *Jefferson Parish*. It seems to me that the only case in which the theory might matter is, curiously, the case in which the *per se* rule has received its strongest articulation, *Jefferson Parish*.

WHETHER PROFESSOR ELHAUGE'S LIFE SUPPORT FOR THE SINGLE MONOPOLY PROFIT THEORY MATTERS MUCH TO THE ACTUAL CASE LAW, THOUGH, IS QUESTIONABLE.

Kodak isn't plausibly a case of fixed proportions. There were thousands of Kodak replacement parts. No matter what Justice Scalia wrote (customers will demand "one part with one unit of service necessary to install the part"), it's hard to imagine a world in which each part that a customer needs would necessitate a separate service call.

Microsoft is more plausibly a fixed proportions case—one operating system, one browser. But that wasn't really true, either, or perhaps it was just not important. Many corporate customers didn't want a browser at all (they didn't want their employees wasting time surfing the web!), so these products were un-complements for them. Microsoft denied them this option but, because the browser was sold at a nominal zero price, these customers paid no more when they were forced to take Internet Explorer and would have paid no less without it. Even for those customers for whom operating systems and browsers were strong complements, though, it's not clear to me that these two programs are used in fixed proportions. There is continuing, but perhaps varied, demand for upgrades of software. Microsoft continued to provide new versions of IE more frequently than it could provide new versions of Windows, which seems to me "unfixes" the proportions in use. But, again, I'm not sure that this is the crux of the competition problem

in the case either, because Microsoft wasn't charging a positive price for IE, so consumers weren't paying more if they stuck with IE through all the upgrades or only some of them. The competition problem, of course, was the exclusionary effect on Netscape, which affected innovation and consumer choice in browsers and which also helped to maintain Microsoft's monopoly in the operating system market[8]

Jefferson Parish is the case that looks closest to Professor Elhauge's fixed proportions/no separate utility exception to the (modified) *per se* rule. One surgery, one anesthesia; patients won't take one without the other. Professor Elhauge suggests that maybe the proportions weren't fixed because the number of days in the hospital can vary and some anesthesiologists visit their patients after surgery to see how they are doing. But, really, if we are ever likely to litigate a case of fixed proportions, this would be it.

Before we desert current law though, we should think about those consumers that antitrust law is supposed to protect. In tying, the protection is from being forced to take a product a consumer doesn't want rather than one the consumer would prefer. What stronger case could there be for consumer choice than a case like *Jefferson Parish*, where the choice that a consumer—a patient—might want to make is the choice of the anesthesiologist who will put you out in surgery and, hopefully, wake you up when it's over.

IV. Conclusion

Professor Elhauge's article deals very usefully with what I have called the "second step" of the one monopoly profit theory, the step that argues we should either be indifferent to ties imposed as a way to price discriminate or hail such ties for expanding output. The article not only carefully shows where we should not be indifferent to the monopoly seller's power to impose the tie, because the price discrimination can harm consumers, but also provides a useful bridge from economic theory to the legal rules that courts should apply in antitrust cases. Feeling confident in the economic prediction, Professor Elhauge can then support what he calls the "quasi-per se" rule, or what I would prefer to call a "structured rule of reason" analysis.

PROFESSOR ELHAUGE'S ARTICLE DEALS VERY USEFULLY WITH WHAT I HAVE CALLED THE "SECOND STEP" OF THE ONE MONOPOLY PROFIT THEORY, THE STEP THAT ARGUES WE SHOULD EITHER BE INDIFFERENT TO TIES IMPOSED AS A WAY TO PRICE DISCRIMINATE OR HAIL SUCH TIES FOR EXPANDING OUTPUT.

In my view, though, the bridge he builds relies too heavily on a single pillar—consumer surplus. Concern for effect on consumer surplus is useful, but it is neither necessary nor sufficient for antitrust policy in general or for tying policy in particular. Indeed, it seems to have led Professor Elhauge to argue that the current approach to tying

should be relaxed for those very rare cases that meet the strict requirements of the single monopoly profit theory. I see no reason to give ground in such cases. Other antitrust policies may still justify applying a structured rule of reason, even in the cases that meet the one monopoly profit theory's restrictive assumptions, thereby shifting to the defendant with market power the burden of proving economic justification for the tie.

---

1 Einer Elhauge, *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit Theory*, Harv. L. Rev. 123 (forthcoming Dec. 2009).

2 I refer to "triangles" rather than the more familiar "rectangle" of consumer surplus lost from monopoly pricing because Professor Elhauge draws our attention to all the consumer surplus available above the market price and out to the *y*-intercept, which creates a large triangle if demand is linear.

3 Werden spells out his views in the article that Professor Elhauge cites, *see* Gregory J. Werden, *Competition, Consumer Welfare, & the Sherman Act*, 9 Sedona Conf. J. 87 (2008), and in a subsequent paper, *see Essays on Consumer Welfare and Competition Policy*, *available at* http://ssrn.com/abstract=1352032 (rev. May 15, 2009). In the latter paper Werden helpfully elaborates on the varying uses of "consumer welfare" in the economics and legal literature.

4 For fuller discussion, *see* Neil W. Averitt & Robert H. Lande, *Using The "Consumer Choice" Approach To Antitrust Law*, 74 Antitrust L.J. 175 (2007). For judicial articulation, *see* Aspen Skiing Co. v. Aspen Highlands Skiing Corp., 472 U.S. 585, 606-07 (1985) (consumer preference for skiing on four mountains rather than three).

5 Joseph Brodley wrote thoughtfully about innovation efficiency in *The Economic Goals of Antitrust: Efficiency, Consumer Welfare, and Technological Progress*. *in* Revitalizing Antitrust in its Second Century 95 (Harry First, Eleanor M. Fox, & Robert Pitofsky, eds., 1991). For the naïve Schumpeterian view, *see* Verizon Communications Inc. v. Law offices of Curtis V. Trinko, LLP, 540 U.S. 398, 407 (2004).

6 International Technologies Consultants, Inc. v Pilkington PLC, 137 F.3d 1382, 1392-93 (9th Cir. 1998).

7 It might matter more in future cases, however, if litigants, attracted by his proposal, devote more effort to making their ties look like ones with fixed proportions.

8 Microsoft did not seem to have raised the one monopoly profit theory as a defense to charges of tying either the browser or the media player to the Windows operating system, but it did raise the theory in the European Commission's case involving Microsoft's refusal to provide interoperability information between Windows and Microsoft's work group server operating system. Microsoft argued that it had no improper incentive to leverage from the PC operating system market into the work group server operating system market because according to the one monopoly profit theory, it could not increase its monopoly profits even if it obtained a monopoly in the second market. The Commission rejected this argument because the two operating systems were not used in fixed proportions. *See* Case COMP/C-3/37.792—Microsoft Corp., Comm'n Decision, 2007 O.J. (L 32) 23, ¶¶ 764-67 (Mar. 24, 2004), *available in full at* http://ec.europa.eu/competition/antitrust/cases/decisions/37792/en.pdf.

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Can Bundled Discounting Increase Consumer Prices Without Excluding Rivals?

A Comment on *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit* by Einer Elhauge

Daniel A. Crane & Joshua D. Wright

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Can Bundled Discounting Increase Consumer Prices Without Excluding Rivals?

*Daniel A. Crane & Joshua D. Wright**

I. Introduction

Since we abhor suspense, we will quickly answer the question our title poses: No. As a general matter, bundled discounting schemes lower prices to consumers unless they are predatory—that is to say, unless they exclude rivals and thereby permit the bundled discounter to price free of competitive restraint. The corollary of this observation is that bundled discounting is generally pro-competitive and pro-consumer and should only be condemned when it is capable of excluding rivals.[1]

We pose and answer this question because it is at the heart of Section VI of Professor Elhauge's provocative draft article which is the subject of this symposium.[2] In Section VI, Professor Elhauge argues that bundled discounting can have "power effects" identical to conventional tying arrangements irrespective of any exclusionary effect on rivals as well as that cost/revenue tests for bundled discounting perversely immunize the worst bundled discounting schemes—those that represent the highest non-exclusionary price increases to consumers.

We disagree with Professor Elhauge on these propositions, as we do on many of the earlier arguments in his draft. At a later date, we will offer a fuller response to his arguments and a qualified defense of a "neo-Chicago" perspective on monopoly leverage, price discrimination, and bundled discounting. Qualified, because we do not believe that monopoly leverage is impossible, that price discrimination is always efficiency-enhancing, or that bundled discounts can never exclude competitors or harm consumers. Rather, we believe that if Chicago overstated its case on each of these points, post-Chicago has far overstated its case on

*Daniel Crane is Professor of Law, University of Michigan and Joshua Wright is Professor, George Mason University School of Law and Department of Economics.

each of these points. Indeed, the best available empirical evidence suggests the frequency of instances of bundled discounts and tying arrangements resulting in harm to consumers as compared to those arrangements improving consumer welfare is very low.[3] Particularly, we believe that:

1. The conditions necessary for monopoly leveraging through tying are narrow and rarely exhibited in real markets and, thus, we should continue to be presumptively skeptical about leverage claims. Further, the theoretical analyses of anticompetitive bundling, tying, and bundled discounts contain highly stylized and restrictive assumptions, assume away efficiency benefits of these practices, and have not generated testable hypotheses supported by empirical tests.

2. The conditions necessary for price discrimination through tie-ins to be output-reducing are rarely exhibited in real markets. Price discrimination should be thought of as competitively neutral in static efficiency terms and frequently, but not always, competitively beneficial in dynamic efficiency terms. More precisely, and contrary to Professor Elhauge's analysis, price discrimination's effects on both static total- and static consumer-welfare are generally ambiguous depending on market conditions. When one takes into account the incentives for price discrimination to intensify price competition and dynamic efficiencies such as the incentive to innovate and offer new products, it becomes clear that sound antitrust policy should view price discrimination as a legitimate and normal part of the competitive process.[4]

3. Bundled discounts only rarely partake of the qualities of tie-ins and they should generally enjoy legal protection unless they are predatory.

INDEED, THE BEST AVAILABLE EMPIRICAL EVIDENCE SUGGESTS THE FREQUENCY OF INSTANCES OF BUNDLED DISCOUNTS AND TYING ARRANGEMENTS RESULTING IN HARM TO CONSUMERS AS COMPARED TO THOSE ARRANGEMENTS IMPROVING CONSUMER WELFARE IS VERY LOW.

For the purposes of this symposium, we tackle only the last of these propositions. In brief, we argue that Elhauge's "power effects" thesis as to bundled discounts rests on a faulty premise—that the monopolist is free to threaten an unlimited price on the monopoly item in the bundle and, consequently, can charge a higher price for the bundle than it could for sales of the goods individually. To the contrary, since a rational monopolist will already have charged the profit-maximizing monopoly price on the monopoly item, its threat to charge a higher price unless the customer accedes to a bundled discount demand is hollow. Execution of such a threat would harm the monopolist, and harm it considerably more than the opposite predatory strategy of cutting prices.

While there may be a few examples of such strategies in the real world, we are skeptical that such strategies occur frequently enough to organize legal rules around them. The economics literature and available evidence supports our skepticism. Bundled discounting law should focus on the paradigmatic threat—that a bundled discounting package will exclude rivals and thereby increase the defendant's monopoly power.

II. Bundled Discounts as Tie-Ins

A practice ostensibly related to tying that has received much attention in the last decade is bundled discounting—where the dominant firm offers customers a discount if they choose to purchase a package of goods or services.[5] There is presently a circuit split over how antitrust law should evaluate such discounts. The U.S. Court of Appeals for the Third Circuit treats them as akin to tying or exclusive dealing arrangements.[6] The U.S. Court of Appeals for the Ninth Circuit treats them as akin to predatory pricing, subject to a discount attribution rule.[7]

Bundled discounts differ nominally from tie-ins insofar as they offer the buyer a choice of either buying the competitive or monopoly products. The Supreme Court has suggested that the offering of the option to buy the two goods unbundled defeats a tying claim, even if the two goods are offered jointly at a lower price.[8] Still, courts have sensibly recognized that the seller's offer to sell the goods unbundled could be a sham concealing a de facto tying arrangement if the unbundled price was set so high that it would not be economically rational for any customer to accept it.[9]

However, when a significant number of buyers choose to disregard the discount offer and instead purchase the goods individually, it is unlikely that the discount offer is coercive.

However, when a significant number of buyers choose to disregard the discount offer and instead purchase the goods individually, it is unlikely that the discount offer is coercive. The volume of the Areeda-Turner treatise on which Elhauge was a co-editor suggests, as a rule of thumb, that only "separate sales [falling] below ten percent presumptively indicate a de facto tie."[10] The treatise further suggests that when separate sales are above ten percent, the package discount should not be treated as tying at all and that the defendant should not be required to justify the discount as cost-justified.[11]

In his current draft, Elhauge rejects the Treatise's position and proposes a new test that would treat non-exclusionary bundled discounts as unlawful tie-ins under specified circumstances. Elhauge would condemn as an unlawful tie those bundled discount offers where the defendant has market power over the "linking product," the "unbundled price for the linking product exceeds the but-for level," and the defendant cannot offer an offsetting efficiency justification.[12]

As an initial matter, we are very skeptical that identifying the "but-for" price of the linking, *i.e.*, monopoly, product will be feasible in most cases. Elhauge admits the "determining the but-for price can be difficult," but asserts—without citing any examples—that internal business documents or regression analyses will often provide evidence of the but-for price.[13] There are many problems with Elhauge's suggestion.

For one, Elhauge assumes a clean before-and-after story where the defendant used to engage in only single-product pricing and then moved to a bundled discount scheme. In our experience, bundled discounts stories are usually far more dynamic than that simplistic two-stage analysis, with constantly shifting pricing and discounting structures, product innovation, cost changes, and industry dynamics making it impossible to determine clean before-and-after figures.

FURTHER, THE SEARCH FOR THE BUT-FOR PRICE IS BOUND TO RUN INTO THE DIFFICULTY THAT, AS BOTH ELHAUGE AND CHICAGO SCHOOL SCHOLARS BELIEVE, BUNDLED DISCOUNTS OFTEN PRODUCE PRICE DISCRIMINATORY EFFECTS.

Further, the search for the but-for price is bound to run into the difficulty that, as both Elhauge and Chicago School scholars believe, bundled discounts often produce price discriminatory effects. A seller with a monopoly over the linking product will often have engaged in some price discrimination even prior to initiating a bundled discount program and will probably do so afterwards. The aggrieved plaintiff may very well be the loser in the shift from a less-efficient to more-efficient price discrimination scheme. From the plaintiff's perspective, the shift may appear to raise prices in the linking product even though average prices fall. It would be anomalous to allow the individual plaintiff's idiosyncratic experience to determine the legality of the discount, but proving the effect on average prices across all buyers may be impossible.

In any event, Elhauge's assumption that dominant firms will be able to increase the linked product price over the but-for price rests on a faulty premise. He asserts that "[b]ecause the defendant is free to set the noncompliant prices at whatever level it wishes, it can set them above the levels that would have prevailed 'but for' the bundling."[14] Thus, Elhauge argues, a package price that nominally offers discounts does not reflect true price reductions at all, but only a concession off a threatened price that is higher than the prices that would have prevailed absent the monopolist's demand for bundling. This assumption frames much of Elhauge's "power effects" arguments about bundled discounts.

The central problem with Elhauge's argument is that the monopolist cannot obtain much leverage by demanding a price above its profit-maximizing monopoly price. Unless the monopolist has been engaging in some form of limit pricing,[15] it has already priced the monopoly product at the level that makes any further price increase unprofitable. Consequently, any threatened price increase on the monopoly product to punish the buyer for failing to purchase the package

would inflict costs on the seller as well as the buyer. The threat to raise the "tying" product's price thus lacks credibility.

Suppose, for example, that the dominant firm enjoys a monopoly over Product A but faces competition for Product B. The profit-maximizing monopoly price for Product A is $10 and the marginal cost of Product B—which is also the price prevailing in the competitive market—is $5. The dominant firm would ordinarily sell the AB combination for $15. Suppose it seeks to leverage its market power from Product A to Product B. Under Elhauge's approach, the dominant firm could obtain a price above $15 by threatening to increase the price of Product A to, say, $12 if buyers refused to pay, say, and $16 for an AB package. But since $10 was the profit-maximizing monopoly price of A, it would be unprofitable for the dominant firm to raise the price to $12. At $12, the dominant firm would face elastic demand and unprofitably lose sales. Hence, the threat to raise price to $12 would be a hollow one, since it would be as unprofitable for the seller as for the buyer.

One might respond that raising the monopoly price above the profit-maximizing level is simply another form of profit sacrifice that monopolists might utilize to discipline the market. Like below-cost pricing, such unprofitably high pricing might allow the monopolist to exclude rivals or engage in wealth-transferring price discrimination strategies which would, in turn, permit the monopolist to recoup the costs of its unprofitable pricing campaign.[16]

But, if below-cost pricing strategies are risky propositions for the monopolist, above-profit maximizing pricing strategies are even more so. When a predator lowers its price below its cost, it expands output, enlarges its market share, steals customers from its rivals, and often brings new customers into the market. One of the reasons that it is difficult to distinguish predatory pricing from pro-competitive promotional pricing is that, even in a competitive market, temporary aggressive price-cutting may have long-run benefits for the price-cutter if it is able to build customer loyalty in its expanded share of the market. Also, expanding the dominant firm's market share may boost its status and prestige in the market. Even if the price-cutting is truly predatory in the sense that the dominant firm would not have undertaken such a strategy unless it expected to be able to recoup its lost profits in a less-competitive market, the enhanced market share and its loyalty-building and status-building effects may be silver linings in the event the predatory campaign fails.

BUT, IF BELOW-COST PRICING STRATEGIES ARE RISKY PROPOSITIONS FOR THE MONOPOLIST, ABOVE-PROFIT MAXIMIZING PRICING STRATEGIES ARE EVEN MORE SO.

Pricing above the profit-maximizing price is just the opposite. The dominant firm must now cede sales to rivals. Those rivals obtain short-run benefits as their own market share expands and may also enjoy the long-run customer loyalty and

prestige enhancements that a predator experiences. Although the supra-monopoly price need only continue long enough to coerce customers to accept the bundle, that may be long enough to shift the market dynamics in favor of rivals. It is unlikely that many firms would frequently run such a risk. If there is one thing that makes sales executives nervous, it is the prospect of their customers experimenting with a rival's product.

We anticipate three objections to this line of argument. First, some may object that the monopolist over Product A does not have to fear diversion of sales to rivals since, by definition, there are no rivals for Product A. But this argument misconceives the nature of competition in two ways. First, a monopoly does not have to mean a 100 percent market share. The dominant firm may very well face some limited competition within the relevant market and those competitors may be positioned to expand production in the event of further price increases by the dominant firm.

There is an even more fundamental economic point. A monopolist's profit-maximizing price occurs in the elastic portion of the demand curve. The reason that any further price increase would be unprofitable is that the marginal customers would begin switching to other products if the defendant increased its price. Hence, by increasing its price above the profit-maximizing level, the defendant would be inviting its customers to divert purchases to adjacent products that were not previously in direct competition with the monopolist's product. In effect, the monopolist's price increase would be encouraging its customers to consider substitutes for the monopolist's product. Most sales managers would not want to run the risk of their customers experimenting with new products and deciding they prefer them to the monopolist's product.

A second line of objection would follow the game theoretic literature on predatory pricing that suggests that dominant firms do not have to incur the costs of actual predatory pricing if they can obtain a reputation as predators and, hence, deter entry by threatening predation.[17] Perhaps the monopolist could occasionally discipline a customer who rejects its bundled offer by raising the stand-alone monopoly price and thereby obtain a reputation as a punitive seller. As Frank Easterbrook demonstrated several decades ago, there are reasons to be skeptical about reputational theories in single-product predation.[18] There are even more reasons to be skeptical of such a theory when the threat is directed at customers rather than rivals. It is one thing to develop a reputation as a punisher of rivals and quite another to develop a reputation as a punisher of customers. Monopolists do not depend on the good will of their rivals, but they do depend on the good will of their customers.

IF THERE IS ONE THING THAT MAKES SALES EXECUTIVES NERVOUS, IT IS THE PROSPECT OF THEIR CUSTOMERS EXPERIMENTING WITH A RIVAL'S PRODUCT.

Finally, one might object that dominant firms can threaten a supra-monopoly price because the customers will not know that it will harm the monopolist. However, informational asymmetries between the buyer and seller are unlikely to allow the seller to bluff the buyer into believing the threat. The buyer has as much information as the seller about its demand elasticity. The buyer knows that at the threatened higher price point it will simply begin substituting other products and that the seller will therefore experience pain as well if it follows through on its threat. The buyer thus has a strong counter-threat to the seller's threat.

POST-CHICAGO HAS NOT YET MADE THE CASE THAT SUPRA-MONOPOLY PRICING THREATS ARE A REALISTIC OR FREQUENT OCCURRENCE.

We do not claim that a monopolist could never coerce customers to accept a bundle by threatening a supra-monopoly price on the tying product. We are simply skeptical that this would happen often enough to craft legal rules designed to prevent it. In the push-and-pull between Chicago and post-Chicago theories, the issue again comes down to evidence.[19] Post-Chicago has not yet made the case that supra-monopoly pricing threats are a realistic or frequent occurrence.

III. Bundled Discounts as True Discounts

In order to be considered a tying arrangement, bundled discounts would have to coerce buyers to forego their preferred buying patterns.[20] If such disguised tying occurs, it is surely in a small percentage of all bundled discounting cases. Bundled discounting is pervasive across competitive markets where market power is not conceivably present and where the practice therefore cannot be coercive. Further, bundled discounts that represent the transmission of savings from economies of scale or scope—which is often the case—are not coercive even in imperfectly competitive markets. Elhauge would permit a cost-justification defense akin to that allowed for commodity price discrimination under the Robinson-Patman Act.[21]

While such justifications should clearly be allowed if bundled discounts are held to be potentially anticompetitive, the post-Chicago School's focus on the buyer's motivations and justifications for bundled discounting schemes often misses the mark. For, in many cases, the buyer rather than the seller initiates the bundled discount scheme, or the buyer and the seller are equally in favor of the contract's bundled pricing structure.

Why would a buyer enter into a contract that made its favorable pricing options contingent upon minimum purchase volumes across multiple product lines? The answer is that the buyer may leverage its buying power across multiple product lines in order to obtain more favorable pricing. Hence, contrary to post-Chicago assertions that bundled discounting is not a true price reduction,

buyer-initiated bundled discounting is often an essential feature in a buyer's strategy to lower procurement costs.

The formal analysis of buyer-initiated bundled discounting follows Klein & Murphy's analysis of retailer-initiated exclusive shelf-space contracts.[22] In Klein & Murphy's model, firms compete for preferred distribution from retailers relative to rival products. The preferred method of distribution often involves retailer exclusivity- or partial exclusivity-commitments in exchange for compensation from manufacturers, such as wholesale price discounts, slotting fees, or even cash payments.[23]

There are two fundamental economic questions raised by this form of competition. The first is whether there is a pro-competitive explanation for the purchase of preferred distribution or exclusivity.[24] The second is whether payment in the form of a discount, in this case a bundled discount, is efficient.[25] We focus on the first question here.[26] The retailers are able to obtain lower prices (or the equivalent) from manufacturers because, by committing that all of their customers will purchase a single brand within the relevant product category (spices, for example), the retailer elasticizes the demand facing the manufacturer.[27] The retailer essentially acts as its customers' bargaining agent, committing the customers to buy in a block instead of picking based on brand preference at the point of sale.[28] Customers lose variety but obtain lower prices.[29]

A similar analysis applies to buyer offers to purchase minimum volumes of a product from a diversified seller across the seller's various product lines. In a cost-free world, the buyer would prefer to pick and choose its brands on a product-by-product basis. However, the buyer might also prefer a price reduction to the option to maintain brand variety. By combining multiple products into a single package purchase, the buyer can credibly signal to the seller that it is foregoing its product variety preferences in exchange for a lower price. By jettisoning its individual variety preferences (or, to disaggregate the buyer, the variety preferences of the purchasing of separate product purchasers within a large organization), the buyer effectively elasticizes the demand facing the seller and can thereby drive the price lower.

Unlike Elhauge's model of threatened supra-monopoly prices, there are abundant real-world examples of buyers pursuing bundled discounting schemes.

Unlike Elhauge's model of threatened supra-monopoly prices, there are abundant real-world examples of buyers pursuing bundled discounting schemes. Consider, for example, the federal government's procurement guidelines on bundling. The guidelines contemplate that federal government buyers may consider making solicitations for bundled contracts in order to lower the price of the acquired goods or services.[30] The guidelines recognize that bundling may have adverse effects on small businesses and therefore requires a finding that the bundling would have "measurably substantial benefits."[31] These include "cost savings or price reductions," "quality improvements that will save time or improve or enhance performance or effi-

ciency," "reduction in acquisition cycle times," "better terms and conditions," and "any other benefits."[32] These "measurably substantial benefits" must generally equal 10 percent of the estimated contract value for contracts worth $75 million or less and at least 5 percent or $7.5 million (whichever is greater) for contracts worth more than $75 million.[33] In sum, the federal government's procurement guidelines call for federal buyers to solicit substantial discounts for entering into bundled contracts.

Similarly, medical supply Group Purchasing Organizations ("GPOs") and Pharmacy Benefit Managers ("PBMs") employ bundled discount strategies to drive prices lower on behalf of their constituencies (usually hospitals and insurance companies).[34] Elhauge's argument that "[b]uyers face a collective action problem that requires a collective action solution through antitrust law"[35] misses the point that GPOs, PBMs, and other buyer cooperatives that strategically employ bundled discounts are organized precisely in order to solve a collective action problem. By collectively committing to trade variety for lower prices, the purchasing organization prevents the seller from exploiting the individual members' variety preferences to obtain higher prices.

While the previous examples have generally focused on power buyers, a significant implication of Klein & Murphy's model is that the buyer's ability to elasticize the demand facing the seller and hence obtain lower prices does not depend on the buyer having monopsonistic power.[36] Hence, even relatively powerless buyers facing relatively powerful sellers may have the ability to bargain for lower prices by committing to purchasing multiple products. Far from being a seller-side power tool, bundled discounting may be a buyer-side power tool.

We do not claim that customer-initiated bundled discount schemes are uniformly beneficial to end consumers. Customer-initiated exclusive dealing may be of greater concern when the customer resells the product downstream and is thus capable of passing on any overcharge imposed by the seller.[37] Some intermediate buyers may tolerate bundled discounts that increase their own profitability even if the long-run effects of such discounts are to exclude competitors and thereby increase prices to end consumers. But that only means that the proper focus on bundled discount law should remain exclusion of rivals. While bundled discounts are not often exclusionary,[38] the possibility that they are disguised predatory discounts—not that they are disguised price increases—should be the focus of the antitrust inquiry.

IV. Conclusion

Professor Elhauge has written a thoughtful and important article that challenges the consensus that seemed to be emerging around a discount attribution test for bundled discounts. Nonetheless, his creative arguments rest on flawed assumptions. Bundled discounts generally benefit consumers and only harm them in the

narrow set of circumstances where they exclude rivals. "Power effects" should not be a concern of bundled discounting law. In future work, we will address other aspects of his paper.

---

1 Whether those rivals should be "equally efficient" to the defendant is a topic that we do not address in this symposium essay because it is not necessary to the narrower topic on which we focus.

2 Einer Elhauge, *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit Theory*, Discussion Paper No. 629, forthcoming 123 Harv. L. Rev. (Dec. 2009), *available at* http://www.law.harvard.edu/programs/olin_center/.

3 *See generally* Bruce H. Kobayashi, *Does Economics Provide A Reliable Guide to Regulating Commodity Bundling By Firms? A Survey of the Economic Literature*, 1(4) J. Competition L. Econ. 707 (2005); David S. Evans & Michael Salinger, *Why Do Firms Bundle and Tie? Evidence from Competitive Markets and Implications for Tying Law*, 22 Yale J. Reg. (2005).

4 *See* Joshua D. Wright, *Missed Opportunities in Independent Ink*, 5 Cato Sup. Ct. Rev. 333, 348-356 (2006) (discussing relationship between price discrimination and welfare measures). We leave for later analysis our objection to Professor Elhauge's claim that antitrust law has committed to a course that would require it to micromanage markets to identify and sanction instances of tying, bundling, and bundled discounts that reduce static consumer welfare. We believe such a policy would be counter-productive for consumers, unadministrable, and run afoul of antitrust law's tolerance of simple monopoly pricing (which obviously reduces static welfare), and would be inconsistent with the Supreme Court's antitrust jurisprudence.

5 The literature includes Philip Areeda & Herbert Hovenkamp, IIIA Antitrust Law ¶ 749 (2008); Daniel A. Crane, *Mixed Bundling, Profit Sacrifice, and Consumer Welfare*, 55 Emory L. J. 423 (2006); Herbert Hovenkamp, *Discounts and Exclusion*, 200 Utah L. Rev. 841; Thomas A. Lambert, *Evaluating Bundled Discounts*, 89 Minn. L Rev. 1688 (2005); Kobayashi, *supra* note 3.

6 LePage's Inc. v. 3M, 324 F.3d 141 (3d Cir. 2003) (en banc).

7 Cascade Health Solutions v. Peachealth, 515 F.3d 883 (2008).

8 Northern Pac. Ry. Co. v. U.S., 356 U.S. 1, (1958) ("Of course where the buyer is free to take either product by itself there is no tying problem even though the seller may also offer the two items as a unit at a single price.").

9 Ortho Diagnostic Sys., Inc. v. Abbott Labs., Inc., 920 F. Supp. 455, 471 (S.D.N.Y. 1996).

10 Phillip E. Areeda, Herbert Hovenkamp, & Einer Elhauge, Antitrust Law: An Analysis of Antitrust Principles and their Application, ¶ 1758b, at 181 (1996).

11 *Id*. at 348.

12 Elhauge, *supra* note 2 at 59, 79.

13 *Id*. at 79.

14 Elhauge, *supra* note 2 at 57.

15 *See generally* Paul Milgrom & John Roberts, *Limit Pricing and Entry Under Incomplete Information: An Equilibrium Analysis*, 50 Econometrica 443 (1982).

16 *See*, *e.g.*, Brooke Group Ltd. v. Brown & Williamson Tobacco Corp., 509 U.S. 209 (1993) (requiring predatory pricing plaintiff to prove that defendant's predation created a dangerous probability that defendant would be able to recoup the costs of predation at a later time).

17 *See generally* Patrick Bolton, Joseph F. Brodley, & Michael H. Riordan, *Predatory Pricing: Strategic Theory and Legal Policy*, 88 Geo. L. J. 2239, 2301-03 (2000).

18 Frank H. Easterbrook, *Predatory Strategies and Counterstrategies*, 48 U. Chi. L. Rev. 264,282-88 (1981).

19 *See* Daniel A. Crane, *Chicago, Post-Chicago and Neo-Chicago*, 75 U. Chi. L. Rev. ___ (2009) (forthcoming); Joshua D. Wright, *Overshot the Mark? A Simple Explanation of the Chicago School's Influence on Antitrust*, 5(1) Competition Pol'y Int'l (2009).

20 *See* Jefferson Parish Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2, 12 (1984) ("Our cases have concluded that the essential characteristic of an invalid tying arrangement lies in the seller's exploitation of its control over the tying product to force the buyer into the purchase of a tied product that the buyer either did not want at all, or might have preferred to purchase elsewhere on different terms.")

21 Elhauge, *supra* note 2 at 79-80.

22 Benjamin Klein & Kevin M. Murphy, *Exclusive Dealing Intensifies Competition for Distribution*, 25 Antitrust L. J. 433 (2008).

23 *Id*. at 433-34. *See also* Benjamin Klein & Joshua D. Wright, *The Economics of Slotting Contracts*, 50 J. L. & Econ. 421 (2005).

24 Klein & Wright, *Id*.

25 Benjamin Klein, *Bundled Discounts as Competition for Distribution*, Global Competition Pol'y (June 2008). *See also* Joshua D. Wright, *Antitrust Law and Competition for Distribution*, 23 Yale J. Reg. 169 (2006) (analyzing bundled discounts as a form of competition for distribution).

26 One reason it may be efficient to purchase exclusive or preferred distribution in the form of a discount on the "monopoly" product is that the seller earns a higher margin on that product than the competitive product. Thus, preferred distribution for the competitive product can be purchased more efficiently by discounting the monopoly product, increasing the manufacturer's profit, and improving consumer welfare. *See* Barry Nalebuff, *Bundling as an Entry Barrier*, 119 Q.J. Econ. 159 (2004).

27 *Id*. at 444.

28 *Id.*

29 *Id.*

30 General Services Administration Acquisition Manual Appendix 519(f) (2004), *available at* http://www.acqnet.gov/GSAM/current/pdf/GSAM.pdf. ("Bundling means consolidating 2 or more procurement requirements for goods or services previously provided or performed under separate contracts into a solicitation of offers for a single contract . . .").

31 *Id.*

32 *Id*.

33 *Id*.

34 *See*, *e.g.*, Herbert Hovenkamp, *Discounts and Exclusion*, 2006 Utah L. Rev. 841, 859-60..

35 Elhauge, *supra* note 2 at 66.

36 Klein & Murphy, *supra* note 22 at 449.

37 *See* Richard M. Steuer, *Customer-Instigated Exclusive Dealing*, 68 Antitrust L. J. 239, 242 (1997). *See also* Robert G. Harris & Lawrence A. Sullivan, *Passing on the Monopoly Overcharge: A Comprehensive Policy Analysis*, 128 U. Pa. L. Rev. 268, 275 (1979) (noting that "in a multiple-level chain of distribution, passing on monopoly overcharges is not the exception: it is the rule").

38 *See* Easterbrook *supra* note 18 at 273 (observing that "as long as victims and customers have rational expectations about the future conduct of the predators, and the predators themselves behave rationally, the intended victim should always be able to offer some package that is more attractive to customers than the monopolist's offer of low prices followed by monopoly prices.").

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Price Discrimination and Welfare

A Comment on *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit* by Einer Elhauge

Barry Nalebuff

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Price Discrimination and Welfare

*Barry Nalebuff**

I. Introduction

Elhauge (2009)[1] provides a wide-ranging article that is impressive both in its clarity and its holistic attack on the practice of bundling and tying. In this commentary, I will focus my attention on one aspect of his presentation, namely the effect of price discrimination via metering and tying on consumer welfare and total welfare.

Elhauge makes the claim that we should not suppose that the total welfare effects of price discrimination are positive. Even if they are, he suggests that this perspective is too narrow; a price-discriminating monopolist will make more money and so may incur greater ex ante costs to secure its market position. And if total welfare still rises after taking these costs into account, Elhauge makes the further argument that antitrust is and should be focused on consumer welfare, not total welfare. In that domain, the presumption should be that price discrimination lowers consumer welfare.

The first claim that (what Elhuage calls ex post) total welfare goes down may be surprising since it runs counter to the intuition that comes from first-degree or perfect price discrimination. Perfect price discrimination is typically thought to achieve the efficient outcome and therefore it raises total welfare. As I discuss below, I think that perspective is too simplistic, as it ignores the real costs associated with implementing a price discrimination system. But, putting that issue aside, it is easy to see why there is a presumption that imperfect price discrimination moves total welfare in the same direction as perfect price discrimination. Elhauge argues that this intuition is unfounded.

*Barry Nalebuff is the Milton Steinbach Professor at Yale School of Management.

This comment provides some models and examples that challenge Elhauge's argument for the case of metering and tying. My primary focus is on the given market structure and thus I do not consider how price discrimination may change the ex ante competition.[2] Using his framework, I redo one of his models using continuous rather than discrete variables. This simplifies the mathematics and allows me to provide conditions under which price discrimination via tying will raise total welfare. I show that total welfare rises in a more general version of his model. I also show that a small amount of price discrimination generally increases total welfare in a model with linear demand. While it is certainly possible for price discrimination via tying to lower total welfare, the results here suggest why this might be the exception to the rule for the case of tying.

WHILE IT IS CERTAINLY POSSIBLE FOR PRICE DISCRIMINATION VIA TYING TO LOWER TOTAL WELFARE, THE RESULTS HERE SUGGEST WHY THIS MIGHT BE THE EXCEPTION TO THE RULE FOR THE CASE OF TYING.

Turning to consumer welfare, here there is more support for Elhauge's presumption that price discrimination is harmful to consumers. Consumer welfare falls in the first set of models and the effect is ambiguous in the second set.

It helps to distinguish between the types of imperfect price discrimination, namely second- and third-degree price discrimination. Under third degree price discrimination, a monopolist can charge a different price to different groups based on some exogenous identifying feature. Thus a firm might offer a lower price to senior citizens (or a higher price to non-senior citizens) that reflects different price elasticities. An early example of such price discrimination (discussed by Pigou[3]) is that the English railroads charged a higher price for transporting copper compared to coal. The service provided by the railroad was the same, but copper had a higher value and thus the producers could be charged more. The cost of this type of price discrimination is that it may lead the monopolist to price some of the highest-value customers out of the market. In the case of a linear demand, only half the highest-value customers will be served. This inefficiency lowers consumer welfare and might lower total welfare.

My first result presents a model in which the total welfare effect of third-degree price discrimination is unambiguously positive. The model is a variation of the model presented by Elhauge, the primary difference being that consumer types are continuous rather than discrete and have no minimum demand. This ends up making an important difference as it implies that absent price discrimination, the monopolist will choose to exclude some consumers from the market. Once the monopolist charges a high enough price to exclude some customers, raising the price further becomes even more attractive as there are no more of these customers to be lost. It is this positive feedback that leads to lower total welfare under the single monopoly price.[4] One surprising result is that total welfare rises under price discrimination even though output (as measured by the tied or metered good) falls.

Under second-degree price discrimination, the firm can't target customers by their type, only by their actions. Thus a monopolist might offer a quantity discount, a Saturday-night stayover discount, or a tied-in sale (such as overpriced ink) to capture more of the surplus from high-value customers and expand the market to low-value customers.[5] While the monopolist charges more to high-value customers, the higher price isn't so much more that they end up being excluded from the market. Because high-value customers can always act like low-value customers, they have the option of taking the same price/quantity choice as the low-value customers. Were they to do so, they would get higher surplus than the low-value types and hence the high-value customers are more likely to participate in the market. While second-degree price discrimination will generally keep the highest-value customers in the market, their demand will be curtailed and so there will be some inefficiency.

One surprising result is that total welfare rises under price discrimination even though output (as measured by the tied or metered good) falls.

The main result I show is that a small amount of metering will lead to an increase in total welfare under a reasonably general set of conditions. The result is general in that it doesn't depend on the distribution of consumers in the market. It is limited in two important regards. First, the result depends on a linear demand specification. Second, the result is only for a small amount of metering (or tied-in sales). A monopolist would generally like to do more than a small amount of price discrimination and the result is silent as to the impact of the profit-maximizing amount of price discrimination.

I have two reasons for looking at this case. Firms may be limited in amount of metering, and thus the impact of small amounts of price discrimination is relevant. Second, the result provides some intuition for why imperfect price discrimination via metering may more generally raise total welfare. We know that price discrimination initially raises welfare and so any subsequent negative impact must be large enough to counter the initial gains.

Why might price discrimination be limited in size? Take the case of ink cartridges. HP can sell them at a price premium, but not without limit. At some point, users can and will figure out how to refill cartridges or buy them on the black market. The ability to enforce a tie is limited.

My starting point is a model in which the welfare effects are unambiguous: Price discrimination increases total welfare and hurts consumers. The weaknesses of this model are clear and it is designed to be a jumping off point for more realistic extensions.

II. Baseline Model

Consumers buy a base good for the purpose of using it to make some exogenous output amount n, where n varies by consumer. The value of each unit of output is v, where v is equal across all consumers. The base good and its output are both produced at zero cost.

Absent price discrimination, the firm has to charge a single price p for the base good. Consumers buy the good if

$$n \geq p/v.$$

Let $F()$ represent the cumulative distribution of n. In the case where F is uniform on [0, 1],

$$\Pi = p * [1 - \frac{p}{v}].$$

The profit-maximizing price is $p=v/2$ and only half the consumers (those with $n \geq 1/2$) purchase the good. More generally, the first-order condition is always positive at $p=0$. Thus some consumers will be excluded from the market and the monopoly price is inefficient.

In contrast, if the monopolist is able to meter or engage in a tied-in sale, the monopolist will set a price of v per unit of output and provide the base unit for free to all customers. The net result is perfect price discrimination. Consumers end up with zero surplus, and the result maximizes total welfare.

This model is a special case in that price discrimination is usually imperfect and thus unable to achieve a fully efficient outcome. Even though it is a special case, this model provides some intuition for the claim that second-degree price discrimination (and metering in particular) will typically increase total welfare and decrease consumer welfare. Firms might not be able to achieve the perfect result, but as long as they get close, the results will be directionally the same.

Elhauge argues that this model is such a special case that any intuition drawn would be misleading. In particular, we should not use this model to draw the presumption that price discrimination increases total welfare. Clearly the model is a special case. A first question to consider is whether it does a good job representing *any* real-world model of preferences. The surprise is that this obviously stylized model does a good job describing the facts in *Independent Ink v. Trident*.[6]

THE SURPRISE IS THAT THIS OBVIOUSLY STYLIZED MODEL DOES A GOOD JOB DESCRIBING THE FACTS IN *INDEPENDENT INK V. TRIDENT*.

Trident manufactures a proprietary printer head that is typically used for high-speed printing. Their printer head might be used to date-stamp boxes along a production line—for example, the sell-by date on a carton of beer. The cost of this printing is truly insignificant when measured as part of the total production

cost. A beer manufacturer is unlikely to adjust its price of beer in response to the cost of ink used to print the sell-by date. Thus it is reasonable to assume that the number of boxes stamped (and hence the demand for ink) is exogenous, at least from the perspective of Trident.

How much is the manufacturer willing to pay for the use of Trident's printer head (and contractually provided ink)? The customer compares the unit cost of Trident's product to that of a rival technology. To the extent that the unit cost of Trident's product is v lower than that of a rival, the customer would be willing to pay v per box stamped.[7] Customers that are using similar production technologies would likely have similar cost savings per unit. Thus it seems like a reasonable first approximation to consider the case where the units demanded n varies exogenously across customers, while the value of each unit is equal to v. In such a world, it is not surprising that the manufacturer would seek to engage in price discrimination via metering or a tied-in sale.

There are two quite restrictive assumptions in the baseline model. The first is that the value of each output unit is constant across all consumers. While that may describe some applications, we should consider how the results change when the value of the output varies across the population. Thus in Model I we retain the assumption that the total demand by a customer of type n is exogenous and equal to n. But the value of each of those n units will vary across the population. For example, the production line for beer might move slower than one for soda and so the incremental value of the Trident printer head could be lower.

A second restrictive assumption in the baseline model is that all units have a constant incremental value up to the exogenous demand n. There is no declining marginal utility of consumption. While that may be appropriate to a commercial application like Trident, for many consumer applications we expect the incremental value of consumption to decline. The analysis of the case with declining marginal utility is presented in Model II.

III. Model I

As before, a consumer buys a base good for the purpose of using it to make some output. Thus a printer has no utility of its own other than through making copies, or a car has no utility other than through the miles it is driven. For clarity of exposition, we will continue to use the example of a printer as the base good and copies as the output.

Different customers have different levels of utility for copies. The value of each copy is distributed across the population uniformly over $[0, A]$. At a price of c per copy, customers with per-copy values over $[c, A]$ would want to make copies, but the surplus created may not justify the purchase of the printer. If the printer is

sold at a price of p, then the customer with value v who has demand for n copies will buy the printer provided

$$n(v - c) \geq p.$$

Finally, the number of copies demanded, n, is distributed uniformly over $[0, \bar{N}]$.

Absent metering or tied-in sales, the monopolist is forced to charge the competitive price (here 0) for the copies. We compare this outcome to the scenario where the monopolist is able to charge the consumer a per-copy fee of c. This may be done via direct metering or via the required purchase of a tied-in good such as ink or a cartridge at an above-market price.

This setup is almost identical to the model considered by Elhauge. The primary difference between our approaches is that he considers the case where the number of copies demanded is discrete (n=1, 2, 3, …). Clearly it is the case that copies are not truly divisible. The assumption that the demand for copies is continuous is a convenience that greatly simplifies the mathematics. For the most part, this assumption has little impact on the results except for the case where the market is concentrated on small n. Even here one should think of copies as being measured in units of 2,000, where the copies are sold via proprietary toner cartridges.[8]

Unlike the analysis of Elhauge, in the continuous model we find that the comparison result does *not* depend on the range of copies demanded. Our results are unchanged if n is uniform over [0,1] or [0,2] or [0,100], while his results vary based on whether there are just two groups (who want either 1 or 2 copies) or more groups, say four, who want either 1, 2, 3, or 4 copies.

Theorem 1: For all $\bar{N}$, Total Surplus is higher under price discrimination than under the single monopoly price; the gain is 4.9 percent.

The proof for this result and all following theorems and corollaries are presented in a mathematical appendix.

Corollary: Base-unit demand increases by forty percent under price discrimination, while the total demand for the complementary product decreases by two percent.

At first glance, it might seem peculiar that total welfare goes up even while demand falls, albeit by only 2 percent. Consumers only care about the base unit (printer) for the use of the complementary product (copies). Thus it would seem that the increase in total welfare combined with a decline in relevant output contradicts the results from Varian[9] and Schmalensee[10] that an increase in output is a necessary condition for welfare to rise under 3rd-degree price discrimination.

The explanation is not due to a difference in the type of price discrimination. While the present model is set up to be 2nd-degree price discrimination, this is a special case where the results of the two types of discrimination coincide. Instead

of charging group "n" a price of $A/2$ per copy, the group could be charged a fixed price of $nA/2$. The same set of customers would buy printers and the same number of copies would be made. The explanation is due to the fact that customers are buying different quantities, which takes us outside of the Schmalensee/Varian framework.

Price discrimination expands the set of high-value low-quantity customers. Consider the case where $\bar{N}$=36, which results in a monopoly price p just above 10A. Absent price discrimination, customers with the highest value per copy (A) and a demand for only nine or ten copies are excluded from the market. It is better to include both these two customers and have them get 19 copies together (a welfare addition of 19A) rather than sell thirty copies to the one customer who values each copy at 0.4A each. The latter customer was willing to buy the printer at a price up to 12A and so purchased the printer absent price discrimination. Under price discrimination, her value per copy is below $A/2$ and so she (and her 30 copies) are excluded from the market, for a welfare loss of 12A. The net gain here is 7A, although total copies purchased falls from 30 to 19.

The numbers work out easier with high values of $\bar{N}$, but that is not essential to the argument. Absent price discrimination, consumers are ranked by the product of their value per copy (v) times n, or $v*n$. Total welfare can't go up with the same number of customers buying fewer total copies, as the welfare-per-customer is maximized under one price. Once we allow the set of customers to expand in number, then it is welfare enhancing to include a large number of customers with high valuations per copy and a low demand for copies. Essentially, that is what happens under price discrimination. The customer base expands by 40% and the average value per copy rises. The total number of copies purchased falls slightly under price discrimination, but not enough to offset the gain in average value.

Theorem 2: For all $\bar{N}$, Consumer Surplus is lower under price discrimination than under the single monopoly price; the loss is 18.7 percent.

While the result on consumer surplus is in accord with the finding of Elhauge, the total surplus result is similar, but not identical. Elhauge finds that total surplus is lower under price discrimination for $\bar{N}$ = 2 and 3 and higher for $\bar{N} \geq 4$.

There are two reasons for the differences in total welfare results. The first is that in the continuous distribution of types leads to somewhat different mathematics than the discrete case. The second difference is that in Elhauge's model, the lowest value of n is 1, not 0. Putting these elements together means that with a continuous distribution of n between [0, $\bar{N}$], at any positive price charged by the monopolist there will be some group that is on the margin of being entirely excluded from the market (specifically, the group with $n=p/A$.) In contrast, with a discrete number of groups, the monopolist may not be on the margin of losing a group. Raising the price will reduce the fraction of each group that buys the

printer but may not change whether an entire group is excluded from the market or not.

The potential to exclude a group gives the single-price monopolist a greater incentive to raise price and this further reduces total surplus. The reason is that once the group has been excluded, an additional increase in price loses fewer consumers and this makes price increases more attractive on the margin.

THE POTENTIAL TO EXCLUDE A GROUP GIVES THE SINGLE-PRICE MONOPOLIST A GREATER INCENTIVE TO RAISE PRICE AND THIS FURTHER REDUCES TOTAL SURPLUS.

Consider the solution to Elhauge's model when A=200 and $\bar{N}$ = 4. There are two candidates for an equilibrium. Under the assumption that all four groups are served in the market, the optimal price ($192) does indeed lead to positive demand from all four groups.[11] The maximum willingness to pay per unit is $200, and so even the customer group with demand for a single unit has some positive demand at a price of $192. Consumer welfare is $84,800, profits are $76,800, and total welfare is $161,600.

The issue is that the fixed-point argument is only a necessary and not sufficient condition for profit maximization. There is another fixed-point solution under a different assumption and one that leads to higher profits. Under the assumption that the customers with demand of 1 unit are all excluded, the profit-maximizing price becomes $277, and so the group with n=1 is indeed excluded.[12]

As the monopolist raises its price from $192 to $200, this reduces profits (as $192 is the profit-maximizing price when all four groups are being served). However, once the monopolist raises its price above $200, there are no more customers from group one to be lost. This makes further price increases more profitable and leads the monopolist to go all the way up to $277. At p=$277, profits are higher at $83,077. The resulting consumer welfare is $55,392 and total welfare is $138,468. Total welfare under price discrimination is higher at $150,000, while consumer welfare is lower at $50,000.

Our results are the same once $\bar{N}$ is large. This is not because the continuous case is the same as large $\bar{N}$. Rather, the fact that Elhauge's model uses 1 as the lowest value rather than 0 becomes much less important when $\bar{N}$ is large. Elhauge finds that price discrimination lowers welfare for $\bar{N}$ = 2 or 3.[13] (The case with just one type leads to the same result with or without price discrimination.) The reduction in total welfare is a result of the assumption that the minimum number of copies demanded is 1 unit (where a unit represents a cartridge or 2,000 copies). Because of the minimum, even with a continuous distribution the monopolist finds that no consumer groups are on the margin. For example, with n distributed uniformly on [1, 3], the monopolist would set a price of $A/\ln(3)=0.91A$, a price that leads to strictly positive demand from the n=1 customer group.

The primary reason why price discrimination raises welfare is that it allows half the consumers of each type to be served, including half the low-value types who might be excluded under a one-price monopoly. If there are no low-value types to be excluded, price discrimination will lead to lower welfare. A necessary condition for total welfare to increase is that price discrimination expand the base-unit sales. Thus price discrimination is put at a disadvantage when the minimum type is n=1 rather than n=0. Once $\bar{N}$ becomes large enough (3.51), some groups will be excluded and total welfare soon thereafter is higher under price discrimination.

Theorem 3: Assume that n is distributed uniformly over $[1, \bar{N}]$. For $\bar{N} > 4.58$, total surplus is higher under price discrimination.

Our results above all rely on a uniform distribution on n over $[0, \bar{N}]$ or $[1, \bar{N}]$. It is possible to make some progress with a more general distribution. For any continuous and positive distribution of n over $[0, \bar{N}]$, price discrimination will increase total sales of the base unit. Of course, this is only a necessary and not sufficient condition to improve total welfare.

Theorem 4: Under the assumption that n has continuous support over $[0, \bar{N}]$, the one-price monopolist restricts sales of the base unit relative to the price discrimination case.

From these results so far, I take away the presumption that price discrimination via metering raises total welfare and lowers consumer welfare. It is possible that price discrimination leads to lower total welfare. This typically arises when there is some minimum demand type and the single-price monopolist chooses to serve all consumer groups in the market.

FROM THESE RESULTS SO FAR, I TAKE AWAY THE PRESUMPTION THAT PRICE DISCRIMINATION VIA METERING RAISES TOTAL WELFARE AND LOWERS CONSUMER WELFARE.

Even if total welfare usually rises, the fact that total welfare might fall under price discrimination is icing on the cake for Elhauge. A presumption or even a demonstration that total welfare rises would not create an antitrust defense. Elhauge is quite clear and explicit that antitrust law is based—and should be based—on consumer welfare, not total welfare. Thus his argument against price discrimination, tied sales, and metering does not depend on this leading to a reduction in total welfare. But others may disagree on how the law is and how it should be in terms of evaluating consumer welfare versus total welfare. I will return to this issue in the conclusion.

I turn now to a model in which customers experience a declining marginal value of output. The results from this model provide further support for the presumption that price discrimination via tying and metering raises total welfare. The effect on consumer welfare is ambiguous.

The main result is the following: Given preferences that lead to linear demands with different intercepts, a small amount of price discrimination always

raises total welfare. This is a general result in that it does not depend on the specific distribution of consumer preferences.

While the result does not tell us what happens when the amount of price discrimination is chosen to maximize monopoly profits, it helps us appreciate that price discrimination at least starts off on the right foot in terms of increasing total welfare. The effect of a small amount of price discrimination is more relevant than it might at first appear because firms may be limited in how much metering or tying they can practically accomplish.

The case of ink or toner cartridges is again instructive. HP might like to completely restrict others from selling toner cartridges that are compatible with their laser printers. They have designed cartridges with a patented shape that blocks entry into the new cartridge market. They have added a chip to some cartridges that detects when the ink is empty and prevents the printer from working even if the once empty cartridge has been refilled. However, if the price of the cartridge gets too high, entrants will find ways to collect and remanufacture the spent cartridges. At some point, the savings become too large and buyers find a way to avoid paying the large mark up.

The case of *Independent Ink v. Trident* also illustrates this issue. According to Independent Ink, Trident was selling refills for a price of $325.[14] This was sufficiently high that buyers were willing to break their contractual agreement with Trident and buy refills from Independent Ink, who could profitably sell refills at $125 to $189. The cost of monitoring and enforcing the contracts with each buyer is substantial. From Trident's perspective it was cheaper to go after the rival manufacturer than each customer.

This cost of monitoring and enforcement may limit how much a monopolist can raise the price of the tied-in product above the competitive level. For that reason, we are interested in the welfare impact of a small amount of price discrimination. The fact that tied-in sales initially improve total welfare may present a defense to those who seek to defend tying on the grounds that it is efficient.

IV. Model II

A consumer buys a base good for the purpose of using it to make some output. Again, for clarity of exposition, we will use the example of a printer as the base good and copies as the output.

Different customers have different levels of utility for copies. Even a customer with a high value of copies will not value all copies equally. For simplicity, we assume that a customer of type a values the q^{th} copy at $a - q$. Here the number of copies is assumed to be a continuous variable. Thus if copies are priced at c, the consumer of type a will demand

$$D(a, p) = a - c$$

copies, if that consumer has bought a copier. Without loss of generality, we assume that the marginal cost of copies is zero and so the price per copy should be interpreted as a markup over cost.[15]

The price per copy is best thought of as a metering device or tied-in sale. For example, if the printer monopolist forces the customer to use its special ink, then we can calculate c as the implied price of the ink per standardized page of printing. While toner cartridges are typically sold in packages of 2,000 copies, over the lifetime of the printer this integer problem should not be an important factor. It might also be possible for the monopolist to meter output directly, either through a counter on the printer or the odometer of a car (as is done with car leases).

The printer is sold for a fixed price, p. Given a base price of p and a per-unit cost of c, a consumer of type a will only buy the printer if her total surplus is weakly positive. The total surplus of consumer type a is $(1/2)(a - c)(a - c) - p$. Thus consumer of type a will make the purchase provided

$$a \geq \sqrt{2p} + c.$$

Total profits for the monopolist consist of the profits from the base sales along with the profits from the tied-in sales. Profits from the sale of each base good are $p - \mu$, where μ is the constant production cost of each base unit. Total profits are thus:

$$\Pi = \int_{a \geq \sqrt{2p} + c} [p - \mu + c(a - c)] f(a)\, da$$

where consumer preferences are distributed according to the atomless density function $f(a)$.

Total surplus is

$$TS = \int_{a \geq \sqrt{2p} + c} [\frac{a^2 - c^2}{2} - \mu] f(a) da.$$

Absent a tied sale, the monopolist is constrained to the competitive price (0) for the tied-in product. Thus the single-price monopolist maximizes profits subject to the constraint that c=0. We denote this profit maximizing price by $p(0)$. More generally, let $p(c)$ be the profit-maximizing base price when the monopolist charges c for the tied good and correspondingly $\Pi(p(c), c)$ are the profits, and $TS(p(c),c)$ is the total surplus.

Theorem 5:

$$p'(c) < -\sqrt{2p}$$

$$d\Pi(p(c), c)/dc > 0$$

$$dTS(p(c), c)/dc > 0$$

The formal proof is in the appendix. It is important to note that the results do not depend on the distribution $f(a)$ other than the requirement that the maximization problem is quasi-concave.

When the monopolist is allowed to engage in a small amount of price discrimination via a tied-in sale, it will strictly want to take advantage of this opportunity: profits go up. Because incremental sales are now more valuable, the monopolist will lower the price of the base good. The base price falls enough so that the net result is a lower price to the marginal consumer, thus expanding total demand and increasing total welfare. (It is also the case that each consumer buys slightly less of the tied-in good, however this has no welfare loss as the incremental unit lost has almost no value.) The total change in consumer surplus is ambiguous. The increased price for the tied-in product is harmful, but the reduction in base price might lead to an overall gain for consumers.

V. Conclusions

Although the models in this paper suggest that price discrimination via metering or tied-in sales will typically increase total welfare, these economic models miss many of the most important costs associated with price discrimination. Price discrimination is not free. Firms spend a large amount to implement the tying practices and consumers spend resources to avoid them. For example, HP spends resources to design a proprietary shape for its toner cartridge and further resources to ensure that spent cartridges cannot be refilled. Trident spends large amounts enforcing its customers' contractual obligations to buy its expensive ink. These costs are inefficiencies that are typically left out of the model.[16]

Even when tying leads to higher total welfare, we should recognize that the primary source of the gain is that the monopolist is being less inefficient.

Further, tying may impose collateral damage costs on the tied-good market. The forced tied sale may make the complementary market less competitive. This could make it more difficult for others to enter the monopolized base-good market.

Even when tying leads to higher total welfare, we should recognize that the primary source of the gain is that the monopolist is being less inefficient. While a gain is a gain is a gain, there is something different about a firm increasing total welfare by inventing a better mousetrap versus being a less inefficient monopolist. We want to reward firms for making better products, and not for becoming better monopolists.

Section II of the Sherman Act prohibits a firm from monopolizing any part of trade or commerce. To the extent that the firm engages in price discrimination, it becomes a more powerful monopoly. Thus even a firm that has earned its

monopoly by being the best product on the market starts to cross the line and monopolize the market when it engages in price discrimination.

A similar perspective is provided in the U.S. DOJ/FTC, Horizontal Merger Guidelines (1992, revised 1997): "The unifying theme of the Guidelines is that mergers should not be permitted to create or enhance market power or to facilitate its exercise." It seems clear that price discrimination is an enhancement of a firm's market power.

While we have been focused on total welfare and consumer welfare, there is no requirement that we pick either as the antitrust standard. These are two extremes. Total welfare equals profits + consumer welfare. More generally, we can look at a weighted sum of profits and consumer welfare. If we place equal weight on profits and consumer welfare, the result is a total welfare standard. If we place zero weight on profits, the result is a consumer welfare standard. In between, there is an infinite range of options. For example, the courts could weight consumer welfare twice as high as profits.

Before we condemn tying, we should reflect on the fact that the same results can be achieved via metering as with tying. Instead of requiring the consumer to use its overpriced ink, the monopolist could simply charge a price per copy. While metering is legal, for a firm with market power, tying is *per se* illegal. Since the effect on consumers and total welfare is the same, there is the question of why the law treats these two cases differently.[17]

One solution would be to harmonize the law to make tying legal or at least subject to a rule-of-reason test. Alternatively, one could make metering (by a firm with market power) subject to antitrust. It could be seen as a violation of the Sherman Act, Section II, as it allows a firm with market power to further monopolize the market.

Elhauge's response (section IV. A) is to emphasize that firms may find it more difficult or expensive to engage in direct metering than a forced tie. He observes that direct metering is uncommon. If it turns out that tying is the easier method of engaging in price discrimination, there is no reason to facilitate this practice.

Of course, there are efficiency reasons to employ metering besides price discrimination, the primary one being risk sharing. A buyer who is unsure if the service will work or the quantity it will require might prefer to pay on a per-unit basis rather than a single upfront price.

THERE ARE AT LEAST TWO ADVANTAGES OF DIRECT METERING OVER TYING.

There are at least two advantages of direct metering over tying. One is that the effect is transparent. A buyer may not appreciate what the true cost is per use when the base good comes with a tied-in sale. For example, many buyers may not know the implied per copy charge that comes with each printer due to the markup on ink

and proprietary toner cartridges. A second concern is that the forced tied sale may make the complementary market less competitive.[18] This could make it more difficult for others to enter the monopolized base-good market.

While I have argued that tying may typically lead to improved social welfare, I want to reiterate that this is not a legitimate justification according to Elhauge (section IV. C).[19] In that light, I hope that this comment will help focus the debate. The reasons to condemn tying are in spite of, not because of, its potential to improve total welfare.

VI. Appendix

Theorem 1: For all $\bar{N}$, Total Surplus is higher under price discrimination than under the single monopoly price; the gain is 4.9 percent.

Proof: Recall that a consumer of type n is interested in n units of output. The value of each unit is constant and distributed in the population uniformly between 0 and A. Thus the total demand from type n customers at price p is $(1/A)(A - p/n)$.

Costs are zero for both the base good and the tied good. Thus at price p, total profits are

$$\Pi = \int_{\frac{p}{A}}^{\bar{N}} \frac{p}{A}[A - \frac{p}{n}]f(n)\,dn.$$

Profits are maximized when

$$\frac{d\Pi}{dp} = \int_{\frac{p}{A}}^{\bar{N}} [1 - \frac{2p}{An}]f(n)\,dn = 0.$$

The uniform distribution of n allows us to replace $f(n)$ by f and then integrate to find a closed-form solution.

$$\frac{d\Pi}{dp} = \{[\bar{N} - \frac{p}{A}] - \frac{2p}{A}[\ln(\bar{N}) - \ln(\frac{p}{A})]\}f = 0.$$

Define $z = \bar{N}A/p$.

$$\frac{d\Pi}{dp} = \frac{p}{A}f\{z - 1 - 2\ln(z)\} = 0 \Rightarrow z \approx 3.51.$$

Thus we have the general solution for p:

$$p \approx \frac{\bar{N}A}{3.51}.$$

At this price, roughly 72 percent of the highest-value customers are served in the market. Overall, only 36 percent of customers are served.

Turning to Total Surplus:

$$TS = \int_{\frac{p}{A}}^{\bar{N}} \frac{n}{A}\frac{A + \frac{p}{n}}{2}(A - \frac{p}{n})f(n)dn$$

$$= \int_{\frac{p}{A}}^{\bar{N}} \frac{n}{2A} [A^2 - (\frac{p}{n})^2] f(n) dn$$

$$= \frac{1}{4} \int_{\frac{p}{A}}^{\bar{N}} [2nA - \frac{2p}{n} \frac{p}{A}] f(n) dn.$$

Using the first-order condition for p, we can substitute A for $2p/n$ in the integral.

$$TS = \frac{1}{4} \int_{\frac{p}{A}}^{\bar{N}} [2nA - p] f(n) dn$$

$$= \frac{1}{4\bar{N}} \{A[\bar{N}^2 - (\frac{p}{A})^2] - p(\bar{N} - \frac{p}{A})\}$$

$$= \frac{\bar{N}A - p}{4}$$

$$p = \frac{\bar{N}A}{3.51} \Rightarrow TS_{NoPD} = \frac{\bar{N}A * (1 - \frac{1}{3.51})}{4}$$

When the monopolist is allowed to engage in price discrimination, it charges a positive price per tied good (cartridge). Since all groups are identical, the monopolist would charge the same price per cartridge to each group. Profits are maximized when the monopolist sells to half the consumers at a price of $A/2$. In that case, the customers who buy have an average valuation of $3A/4$ and buy $\bar{N}/2$ units:

$$TS_{PD} = \frac{1}{2} \frac{3A}{4} \frac{\bar{N}}{2} = \frac{3A\bar{N}}{16} .$$

It then follows that total surplus is almost 5 percent higher under price discrimination:

$$\frac{TS_{PD}}{TS_{NoPD}} = \frac{3/16}{(1 - \frac{1}{3.51})/4} \Rightarrow \frac{3}{(4 - \frac{4}{3.51})} = 1.049.$$

Corollary: Base-unit demand increases by forty percent under price discrimination, while the total demand for the complementary product decreases by two percent.

Proof: Absent price discrimination, demand for base units is

$$D(p) = \frac{1}{2} \int_{\frac{p}{A}}^{\bar{N}} f(n) dn = \frac{\bar{N} - \frac{p}{A}}{2\bar{N}} = 0.358.$$

With price discrimination, demand for base units is 40 percent higher at ½.

Absent price discrimination, the number of copies sold is

$$\int_{\frac{p}{A}}^{\bar{N}} \frac{n}{A} [A - \frac{p}{n}] f(n) dn = \int_{\frac{p}{A}}^{\bar{N}} [n - \frac{p}{A}] f(n) dn$$

$$= \frac{1}{2\bar{N}} [\bar{N} - \frac{p}{A}]^2 = \frac{\bar{N}}{2} [1 - \frac{p}{\bar{N}A}]^2 = \frac{\bar{N}}{2} * 0.511.$$

With price discrimination, half the customers are served. Since the average customer demands $\bar{N}/2$ copies, the number of copies sold is 2.29% lower:

$$\frac{\bar{N}}{2} * \frac{1}{2}.$$

Theorem 2: For all $\bar{N}$, Consumer Surplus is lower under price discrimination than under the single monopoly price; the loss is 18.7 percent.

Proof: Under price discrimination, half the consumers in each group buy and get $3A/4 - A/2 = A/4$ of surplus on $\bar{N}/2$ units. Thus total consumer surplus is

$$CS_{PD} = \frac{1}{2}\frac{A}{4}\frac{\bar{N}}{2} = \frac{A\bar{N}}{16}.$$

Absent price discrimination, consumer surplus is

$$\begin{aligned} CS_{NoPD} &= \int_{\frac{p}{A}}^{\bar{N}} \frac{n}{A}\frac{A-\frac{p}{n}}{2}(A-\frac{p}{n})f(n)dn \\ &= \int_{\frac{p}{A}}^{\bar{N}} \frac{n}{2A}(A-\frac{p}{n})^2]f(n)dn \\ &= \frac{1}{2A}\int_{\frac{p}{A}}^{\bar{N}} (nA^2 - 2pA + \frac{p}{n}p)f(n)dn \\ &= \frac{1}{2}\int_{\frac{p}{A}}^{\bar{N}} (nA - 2p + \frac{p}{n}\frac{p}{A})f(n)dn. \end{aligned}$$

Using the first-order condition for p, we have

$$\begin{aligned} CS_{NoPD} &= \frac{1}{2}\int_{\frac{p}{A}}^{\bar{N}} (nA - \frac{3}{2}p)f(n)dn \\ &= \frac{1}{4\bar{N}}[A(\bar{N}^2 - (\frac{p}{A})^2) - 3p(\bar{N} - \frac{p}{A})] \\ &= \frac{1}{4}[A\bar{N} + 2\frac{p^2}{A\bar{N}} - 3p]. \end{aligned}$$

Given that $p \approx \bar{N}A/3.51$, consumer surplus is cut by 18.7 percent under price discrimination:

$$CS_{NoPD} \approx \frac{A\bar{N}}{4}[1 + \frac{2}{3.51^2} - \frac{3}{3.51}] = \frac{A\bar{N}}{13}$$

$$\frac{CS_{PD}}{CS_{NoPD}} = \frac{\frac{A\bar{N}}{16}}{\frac{A\bar{N}}{13}} = \frac{13}{16} = 1 - 0.187.$$

Theorem 3: Assume that n is distributed uniformly over $[1, \bar{N}]$. For $\bar{N} > 4.58$, total surplus is higher under price discrimination.

For $p \le A$, all types have positive demand, and the profit-maximizing price satisfies:

$$p = \frac{(\bar{N}-1)A}{2\ln(\bar{N})}.$$

It then follows that $p \leq A$ only if $\bar{N} \leq 3.51$. (It can be checked that $p > A$ are not profit maximizing.) For $\bar{N}$ in this range, total surplus will be higher absent price discrimination. This is because half the market will be served in both cases, but the allocation of consumers is more efficient with a single price.

Once $\bar{N} > 3.51$, we are back to our earlier model in which some consumer types will be excluded from the market by the high price. Under one price, the calculation of consumer surplus is exactly as before except that the density is now $1/(\bar{N} - 1)$ rather than $1/\bar{N}$:

$$TS = \frac{\bar{N}}{\bar{N}-1}\,\frac{A\bar{N}-p}{4} = \frac{\bar{N}}{\bar{N}-1}\,A\bar{N}\,\frac{2.51}{14.04}.$$

Turning to the case of price discrimination, the only difference is that the demand for cartridges varies from 1 to $\bar{N}$, rather than 0 to $\bar{N}$. Hence the average demand is $(\bar{N}+1)/2$.

$$TS_{PD} = \frac{3A(\bar{N}+1)}{16}$$

$$\frac{3A(\bar{N}+1)}{16} > \frac{\bar{N}}{\bar{N}-1}\,A\bar{N}\,\frac{2.51}{14.04} \text{ iff}$$

$$1.05(\bar{N}+1) > \frac{\bar{N}^2}{\bar{N}-1} \text{ iff}$$

$$0.05\bar{N}^2 > 1.05 \text{ or } \bar{N} > \sqrt{21} = 4.58.$$

Theorem 4: Under the assumption that n has continuous support over $[0, \bar{N}]$, the one-price monopolist restricts output relative to the price discrimination case.

Proof: Absent price discrimination, demand is

$$D = \int_{\frac{p}{A}}^{\bar{N}} \frac{1}{\bar{A}}\left[A - \frac{p}{n}\right] f(n)\,dn = \int_{\frac{p}{A}}^{\bar{N}} f(n)\,dn + pD'(p)$$

$$= \int_{\frac{p}{A}}^{\bar{N}} f(n)\,dn - D(p) \text{ at the profit-maximizing } p$$

$$= \frac{1}{2}\int_{\frac{p}{A}}^{\bar{N}} f(n)\,dn < \frac{1}{2}$$

where the strict inequality follows from the fact that $p > 0$ and so some consumers will be excluded from the market. The result then follows as the price-discriminating monopolist sells to precisely half of each customer group.

Theorem 5:

$$p'(c) < -\sqrt{2p}$$

$$d\Pi(p(c), c)/dc > 0$$

$$dTS(p(c), c)/dc > 0$$

Proof: The first-order condition that determines the optimal base price is:

$$\frac{d\Pi}{dp} = 1 - F(\sqrt{2p} + c) - \frac{1}{\sqrt{2p}} f(\sqrt{2p} + c)(p - \mu + c\sqrt{2p}) = 0.$$

Note that at $c = 0$, $\frac{d\Pi}{dp} > 0$ at $p = \mu$, so $p(0) > \mu$. This first-order equation implicitly defines $p(c)$:

$$\frac{d^2\Pi}{dp^2}\frac{dp}{dc} + \frac{d^2\Pi}{dpdc} = 0.$$

The two parts of the equation are

$$\frac{d^2\Pi}{dpdc} = -2f(\sqrt{2p} + c) - f'(\sqrt{2p} + c)(\tfrac{p-\mu}{\sqrt{2p}} + c)$$

$$\left.\frac{d^2\Pi}{dpdc}\right|_{c=0} = -2f(\sqrt{2p}) - \tfrac{p-\mu}{\sqrt{2p}} f'(\sqrt{2p})$$

$$\frac{d^2\Pi}{dp^2} = -\frac{3}{2}\frac{1}{\sqrt{2p}} f(\sqrt{2p} + c) - \frac{\mu}{2\sqrt{2}p^{3/2}} f(\sqrt{2p} + c) - \frac{1}{\sqrt{2p}} f'(\sqrt{2p} + c)(\tfrac{p-\mu}{\sqrt{2p}} + c)$$

$$\left.\frac{d^2\Pi}{dp^2}\right|_{c=0} = -\frac{3}{2}\frac{1}{\sqrt{2p}} f(\sqrt{2p}) - \frac{\mu}{2\sqrt{2}p^{3/2}} f(\sqrt{2p}) - (\frac{p-\mu}{2p}) f'(\sqrt{2p}).$$

Thus $p'(c)$ at $c=0$ is

$$\left.\frac{dp}{dc}\right|_{c=0} = \frac{2f(\sqrt{2p}) + \frac{p-\mu}{\sqrt{2p}} f'(\sqrt{2p})}{\frac{-3}{2\sqrt{2p}} f(\sqrt{2p}) - \frac{\mu}{2\sqrt{2}p^{3/2}} f(\sqrt{2p}) - \frac{p-\mu}{2p} f'(\sqrt{2p})}$$

$$= -\sqrt{2p}\left[\frac{4f(\sqrt{2p}) + \frac{\sqrt{2}(p-\mu)}{\sqrt{p}} f'(\sqrt{2p})}{(3 + \frac{\mu}{p}) f(\sqrt{2p}) + \frac{\sqrt{2}(p-\mu)}{\sqrt{p}} f'(\sqrt{2p})}\right]$$

$$= -\sqrt{2p}\left[1 + \frac{f(\sqrt{2p})(1 - \frac{\mu}{p})}{(3 + \frac{\mu}{p}) f(\sqrt{2p}) + \sqrt{2p} f'(\sqrt{2p})}\right] < -\sqrt{2p},$$

as $p(0) > \mu$ and the denominator is positive by the local concavity of the profit function. Turning to the effect on profits:

$$\frac{d\Pi}{dc} = \frac{\partial\Pi}{\partial p}\frac{dp}{dc} + \frac{\partial\Pi}{\partial c} = \frac{\partial\Pi}{\partial c}$$

$$\Pi = \int_{a \geq \sqrt{2p}+c} [p - \mu + c(a - c)] f(a) da$$

$$\left.\frac{d\Pi}{dc}\right|_{c=0} = \int_{a \geq \sqrt{2p}+c} a f(a) da - [p(0) - \mu] f(\sqrt{2p}).$$

Employing the first-order condition that determines the optimal p, we have

$$\left.\frac{d\Pi}{dc}\right|_{c=0} = \int_{a \geq \sqrt{2p}} a f(a) da - (\sqrt{2p})(1 - F(\sqrt{2p}))$$

$$= \int_{a \geq \sqrt{2p}} (a - \sqrt{2p}) f(a) da > 0.$$

Finally, the effect on Total Surplus is positive:

$$TS = \int_{a \geq \sqrt{2p}+c} [(a - c) \frac{a + c}{2} - \mu] f(a) da$$

$$= \int_{a \geq \sqrt{2p}+c} [\frac{a^2 - c^2}{2} - \mu] f(a) da$$

$$\frac{dTS}{dc} = \frac{\partial TS}{\partial p} \frac{dp}{dc} + \frac{\partial TS}{\partial c}$$

$$= -c[1 - F(\sqrt{2p} + c] - \frac{(\sqrt{2p} + c)^2 - c^2 - 2\mu}{2} f(\sqrt{2p} + c)\left[1 + \frac{1}{\sqrt{2p}} \frac{dp}{dc}\right]$$

$$\left.\frac{dTS}{dc}\right|_{c=0} = -(p - \mu) f(\sqrt{2p})[1 + \frac{1}{\sqrt{2p}} \frac{dp}{dc}] > 0$$

as

$$\frac{dp}{dc} < -\sqrt{2p}. \blacktriangleleft$$

---

1 Einer Elhauge, *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit Theory*, 123 HARV. L. REV. (forthcoming Dec. 2009).

2 Going forward, for simplicity I will simply refer to social welfare, which is what Elhauge considers to be ex post social welfare.

3 A.C. PIGOU, THE ECONOMICS OF WELFARE (1920).

4 The model leads to an unusual use of third-degree price discrimination, namely one that comes out exactly the same as metering or second-degree price discrimination. In the model, the price-discriminating monopolist is able to charge a different price to each consumer group based on the group's exogenous demand for the tied good (copies). Normally, under third-degree price discrimination, a consumer type n is charged p_n. Here, all consumer groups are identical other than in their demand for copies and the optimal price-discrimination charge increases linearly with the type. Thus a consumer group who wants n copies is optimally charged $nA/2$ for a printer. This is 3rd-degree price discrimination as the price varies based on the group type where type n is taken to be exogenous and observable. In theory, group n could make any number of copies for the fixed price of $nA/2$, but they are only interested in n copies. Charging $A/2$ per copy leads to the exactly same result when the demand for copies is exogenous (as the customer of type n ends up buying n copies and paying $nA/2$), even if unobservable. In this case, the monopolist doesn't have to know which consumer is which in setting

the price. Since the demand for copies is exogenous and not sensitive to the price (assuming that the customer finds it worthwhile to buy a printer), there is no issue of self-selection. Thus the monopolist can achieve the third-degree price discrimination result using a linear two-part tariff or metering. This would not be true if the distribution of valuations were different according to the number of copies desired. It would also not be true of the customer type who wanted n copies had some declining marginal valuation and would consider buying fewer than *n* depending on the pricing.

5 Although it may be a dispute over semantics, Elhauge is incorrect when he seeks to classify metering ties as third-degree price discrimination (in footnote 82). As discussed in the previous note, his model is the exceptional case where the two approaches lead to the same result. Elhauge writes "The categorization is a bit ambiguous because while tying does present all buyers with the same price schedule (like second-degree price discrimination), it also effectively charges buyers higher prices for the tying product if they likely value it more (like third-degree price discrimination)." The fact that high-value buyers pay more has nothing to do with the classification type. Second-degree price discrimination arises whenever a monopolist can't tell one consumer type from another and is thus limited to a price schedule that induces self-selection. If all consumers are charged the same price per copy, this is 2nd-degree price discrimination. If consumers are charged different prices for the printer based on their observable and exogenous type (e.g., business or residential), this would be third-degree price discrimination.

6 Illinois Tool Works Inc. v. Independent Ink, Inc., 547 U.S. 28.

7 If, for example, Trident's printer is faster or more reliable, this would lead to a production costs savings of some amount *v* per box.

8 Customer demand can be continuous even if the units sold are discrete. Thus a customer might ideally want to purchase 2,200 copies or 1.1 toner cartridges. With discrete units,, the customer will be forced to choose between buying one or two cartridges, and therefore cutting back his printing to 2,000 or expanding it to 4,000. Neither I nor Elhauge consider such a model.

9 Hal R. Varian, *Price Discrimination and Social Welfare*, 75 Amer. Econ. Rev. 4, 870-75 (1985).

10 Richard Schmalensee, *Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination*, 71 Amer. Econ. Rev. 1, 242-247 (1981).

11 The optimal price under the assumption that consumers with demands 1, 2, 3, and 4 are served is $4A/[2*(1+1/2+1/3+1/4)] = 24A/25 < A$, so all groups are served. In the numerical example, A=200 and so p=192.

12 The optimal price under the assumption that consumers with demands 2, 3, and 4 are served is $3A/[2*(1/2+1/3+1/4)]=18A/13$. Here $A < 18A/13 < 2A$ which implies that groups 2, 3, and 4 are all partially served, while the first group is excluded. In the numerical example, A=200 and so p=276.9.

13 The Elhauge case with $\bar{N}$=2 or 3 is similar to earlier results from Robinson, *see* Joan Robinson, The Economics of Imperfect Competition (1933) and Richard Schmalensee, *Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination*, Am. Econ. Rev. 71 (1981): 242-247. In these models, demand from each customer group is linear. If all consumer groups are served, the monopolist sells to half the customers in the market, the same fraction as under price discrimination. Since the allocation is better under a single price, price discrimination is inefficient.

14 *See* Barry Nalebuff, Supreme Court Amici Curiae in support of Independent Ink, Inc. in Illinois Tool Works, Inc. v. Independent Ink, Inc., 2005 (with Ian Ayres and Lawrence Sullivan).

15 If there is a marginal cost, then change a to a' where the a' is reduced by the marginal cost. Since the result does not depend on the distribution of a, replacing a by a' has no effect.

16 These costs are not limited to tying. Think of all the resources that airlines employ to price discrimination against business customers; these include advance booking discounts, Saturday-night stayover requirements, and frequent-flyer rewards. Then think of all the strategies that business consumers use to avoid being subject to price discrimination, including back-to-back ticketing, phantom returns, and staying over the weekend. For more on this subject, *see* Barry Nalebuff, *Exclusionary Bundling*. 50 Antitrust Bull. 3, 321-370 (2005).

17 In some cases with commercial customers, direct metering may lead to a violation of the Robinson Patman Act.

18 The two points are discussed further in Barry Nalebuff, *Unfit to Be Tied: An Analysis of Trident v. Independent Ink (2006)*, in The Antitrust Revolution, 5th edition, 365-88 (J. Kwoka & L.White eds., 2009).

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

The Undead?

A Comment on *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit* by Einer Elhauge

Paul Seabright

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

The Undead? A Comment on Professor Elhauge's Paper

*Paul Seabright**

I. Introduction

Professor Einer Elhauge has written a paper whose title (*Tying, Bundled Discounts, and the Death of the Single Monopoly Profit Theory*[1]) announces its large ambition—to drive a stake through the heart of the Chicago School's Single Monopoly Profit theory. Perhaps I watch too many scary movies, but even after watching his valiant efforts I still sense an uncanny presence, as though the creature will continue to haunt competition policy in spite of his assurances. In this note I want to explain why I think the creature may have more resilience than he has anticipated. Its resilience matters: Professor Elhauge's arguments are used to motivate a vision of the priorities for antitrust enforcement that may be seriously misguided if his optimism is unfounded.

Economic theories are useful ways to think about the world, but only if used in conjunction with empirical evidence. A theory is just a way to organize the evidence we have: It tells us that if certain conditions hold then certain other conditions will hold as well, and it is useful only if we have independent evidence that the first set of conditions holds.[2] We have known for some years now that the Single Monopoly Profit theory is not true always and everywhere and that, therefore, tying and bundling could be used anticompetitively.[3] What matters is whether we can identify in practice when such conditions hold, and whether we have evidence that such conditions hold often enough for anticompetitive tying to be considered a frequent occurrence rather than a relatively rare exception to a more general Chicago rule.

It is now well known that tying can enable a firm with market power to practice price discrimination between different kinds of consumers. This may have

*Paul Seabright is Professor of Economics at the Toulouse School of Economics (GREMAQ/IDEI) and Research Fellow of the Centre for Economic Policy Research (CEPR).

positive or negative effects on consumer welfare according to circumstances. It can also, under conditions developed extensively in Professor Elhauge's paper in an example involving printers and scanners, allow such a firm to extract more of the surplus from multi-unit buyers. And it can sometimes be used profitably to extend market power into an adjacent market, either by evicting (or preventing entry by) a rival, or by weakening the rival (for instance, by raising its costs) so that it competes less effectively.

I want to make three main points. First, it is an empirical question whether the conditions under which tying can be anticompetitive are frequent or rare. But Professor Elhauge offers us no empirical evidence, instead relying on his own intuitions about the kinds of circumstances that are likely or not. Second, the example he develops at length to show that tying can extract more surplus from both tying and tied markets is a bizarre one, resting on a type of tying that is extremely rare and of doubtful feasibility; his argument is not generalizable to more normal cases. Third, he has failed to take account of the ubiquity of assembly operations in a modern industrial economy, a very large number of which are entirely harmless although his diagnostic tools would consider them presumptively suspect. Overall, the implication he draws that "Even without a substantial foreclosure share, tying by a firm with market power generally increases monopoly profits and harms consumer and total welfare, absent offsetting efficiencies" is both unjustified as science and impractical as policy.

First, it is an empirical question whether the conditions under which tying can be anticompetitive are frequent or rare. But Professor Elhauge offers us no empirical evidence, instead relying on his own intuitions about the kinds of circumstances that are likely or not.

II. The Need for Empirical Evidence

For an argument constructed largely from theoretical examples, Professor Elhauge's paper contains a large number of words such as "likely" (56 instances), "probably" (7 instances), "generally" (45 instances), "often" (21 instances) and "usually" (22 instances). There is even one charming instance where he writes that a particular condition is "probably usually" met.[4] These are used to buttress a large number of empirical assertions, many of them highly controversial. Yet I have been unable to find in the paper a single instance of the use of these terms which is supported by a careful empirical study. Perhaps the most striking case concerns the welfare implications of price discrimination, which are well known to be ambiguous.[5]

These welfare implications are ambiguous for two main reasons. First, compared to uniform pricing by a firm with market power, a price discriminating firm will charge higher prices to some buyers and lower prices to others. The consumer welfare effect will require balancing the harm to the first group against the

benefit to the second group. Second, price discrimination often increases profits (which is why firms do it), and it may be that these profits can offset some degree of net harm to consumers, even if profits carry lower weight than consumer surplus. The cases in which it does not increase profits involve either the intensification of competition by discrimination (to consumers' benefit) or a monopolist's commitment problems over time (where price discrimination likewise benefits consumers to the monopolist's detriment).

I am not aware of any empirical study that has tried to investigate whether, in a modern economy overall, the conditions under which price discrimination increases welfare are more likely than those under which they reduce it. Professor Elhauge does not cite any. This does not, however, deter him from claiming the support of the economic literature for the conclusion that "imperfect price discrimination likely decreases consumer welfare."[6] This is a travesty of what the literature says: It has shown conditions under which imperfect price discrimination lowers consumer and total welfare, and it is Professor Elhauge's own assertion—based on generalization from particular examples with such simplifications as linear demand schedules[7]—that these conditions are "likely" to hold in tying cases.[8]

I AM NOT AWARE OF ANY EMPIRICAL STUDY THAT HAS TRIED TO INVESTIGATE WHETHER, IN A MODERN ECONOMY OVERALL, THE CONDITIONS UNDER WHICH PRICE DISCRIMINATION INCREASES WELFARE ARE MORE LIKELY THAN THOSE UNDER WHICH THEY REDUCE IT.

Not knowing of empirical studies to the contrary either, I cannot know whether Professor Elhauge's intuitions are reliable. But neither can he. And I can suggest some reasons why we would be unwise to trust his intuitions as the last word on the matter. First of all, it is easy to think of common cases where price discrimination is likely to enable groups of consumers to be served who might otherwise be served little or not at all. These cases will increase overall consumer welfare since they benefit these groups while leaving more or less unaltered the conditions under which the rest of the market is served. End-of-season clothing sales, cheap train tickets for seniors, educational discounts on software for students, sales of low-priced pharmaceuticals to developing countries, pre-paid mobile phone tariffs, children's prices in restaurants and movie theatres; the list is long (and most of these cases are popular even among people who think that, in the abstract, price discrimination is a bad thing). Price discrimination of this kind almost certainly enhances consumer welfare: If pharmaceutical companies had to charge identical prices in the United States to those they charge in Bangladesh, who can doubt that they would simply withdraw from the Bangladesh market, with no beneficial impact on their pricing in the United States to compensate? I do not know whether the kind of price discrimination made possible by tying is more like these cases or more like the welfare-reducing cases, but I am sure that argument by analogy with textbook examples using linear demand is not the way to settle the question.

A second reason for caution is that Professor Elhauge claims that producer surplus should essentially be given zero weight in social welfare, even though most of the arguments he gives for this conclusion (such as the higher average income of shareholders when compared to consumers[9]) imply that they should be given a lower weight but still one greater than zero. He asserts—again without any empirical backing—that "any additional monopoly profits reaped by tying will be dissipated by the costs of competing to obtain market power."[10] That there is some such dissipation is not seriously disputed by economists, but there are also beneficial effects on innovation of competition to obtain market power, as is recognized in the patent system. It is an empirical question what the net impact of these countervailing forces will be. There is a large literature trying to measure such effects, with far from conclusive findings (though several scholars have found "U-shaped" results, with some degree of market power being more beneficial to innovation and growth than either complete monopoly or a high degree of competition).[11] There are also harmful effects of monopoly other than those Professor Elhauge mentions, such as the dissipation of monopoly rents through high production costs.[12] But their overall impact on the social value of producer profits remains an empirical question. Professor Elhauge does his readers no service by claiming that his own intuition can be substituted for empirical research.

My unscientific impression is that most economists would consider that a world in which all price discrimination was forbidden would have lower total welfare than a world in which all price discrimination was permitted. Their main ground would probably be that price discrimination of some kind is so pervasive (try thinking of industries where firms never give discounts to loyal customers), and that so many firms have some levels of fixed costs which they need to recover by pricing even a little above marginal cost, that innovation will be increased for given cost to consumers if firms can do this in ways that are responsive to differential price elasticities. My (again unscientific) impression is also that this reasoning is correct. I have less clear intuitions about the effects purely on static consumer welfare, which might, on average, go either way. But I am not interested in persuading readers that my intuitions are more reliable than Professor Elhauge's. Choosing one scholar's intuitions over another's is not the way in which this question should be settled.

Choosing one scholar's intuitions over another's is not the way in which this question should be settled.

III. Printers and Scanners

In pages 8-14 of his paper, Professor Elhauge develops an example of tying which is designed to show "the leveraging of one monopoly profit into two monopoly profits that the single monopoly profit theory said was impossible."[13] His example involves two goods, printers and scanners, demand for which is independent. Buyers (who are identical) also buy multiple units for which their willingness to pay is declining with the number of units bought, so that each buyer in effect has

a downward sloping demand curve. The fact that buyers are identical means that this is not a story about tying facilitating price discrimination, but a case—indeed a challenging one for the Chicago doctrine—in which monopoly rent in one market is independent of monopoly rent in the other. Printers are supplied monopolistically, scanners are supplied competitively. So far, so good.

Now comes the strange part. "The printer monopolist can often extract this individual consumer surplus," writes Professor Elhauge, "by refusing to sell its printers at the monopoly price to buyers unless they also agree to buy *all* their scanner requirements from the printer monopolist."[14] This is a tying requirement such as the world has rarely seen outside of gangster life. A normal tie would say "if you buy a printer you must buy a scanner with it," but would leave you able to buy any subsequent scanners from the competitive supplier. This would leave you still facing the competitive marginal cost for scanners. And your marginal cost for the printer would have been raised by the monopolist's margin on scanners, since every extra printer you buy means you must buy one more scanner at the monopoly price before being free to buy at the competitive price. Thus the tie would lower your marginal willingness to pay for printers exactly as the one monopoly profit theory says it would. Nor does the argument depend on there being one scanner sold per printer: any fixed number of overpriced scanners that must be bought with printers would still raise the implicit marginal cost of printers.

THIS IS A TYING REQUIREMENT SUCH AS THE WORLD HAS RARELY SEEN OUTSIDE OF GANGSTER LIFE.

Except where the two goods are technologically complementary, a circumstance I shall consider in a moment, it is hard to see how any tie that forced the buyer of the monopoly good to buy from the monopolist all subsequent supplies of the competitive good could possibly be enforced without illegal coercion. How can the monopolist possibly know whether the buyer has bought more scanners than printers? Even if the monopolist could know, what could stop the buyer of the printer from setting up a separate subsidiary that buys and operates its scanners? It would be like a heart surgeon who is the only one capable of curing your heart condition insisting that you should thereafter never drink in any bar but the one run by his shady brother. Or like Microsoft insisting that when you use its operating system you must also buy from it, not its browser (which is a complementary good) but also all your future supplies of coffee or Scotch whisky at monopoly prices. Many monopolists might dream of such powers but they are unenforceable in fact and in law, and the kinds of tying contracts discussed in competition cases bear no resemblance to them.

There is only one circumstance in which the monopolist can realistically enforce such a tie. That is where the monopoly good is technologically complementary to the competitively supplied good in such a way as to make useless (or more generally to lower the value of) any version of the latter supplied by a competitor. The classic instance is in aftermarkets, such as for replacement parts.

Here the tie may say "if you buy a printer you must buy your cartridges from us." But this is enforceable only to the extent that printer cartridges are technologically complementary to printers, so that using rival cartridges is either impossible or liable to pose a risk to the operation of your printer, either directly or by invalidating its guarantee. And because they are technologically complementary, they will be economically complementary, so their demand will not be independent. Then the tie will lower the willingness to pay for the printer, just as the one monopoly profit theory claims it would.

> THERE IS ONLY ONE CIRCUMSTANCE IN WHICH THE MONOPOLIST CAN REALISTICALLY ENFORCE SUCH A TIE.

To summarize, Professor Elhauge's printers-and-scanners example relies on two conditions—namely independent demand for the two goods, and an enforceable tie obliging the purchaser of the monopoly good to buy all future supplies of the competitive good at monopoly prices—which are inconsistent with each other except in wildly implausible circumstances. The example, ingenious as it is, tells us nothing about the welfare implications of tying in general.

IV. The Ubiquity of Tying in a Modern Economy

Professor Elhauge writes at several points as though tying is an egregious and mostly conspicuous exception to the normal law-abiding behavior of modern firms. He is prepared to allow tying if offsetting efficiencies can be demonstrated, and the fact that this places the burden of proof on the firm suggests he thinks that cases where there are efficiencies are likely to be unusual.[15] He exempts, also under some conditions, products that are used in fixed ratios and lack separate utility, and he appears to consider that this caveat will remove the risk that assembled products might mistakenly be viewed as ties.

However, these two suggestions radically underestimate the extent to which vast numbers of firms in a large range of industries have business models that are built around the assembly for their customers of component products, many of which have separate utility. Newspapers contain bundles of articles, television channels contain bundles of programs, software packages contain bundles of features, travel service packages contain bundles of holiday trips, electronic goods contain bundles of components (such as memory cards in computers, speakers in television sets, and earphones supplied with MP3 players), restaurant menus contain bundles of dishes, prepared meals contain bundles of ingredients, websites contain bundles of contributions, cars contain bundles of features. Guitars typically come supplied with strings and cameras with memory cards, though buyers can, and often do, substitute other versions for the pre-supplied ones. GPS devices come with pre-installed maps and mobile phones with pre-installed ring tones; all of these have separately marketed substitutes. Hotel rooms come equipped with minibars, and hotel bathrooms with shampoo. Lamps come with

bulbs and cars come with car radios. The list is endless. In some cases the market power of the sellers is negligible, but this is far from true for all of them. And even so, what are the implications for firms that acquire market power in industries where bundling is the norm? Should an entire business model become suspect because Professor Elhauge's intuitions tell him that tying is "generally" harmful?

In February 2009 the low cost airline Ryanair caused widespread derision in the press and among customers when it announced that it was considering charging customers for the use of toilets in its aircraft.[16] This service had previously been bundled with the air ticket. Many airlines have significant market power on individual routes: Is public policy seriously to consider obliging them all to follow Ryanair's example on those routes? Professor Elhauge might reply that this is obviously not a serious case, and no antitrust enforcement time or energy would be wasted pursuing cases such as these. Unfortunately, though, reasonable people do not agree on which tying examples are serious and which are not. Some people could not seriously imagine that Microsoft could be reproached for upgrading the features in the browser that is bundled with its operating system, given that rival browsers are downloadable easily for free; others consider this a very serious problem indeed. So long as antitrust doctrine presumptively prohibits, on the part of firms with significant market power, practices that are extremely widespread throughout every part of a sophisticated modern economy, the choice of enforcement priorities will depend on the idiosyncratic perception of any antitrust official with time to spare and a reputation to make. One does not have to be a cheerleader for Chicago School economics (and I am not) to think that is not a desirable direction in which to move antitrust in the 21st century.

THE CHICAGO DOCTRINE OF ONE MONOPOLY PROFIT MAY NOT EXACTLY BE STALKING THE NIGHT LOOKING FOR FRESH BLOOD, BUT FOR THE TIME BEING IT REMAINS DEFIANTLY UNDEAD.

In short, we need a more precise and empirically better grounded view of the circumstances under which tying by firms with market power is harmful to competition than Professor Elhauge's paper has given us. The Chicago doctrine of one monopoly profit may not exactly be stalking the night looking for fresh blood, but for the time being it remains defiantly undead.

---

1 Einer Elhauge, *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit Theory*, HARV. L. REV. 123 (forthcoming Dec. 2009).

2 Sometimes we may have no direct evidence about the first set, but infer indirect evidence from the fact that some of the second set of conditions hold, and use this to make further inferences about the rest of the second set.

3 A useful survey of reasons is in Jean Tirole, *The Analysis of Tying Cases: A Primer*, 1 COMPETITION POL'Y INT'L 1, 1-25 (Spring 2005).

4 Elhauge, *supra* note 1 at 12.

5 *See* Hal Varian, *Price Discrimination and Social Welfare*, 75 Amer. Econ. Rev 4, 870-875 (1995); Mark Armstrong, *Recent Developments in the Economics of Price Discrimination*, Advances in Economics and Econometrics: Theory and Applications, Blundel & Persson, eds, (2006); Mark Armstrong & John Vickers, *Welfare Effects of Price Discrimination by a Regulated Monopolist*, 22 Rand, 4, 571-581 (1991).

6 Elhauge, *supra* note 1, 2. All page references are to this paper unless otherwise specified.

7 Even with linear demand there may be good arguments for allowing price discrimination because of effects on innovation; *see* Theon van Dijk, *Innovation incentives through third-degree price-discrimination in a model of patent breadth*, 47 Econ. Letters, 3-4, 431-435 (1995).

8 Professor Elhauge's precise claim is that "the economic literature proves that price discrimination always decreases total welfare unless it affirmatively increases output" (Elhauge, *supra* note 1 at 34). While correct, this is phrased in such a way as to imply that increasing output is an unusual thing for price discrimination to do. Professor Elhauge provides no empirical arguments to support this view.

9 Elhauge, *supra* note 1 at 41. Although this claim is probably correct, it does not imply that all or even most shareholders are wealthy. Many individuals of modest means are shareholders through retirement plans.

10 *Id*. at 40.

11 *See* Philippe Aghion, Nick Bloom, Richard Blundell, Rachel Griffith, & Peter Howitt, *Competition and Innovation: an inverted-U relationship*, Q. J. Econ. 120, 721-728 (2005); Wendy Carlin, Mark Schaffer & Paul Seabright, *A Minimum of Rivalry: Evidence from Transition Economies on the Importance of Competition for Innovation and Growth*, Berkeley Electronic Press Contributions to Econ. Analysis & Pol'y 3, 1284 (2004).

12 *See* Charles Ng & Paul Seabright, *Competition, Privatisation and Productive Efficiency: Evidence from the Airline Industry*, 111 Econ. J. 473, 591-619 (2001).

13 Elhauge, *supra* note 1, at 11.

14 *Id*. at 9, emphasis added.

15 In a similar vein, Professor Elhauge argues that defendants should be entitled escape a quasi *per se* prohibition by proving that price discrimination increases output. This way of placing the burden of proof implies he thinks output-increasing instances of price discrimination are the exception not the norm. As noted above (note 8), this presumption has not been established by any empirical argument in his paper.

16 *See Pilots Aghast at Ryanair Toilet Charge*, The Times, 27 February 2009, *available at* www.timesonline.co.uk/tol/travel/news/article5815088.ece.

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

The AT&T Case: A Personal View

Thomas E. Kauper

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

The AT&T Case: A Personal View

*Thomas E. Kauper**

The *AT&T* case,[1] asserting that the company had acted in violation of the antitrust laws and seeking its dissolution, was filed under my direction in 1974, and culminated in a consent decree[2] that brought the largest dissolution in American antitrust history. From the outset the case presented a host of institutional, regulatory, procedural, and substantive issues that continue to plague antitrust enforcement agencies, courts, and economic policy makers both here and abroad. It also had the elements of a soap opera, with a degree of suspense, a bit of anger, some embarrassment, a lot of courage, a large cast of characters, intra-agency battling, and leaks to reporters. This brief paper addresses the case in personal terms, with an emphasis on why and how the case was filed, along with an assessment of its consequences, with some history and a few anecdotes thrown in.

*Professor of Law Emeritus, University of Michigan Law School. Professor Kauper served as Assistant Attorney General in charge of the Antitrust Division, United States Department of Justice, from 1972 through 1976. This paper was originally presented as an after dinner talk at the Newport Summit Conference of LECG in June 2008. The paper is a revision and expansion of those remarks. This is not a research paper. It represents, for the most part, the author's own recollections of both events and of conversations with others involved in the case, particularly with William Baxter, who was Assistant Attorney General when the case was ultimately settled. Because these are personal recollections and views, documentations and citations have been kept to a minimum.

The *AT&T* case,[1] asserting that the company had acted in violation of the antitrust laws and seeking its dissolution, was filed under my direction in 1974. It culminated in a consent decree[2] that brought the largest dissolution in American antitrust history. This brief paper addresses the case in personal terms, with an emphasis on why and how the case was filed, along with an assessment of its consequences, with some history and a few anecdotes thrown in.

From the outset the case presented a host of institutional, regulatory, procedural, and substantive issues that continue to plague antitrust enforcement agencies, courts, and economic policy makers both here and abroad. It also had the elements of a soap opera, with a degree of suspense, a bit of anger, some embarrassment, a lot of courage, a large cast of characters, intra-agency battling, the intervention of Watergate, leaks to reporters, shareholder protests, but, I am afraid, with little interest that could be called romantic. It was a case with a long history, a history in a sense going back to 1913, where I will begin in a moment.

But first let me list the several issues I will address. Why was the case filed in the first place, and could it have been filed and won today? Did the case accomplish anything that modern technological development and the market would not have accomplished anyway? Should we simply have substantially deregulated and left it to the market without antitrust intervention at all? In short, was the case pointless? Did the case result in any significant development of Section 2 of the Sherman Act?[3] If not, and I do not think it did, what other lessons can we learn from it?

From the outset, the decision to file the case, and subsequently the entry of the decree, was severely criticized on a number of grounds. The United States had the best telephone system in the world (probably true in 1974) so why mess with it? Shareholders (who seemed to be about half the population of the United States) who relied on AT&T's dividends would be badly hurt (not true as it turned out). Consumers would be confused as to source of service (as they undoubtedly were for awhile) and would not receive the benefits of lower prices. Moreover, many consumers would not want choice—reliance on Ma Bell was easy (this proved to be true for at least a significant number of consumers). Still others expressed the view that the case was nothing more than a power struggle between an entrenched Justice Department bureaucracy and a comparable bureaucracy at AT&T.[4]

The United States had the best telephone system in the world (probably true in 1974) so why mess with it?

Finally, the case and settlement were criticized on the grounds that it was not based upon a single, coherent philosophy or a genuine, reasoned consensus or a farsighted public policy strategy.[5] In one sense, there is merit to this criticism. When we filed the case, we did not have a complete telecommunications plan with defined roles for free markets and economic regulation and a clear sense of

technology as it would evolve. I am confident that Bill Baxter[6] had no comprehensive scheme in mind when he pushed for the settlement.

But the criticism, I believe, misses the point. The antitrust laws in the United States stand on their own. There is no process for bringing antitrust cases into some overreaching public policy making mechanism. The Sherman Act seeks the preservation of markets, absent some clear direction from Congress to the contrary. We did not believe such a determination had been made by Congress. As we viewed it, the case was largely about opening telecommunications markets to the rapid technological change that was occurring. It was our expectation that the market would do the rest.

AS WE VIEWED IT, THE CASE WAS LARGELY ABOUT OPENING TELECOMMUNICATIONS MARKETS TO THE RAPID TECHNOLOGICAL CHANGE THAT WAS OCCURRING.

This, I assume, is the expectation in any government antitrust litigation. Establishment of some amended and newly created regulatory regime would not have been possible then or in the foreseeable future. No one knew where new technology would take us. Indeed, it is not clear that we know yet. But the regulatory regime as it existed in 1974 did not extend to everything in the case, and in any event, as those charged with its administration asserted, it was failing. Although clearly as many uncertainties should be eliminated as possible, antitrust cases rest on the belief that markets work.

To understand the thinking that led to the case, we must go back in time. AT&T was the result of a series of consolidations following the creation by Alexander Bell and others of the Bell Telephone Co. Until 1894, all local exchanges operated under license from Bell. When the basic patent expired in that year, the number of local exchanges expanded dramatically. Meanwhile the beginning of long distance transmissions was underway through AT&T, which initially was a subsidiary of Bell until the ownership structure was reversed about 1900. With the burgeoning of independent local exchanges, the first interconnection issues began to arise as exchanges sought ways to connect to each other. In 1913, the government accepted the basic premise that the phone network could operate most efficiently as a regulated monopoly, and took from AT&T a commitment—the so-called "Kingsbury Commitment"—that it would connect otherwise independent exchanges through its network.[7] This early set of interconnection issues was the reverse of those at issue in the 1974 case, where one major issue was connection of independent long distance providers to local exchanges, virtually all of which were, by the 1950s, controlled by the Bell operating companies.

In 1949, the Justice Department filed an action under the Sherman Act seeking divestiture by AT&T of its manufacturing arm, the Western Electric Company. The case was settled in 1956 with a consent decree prohibiting

AT&T, *inter alia*, from engaging in any line of business that was not part of its regulated telecommunications business or work for the government.[8]

My own thinking about AT&T began with the 1956 decree. The decree was agreed to under somewhat peculiar circumstances in a private meeting between AT&T and the Attorney General (Brownell) at a resort hotel away from Washington.[9] But more importantly, the decree was, in my judgment, profoundly anticompetitive. Prohibition of entry by AT&T into new markets made little sense to me; a feeling that grew as AT&T had to seek approvals for business activities about which the decree raised questions. In essence, the Department was regulating the lines of business available to AT&T. I have had an aversion to regulatory decrees ever since. The most obvious adverse effect was to keep AT&T out of the computer business. Indeed, one may wonder whether the government's ill-fated IBM case would ever have seen the light of day but for the 1956 decree.

The 1956 decree needed to be re-examined. This would require a new investigation into AT&T's conduct with respect to equipment and the relationships among AT&T, its operating companies, and Western Electric. A full investigation would also have to deal with the impact these relationships had on the rapid degree of technological change then taking place, much of it originating from firms outside the AT&T system. Complaints about the inability to connect equipment of outside manufacturers were even made to the Justice Department by Bell operating company officials. The investigation was authorized in 1973.

COMPLAINTS ABOUT THE INABILITY TO CONNECT EQUIPMENT OF OUTSIDE MANUFACTURERS WERE EVEN MADE TO THE JUSTICE DEPARTMENT BY BELL OPERATING COMPANY OFFICIALS.

A second investigation was then already underway, born as a result of AT&T's refusal to interconnect potential rivals in the long distance market, particularly MCI, to the local Bell operating companies. Without such interconnection, MCI—developing long distance capability through microwave transmission—could not reach local telephone subscribers. The role of MCI in the investigation has been disputed. It had taken its grievances to Congress, the Federal Communications Commission ("FCC"), and state regulators, without much success.[10] But the denial of interconnection to MCI and, subsequently, other potential long distance competitors, raised serious antitrust issues. So too did the refusal by AT&T to permit customers to connect their own terminal equipment to AT&T lines.

This issue had been fought before the FCC, leading to the *Carterfone*[11] decision by the Commission, a decision invalidating the AT&T tariff that prohibited so-called foreign attachments. But AT&T continued to resist, and the FCC seemed unable to keep up with each variant AT&T threw up. These interconnection issues were driven by technological change that AT&T had, to this point, managed to keep at bay. Even the FCC conceded that it seemed unable

effectively to regulate AT&T. In conversations FCC commissioners and staff took the position that AT&T was "unregulatable," and that the only people who fully understood AT&T were employed by it.[12]

These two investigations proceeded apace and were ultimately joined into one. The Antitrust Division of the Department of Justice ("DOJ" or "Division") did not have a great deal of economic expertise when the investigation began and was frustrated by difficulties in getting outside consultants because so many economists had ongoing relationships with Bell Labs. The newly created Economic Policy Office, with its coterie of industrial organization economists, was just coming into being. But the expertise was found, and by early fall of 1974 the staff recommended a complaint charging AT&T with violating Section 2 of the Sherman Act. The charges included obstructing sales of telecommunications equipment, particularly switching equipment, to the local Bells; similarly obstructing attachment of customer-owned equipment to the AT&T system; and denying interconnection with the local Bells to potential long distance competitors.

The complaint asked for the dissolution of AT&T, with the separation of Western Electric, the local Bells, and AT&T and its Long Lines Division. Such dissolution, it was believed, would be far more effective than various kinds of regulatory decrees a court might impose. (As it happened, however, even after the break-up, Judge Greene ended up regulating some elements of the former AT&T business.)

As we drew to the close of the investigation, the case became complicated by external events. Information was leaked (a chronic bureaucratic problem), to the point where I received a call from Jack Anderson, Washington's most dreaded columnist, who clearly had in front of him a copy of the staff draft of a memorandum to the Attorney General and a full copy of the draft complaint. My "no comments" seemed to have little impact. We decided to hold the case up for awhile until the smoke cleared. We never did learn the identity of the leaker.

But we were being overtaken by other external events. The Watergate scandal had reached a crisis point. Attorney General Kleindienst was dismissed and replaced by Elliott Richardson. Richardson apparently did inform President Nixon of the ongoing AT&T investigation, which Richardson fully supported, but by early 1974 the White House was in total disarray. It is unlikely that the AT&T investigation was anywhere on the President's radar screen. Indeed, to those of us who had any dealings with the White House, it appeared that there was no effective presidency at all. Then came the Saturday Night Massacre, born of the President's desire to fire the Watergate Special Prosecutor. Elliott Richardson refused the President's request and was fired. The Deputy Attorney General also refused and met the same fate. Robert Bork, who supported the case, became Acting Attorney General and, ultimately and critically for us, Senator William Saxbe was named Attorney General. In the meantime, President Nixon resigned and Gerald Ford assumed the presidency.

Continuity in these circumstances was difficult. We began keeping our briefing material in loose leaf binders. It was hard to know who knew what, or had said what to whom. As a result of Watergate, credibility of Department attorneys was at an all-time low. It was at this point that my recommendation to file the *AT&T* case moved to the Attorney General's office. We advised AT&T that we had recommended a case. At their request, we set up a meeting for November 20, 1974 with the Attorney General to give AT&T counsel the opportunity to present their arguments against the case to him. I assured them that they would be heard before any final decision about filing was made.

Continuity in these circumstances was difficult. We began keeping our briefing material in loose leaf binders. It was hard to know who knew what, or had said what to whom. As a result of Watergate, credibility of Department attorneys was at an all-time low.

One of the lessons I learned that day was that things are not always what they seem, or you would like them to be. Saxbe had been given a lengthy memorandum about the case. He was briefed first thing in the morning, the briefing ending with the statement that the meeting was simply to hear AT&T's arguments, and that I and others on the Division would meet with him subsequently to make a final decision on filing. The meeting began with AT&T's counsel asking Saxbe about his state of mind, so that he could address Saxbe's concerns. Saxbe's answer shocked everyone. He simply said "I intend to file an action against you."

This was not what we anticipated nor what counsel for AT&T expected. I had personally promised them a meeting with the Attorney General *before* any final decision was made, a promise that had now been broken. AT&T's counsel had every reason for anger. Notification was given to the Securities and Exchange Commission, and trading in AT&T stock was suspended. Following a recess and a brief meeting with the Attorney General and those of us from the Antitrust Division involved, the decision was made to file the case early that afternoon. And so the case was filed on an earlier date than originally intended. Attorney General Saxbe left immediately to go hunting. President Ford was traveling in Japan. The process to break up AT&T was formally underway.[13] The case was filed the same afternoon, starting the process that led to the breakup of Ma Bell.

Why, in the end, did we file the case? What did we expect (or hope) to achieve? The obvious answer is that AT&T's conduct was subject to the antitrust laws; that it violated those laws; and that its anticompetitive conduct required the breakup of the company. This is the so-called law enforcement answer, and by that measure the case was a success. But while obvious, the answer is too simple. In the end, the case was about breaking the hold of AT&T on technological development while frustrating others' efforts to enter markets in which AT&T had long been the entrenched incumbent, protected in part by a regula-

tory regime that was, in our minds, irrelevant to some of AT&T's conduct and which, in any event, was failing.

The refusal to provide local exchange interconnections to potential long distance rivals, the frustration of the attachment of user-owned terminal and other equipment to the AT&T system, the pressure on the operating companies to utilize only equipment manufactured by Western Electric, and the cross-subsidization running from regulated markets to unregulated markets (a particular concern of William Baxter, the Assistant Attorney General who ultimately was responsible for the final decree),[14] were all of a piece. All involved what we viewed as artificial barriers to entry and the frustration of technological development. We firmly believed that free markets would do better and would, in the long run, bring greater consumer choice and lower prices. Whether the case succeeded in these respects is a subject to which I will return.

With the filing of the case, it proceeded through discovery and trial before Judge Harold Greene. Between filing and settlement four different Assistant Attorneys General kept the case going, and remained committed to it. Such continuity has been a hallmark of the Antitrust Division's history. While the case ultimately was settled with the far-reaching dissolution decree with which we are all familiar, there were opinions written by Judge Greene dealing with the motions to dismiss filed by AT&T. Relatively early on Judge Greene rejected the defense argument that exclusive jurisdiction over the matters raised in the complaint was in the Federal Communications Commission and that therefore an antitrust court lacked the authority to proceed.[15]

I believed then, and I continue to believe, that this was the central issue in the case, the make or break point. In very broad terms, the motion to dismiss went to whether all the claims raised in our complaint should continue to be handled by a regulatory agency—an agency that had itself recognized its inability effectively to regulate AT&T in the face of fast moving technological change—or whether the antitrust laws should be used to bring about a more market-oriented approach to the future development of the American telecommunications system.

THE INTERFACE BETWEEN FREE- AND REGULATED MARKETS STILL REMAINS A PRIMARY ISSUE IN TELECOMMUNICATIONS EVEN TODAY.

In legal terms the issue was not simple. The interface between free- and regulated markets still remains a primary issue in telecommunications even today. The role of antitrust in these markets today is unclear, particularly given the Supreme Court's predilection, as seen in its *Trinko* decision, in the direction of regulatory controls and away from antitrust, with "its considerable difficulties."[16]

In resolving the exclusive jurisdiction issue, Judge Greene was not asked what public policy should be. Rather, the inquiry was how Congress had resolved these

issues in the Communications Act, where there was no express antitrust immunity provided. His examination of the "relatively weak regulatory controls"[17] that might be applied to AT&T's conduct, as well as the fact that some of the alleged conduct was not subject to regulatory controls at all, led him to conclude that there was no implied repeal of the antitrust laws intended. So the biggest hurdle to the government case had been overcome.

The case moved on to trial of the substantive antitrust issues (where rightly or wrongly the government was convinced its case would easily withstand attack). While Judge Greene resolved the jurisdictional issues in favor of antitrust, one of the lessons learned from *AT&T* is that the case was but one step in what has become a long journey through the regulatory-antitrust interface. The case, and the restructuring it brought about, required policy makers to reconsider the role of direct regulation—indeed it forced such reconsideration—but it was hardly a definitive resolution. Competition in these markets have brought radical changes; changes that, in turn, have required almost continuous re-examination and searches for effective solutions to the new problems dissolution brought—problems Judge Greene could hardly have foreseen.

In any event, disposition of this initial critical motion brought the case to trial. If discovery and the trial teach us anything, it is that judges matter. In the government's case against IBM,[18] a case that was in a sense tainted from the beginning,[19] discovery was protracted, disorganized, and bitter. Trial was laborious with very little judicial direction. As one of the Department's trial lawyers observed to me, it was "not a trial but an institution."[20] There were a number of reasons, but much can be laid on the judge.

I said in an interview shortly before I left Justice that while the issues in both AT&T and IBM were important, it might well be that the primary question would be whether such cases could be tried to a conclusion at all.

In contrast, Judge Greene streamlined discovery, and kept a tight control on witnesses, their testimony, and other elements of the trial process. Filed more than five years after *IBM*, trial in the *AT&T* case was nearing the end when *IBM* was dismissed, still dragging along in trial. And the process came off very well compared to the two other big cases of the day, the FTC's case against the cereal and petroleum industries.[21] I said in an interview shortly before I left Justice that while the issues in both *AT&T* and *IBM* were important, it might well be that the primary question would be whether such cases could be tried to a conclusion at all. The *IBM* trial seemed to suggest they could not be. But *AT&T* convincingly established that, with good judicial management, such cases could be tried efficiently. That is one of the great legacies of the case.

At the conclusion of the government's case AT&T filed its second motion to dismiss, this time asserting that the government had failed in its case in chief.

Judge Greene rejected the motion in a strongly worded opinion, concluding that on all of the elements of the case the government demonstrated that the Bell System "had violated the antitrust laws."[22] This conclusion was so boldly stated that AT&T objected they had been found guilty without ever having presented its case in rebuttal. By this time, Judge Greene was well aware that there was a strong effort being made within the executive branch to get the President to order a dismissal or, failing that, to find a way to settle the case without substantial divestiture.[23] There is reason to believe that Judge Greene's opinion was meant to strengthen the position of the government within the councils of the executive branch. It would be more difficult to order dismissal of a case that had already withstood a motion to dismiss then one where there had been no ruling.

The opinion is of interest today because it is the only major substantive ruling in the case. Given the court's rulings, it raises the obvious question whether the outcome would have been the same had today's governing standards been applied in 1981, or even in 1974 when the case was filed. The answer is far from clear. After reconfirming his ruling on jurisdiction, Judge Greene concluded without extensive discussion that AT&T did in fact have, and long had had, monopoly power in several defined telecommunications markets, a ruling I believe would have been made even under today's standards. AT&T had, after all, long described itself as a kind of benevolent monopoly.

The treatment of conduct is more debatable. As to prohibition of the attachment of customer-owned equipment to the AT&T system, the court relied rather loosely on the *Terminal Railroad* case and several decisions relying on something at least akin to the essential facility doctrine.[24] It found that there was an adequate showing that AT&T lacked any reasonable business justification for its actions. On interconnection of rival inter-city carriers to local exchanges, Judge Greene was more explicit in his reliance on *Terminal Railroad* and the bottleneck monopoly and/or essential facility doctrines (noting that the conduct could also be described as monopoly leveraging).

He deferred ruling on whether compliance with standards of the Communications Act would be a defense to a claim of antitrust violation. Judge Greene was more cautious with respect to claims of cross subsidization from regulated to unregulated markets, the so-called Baxter theory. After discussion of whether predatory pricing standards (and particularly the Areeda-Turner test)[25] should be applied, he ultimately left that legal issue for subsequent resolution. Finally, with respect to the Western Electric equipment issues—the barriers imposed on operating companies with respect to use of non-Western equipment—the court concluded that the issue went well beyond simple vertical integration since the barriers and incentives employed by AT&T were not the result of vertical integration alone.

What would we make of this today? In *Trinko* the Supreme Court pronounced that it has never approved the essential facility doctrine.[26] It has applied the

below-cost standard adopted in *Brooke Group*[27] to a variety of pricing actions.[28] Vertical integration and its necessary consequences are likely to be viewed more favorably than twenty-five years ago. At the same time, the opinion was for the most part consistent with antitrust precedent of its time.

Would the outcome today be the same? Given the current views on essential facilities and vertical integrations that seem to prevail today, would the Department even file the case? While there is no obvious answer to these questions, I remain convinced that AT&T's conduct was anticompetitive and should have been challenged. In substantive terms the case today would have been more difficult. And it would have been yet more difficult given *Trinko*'s seeming preference for regulatory solutions to interconnection problems, although the Court in *Trinko* was confronted with a far more detailed, comprehensive, and crafted regulatory regime than existed in 1974.

In the end, the case settled and Judge Greene never actually ruled on the merits. But the opinion on the motion to dismiss played a major role in the outcome, for three reasons. First, I believe it finally convinced AT&T that it was more likely than not to lose the case at the trial's conclusion. Second, it strengthened the hand of the Department in any settlement negotiations. And, as noted, it made it far more difficult for officials in the Executive Branch *outside* the Department to secure a dismissal of the case. For by the end of the government's case, pressures were mounting to bring the case to an end without the breakup of AT&T. The case was, in fact, being fought on a different front.

From the outset, AT&T and others had sought solutions to the case outside the courtroom. But on the legislative side, its attempt to deal with some elements of the case through extensive amendments to the 1934 Communications Act died in the bowels of the House Antitrust Subcommittee. A settlement that would have required partial divestiture—specifically of Pacific Telephone, two smaller local companies, and 40 percent of Western Electric—along with a detailed agreement on interconnection with other long distance companies was nearly agreed upon on the eve of trial, as trial preparations were proceeding.[29] Judge Greene set the trial date back to permit finalization of the proposed settlement. The settlement was in the hands of Sandy Litvack, the then Division chief, whose two superiors were recused. In the end, the deal fell through. Litvack was departing, and William Baxter, the incoming Assistant Attorney General, found the deal unacceptable. Baxter had publicly supported the case and the relief originally proposed.

Baxter took office with the Reagan administration. Despite his commitment to the case, which he reaffirmed publicly, several incoming cabinet members (most notably the Secretaries of Commerce and Defense) had publicly called for its dismissal.[30] Indeed, during his campaign, President Reagan had offhandedly criticized the case.[31] As the trial began, AT&T officers were seeking the assistance of these

and other officials to get the case dropped. A cabinet level task force, without the participation of the Justice Department, recommended dismissal by the President.[32] The Attorney General, William French Smith, was recused.

So when the day came to meet with the cabinet and President, Baxter was basically on his own. (In fact, had Attorney General Smith not been recused, he likely would have dismissed the case on his own—once again, the quirk of recusal may have had a dramatic impact.) The matter was left hanging. When the proposal to dismiss came before James Baker, the newly appointed White House Chief of Staff, the process slowed down, though the cabinet committee tried hard to get the President to act before Judge Greene ruled on the motion to dismiss. But Baxter refused to budge, and Baker was nervous about the political fallout of a presidentially-directed dismissal, apparently referring to a fear exacerbated by Judge Greene's expressed concern about administration meddling.[33] So perhaps Watergate saved the day again. Then came his opinion on the motion to dismiss, and all hopes for intervention was lost.

So perhaps Watergate saved the day again.

It was a courageous and, as it turned out, politically skillful stand by Bill Baxter. The last legislative efforts simultaneously failed. In the end, AT&T made the basic decision to break itself up in accord with a reorganization plan it had initially prepared, and to accept provisions requiring equal access by long distance carriers to local exchanges. The 1956 decree was formally abrogated. There was high drama in the negotiations but there is not time for that here. But to add to the drama, Justice announced the dismissal of the IBM case the same day the deal in the *AT&T* case was announced.[34] The Department's two big cases effectively ended. Baxter was correct in dismissing *IBM*. It was a case with an aura of illegitimacy, filed on the last day of the Johnson administration. My predecessor, angered by its filing, did little to move the case along. I made the unfortunate decision to put the case to trial. For all that went right in the trial of *AT&T*, we can and have learned from all that went wrong on *IBM*.

The *AT&T* case did not of course end with the decree. Details of the reorganization were left largely to AT&T. And there were hundreds of public comments to be dealt with. Judge Greene made decision after decision that had a significant impact on the industry (some quite ill-advised). The operating companies were kept out of the long distance market; they were to be in essence "quarantined." This may have been ideologically pure, consistent with Baxter's keeping of regulated- and unregulated markets separate, but I am not sure it was wise. For a number of years Greene continued to make rulings that became more and more regulatory, ultimately provoking legislative change, most recently embodied in the Telecommunications Act of 1996.

So what do we learn from *AT&T*, and what was its effect?

1. First, we learned that such a case can be tried. The trial procedures and methods used to control discovery and trial worked, and became the model for the relatively expeditious trial of the *Microsoft* case.[35]

2. Second, judges truly matter, as any comparison with *IBM* demonstrates. Judge Greene was prepared to organize and push the parties, and it worked. He was a quick learner. He may have been driven by a desire to build his reputation, but that ambition served everyone well.

3. Third, we began to get a better handle on the use of economists in both the Division and at trial. This was a transition time for the role of economists at the Division, with the new Economic Policy Office just coming into being. The *AT&T* case was an immediate challenge.

4. Fourth, time and again we learned that in litigation, as in life generally, things are not always as they appear. The trial proceeded apace while, largely unbeknownst to the trial staff, the real forum was the White House. It was at that level that the case was ultimately won.

5. Fifth, we also learned that presidential involvement in an antitrust case, while surely legitimate, is almost never likely to occur. In *AT&T*, virtually the entire cabinet and most likely the President as well agreed the case should be dismissed. Yet the fear of political repercussions caused the President to stay his hand.

6. Sixth, we learned that actors matter. What if there had been no Watergate and no Attorney General Saxbe? What if President Ford had been informed of the case in advance of its filing? What if Sandy Litvack's superiors had not been recused, or if Attorney General Smith had not been recused, leaving Bill Baxter to act on his own? What if the Assistant Attorney General had been someone other than William Baxter, or the White House Chief of Staff had not been James Baker? We will never know, but any change in the cast of characters could have affected the outcome.

We did not, it seems to me, learn much about substantive standards under Section 2 of the Sherman Act. Judge Greene's opinion was not final, but might not have survived *Trinko*. It is the cross-subsidization issue that today is of the greatest interest, referencing the yet-to- develop sacrifice standard. But the whole cross-subsidy argument was never resolved. Little was said about general Section 2 standards, but it has always seemed to me that in bench trials verbalization of general standards matters little. Nor did we learn much about the mechanics, as opposed to the appropriateness, of divestiture. This was a unique case. The remedy was by consent, representing AT&T's judgment that it likely would lose and wanted to play a major role in restructuring. So the court itself did not make the decision on basic relief, and it is not altogether clear that it would in fact have ordered divestiture even had the court found on the merits

against AT&T. It was AT&T that drew up the basic reorganization plan. Moreover, AT&T was structured in a way that clearly facilitated dissolution. It is highly unlikely that this set of circumstances will ever be seen again.

What did we learn about the appropriate roles of antitrust and direct economic regulation in the telecommunications market? Two things seem clear. First, the regulatory structure as it existed in 1974 was inadequate to meet, in a timely fashion, the challenges of an explosion in technology. Second, the direct regulatory role played by Judge Greene in administration of the antitrust decree was inappropriate, undesirable, and equally ineffective in dealing with the larger issues being presented (even though Judge Greene may have had little choice but to fill the vacuum in policy implementation that existed following the decree's entry).

Beyond that, we may not have learned much. The 1996 Act attempted to redefine the antitrust regulatory interface by legislatively mandating steps to open local markets. It has not been an overwhelming success. So the debate on these questions goes on, and will do so for the foreseeable future. The *AT&T* case was but a step along the road. Finally, there was one more important lesson. If you are going to file a case as politically charged as *AT&T*, do it in the wake of a Watergate scandal and while the President is outside the country.

What then was the effect of the case? Could or would competitive markets here have come into being simply as a result of technological and market changes without antitrust intervention at all? And even if such intervention was appropriate, was the dissolution of AT &T a necessary remedy?

The immediate effects of the decree were shareholder anger and consumer confusion. It did not take us long to figure that out. There were also surprises. A number of executives of the Bell operating companies were pleased. One was actually heard to assert the famous Martin Luther King line "free at last."[36] Most shareholders ultimately prospered and, over time, consumer confusion dissipated. Over the decade that followed, consumer choices (at least for long distance service) expanded and—I think most would agree—consumer prices, adjusted for inflation, dropped. Technology-driven changes came even faster than we envisioned. While there are many reasons for this, the breakup played at least some part.

A number of executives of the Bell operating companies were pleased. One was actually heard to assert the famous Martin Luther King line "free at last."

There was another impact that no one envisioned. In foreign markets, particularly in Europe, where telephone systems were state-owned or in the hands of monopolists, the *AT&T* case contributed to privatization and the opening of markets simply by provoking some of these countries to look to the opening of markets in the United States.

In short, it seems to me that the historical record demonstrates that the case accomplished most of what we believed it would and more besides. But it was not any kind of final solution. Technological change came too fast and brought a myriad of new problems to the fore. The changes worked by the decree were nothing but the first steps. There are many more to be taken.

The question remains whether the case, with all its time and expense, was either unnecessary or futile. It could be argued the case was unnecessary because technological change could not be held back and would have worked to open markets even without the breakup, or because some less disruptive remedy—either in an antitrust court or in some regulatory process—could have affected the same outcome with far less disruption or expense. Or it could be argued it was futile in the sense that the industry, through a series of mergers and consolidations, has returned to the highly-concentrated markets that existed before the case was filed. AT&T, it is said, has simply recreated itself.

This last argument I find specious. It is true that concentration levels have been increasing across a spectrum of technologies. But it is a different, far more competitive set of markets. To be sure, vigilance is required to assure that they remain so. But we are nowhere near the entrenched monopoly of AT&T in 1974. Would technological change itself have brought competitive markets over time? In my view, it is at least clear that it would have taken far longer and would have required dramatic regulatory change. Had it been left to the FCC with the statutory authority it had in 1974, I see no reason to believe change would have come faster, at less expense, or more effectively.

The most difficult question for me is whether some less costly and disruptive remedy in the antitrust case could have achieved the same ends. I simply do not know whether a court-mandated open interconnection requirement, coupled with some equipment divestiture and sale of assets to a new company, would have been sufficient. Assistant Attorney General Litvack was close to such a settlement but Bill Baxter found it unacceptable. Whatever the logic, the die was cast.

IT, OF COURSE, REQUIRED AN ACT OF FAITH IN THE OPERATION OF OPEN MARKETS. BUT, IN THE END, DOES NOT ALL OF ANTITRUST?

In the end, and with the benefit of hindsight, the case acted as a catalyst that both facilitated rapid technological change and brought new regulatory regimes into being. It, of course, required an act of faith in the operation of open markets. But, in the end, does not all of antitrust?

1 United States v. American Telephone & Telegraph Co, Civil Action No. 74-1698 (D.D.C. 1974).

2 The consent decree was in the form of a modification of the 1956 consent decree that ended earlier litigation against AT & T. The decree may be found at 1982-2 CCH Trade Cas. &64,900 (D.D.C.).

3 15 U.S.C. §2.

4 *See e.g.*, Steve Coll, The Deal of the Century: The Breakup of AT&T 373 (1986).

5 Id. at 369.

6 William F. Baxter was the Assistant Attorney General in Charge of the Antitrust Division who negotiated the settlement agreement.

7 United States v. American Telephone & Telegraph Co., Decrees and Judgments in Federal Antitrust Cases 483, 497 (D.Ore. 1914) (consent).

8 The 1956 decree may be found at 1956 CCH Trade Cas. 68,246 (D.N.J.).

9 *See* Steve Coll, *supra* note 4, at 59, noting that Attorney General Brownell met privately with representatives of AT&T "at a West Virginia resort" and referring to the settlement process as a "scandal."

10 Space does not permit a recitation of the history of the involvement of MCI and its chairman, William McGowan, in the process leading up to the filing of the AT&T case in 1974. This involvement is chronicled in considerable detail in Steve Coll, *supra* note 4, at 11-52, 200-210.

11 In re Use of the Caterfone Device in Message Toll Telephone Service, 13 F.C.C. 2d 420 (1968). For an extended discussion of *Caterfone* and its implications by one of the FCC Commissioners involved, see Nicholas Johnson, *My Story*, 25 Santa Clara Computer and High Tech L.J. 677 (2009).

12 This was the view of Dean Burch, Chairman of the FCC from 1969 to 1974, in several private conversations. Others on the staff of the FCC shared this view. *See* Steve Coll, *supra* note 4, at 373.

13 The events of November 20, 1974 are recited in far greater detail in Steve Coll, *supra* note 4, at 66-71. Most of Coll's information about those events came, I believe, from Keith Clearwaters, a deputy in the Antitrust Division who took part in all of the meetings on that day. While Coll's report of the dialogue during the meetings seems somewhat overblown, the facts as he recited them are accurate.

14 Baxter had expressed this concern throughout the negotiations over the decree.

15 United States v. American Telephone & Telegraph Co., 461 F.Supp. 1314 (D.D.C. 1978). In an earlier ruling made before Judge Greene was given the case, Judge Waddy had reached the same conclusion. United States v. American Telephone & Telegraph Co., 427 F.Supp. 57 (D.D.C. 1976).

16 Verizon Communications Inc. v. Law Offices of Curtis v. Trinko, LLP, 540 U.S. 398, 414 (2004). In *Trinko*, the Court narrowly read its precedents with respect to the duty of monopolists to deal with their rivals and declined to expand this duty, suggesting that such issues were better left to the regulatory regime established by the Telecommunications Act of 1996, 47 U.S.C. §251.

17 461 F.Supp., at 1328.

18 United States v. IBM Corp., Dkt.No. 69-Civ.-200 (S.D.N.Y.) (complaint filed January 17, 1969).

19 The case was viewed as tainted in the minds of some because it was filed on virtually the last day of the Johnson administration, a legacy left to the incoming Nixon administration to deal with.

20 *See also* Donald I. Baker, *Government Enforcement of Section Two*, 61 Notre Dame L.Rev. 898, 911 (1986) ("The case simply became unwieldy beyond anyone's worst expectations of antitrust nightmare"). Baker headed the Antitrust Division during some of the IBM litigation.

21 Kellog Co., FTC Docket No. 8883 (1972), *dismissed*, 3 CCH TRADE REG.REP. §21,899 (FTC 1982) (cereals); Exxon Corp., FTC Docket No. 8934 (FTC 1973), *dismissed*, 3 CCH TRADE REG.REP.§21,866 (FTC 1981) (petroleum). Both cases were protracted and characterized by discovery and procedural hassles.

22 United States v. American Telephone & Telegraph Co., 524 F.Supp. 1336, 1381 (D.D.C. 1981).

23 *See* note 33 infra.

24 Terminal R.R. Ass'n of St. Louis v. United States, 224 U.S. 383 (1912). Judge Greene also relied upon Hecht v. Pro-Football, Inc., 570 F.2d 982 (D.C. Cir. 1977). The essential facilities doctrine, as he perceived it, requires a monopolist controlling a facility access to which is essential to rivals in order to compete, to provide such access, if feasible, on a reasonable non-discriminatory basis, 524 F.Supp., at 1351, 1360.

25 Phillip Areeda & Donald F. Turner, *Predatory Pricing and Related Practices Under Section 2 of the Sherman Act*, 88 HARV.L.REV. 697 (1975) (adopting a below-average cost standard to be used in identifying pricing that is predatory).

26 Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, LLP, 540 U.S. 398, 411 (2004). The Court found it unnecessary to reach the question since access could be obtained through the FCC and compelled access was therefore not "essential."

27 Brooke Group Ltd. v. Brown & Williamson Tobacco Corp., 509 U.S. 209, 222-3 (1993) (to establish predatory pricing that violates Section 2 of the Sherman Act prices must be below "an appropriate measure" of costs and the monopolist the probability of recouping what it lost because of its below cost strategy.

28 *See* Weyerhaueuser v. Ross-Simmons Hardwood Lumber Co., __ U.S. __, 127 S.Ct. 1609 (2007) (predatory bidding); Pacific Bell Telephone Co. v. linkLINE Communications, Inc., __ U.S. __ (2009).

29 STEVE COLL, *supra* note 4, at 148-160, 172-179.

30 Secretary of Defense Casper Weinberger and Secretary of Commerce Malcolm Boldridge had both very publicly called for dismissal of the case. *See* James B. Stewart, *Whales and Sharks*, THE NEW YORKER 40 (February 15, 1983). These two cabinet secretaries led the attack on the case, an attack leading to the appointment of a cabinet level task force to consider the matter. *See* STEVE COLL, *supra* note 4, at 211-223.

31 STEVE COLL, *supra* note 4 at 185.

32 *Id.*, at 185. The brief version in this paper of the events leading up to the President's decision not to order dismissal of the case is based in large part on conversations with William Baxter. They are spelled out in far greater detail (and accurately I believe) in STEVE COLL, *supra* note 4, at 211-229, 239-253.

33 References were made at a meeting in the White House to Dita Beard, the IT&T lobbyist who played a significant role in the alleged antitrust scandal concerning the settlement of an antitrust case against IT&T. *See* James B. Stewart, *supra* note 30, at 40. The investigation of the charges made White House officials very nervous about any contacts with the Antitrust Division.

34 United States v. International Business Machines Corp., No. 69-Civ.-2001 (S.D.N.Y.) (complaint filed January 17, 1969). The stipulation of dismissal was filed on January 8, 1982. In dismissing the case, Assistant Attorney General Baxter stated that he had found the case to be "without merit" and that there was "little prospect of victory." *See* In re International Business Machines Corp., 687 F.2d 591, 594 (2d Cir. 1982).

35 United States v. Microsoft Corp., 253 F.3d 34 (D.C. Cir. 2001). While critical of the trial judge in some respects, the appellate court commented favorably on the handling and expedition of the case by the trial court.

36 Quoted in Steve Coll, *supra* note 4, at 321.

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Introduction to Harberger's *Monopoly and Resource Allocation*—The Pioneering Article on Deadweight Loss and Empirical Measurement of the Social Costs of Monopoly

Hill B. Wellford

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Introduction to Harberger's *Monopoly and Resource Allocation*—The Pioneering Article on Deadweight Loss and Empirical Measurement of the Social Costs of Monopoly

*Hill B. Wellford**

I. Introduction

Arnold Harberger's 1954 article, *Monopoly and Resource Allocation*,[1] brought empirical analysis of the social costs of monopoly into the mainstream of antitrust work. In the mid-twentieth century, the dominant mode of monopoly analysis in the United States (and therefore worldwide) was structural rather than empirical, and that structural approach supported a highly interventionist antitrust regime. Harberger's 1954 article broke with the then-current economic orthodoxy and set monopoly research on a path that would lead to a strong shift toward empiricism and the development of a more cautious approach for antitrust enforcement. The article is famous for bringing monopoly deadweight loss analysis into the mainstream, graphically represented (see page 283 of the reprint that follows) as the "deadweight loss triangle" familiar to all modern students of antitrust; so much so, in fact, that deadweight loss triangles are now

*Partner, Bingham McCutchen LLP, Washington, DC. I thank Joseph Matelis for his helpful comments. The views expressed here are my own, not those of my firm.

known as "Harberger triangles."[2] But it was Harberger's final estimate of the social costs of monopoly that was the bombshell in this work.

Harberger concluded that the aggregate social costs of monopoly in the U.S. were tiny: only about 0.1 percent of economic output, costing the average American about $48 in today's dollars. Although Harberger did not say so explicitly—the word "antitrust" does not appear in his paper—this conclusion suggested that antitrust enforcement should be ratcheted back, and even called into question whether antitrust enforcement should be attempted at all.

But it was Harberger's final estimate of the social costs of monopoly that was the bombshell in this work.

As a professor of economics at the University of Chicago from 1953 to 1982, Harberger focused his career on the economics of public finance and taxation, and he mostly left the specifics of the antitrust debate that blossomed in the 1960s and 70s to other scholars who focused on antitrust. As a result, it is possible to meet antitrust lawyers today who do not know Harberger's name; however, every modern antitrust lawyer uses tools and, if policy oriented, participates in debates that can be traced directly to Harberger, and particularly to his 1954 article. What follows below is a reminder of why Harberger deserves a re-reading. This introduction is organized into three short sections: a summary of the state of monopoly economics at the time Harberger published *Monopoly and Resource Allocation*; the paper's key points; and the paper's role in shaping monopoly economics and antitrust practice as we know them today.

II. Structural Analysis and the Economic Orthodoxy before Harberger

To understand why *Monopoly and Resource Allocation* was so revolutionary, one must recall the state of monopoly economics and antitrust thinking of the mid-Twentieth century United States. To a modern student of antitrust, for whom the Chicago School is familiar and *Von's Grocery*[3] is a kind of epithet, it may be difficult to imagine a time when structural analysis was dominant. But dominant it was. Herbert Hovenkamp explained the mid-century mindset at length in his *Introduction to the Neal Report* (in the Spring issue of this magazine).[4] As Hovenkamp observed, economists and law professors had spent the first fifty years of the Twentieth Century creating an elaborate theoretical body of work eventually known as the "structure-conduct-performance" (S-C-P) paradigm. The most elegant and most tested model of industrial economics of its time,[5] the S-C-P paradigm represented the high point of structuralism. According to the paradigm, concentration (structure) powerfully influenced conduct, with increases in concentration almost inevitably causing decreases in competition (conduct); less competition almost inevitably led to decreased efficiency and social welfare (per-

formance); and therefore one could effectively delete the middle step and state that structure equals performance. Since the middle step regarding conduct could be ignored (this was the "disappearing middle" in the language of the day), economists, it was assumed, need not evaluate competitive behavior directly. Economists using these structural methods had concluded by the 1950s "that some 20 to 30 to 40 per cent" of the U.S. economy was "effectively monopolized,"[6] and that social welfare losses were correspondingly large.

Ultimately, structuralism and the S-C-P paradigm found their way into the Neal Report,[7] a report on competition in the U.S. economy commissioned by President Lyndon Johnson in 1967 and published in 1969, that suggested reforms to the antitrust laws. The Neal Report is an excellent single source for anyone curious about the economic orthodoxy against which Harberger was working. The Report observed with alarm that "industries in which four or fewer firms account for more than 50 percent of output produce nearly 24 percent" of the total value of manufactured products in the U.S., and stated that "[a]n impressive body of economic opinion and analysis supports the judgment that this degree of concentration precludes"—not reduces, but *precludes*—"effective market competition [.]"[8] The Report proposed a Concentrated Industries Act under which the Department of Justice would "search out" concentrated industries—defined as those in which the four largest firms' combined market share exceeded 70 percent—and order divestitures so that no firm would have a market share above 12 percent. Even a firm with a 15 percent market share would see "steps to reduce" its share under this law.[9] And the Report even took aim at the patent system, stating that "patents are one of the principal sources of monopoly power" and calling for legislation "to establish the principle that a patent which has been licensed to one person shall be made available to all other qualified applicants on equivalent terms."[10] Truly, this was a different model of antitrust than today's: the markets seen as "precluding" competition in the Neal Report could have Herfindahl-Hirschman Index (HHI) scores as low as 650, well under the 1000 HHI value that the U.S. government's Horizontal Merger Guidelines regard as "unconcentrated,"[11] and in which mergers now have a virtual safe harbor.[12]

By the time of the Neal Report's publication in 1969—although one would not realize this from the Report itself—the consensus surrounding structural economics was breaking up. The Report served simultaneously as structuralism's culmination and its last gasp. Hovenkamp's observation on this point cannot be improved, so I will simply quote it:

> "The tragedy of the Neal Report is that the model it represented was just on the verge of complete, catastrophic replacement. . . . Indeed, the publication of the Neal Report played no small part in instigating a massive reaction among younger academics that eventually cast the S-C-P paradigm onto the dung heap of defunct economic doctrines."[13]

That massive reaction was led by a small number of scholars dedicated to antitrust, including one who served on the Neal commission itself: Robert Bork, who had written the seminal article *The Crisis in Antitrust* (1963),[14] wrote a strong dissent to the Neal Report, and later published *The Antitrust Paradox* (1978). But although the reaction came to prominence in the 1960s and 70s, it would be a mistake to imagine it bursting onto the scene without precedent, as if a new Athena had sprung forth fully formed from the side of Bork's head. The reaction was built on a foundation laid by Harberger.

III. The Key Points of *Monopoly and Resource Allocation*

So what exactly is so different about *Monopoly and Resource Allocation*—what did Harberger do that was against the structuralist orthodoxy of his time? Four things: he directly asked whether it was possible to, in his words, "try to get some quantitative notion of the allocative and welfare effects of monopoly,"[15] in particular in U.S. manufacturing; he made a graphical representation of the deadweight loss triangle; he used an empirical estimate of that deadweight loss to answer his question; and, when the loss appeared to be very small, he stated this conclusion:

> "[I]t seems to me that the monopoly problem does take on a rather different perspective in light of the present study. Our economy emphatically does not seem to be monopoly capitalism in big red letters."[16]

The last point was certainly revolutionary; it surprised even Harberger, who said, "I must confess I am amazed at this result."[17] But the first three points were no less groundbreaking, at least as a matter of academic inquiry.

SO WHAT EXACTLY IS SO DIFFERENT ABOUT *MONOPOLY AND RESOURCE ALLOCATION*—WHAT DID HARBERGER DO THAT WAS AGAINST THE STRUCTURALIST ORTHODOXY OF HIS TIME? FOUR THINGS.

Taking these points in order, one begins with the surprising observation (to a modern student of antitrust) that before Harberger, academics did not even try to estimate the magnitude of monopoly welfare loss economy-wide. Harberger's estimate was the first.[18] Why was there so fundamental a hole in the literature? The answer seems to be both that it was assumed to be extremely difficult to do so, and that it was assumed to be unnecessary—few doubted that monopoly losses were quite severe. Harberger himself observes

that "I never really tried to quantify my notions of what monopoly misallocations amounted to, and I doubt that many other people have."[19]

A subtler answer may be that a sort of feedback loop was at work. Prominent academics said that only structural analyses, not empirical estimates, were feasible and necessary; so judges entertained only structural arguments; so lawyers employed only structural expert witnesses, not empiricists, in important cases; and so empiricists never became prominent in antitrust academia. This may help explain why it fell to Harberger, an obscure (to antitrust experts) economist at Chicago focusing on tax matters, to create a revolution under the very noses of his antitrust colleagues. The Neal Report is, after all, named for commission chairman Phil C. Neal, then Dean of the University of Chicago Law School. Harberger became a grandfather of what came to be known as the empiricist Chicago School but it is worth noting that the actual school in Chicago in 1954 was quite friendly to structuralism.

THIS MAY HELP EXPLAIN WHY IT FELL TO HARBERGER, AN OBSCURE (TO ANTITRUST EXPERTS) ECONOMIST AT CHICAGO FOCUSING ON TAX MATTERS, TO CREATE A REVOLUTION UNDER THE VERY NOSES OF HIS ANTITRUST COLLEAGUES.

After asking the hitherto unexamined question, Harberger set about using the deadweight loss triangle to answer it. Harberger in 1954 was not the first to draw such a figure. Deadweight loss triangles (under various names) had been known at least since the 1840's work of a French engineer named Jules Dupuit, who used them to measure the consumer benefits of public works.[20] Others used them over the intervening century to evaluate the loss due to many distortions, including taxes, which is almost certainly how they came to be on Harberger's mind.[21]

Harberger did not appear to believe that his use of deadweight loss triangles was revolutionary; he introduced a triangle without fanfare in his Figure 1, and never called it a deadweight loss triangle or gave it a name of any kind in the 1954 article.[22] But it would be a mistake to minimize Harberger's innovation just because the basic idea of the triangle was already known; almost no economists were measuring deadweight loss triangles empirically in Harberger's time, and none were using them to estimate monopoly effects.[23] This was an important omission: without such estimates, it was impossible to offer reliable answers to important questions about monopoly distortions, and antitrust economics lacked the empirical grounding that later facilitated rapid progress.[24]

Harberger's empirical findings in *Monopoly and Resource Allocation* are best taken directly by reading the article, of course, but they can be summarized briefly. First, Harberger looked for a time when economic data was relatively well kept and economic shocks were relatively few. This was no easy task to an academic working in the early 1950s—the United States had seen three major wars

and a Great Depression in just the past fifty years—but Harberger was able to find a suitable period in the late 1920s.

He averaged rates of economic return over a five-year period (to further smooth out temporary distortions) for 73 manufacturing industries, assumed that the average rate of profit was the competitive profit, measured how each of the industries deviated from that competitive profit, and then took that deviation as the amount that "prices in each industry were 'too high' or 'too low' when compared with those that would generate an optimal resource allocation."[25] He then applied a formula to determine the amount that consumer welfare would increase or decrease if each industry either acquired or divested itself of the appropriate amount of resources to remove the distortions he found; he expanded that figure to cover the whole economy (not just the sectors that he directly measured); and he got "what we really want: an estimate of by how much consumer welfare would have improved if resources had been optimally allocated throughout American manufacturing in the late twenties."[26] He then applied several reductive factors, since this number was a measure of all distortion, not merely monopoly distortion; however, he applied the reduction conservatively, meaning that "in short, [he] labored at each stage to get a big estimate of the welfare loss [.]"[27] Even so, he said, "we come out at the end with less than a tenth of a per cent of the national income."[28]

Harberger was cautious about his results. He acknowledged that some factors may have caused him to underestimate the harm (although others, he noted, may have caused him to overestimate it). He declared that he did not mean to minimize the effects of monopoly: "a tenth of a per cent of the national income is still over 300 million [in 1954] dollars,"[29] or about $14.29 billion today. And he was at pains to admit that he did not examine certain ancillary effects; for example, he decided not to take on the task of analyzing the redistributions of income that arise when monopoly is present.[30] "All I want to say here," he wrote, "is that monopoly does not seem to affect aggregate welfare very seriously through its effect on resource allocation."[31] Harberger did not call for changes to antitrust practice—as previously mentioned, the word "antitrust" never appears in *Monopoly and Resource Allocation*—and in fact, in the 1954 article, he did not call for policy changes of any kind. Then again, with these results, he did not need to.

HARBERGER'S ARTICLE SEEMS TO TODAY'S READER TO BE SURPRISINGLY MODERN: IT PRESENTS EMPIRICISM AS A GIVEN, NOT AS SOME TYPE OF NEW AND UNTRIED DEVICE.

One final note about the article itself. Unlike the Neal Report, which Hovenkamp described as "a trip to another world," Harberger's article seems to today's reader to be surprisingly modern: it presents empiricism as a given, not as some type of new and untried device. True, the writing may appear old-fashioned: the article proceeds in a conversational, almost folksy style more suited to the first half of the Twentieth Century than the second, making the reader feel as if he or she were seated in a

winged chair before a fireplace during one of Chicago's brutal winters, casually bantering with a colleague over some minor academic point.

There is no hint from Harberger's tone that he was shaking the entire foundation upon which early- and mid-century antitrust practice was based. To the modern reader who knows what became of this article, the disconnect between tone and substance is a bit shocking; it is as if the professor has offered a tumbler of aged scotch and, after accepting, one discovers the glass to contain a hand grenade. In the final analysis, Harberger is revealed as a master of modesty and understatement. Modern academics, seeking as they would even the smallest measure of Harberger's renown, might want to take note.

IV. Deadweight Loss and Harberger's Thesis in Modern Antitrust Practice

Harberger may not have made policy prescriptions in his 1954 article but he was indeed motivated by policy. And he appears to have been a bit frustrated that policy changes take time, as they did in the area of antitrust. As proof, look no further than the facts that *Von's Grocery* was a 1966 decision (12 years after *Monopoly and Resource Allocation*) and the Neal Report was published in 1969 (15 years after). By 1964, Harberger was calling explicitly for policy to catch up to the new empiricism in monopoly economics:

> "The measurement of deadweight losses is not new to economics by any means. It goes back at least as far as Dupuit.... Nonetheless I feel that the profession as a whole has not given to the area the attention that I think it deserves. We do not live on the Pareto frontier, and we are not going to do so in the future. Yet policy decisions are constantly being made which can move us either toward or away from that frontier. What could be more relevant to a choice between policy A and policy B than a statement that policy A will move us toward the Pareto frontier in such a way as to gain for the economy [a wealth effect greater than] policy B ...?"[32]

Eventually, other economists did catch up and, with them, policymakers. The Neal Report quickly became a dead letter, due in part to the influence of Harberger's work. As various scholars examined both Harberger's specific results and his approach, a debate ensued, and in general his results regarding deadweight loss effects proved robust. The debate ranged across several disciplines—from antitrust to corporate income tax—but Harberger's work survived, in part because Harberger made conservative estimates and in part because many aspects

of alternative calculations and methodological specifications tended to cancel each other.[33] For work that supported the thesis of *Monopoly and Resource Allocation*, see F.M. Scherer's *Industrial Market Structure and Economics* (2d ed. 1980) and studies by Schwartzman, Siegfried, Tiemann, & Worcester.[34] Other studies found greater or lesser welfare effects in different time periods, but this author is not aware of any well-respected study of the U.S. economy that finds a different and larger effect sufficient to support the highly interventionist antitrust approach that prevailed in 1954. (Note that this may not be as true for other nations' economies; for example, Jenny & Weber in 1983 found that the deadweight loss in France might be as high as 7.3 percent.[35])

The larger effect of Harberger's article has been to reframe the terms of antitrust work, both as a matter of case practice and policy debate. On the case practice side, any practitioner knows that in most mergers, single firm conduct, and rule of reason cases, empirical analysis of welfare effects is mandatory. *Per se* rules still exist in antitrust law, and structural analysis still has its place in the initial screens applied by the U.S. Horizontal Merger Guidelines, but in other situations, empirical analysis of welfare effects is often dispositive. As this author has previously explained in greater detail, empirical welfare economics has become almost synonymous with antitrust economics, and antitrust economics has transformed U.S. antitrust law into an "effects based" (outcome based) system via its adoption in landmark Supreme Court decisions. So much so, in fact, that the Supreme Court—having become comfortable with such economics through its antitrust jurisprudence—now appears to be using welfare economics to transform other areas of the law as well.[36]

On the policy side, Harberger's work and subsequent similar studies forced defenders of antitrust to react, and now form the background against which academics measure arguments over the proper level of antitrust enforcement. That debate has not been wholly negative for the antitrust side. True, some have concluded that the antitrust flame is not worth the candle, and that the Sherman Act should be repealed. Most, however, have concluded that while antitrust should be less interventionist than its 1950s model, antitrust law is still meaningful.

> If anything, the adversity represented by Harberger's thesis has made antitrust's defenders smarter and stronger.

If anything, the adversity represented by Harberger's thesis has made antitrust's defenders smarter and stronger. Instead of resisting welfare economics, they have embraced and co-opted it. They have focused the most enforcement effort where the chance of false positives (unmerited enforcement) is least, using policies such as the "antitrust hierarchy," which devotes enforcement resources in descending order to cartels, merger enforcement, and non-merger civil conduct. And they have developed subtler arguments, such as taking the position that deadweight loss should not be the only concern of antitrust law: wealth distribution distortions, rent seeking distortions, and reductions to dynamism and technological

innovation, they have claimed, are difficult to measure via Harberger's method but nonetheless crucial.[37] Such debate is beyond the scope of this introduction. For now, it is enough to observe that the tools and debate of modern antitrust practice can be traced in important ways back to *Monopoly and Resource Allocation*, and that the article is well worth a read by the many antitrust lawyers who came of age after its revolutionary ideas had become the mainstream.

---

1 Arnold C. Harberger, *Monopoly and Resource Allocation*, 44 Am. Econ. Rev. 77 (1954), reprinted at 283 in this issue of Competition Policy International (page citations hereinafter are to the CPI reprint).

2 *See generally* James R. Hines, Jr., *Three Sides of Harberger Triangles*, NBER Working Paper Series, Working Paper 6852 (1998) (available via SSRN). "Harberger triangles" are so called due to the author's entire body of work but the 1954 article and another in 1971 (Arnold C. Harberger, *Three Basic Postulates for Applied Welfare Economics*, 9 J. Econ. Lit. 785 (1971)) are most responsible for this appellation.

3 United States v. Von's Grocery Co., 384 U.S. 270 (1966). In this now-infamous case, the merger of the two largest grocery companies in a market was found to violate Clayton Act section 7, despite the fact that their combined market share was only 7.5 percent. Equally famously, Justice Stewart, in dissent, observed that in such cases brought under section 7, "the sole consistency that I find is that in litigation under S7, the government always wins." *Id.* at 301.

4 Herbert Hovenkamp, *Introduction to the Neal Report and The Crisis in Antitrust*, 5 Competition Policy International 217 (2009).

5 *Id.* at 219.

6 Harberger, Monopoly and Resource Allocation at 283.

7 Report of the White House Task Force on Antitrust Policy (May 27, 1969), originally published at 115 Cong. Rec. 11, 13890, reprinted at 5 Competition Policy International 227 (2009) (page citations hereinafter are to the CPI reprint). The report is called the Neal Report after its chairman, Phil C. Neal, then Dean of the University of Chicago Law School.

8 *Id.* at 228.

9 *Id.* at 237.

10 *Id.* at 230.

11 U.S. Dep't of Justice & Fed. Trade Comm'n, Horizontal Merger Guidelines 15 (1992 & Rev. 1997), available at http://www.usdoj.gov/atr/public/guidelines/hmg.htm.

12 Hovenkamp, Introduction to the Neal Report (supra note 4) at 218.

13 *Id.*

14 Robert Bork & Ward Bowman, *The Crisis in Antitrust*, Fortune (Dec. 1963); and 65 Col. L. Rev. 363 (1965).

15 Harberger, Monopoly and Resource Allocation at 283.

16 *Id.* at 290.

17 *Id*.

18 Robert H. Lande, *Wealth Transfers as the Original and Primary Concern of Antitrust: the Efficiency Interpretation Challenged*, 50 HASTINGS L. J. 871, 879 n. 32 (1999).

19 Harberger, Monopoly and Resource Allocation at 290.

20 Hines, *Three Sides of Harberger Triangles* (*supra* note 2) at 3-4, discussing A. Jules E.J. Dupuit, De la Measure de l'Utilité des Travaux Publics, ANNALES DES PONTS ET CHAUSSÉES, 2d ser., 8 (1844); translated by R.H. Barback as On the Measurement of the Utility of Public Works, INTERN. ECON. PAPERS 2 (1952); reprinted in KENNETH J. ARROW & TIBOR SCITOVSKY EDS., READINGS IN WELFARE ECONOMICS 255 (1969).

21 *See id.* at 4-22 (discussing the history of deadweight loss analysis, primarily in tax economics).

22 *See* Harberger, Monopoly and Resource Allocation at 284.

23 *See* Hines, Three Sides of Harberger Triangles, supra n. 20, at 24.

24 *See id.*

25 Harberger, Monopoly and Resource Allocation at 285.

26 *Id.* at 287.

27 *Id.* at 290.

28 *Id.*

29 *Id.* at 287.

30 *Id.*

31 *Id.* (emphasis added).

32 Arnold C. Harberger, The Measurement of Waste, 54 AM. ECON. REV. 58, 58-59 (1964).

33 Hines, *Three Sides of Harberger Triangles*, *supra* n. 20, at 12.

34 David Schwartzman, *The Burden of Monopoly*, 68 J. POL. ECON. 627 (1960); John J. Siegfried & Thomas K. Tiemann, *The Welfare Costs of Monopoly: An Inter-Industry Analysis*, 12 ECON. INQUIRY 190 (1974); Dean A. Worcester, Jr., *New Estimates of the Welfare Loss to Monopoly, United States: 1956-1969*, 40 S. ECON. J. 234 (1973).

35 Frédérick Jenny & André-Paul Weber, *Aggregate Welfare Loss Due to Monopoly Power in the French Economy: Some Tentative Estimates*, 32 J. INDUSTRIAL ECON. 113 (1983).

36 *See* Hill B. Wellford, *Is the Supreme Court Importing Antitrust Economics into Patent Law? A Different Look at eBay, MedImmune, KSR, and Quanta Computer*, GLOBAL COMPETITION POLICY (Mar. 2009 rel. 2).

37 *See*, *generally*, H. Hovenkamp, *Antitrust Policy and the Social Cost of Monopoly*, 78 IOWA L. REV. 371 (1993) and F.M. Scherer, *Antitrust, Efficiency, and Progress*, 62 N.Y.U. L. REV. 998 (1987).

CPI COMPETITION POLICY INTERNATIONAL

Volume 5 | Number 2 | Autumn 2009

Monopoly and Resource Allocation

Arnold C. Harberger

Published in Competition Policy International (print ISSN 1554-0189, online ISSN 1554-6853), Autumn 2009, Vol. 5, No. 2. For articles and more information, visit www.globalcompetitionpolicy.org.

Classic Reprint: Monopoly and Resource Allocation[1]

*Arnold C. Harberger**

One of the first things we learn when we begin to study price theory is that the main effects of monopoly are to misallocate resources, to reduce aggregate welfare, and to redistribute income in favor of monopolists. In the light of this fact, it is a little curious that our empirical efforts at studying monopoly have so largely concentrated on other things. We have studied particular industries and have come up with a formidable list of monopolistic practices: identical pricing, price leadership, market sharing, patent suppression, basing points, and so on. And we have also studied the whole economy, using the concentration of production in the hands of a small number of firms as the measure of monopoly. On this basis we have obtained the impression that some 20 or 30 or 40 per cent of our economy is effectively monopolized.

In this paper I propose to look at the American economy, and in particular at American manufacturing industry, and try to get some quantitative notion of the allocative and welfare effects of monopoly. It should be clear from the outset that this is not the kind of job one can do with great precision. The best we can hope for is to get a feeling for the general orders of magnitude that are involved.

I take it as an operating hypothesis that, in the long run, resources can be allocated among our manufacturing industries in such a way as to yield roughly constant returns. That is, long-run average costs are close to constant in the relevant range, for both the firm and the industry. This hypothesis gives us the wedge we need to get something from the data. For as is well known, the malallocative effects of monopoly stem from the difference between marginal cost and price, and marginal costs are at first glance terribly difficult to pin down empirically for

*Arnold C. Harberger is Professor of Economics at the University of California in Los Angeles. He is also the Gustavus F. and Ann M. Swift Distinguished Service Professor Emeritus at the University of Chicago

a wide range of firms and industries. But once we are ready to proceed on the basis of constant average costs, we can utilize the fact that under such circumstances marginal and average costs are the same, and we can easily get some idea of average costs.

But that does not solve all the problems, for cost and profit to the economist are not the same things as cost and profit to the accountant, and the accountants make our data. To move into this question, I should like to conjure up an idealized picture of an economy in equilibrium. In this picture all firms are operating on their long-run cost curves. The cost curves are so defined as to yield each firm an equal return on its invested capital, and markets are cleared. I think it is fair to say that this is a picture of optimal resource allocation. Now, we never see this idyllic picture in the real world, but if long-run costs are in fact close to constant and markets are cleared, we can pick out the places where resources are misallocated by looking at the rates of return on capital. Those industries which are returning higher than average rates have too few resources; and those yielding lower than average rates have too many resources. To get an idea of how big a shift of resources it would take to equalize profit rates in all industries, we have to know something about the elasticities of demand for the goods in question. In Figure 1, I illustrate a hypothetical case. The industry in question is earning 20 per cent on a capital of 10 million dollars, while the average return to capital is only 10 per cent. We therefore build a 10 per cent return into the cost curve, which leaves the industry with 1 million in excess profits. If the elasticity of demand for the industry's product is unity, it will take a shift of 1 million in resources in order to expand supply enough to wipe out the excess profits.

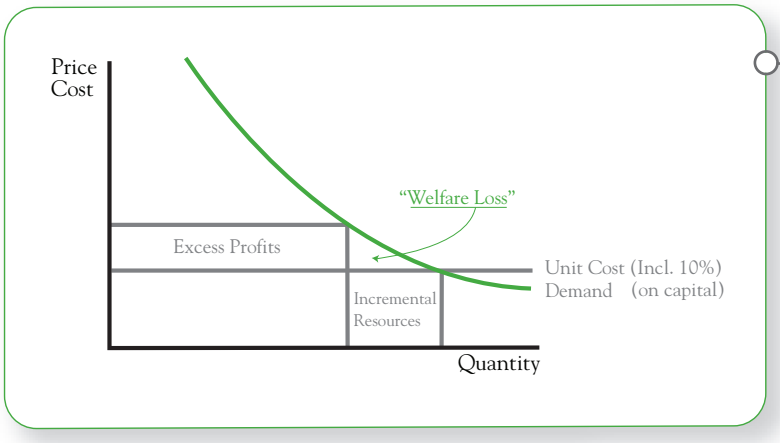


Figure 1

The above argument gives a general picture of what I have done empirically. The first empirical job was to find a period which met two conditions. First, it

had to be reasonably close to a long-run equilibrium period; that is, no violent shifts in demand or economic structure were to be in process. And second, it had to be a period for which accounting values of capital could be supposed to be pretty close to actual values. In particular, because of the disastrous effect of inflation and deflation on book values of capital, it had to be a period of fairly stable prices, which in turn had been preceded by a period of stable prices. It seemed to me that the late twenties came as close as one could hope to meeting both these requirements.

The late twenties had an additional advantage for me—because my choice of this period enabled me to use Professor Ralph C. Epstein's excellent study, *Industrial Profits in the United States* (National Bureau of Economic Research, 1934), as a source of data. Professor Epstein there gives, for the years 1924-28, the rates of total profit to total capital for seventy-three manufacturing industries, with total capital defined as book capital plus bonded indebtedness and total profit defined as book profit plus interest on the indebtedness. To get rid of factors producing short-period variations in these rates of return, I average the rates, for each industry, for the five-year period. The results are given in column 1 of Table 1 [See Appendix]. The differences among these profit rates, as between industries, give a broad indication of the extent of resource malallocation in American manufacturing in the late twenties.

Column 2 presents the amount by which the profits in each industry diverged from what that industry would have obtained if it had gotten the average rate of profit for all manufacturing industry. In column 3, these excesses and shortages of profit are expressed as a per cent of sales in the industry. By analogy with Figure 1, you can see that this column really tells by what percentage prices in each industry were "too high" or "too low" when compared with those that would generate an optimal resource allocation.

Now suppose we ask how much reallocation of resources it would take to eliminate the observed divergences in profit rates. This depends, as you can see in Figure 1, on the demand elasticities confronting the industries in question. How high are these elasticities? It seems to me that one need only look at the list of industries in Table 1 in order to get the feeling that the elasticities in question are probably quite low. The presumption of low elasticity is further strengthened by the fact that what we envisage is not the substitution of one industry's product against all other products, but rather the substitution of one great aggregate of products (those yielding high rates of return) for another aggregate (those yielding low rates of return). In the light of these considerations, I think an elasticity of unity is about as high as one can reasonably allow for, though a somewhat higher elasticity would not seriously affect the general tenor of my results.

Returning again to Figure 1, we can see that once the assumption of unit elasticity is made the amount of excess profit measures the amount of resources that

must be called into an industry in order to bring its profit rate into line. When I say resources here I mean the services of labor and capital plus the materials bought by the industry from other industries. In many ways it seems preferable to define resources as simply the services of labor and capital. This could be done by applying to the value added in the industry the percentage of excess profits to sales. The trouble here is that adding to the output of industry X calls resources not only into that industry but also into the industries that supply it. And by the time we take all the increments in value added of all these supplying industries that would be generated by the initial increase in output of industry X, we come pretty close to the incremental value of sales in industry X. Of course, the movement to an optimal resource allocation entails some industries expanding their output, like X, and others, say Y, contracting their output. If we really traced through the increments to value added which are required in their supplying industries, say Z, we would often find that there was some cancellation of the required changes in the output of Z. Hence by using sales rather than value added as our measure of resource transfer, we rather overstate the necessary movement.

Keeping this in mind, let us return to the data. If we add up all the pluses and all the minuses in column 2, we find that to obtain equilibrium we would have to transfer about 550 million dollars in resources from low-profit to high-profit industries. But this is not the end. Those of you who are familiar with Epstein's study are aware that it is based on a sample of 2,046 corporations, which account for some 45 per cent of the sales and capital in manufacturing industry. Pending a discussion of possible biases in the sample a little later, we can proceed to blow up our 550 million figure to cover total manufacturing. The result is 1.2 billion. Hence we tentatively conclude that the misallocations of resources which existed in United States manufacturing in the period 1924-28 could have been eliminated by a net transfer of roughly 4 per cent of the resources in manufacturing industry, or 1 1/2 per cent of the total resources of the economy.

Now let us suppose that somehow we effected these desired resource transfers. By how much would people be better off? This general question was answered in 1938 for an analogous problem by Harold Hotelling.[2] His general formula would be strictly applicable here if all our industries were producing products for direct consumption. The question thus arises, how to treat industries producing intermediate products. If we neglect them altogether, we would be overlooking the fact that their resource shifts and price changes do ultimately change the prices and amounts of consumer goods. If, on the other hand, we pretend that these intermediate industries face the consumer directly and thus directly affect consumer welfare, we neglect the fact that some of the resource shifts in the intermediate sector will have opposing influences on the prices and quantities of consumer goods. Obviously, this second possibility is the safer of the two, in the sense that it can only overestimate, not underes-

timate, the improvement in welfare that will take place. We can therefore follow this course in applying the Hotelling formula to our data. The results are shown in column 4 of Table 1. This gives, opposite each industry, the amount by which consumer welfare would increase if that industry either acquired or divested itself of the appropriate amount of resources. The total improvement in consumer welfare which might come from our sample of firms thus turns out to be about 26.5 million dollars. Blowing up this figure to cover the whole economy, we get what we really want: an estimate of by how much consumer welfare would have improved if resources had been optimally allocated throughout American manufacturing in the late twenties. The answer is 59 million dollars—less than one-tenth of 1 per cent of the national income. Translated into today's national income and today's prices, this comes out to 225 million dollars, or less than $1.50 for every man, woman, and child in the United States.

Before drawing any lessons from this, I should like to spend a little time evaluating the estimate. First let us look at the basic assumption that long-run costs are constant. My belief is that this is a good assumption, but that if it is wrong, costs in all probability tend to be increasing rather than decreasing in American industry. And the presence of increasing costs would result in a lowering of both our estimates. Less resources would have to be transferred in order to equalize profit rates, and the increase in consumer welfare resulting from the transfer would be correspondingly less.

On the other hand, flaws in the data probably operate to make our estimate of the welfare loss too low. Take for example the question of patents and good will. To the extent that these items are assigned a value on the books of a corporation, monopoly profits are capitalized, and the profit rate which we have used is an understatement of the actual profit rate on real capital. Fortunately for us, Professor Epstein has gone into this question in his study. He finds that excluding intangibles from the capital figures makes a significant difference in the earnings rates of only eight of the seventy-three industries. I have accordingly recomputed my figures for these eight industries.[3] As a result, the estimated amount of resource transfer goes up from about 1 1/2 per cent to about 1 3/4 per cent of the national total. And the welfare loss due to resource misallocations gets raised to about 81 million dollars, just over a tenth of 1 per cent of the national income.

There is also another problem arising out of the data. Epstein's sample of firms had an average profit rate of 10.4 per cent during the period I investigated, while in manufacturing as a whole the rate of return was 8 per cent. The reason for this divergence seems to be an overweighting of high-profit industries in Epstein's sample. It can be shown, however, that a correct weighting procedure would raise our estimate of the welfare cost of equalizing profit rates in all industries by no more than 10 million dollars.[4]

Following is a breakdown of the adjustment for the 8 industries in question:

Figure 2

Industry	Adjusted Profit Rate	Adjusted Rate of Excess Profit	Adjusted Amount of Excess Profits (Millions)	Adjusted Welfare Loss (Millions)
Confectionery	21.1	10.7	11	.530
Tobacco	19.0	8.6	66	2.225
Men's clothing	14.9	4.5	5	.068
Stationery	8.8	—	—	—
Newspaper publishing	27.9	17.5	67	5.148
Proprietary preparations	27.8	17.4	42	4.121
Toilet preparations	50.8	40.4	6	1.400
Printing Machinery	12.9	2.5	2	.064
			199	13.556
Less previous amount of excess profit or welfare loss			–100	–3.845
Net adjustment			99	9.711

Finally, there is a problem associated with the aggregation of manufacturing into seventy-three industries. My analysis assumes high substitutability among the products produced by different firms within any industry and relatively low substitutability among the products of different industries. Yet Epstein's industrial classification undoubtedly lumps together in particular industries products which are only remote substitutes and which are produced by quite distinct groups of firms. In short, Epstein's industries are in some instances aggregates of subindustries, and for our purposes it would have been appropriate to deal with the subindustries directly. It can be shown that the use of aggregates in such cases biases our estimate of the welfare loss downward, but experiments with hypothetical examples reveal that the probable extent of the bias is small.[5]

Thus we come to our final conclusion. Elimination of resource misallocations in American manufacturing in the late twenties would bring with it an improvement in consumer welfare of just a little more than a tenth of a per cent. In present values, this welfare gain would amount to about $2.00 per capita.

Now we can stop to ask what resource misallocations we have measured. We actually have included in the measurement not only monopoly misallocations but also misallocations coming out of the dynamics of economic growth and development and all the other elements which would cause divergent profit rates

to persist for some time even in an effectively competitive economy. I know of no way to get at the precise share of the total welfare loss that is due to monopoly, but I do think I have a reasonable way of pinning our estimate down just a little more tightly. My argument here is based on two props. First of all, I think it only reasonable to roughly identify monopoly power with high rates of profit. And secondly, I think it quite implausible that more than a third of our manufacturing profits should be monopoly profits; that is, profits which are above and beyond the normal return to capital and are obtained by exercise of monopoly power. I doubt that this second premise needs any special defense. After all, we know that capital is a highly productive resource. On the first premise, identifying monopoly power with high profits, I think we need only run down the list of high-profit industries to verify its plausibility. Cosmetics are at the top, with a 30 per cent return on capital. They are followed by scientific instruments, drugs, soaps, newspapers, automobiles, cereals, road machinery, bakery products, tobacco, and so on. But even apart from the fact that it makes sense in terms of other evidence to consider these industries monopolistic, there is a still stronger reason for making this assumption. For given the elasticity of demand for an industry's product, the welfare loss associated with that product increases as the square of its greater-than-normal profits. Thus, granted that we are prepared to say that no more than a third of manufacturing profits were monopoly profits, we get the biggest welfare effect by distributing this monopoly profit first to the highest profit industries, then to the next highest, and so on. When this is done, we come to the conclusion that monopoly misallocations entail a welfare loss of no more than a thirteenth of a per cent of the national income. Or, in present values, no more than about $1.40 per capita.

Before going on, I should like to mention a couple of other possible ways in which this estimate might fail to reflect the actual cost of monopoly misallocations to the American consumer. First, there is the possibility that book capital might be overstated, not because of patents and good will, but as a result of mergers and acquisitions. In testing this possibility I had recourse to Professor J. Fred Weston's recent study of mergers. He found that mergers and acquisitions accounted for only a quarter of the growth of seventy-odd corporations in the last half-century (*The Role of Mergers in the Growth of Large Firms*, pages 100-102). Even a quite substantial overstatement of the portion of their capital involved in the mergers would thus not seriously affect the profit rates. And furthermore, much of the merger growth that Weston found came in the very early years of the century; so that one can reasonably expect that most of the assets which may have been overvalued in these early mergers were off the books by the period that I investigated.

The second possibility concerns advertising expenditures. These are included as cost in accounting data, but it may be appropriate for our present purpose to include part of them as a sort of quasi-monopoly profit. I was unable to make any systematic adjustment of my data to account for this possibility, but I did make a cursory examination of some recent data on advertising expenditures. They sug-

gest that advertising costs are well under 2 per cent of sales for all of the industries in Table 1. Adjustment of our results to allow for a maximal distorting effect of advertising expenditures would accordingly make only a slight difference, perhaps raising our estimate of the welfare cost of monopoly in present values to $1.50 per capita, but not significantly higher.[6]

I should like now to review what has been done. In reaching our estimate of the welfare loss due to monopoly misallocations of resources we have assumed constant rather than increasing costs in manufacturing industry and have assumed elasticities of demand which are too high, I believe. On both counts we therefore tend to overstate the loss. Furthermore, we have treated intermediate products in such a way as to overstate the loss. Finally, we have attributed to monopoly an implausibly large share—33 1/3 per cent—of manufacturing profits, and have distributed this among industries in such a way as to get the biggest possible welfare loss consistent with the idea that monopolies tend to make high profits. In short, we have labored at each stage to get a big estimate of the welfare loss, and we have come out in the end with less than a tenth of a per cent of the national income.

I must confess that I was amazed at this result. I never really tried to quantify my notions of what monopoly misallocations amounted to, and I doubt that many other people have. Still, it seems to me that our literature of the last twenty or so years reflects a general belief that monopoly distortions to our resources structure are much greater than they seem in fact to be.

Let me therefore state the beliefs to which the foregoing analysis has led me. First of all, I do not want to minimize the effects of monopoly. A tenth of a per cent of the national income is still over 300 million dollars, so we dare not poohpooh the efforts of those—economists and others—who have dedicated themselves to reducing the losses due to monopoly. But it seems to me that the monopoly problem does take on a rather different perspective in the light of present study. Our economy emphatically does not seem to be monopoly capitalism in big red letters. We can neglect monopoly elements and still gain a very good understanding of how our economic process works and how our resources are allocated. When we are interested in the big picture of our manufacturing economy, we need not apologize for treating it as competitive, for in fact it is awfully close to being so. On the other hand, when we are interested in the doings of particular industries, it may often be wise to take monopoly elements into account. Even though monopoly elements in cosmetics are a drop in the bucket in the big picture of American manufacturing, they still mean a lot when we are studying the behavior of this particular industry.

Finally I should like to point out that I have discussed only the welfare effects of resource misallocations due to monopoly. I have not analyzed the redistributions of income that arise when monopoly is present. I originally planned to discuss this redistribution aspect as well, but finally decided against it. All I want to

say here is that monopoly does not seem to affect aggregate welfare very seriously through its effect on resource allocation. What it does through its effect on income distribution I leave to my more metaphysically inclined colleagues to decide. I am impelled to add a final note in order to forestall misunderstandings arising out of matters of definition. Resource misallocations may clearly arise from causes other than those considered here: tariffs, excise taxes, subsidies, trade-union practices, and the devices of agricultural policy are some obvious examples. Some of these sources of misallocation will be discussed in a forthcoming paper. Suffice it to say here that the present paper is not concerned with them.

Appendix

TABLE 1

Industry	(1) Rate of Profit on Capital (1924–28)	(2) Amount by Which Profits Diverged from "Average" (Millions)	(3) Column (2) as Per Cent of Sales	(4) Welfare Cost of Divergence in Column (2) (Millions)
Bakery products	17.5%	$17	5.3%	$.452
Flour	11.9	1	0.4	.002
Confectionery	17.0	7	6.1	.215
Package foods	17.9	7	3.3	.116
Dairying	11.8	3	0.7	.010
Canned goods	12.4	1	0.6	.003
Meat packing	4.4	−69	−1.7	.596
Beverages	5.8	−2	−4.0	.080
Tobacco	14.1	27	0.3	.373
Miscellaneous foods	8.1	−13	−2.4	.164
Cotton spinning	10.0	−0	0	0
Cotton converting	8.0	−1	−0.6	.008
Cotton weaving	4.7	−15	−5.5	.415
Weaving woolens	2.6	−16	−9.5	.762
Silk weaving	7.9	−3	−2.3	.035
Carpets	9.8	−1	−1.3	.006
Men's clothing	11.4	1	0.5	.002
Knit goods	12.9	3	1.9	.028
Miscellaneous clothing	13.1	1	1.1	.006
Miscellaneous textiles	9.2	−2	−0.9	.008
Boots and shoes	15.8	9	3.8	.172
Miscellaneous leather products	7.7	−3	−2.1	.032
Rubber	7.6	−23	−2.5	.283
Lumber manufacturing	7.8	−6	−3.9	.118
Planing mills	13.1	1	3.2	.016
Millwork	7.3	−1	−2.9	.014
Furniture	13.4	2	2.2	.022
Miscellaneous lumber	12.9	4	1.7	.034
Blank paper	6.6	−17	−6.2	.524
Cardboard boxes	13.6	2	3.1	.031
Stationery	7.5	−2	−3.0	.030
Miscellaneous paper	9.3	−1	−1.1	.005
Newspapers	20.1	37	8.5	1.570
Books and music	14.6	2	4.3	.042
Miscellaneous printing and publishing	18.6	1	5.6	.028
Crude chemicals	10.2	−0	0	0
Paints	14.6	5	3.3	.082
Petroleum refining	8.4	−114	−3.6	2.032
Proprietary preparations	20.9	25	11.7	1.460
Toilet preparations	30.4	3	15.0	.225
Cleaning preparations	20.8	15	5.5	.413
Miscellaneous chemicals	15.6	45	8.8	.197
Ceramics	10.8	1	1.0	.005
Glass	13.5	4	2.6	.052
Portland cement	14.3	10	8.4	.420
Miscellaneous clay and stone	17.6	14	8.0	.560
Castings and forgings	5.6	−234	−7.7	8.994
Sheet metal	10.5	0	0	0
Wire and nails	11.6	1	1.2	.006
Heating machinery	13.3	3	1.6	.024
Electrical machinery	15.7	48	5.3	1.281
Textile machinery	13.6	3	6.1	.092
Printing machinery	9.7	−0	0	0
Road machinery	17.3	10	6.8	.374
Engines	13.7	2	5.9	.059
Mining machinery	11.0	1	0.7	.004
Factory machinery	11.7	33	3.0	.045
Office machinery	16.1	7	5.6	.194
Railway equipment	6.0	−24	−9.6	1.148
Motor vehicles	18.5	161	4.4	3.878
Firearms	12.9	1	2.0	.010
Hardware	12.8	8	2.3	.092
Tools	11.6	1	1.1	.006
Bolts and nuts	15.4	1	3.1	.016
Miscellaneous machinery	12.6	3	2.2	.032
Nonferrous metals	11.9	15	1.4	.106
Jewelry	10.6	0	0	0
Miscellaneous metals	12.5	14	2.0	.140
Scientific instruments	21.2	20	11.6	1.163
Toys	15.0	1	3.2	.016
Pianos	9.9	−0	0	0
Miscellaneous special manufacturing	12.0	4	1.4	.027
Job printing	13.8	4	2.2	.044

Col. (1)—from Ralph C. Epstein, *Industrial Profits in the United States* (N.B.E.R., 1934), Tables 43*D* through 53*D*. Entries in column (1) are the arithmetic means of the annual entries in the source tables.

Col. (2)—divergences in the profit rates given in column (1) from their mean (10.4) are here applied to the 1928 volume of capital in each industry. Total capital is the sum of book capital (Epstein, Appendix Table 6*C*) plus bonded debt (Epstein, Appendix Table 6*D*).

Col. (3)—1928 figures were used for sales (Epstein, Appendix Table 6*A*).

Col. (4)—measures the amount by which consumer "welfare" fell short of the level it would have attained if resources had been so allocated as to give each industry an equal return on capital. It assumes that the elasticity of demand for the products of each industry is unity and approximates the area designated as "welfare loss" in Figure 1.

1 Arnold C. Harberger, Papers and Proceedings of the Sixty-sixth Annual Meeting of the American Economic Association. 44 Am. Econ. Rev. 2, 77-87 (May, 1954). The following footnote is from the original: "I am indebted to my colleagues D. Gale Johnson, H. Gregg Lewis, and George S. Tolley for stimulating discussions and comments during the preparation of this paper. They are, of course, not responsible for errors that may remain."

2 Harold Hotelling, *The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates*, Econometrica, July, 1938, pp. 242-269. The applicability of Hotelling's proof to the present problem can be seen by referring to p. 252 ff. He there indicates that he hypothecates a transformation locus which is a hyperplane. This is given us by our assumption of constant costs. He then inquires what will be the loss in moving from a point Q on the hyperplane, at which the marginal conditions of competitive equilibrium are met, to a point Q' at which these conditions of competitive equilibrium are not met. At Q' a nonoptimal set of prices P' prevails. These are, in our example, actual prices, while the equilibrium price-vector P is given by costs, defined to include normal profits. Hotelling's expression for the welfare loss in shifting from Q from Q to Q' is $\frac{1}{2} \Sigma dp_i dq_i$ where p_i and q_i are the price and quantity of the i-th commodity. We obtain this by defining our units so that the cost of each commodity is $1.00. The equilibrium quantity of each commodity under the assumption of unit elasticities is then equal to the value of sales of that commodity. If we call r_i the percentage divergence of actual price from cost, we may write the total welfare loss due to monopoly as $\frac{1}{2} \Sigma r_i^2 q_i$ if the elasticities of demand are unity, and as $\frac{1}{2} \Sigma r_i^2 q_i k_i$, if the elasticities of demand are k_i. In column *1* of Table 1, I attribute to each commodity a welfare loss equal to $\frac{1}{2} r_i^2 q_i$. This measure of the welfare loss due to monopoly abstracts from distributional considerations. Essentially it assumes that the marginal utility of money is the same for all individuals. Alternatively, it may be viewed as measuring the welfare gain which would occur if resources were shifted from producing Q' to producing Q, and at the same time the necessary fiscal adjustments were made to keep everybody's money income the same.

3 Epstein, *op. cit.*, p. 530.

4 Epstein's results in samples from small corporations (not included in his main sample) indicate that their earnings rates tend to be quite close, industry by industry, to the earnings rates of the large corporations in the main sample. This suggests that the average rate of profit in the main sample (10.3 per cent) was higher than the average for all industry (8 per cent) because high-profit industries were overweighted in the sample rather than because the sampled firms tended to be the high-profit firms within each industry. The overweighting of high-profit industries affects our estimate of the welfare cost of resource misallocations in two ways. First, quite obviously, it tends to overstate the cost by pretending that the high-profit industries account for a larger share of the aggregate product of the economy than they actually do. Second, and perhaps not so obviously, it tends to understate the cost by overstating the average rate of profit in all manufacturing, and hence overstating the amount of profit which is "built in" to the cost curves in the present analysis. The estimated adjustment of 10 million dollars presented in the text corrects only for this second effect of overweighting and is obtained by imputing as the normal return to capital in the Epstein sample only 8 per cent rather than 10.4 per cent and recomputing the welfare costs of resource misallocations by the method fol1owed in Table 1. It takes no account of the first effect of overweighting, mentioned above, and thus results in an overstatement of the actual amount of welfare cost.

5 The extent of the bias is proportional to the difference between the average of the squares of a set of numbers and the square of the average, the numbers in question being the rates of excess profit in the subindustries. Consider an industry composed of three subindustries, each of equal weight. Assume, for an extreme example, that the rates of excess profit (excess profit expressed as a per cent of sales) are 10 per cent, 20 per cent, and 30 per cent in the three subindustries. The average rate of excess profit of the aggregate industry would then be 20 per cent, and, by our procedure, the estimate of the welfare loss due to that industry would be 2 per cent of its sales. If we had been able to deal with the hypothetical subindustry data directly, we would have estimated the welfare loss associated with them at 2 1/3 per cent of the aggregate sales.

6 I was unable similarly to take account of selling costs other than advertising expenditures, even though some of such costs may be the price paid by firms to enhance market control or monopoly position. In principle, clearly, some share of selling costs should be taken into account, and it is a limitation of the present study that no adjustment for such costs was possible. Scrutinizing Table 1, however, I should suggest that such selling costs are important in only a few of the industries listed, and that an allowance for them would almost certainly not alter the general order of magnitude of the estimates here presented. It should be pointed out, also, that the general conclusions reached in this paper are not closely dependent on the precise data used. Suppose, for example, that we had observed the following situation: industries accounting for half the output of American manufacturing were charging prices which yielded them a 10 per cent "monopoly profit" on sales, while the remainder of industries earned a constant rate of profit on capital (here called normal profit) but no more. If we were, in this situation, to reallocate resources so as to equalize profit rates in all industries, the prices of competitive products would rise and those of monopolistic products would fall. If demand for the product of each sector were assumed to be of unit elasticity, we would estimate the gain in welfare incident upon the reallocation of resources at .125 per cent of total industrial sales. This would be just about a tenth of a per cent of the national income if the ratio of manufacturing sales to national income approximated its 1921-28 figure. The estimated welfare gain is obtained as follows: Under our elasticity assumption, prices would rise by 5 per cent in the competitive sector and fall by 5 per cent in the monopolistic sector, and quantities would change inversely by an equal percentage. Taking 100 as the aggregate sales of manufacturing, the change in output in each sector will be 2.5, and taking 1 as the index of initial prices in each sector, the change in price in each sector will be .05. According to the Hotelling formula, the welfare gain coming from each sector will be 1/2 (2.5) (.05), and when these gains are added together the aggregate gain turns out to be .125.