CATO PAPERS ON PUBLIC POLICY

VOLUME 1 • 2011

CATO INSTITUTE

CATO PAPERS ON PUBLIC POLICY

Volume 1, 2011

CATO PAPERS ON PUBLIC POLICY

Volume 1, 2011

CATO INSTITUTE

Washington, D.C.

Cato Papers on Public Policy, ISBN-10: 1-935308-48-3; ISBN-13: 978-1-935308-48-5. *CPPP* is published annually by the Cato Institute, a nonprofit, nonpartisan 501(c) (3) research organization based in Washington, D.C.

Correspondence regarding subscriptions, changes, of address, procurement of back issues, advertising and marketing matters, and so forth, should be addressed to:

Publications Department
Cato Institute
1000 Massachusetts Ave., N.W.
Washington, D.C. 20001

All other correspondence, including requests to quote or reproduce material, should be addressed to the editor.

To subscribe to *CPPP*, visit www.cato.org/store or telephone (800) 767-1241.

Printed in the United States of America.

Cato Institute
1000 Massachusetts Ave., N.W.
Washington, D.C. 20001
www.cato.org

Publications Director: *David Lampo*
Marketing Director: *Robert Garber*
Circulation Manager: *Alan Peterson*
Cover: *Jon Meyers*

Contents

An Introduction to the *Cato Papers on Public Policy* and Annual Conference

This is the first volume of what will be an annual *Cato Papers on Public Policy*. The goal of the *Papers* and the associated annual conference is to produce new, high-quality research on public policy and to make this research available to a broad audience consisting of academics, policymakers, and journalists.

The research intends to fill a gap in the work that addresses the pros and cons of government policies. Academics produce significant research that analyzes public policies, but much of that work is abstract, technical, and not immediately relevant to real-world policy debates. The *Cato Papers* will evaluate significant economic or social policies using the techniques of modern economics. The papers will be less technical, on average, than a standard journal article, but more technical than a typical policy analysis. In a nutshell, the papers will aim to produce research that employs modern economic methodology but that is firmly focused on what policies are beneficial for the economy and society.

Jeffrey Miron
Editor, *Cato Papers on Public Policy*
Senior Fellow, Cato Institute
Senior Lecturer and Director of Undergraduate Studies, Harvard University

Can the Treasury Exempt Its Own Companies from Tax? The $45 Billion GM NOL Carryforward

J. Mark Ramseyer
Eric B. Rasmusen

ABSTRACT

To discourage firms from buying and selling tax deductions, Section 382 of the tax code limits the ability of one firm to use the ''net operating losses'' (NOLs) of another firm that it acquires. Under the Troubled Asset Relief Program, the U.S. Treasury lent a large amount of money to General Motors. In bankruptcy, it then transformed the debt into stock. GM did not make many cars anyone wanted to buy, but it did have $45 billion in NOLs. Unfortunately for the Treasury, if it now sold the stock it acquired in bankruptcy, it would trigger Sec. 382. Foreseeing this, the market would pay much less for its stock in GM.

Treasury solved this problem by issuing a series of notices in which it announced that the law did not apply to itself. Sec. 382 says that the NOL limits apply when a firm's ownership changes. That rule would not apply to any firm bought with TARP funds, declared Treasury. Notwithstanding the straightforward and all-inclusive statutory language, GM could use its NOLs in full after Treasury sold out. The Treasury issued similar notices about Citigroup and AIG.

Treasury had no legal or economic justification for any of these notices, but the press did not notice. Precisely because they involved such arcane provisions of the corporate tax code, they largely escaped public attention. The losses to the public fisc were not minor—they cost the country billions of dollars in tax revenue. That the effect could be so large and yet so hidden illustrates the risk involved in this kind of tax manipulation. The more difficult the tax rule, the more easily the government can use it to hide the cost of its policies and subsidize favored groups. We suggest that Congress give its members standing to challenge unlegislated tax law changes in court.

J. Mark Ramseyer is the Mitsubishi Professor of Japanese Legal Studies at Harvard Law School. Eric B. Rasmusen is the Dan R. and Catherine M. Dalton Professor in the Department of Business Economics and Public Policy of the Kelley School of Business at Indiana University.

We thank William Allen, Andrew Atkeson, Frank Buckley, Michael Doran, Sally James, Victor Fleischer, Michael Schler, and participants in seminars at the online Cyprosia, the Cato Institute, and the Harvard Law School for their many comments, whether positive or negative. We do not imply that any of these generous readers agree with our conclusions.

Can the Treasury Exempt Its Own Companies from Tax? The $45 Billion GM NOL Carryforward

''Dona clandestina sunt semper suspiciosa.''[1]

1. INTRODUCTION

Year after year, General Motors lost money—enormous sums of money. It designed cars. It built cars. But no one wanted to buy the cars. Over time, it accumulated huge operating losses (''net operating losses,'' or NOLs). The tax code let GM carry forward these NOLs into the future. It let the firm save the losses for that day in the future when it would once again sell cars that people wanted.

The day never came. Instead, in June 2009 GM (call it ''Old GM'') declared bankruptcy. It filed under Chapter 11 of the Bankruptcy Code and sold its assets to a new shell (''New GM'') in a transaction governed by Section 363 of the Code. Old GM's shareholders lost

[1] ''Secret gifts are often suspicious.'' From Sir Edward Coke, *Twyne's Case*, 3 Coke, 80 b (Star Chamber, 1602), in *Cases on the Law of Bankruptcy: Including the Law of Fraudulent Conveyances*, ed. E. Holbrook and R. W. Aigler, 153–157 (Chicago: Callaghan, 1915). *Twyne's Case* was about a fraudulent conveyance by an insolvent debtor to a friendly creditor.

Another passage from the case will be apt when we consider the relationship between statute and regulation:

> To one who marvelled what should be the reason that Acts and statutes are continually made at every Parliament without intermission, and without end; a wise man made a good and short answer, both of which are well composed in verse.
>
> *Quaeritur, ut crescitunt tot magna volumina legis?*
>
> *In promptu causa est, orescit in orbe dolu.*
>
> [In our inexpert translation: ''It might be asked why such a large amount of law grows? The basic reason is that the world's evil has grown.'']
>
> And because fraud and deceit abound in these days more than in former times, it was resolved in this case by the whole Court, that all statutes made against fraud should be liberally and beneficially expounded to suppress the fraud.

their investment. They did not receive stock in New GM. Instead, Old GM's creditors became New GM's stockholders: the U.S. Treasury (with 61 percent), the auto unions, and Canada swapped debt claims against Old GM for equity stakes in New GM. Other Old GM creditors acquired a 10 percent stake in New GM as well. In the fall of 2010, the Treasury re-sold a large amount of its New GM shares to the public, cutting its share to 26 percent.

New GM has the factories, offices, designs, and some of the workers that Old GM had. It also acquired some $18 billion worth of Old GM's NOLs.[2] New GM could not use them to reduce its tax liability immediately, since it was losing money. But in 2010, New GM did turn a profit and presumably will use its NOLs to avoid corporate income tax on that profit (Bunkley 2011).

Ordinarily, when one company buys another's assets, it does not acquire its tax losses too. But the sale from Old GM to New GM qualified as a tax-free ''reorganization'' under Sec. 368 of the tax code: neither Old GM nor New GM incurred a tax liability, New GM entered Old GM's assets on its books with Old GM's ''adjusted basis,'' and New GM acquired Old GM's NOLs.

The problem involved Treasury's plans to sell the shares it took in New GM. If the combined equity stake of any group of shareholders in a ''loss corporation'' like New GM climbs by more than 50 percentage points, Sec. 382 of the tax code limits the firm's ability to use those accumulated NOLs. Given Treasury's large stake in New GM, if it sold its entire stake to the public, those new owners would raise their combined interest by 50 points. New GM would then lose its ability to avoid taxes on future income.

[2] The losses themselves were $45 billion; their book value as an asset is listed as $18 billion. We will use the figure $18 billion even though it is too high because standard accounting rules for tax assets are absurdly inaccurate.

They are inaccurate for two reasons: First, Generally Accepted Accounting Principles require companies to not discount for the time value of money. If a company expects to save $1 million in taxes in 16 years using deferred tax losses, it records that as a current tax asset worth $1 million, even though the present discounted value (at 5 percent interest) is only $458,000. Second, even if there is a good chance that the company will never make a profit again, it records the full amount if ''it is more likely than not'' that the company will someday make enough profit. Thus, if the company just mentioned estimated that its chances of failure before 16 years from now are a mere 49 percent, it would still record the $1 million as $1 million, not $510,000 or $233,580. For a critical view of this rule, see J. E. Ketz, ''Deferred Income Taxes Should Be Put to Rest,'' *SmartPros*, March 2010.

To solve this problem, the Treasury issued a series of notices. The Sec. 382 rules, it declared, would not apply to itself. When it sold its shares in New GM, the new owners might increase their ownership stake by 50 percentage points, but they would not trigger the Sec. 382 limits. The tax code offered no exception for government-owned shares, and the Treasury did not purport to find one. Instead, it just declared that the law did not apply.[3]

The notices also apply to two other companies, AIG and Citigroup. Both of these companies had ownership changes over 50 percent as a result of the Troubled Asset Relief Program and would ordinarily, as in bankruptcy, lose their NOLs. If they retain them, that reduces the apparent (but not real) cost of the bailout because the government can resell its shares at a higher price.

Through these notices, Treasury accomplished two highly political goals:

- It disguised (by billions of dollars) the true cost of the bailouts of GM and other firms.
- It routed funds (again, several billion dollars) to the administration's supporters at the UAW.

Ordinarily, if an administration wildly misstates the cost of its policies or routes public funds to its friends, the press notices and complains. In this case, it did not. The press missed the manipulation precisely because it involved such a complex and highly arcane provision of the tax code. The more obscure the law, in other words, the greater the risk of political manipulation: precisely because its strategy involved such an *abstruse* corner of the law, the administration was able to hide its politicized policies from the public.

We do not address the wisdom of the bailouts themselves. Neither do we ask whether firms should be able to carry forward operating losses, whether they should be able to reorganize tax-free, or why the United States has a corporate income tax at all.[4] These are all

[3] The last of the notices was Internal Revenue Service Notice 2010-2, ''Application of Section 382 to Corporations Whose Instruments Are Acquired and Disposed of by the Treasury Department under Certain Programs Pursuant to the Emergency Economic Stabilization Act of 2008,'' *Internal Revenue Bulletin* 251.

[4] Two recent articles on the incidence and distortions due to the corporate income tax are Harberger (2008) and Kotlikoff and Miao (2010). Auerbach, Devereux, and Simpson (2010) survey the pros and cons of corporate income taxes and the various ways to structure them. Their unavoidable complexity, of which the present paper's subject is just one example, is one strong argument against corporate income taxes.

interesting questions, but we have quite enough to do addressing the topic of selective tax relief through executive decree. Rather than explore these larger questions, we focus on the propriety of the Treasury's manufacturing a tax break to distribute and hide government largesse. More generally, we focus on the wisdom of giving a president the ability to invent a tax deduction for his political supporters without a need to answer to the courts or Congress.

1.1 The Bad Man and the Law

Recall Justice Holmes's description of the law as being the prediction of the ''Bad Man'' about whether a judge would stop him:

> If you want to know the law and nothing else, you must look at it as a bad man . . . who cares only for the material consequences which such knowledge enables him to predict, not as a good one, who finds his reasons for conduct, whether inside the law or outside of it, in the vaguer sanctions of conscience. . . . If we take the view of our friend the bad man, we shall find that he does not care two straws for the axioms or deductions, but that he does want to know what the Massachusetts or English courts are likely to do in fact. I am much of this mind. The prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law. (Holmes 1897)

If a president is Holmes's Good Man, he will obey the Constitution because it is the Constitution. The Treasury gave General Motors an illegal tax break. As a Good Man, he will read our article, feel remorse, and fire everyone involved.

If a president is Holmes's Bad Man, on the other hand—and public choice theory suggests that it is Bad Men who have the best chance of being elected—he will obey the Constitution only when a court can make him obey it.[5] If he hears of our article, he will ignore it. As a lawyer, he knows that nobody has standing to challenge someone else's tax benefits in court. Thus, his ''prophecy about what a court will do'' is easy: nothing. The courts will reject any challenge for lack of standing, whatever the merits of a claim might be.

[5] A politician who upholds his personal principles and resists the will of the median voter or leaves untouched the less honorable tools of political competition will, *ceteris paribus*, lose votes and lose elections. For more explanation, see Ramseyer (1995).

Only potential bad publicity would worry a Bad Man president. But publicity he can skirt by giving the funds through opaque provisions of the tax code. Publicity he can skirt by (take a deep breath) declaring an exemption from the application of Sec. 382 of the tax code to limits on carryforwards of NOLs following a sale under Sec. 363 of the Bankruptcy Code that uses preferred stock, credit bidding, and warrants by one company named GM to a different company also named GM. If the administration gave a billion dollars in cash to its supporters, the press would notice. If it gives it through the obscure details of the corporate tax code, the press will fall asleep.

In the article that follows, we explain the intricacies of the tax break (Section 2). We discuss the law involved (Section 3). If you think all presidents are Good Men, you may stop reading at that point. After all, following the Constitution is just a matter of understanding it. We explain it, you understand it, end of story. Lest some presidents be Bad Men, however, we conclude by exploring procedural reforms Congress might adopt to prevent a recurrence of what happened with GM.

2. WHAT HAPPENED

General Motors was a public corporation with much unsecured debt, including $21 billion owed to the UAW Trust on behalf of retired workers and $27 billion owed to bondholders. None of these stakeholders was senior enough to see much return if the company liquidated in pieces. Probably, none would see much return even if the firm found a buyer for the whole company.

The senior creditors were a diverse lot. The U.S. Treasury had a secured interest in $19.4 billion from TARP loans and $30.1 billion in other loans. The Canadian government held secured claims of $9.2 billion. Government senior debt thus totaled $58.7 billion. Private creditors held another $5.9 billion in secured loans.

GM filed for bankruptcy under Chapter 11 of the Bankruptcy Code. To restructure its finances, it then negotiated a sale under Sec. 363 of the Code. For this transaction, it formed a new shell, New GM. Old GM then sold its assets to New GM. In exchange for its $21 billion *un*secured debt to Old GM, the UAW Trust received 17.5 percent of the common stock of New GM, $6.5 billion in preferred stock, and $2.5 billion in debt. In exchange for their $27 billion of unsecured debt, the other junior creditors received 10 percent of the

common stock of New GM and warrants for another 15 percent. The private secured creditors (the $5.9 billion claim) were paid in full. The Canadian government received 12 percent of the New GM common stock, and the U.S. Treasury received interests detailed shortly below.

To consider the stakes involved, note that in December 2010, New GM had stock worth $54.4 billion and liabilities of $12.9 billion (Ceraso, Moffatt, and Pati 2010), for a total asset value of $67.3 billion. In effect, the sale price in the 363 offer was:

- $58.7 billion in senior credit claims,
- $5.9 billion paid to private secured creditors,
- $5.4 billion in stock (10 percent of $54.4 billion), and
- a portfolio of harder-to-value warrants.

This yields a total of $67 billion plus warrants (Warburton 2010, p. 536).

Apparently, the 363-sale buyers paid $67 billion plus the warrant value for assets worth $67.3 billion. That seems a remarkably high price considering that no other bidder loomed on the horizon. The bankruptcy judge deserves praise for extracting so much value for Old GM's creditors.

This $67.3 billion in asset value is not the net benefit to the 363-sale buyers or the senior creditors, however. That benefit depends on who owns the New GM equity and debt. Old GM's private secured creditors received $5.9 billion in cash for their $5.9 billion in debt. The Canadian government gave up its $9.2 billion in Old GM debt but took a 12 percent stake in the common stock (worth $0.12 \times \$54.4$ billion $= \$6.5$ billion) plus $0.4 billion in preferred stock and $1.3 billion in debt in New GM—for a total value of $8.2 billion.

The most glaring anomaly involved the UAW. The union's trust gave up *un*secured claims of $21 billion and received:

- 17.5 percent of the stock of New GM worth ($0.175 \times \$54.4$ billion $=$) $9.5 billion,
- $6.5 billion in preferred stock, and
- $2.5 billion in debt,

for a total of $18.5 billion. Given that the UAW Trust had been a junior creditor, this was a very good deal. By contrast, the other

unsecured creditors gave up claims of $27 billion and received only 10 percent of the common stock and warrants.

Recall that the U.S. Treasury held secured debt totaling $49.5 billion. In exchange for its claims, it took 61 percent of the stock in New GM (stock worth 0.61 × $54.4 billion = $33.2 billion), $2.1 billion in preferred stock, and a $6.7 billion debt claim against New GM. All told, it received compensation of $42 billion.

Focus on the U.S. government. Through the Sec. 363 sale, it—apparently—lost ($49.5 billion − $42 billion =) $7.5 billion. Anyone who loses only ($7.5 billion ÷ $49.5 billion =) 15 percent on a $49.5 billion loan to a failing firm does well indeed. Yet appearances deceive. The government also gave GM investors $45 billion in NOLs. If the 363 sale had not gone through or the sale had been made to some outside buyer, these NOLs would have disappeared. The book value of these NOLs is $18 billion.

To be sure, Treasury was giving tax breaks partly to itself, and the book value of the NOLs exceeds their market value since it would take some years before GM could exhaust them. If the market value of the NOLs were, say, $12 billion (a little under the estimate of the stock analysts that we cite in Section 2.1 below), then that $12 billion was incorporated into the $54.4 billion equity value of the New GM, and we have overestimated the overall value of the deal for the Treasury. Of its $33.2 billion in stock, $7.32 billion (= 0.61 × $12 billion) was a tax gift to itself.

More simply, consider the $12 billion worth of NOLs an additional loss to the Treasury. In effect, the Treasury lent GM $49.5 billion and lost ([$7.5 billion + $12 billion] ÷ $49.5 billion =) 39 percent. If only Treasury could have inserted a further secret $20 billion of assets into New GM, New GM's stock price would have been so high that Treasury would have appeared to make a profit from the entire affair.

2.1 As GM Told It

Here is how GM describes its tax situation:

> We recorded valuation allowances against certain of our deferred tax assets, which under ASC 852 also resulted in goodwill. (General Motors 2010, p. 82)
>
> In July 2009 with U.S. parent company liquidity concerns resolved in connection with the Chapter 11 Proceedings and

> the 363 Sale, to the extent there was no other significant negative evidence, we concluded that it is more likely than not that we would realize the deferred tax assets in jurisdictions not in three-year adjusted cumulative loss positions.
>
> Refer to Note 22 to our audited consolidated financial statements for additional information on the recording of valuation allowances. (General Motors 2010, p. 138)

Table 1 from New GM's securities filings (p. F-121 of its Form 8-K) shows that New GM claimed to inherit over $18 billion in tax carryforwards from Old GM.[6] Stock analysts wrote:

> We calculate an NPV of GM's deferred tax assets at $17.2bn of which $4bn is related to pension contributions and more than $13bn related to accumulated NOLs and tax credits including R&D credits. (Morgan Stanley 2010)

and

> Via a special regulation, GM's highly valuable US tax assets (worth $18.9B in the US at 09-end) were left intact. . . . Our Dec-2011 price target assumes a present value of $12.4B of (2011-ending) non-European global tax assets. . . . Present-valuing the $18.6B face value figure using a 12 percent discount rate (Ford is 8 percent; we use 12 percent for GM to reflect the lower mix of debt in its cap structure), we arrive at a PV for global economic tax assets ex. Europe of $12.4B at 2011-end. (J. P. Morgan 2010)

Thus, stock analysts were well aware of the existence and value of the NOLs, though they estimated their economic value at lower than their accounting value. This is an important element of the political economy of the situation. It was crucial both that the general public not realize that New GM's value was inflated by the taxes that the Treasury had agreed in advance to forgive, and that stock analysts did understand it. If the analysts missed the point, then when the government sold its GM stock, it would have received a much lower price. It would have given away government revenue,

[6] Not all these tax carryforwards were necessarily NOLs, strictly speaking. They may also include ''built-in losses'' on assets that declined in value and unused tax credits.

Table 1
Components of GM's Temporary Differences and Carryforwards That Give Rise to Deferred Tax Assets and Liabilities

	Successor		Predecessor	
	December 31, 2009 Deferred Tax		December 31, 2008 Deferred Tax	
	Assets	Liabilities	Assets	Liabilities
Postretirement Benefits Other Than Pensions	$4,194	—	$11,610	—
Pensions and Other Employee Benefit Plans	$8,876	$406	$16,171	$8,648
Warranties, Dealer and Customer Allowances, Claims and Discounts	$3,940	$75	$6,682	$90
Property, Plants, and Equipment	$7,709	$278	$7,429	$3,197
Intangible Assets	$1,650	$4,984	$780	—
Tax Carryforwards	$18,880	—	$18,080	—
Miscellaneous U.S.	$5,844	$1,269	$8,122	$288
Miscellaneous non-U.S.	$3,306	$1,944	$3,485	$773
Subtotal	$54,399	$8,956	$72,359	$12,996
Valuation Allowances	$(45,281)	—	$(59,777)	—
Total Deferred Taxes	$9,118	$8,956	$12,582	$12,996
Net Deferred Tax Assets (liabilities)	$162		$(414)	

Source: General Motors Form 8-K (2010), p. F-121.

but without disguising the cost of its bailout—approximately halving it from $24 billion to $12 billion (Terlap 2011).

2.2 Other Firms

Although we focus on GM, Treasury gave legally unauthorized NOLs to two other firms as well. As with GM, it did this by issuing TARP-specific notices about the availability of NOLs. Citigroup, for example, claimed ''tax assets'' of $46.1 billion at the end of 2009. In June 2009, Citigroup and the Treasury agreed to exchange the government's preferred stock for common stock. The government acquired a 33.6 percent ownership stake. In December 2009, Citigroup raised $20.3 billion by issuing about 24 percent new common stock, so Citigroup had passed the threshold for a 50 percent ownership change. In 2010, Treasury sold all of its 7.7 billion shares of common stock for $31.85 billion, a gain of $6.85 billion. According to Citigroup:

> The common stock issued pursuant to the exchange offers in July 2009, and the common stock and tangible equity units issued in December 2009 as part of Citigroup's TARP repayment, did not result in an ownership change under the Code. (Murphy 2010)

By ''ownership change,'' it referred to the Sec. 382 rule detailed in Section 3 below. It based its claim that the section did not apply to it on the Treasury's notices.

For Citigroup, the NOLs had additional importance because of its status as a bank. Banks must worry about regulatory capital requirements. As Davidson (2011) explains:

> Banks hold NOLs as deferred tax assets (DTA's). DTA's, in turn, constitute a portion of a bank's tier 1 capital. Were Citigroup to have lost its ability to use its NOLs, it might have had to write down its tier 1 capital.

A footnote adds:

> 12 C.F.R. sec. 225 at appendix A.II.A.1. NOLs may constitute up to 10 percent of tier 1 capital, to the extent that the institution ''is expected to realize [a tax deduction by their use] within one year . . . based on its projections of future taxable revenue for that year.''

After many travails, in January 2011 AIG completed a reorganization that gave Treasury 92.1 percent of its common stock. AIG claimed ''Deferred tax assets: Losses and tax credit carryforwards'' of $26.2 billion at the end of 2009. It claimed other valuable tax attributes as well,[7] including ''Unrealized loss on investments'' of $8.7 billion (AIG 2009, p. 334). These, too, hinged on notices exempting the firm from the coverage of Sec. 382. AIG acted on the assumption that it had not yet had an ''ownership change'' for tax purposes. It was worried enough about a private-market 50 percent ownership change that would trigger Sec. 382, however, that it installed a poison pill to prevent large share purchases.

3. THE LAW

In fact, the law—arcane in the extreme—does not grant New GM the NOLs it claims if the government sells its shares. Neither does it grant Citigroup and AIG any right to the tax assets they claimed. To be sure, the law lets the GM NOLs survive the Sec. 363 sale in bankruptcy, as we will show. To that extent, New GM did inherit the NOLs. It can continue to use them, however, only so long as the Treasury holds its stock. Once Treasury sells its shares to the public, New GM should by statute lose its access to most if not all of the loss carryforwards.

New GM did claim the NOLs and the Treasury concurred. For 2010, New GM had access to the losses because the government had not yet sold enough of its stock. But once it sells, New GM will be able to claim the losses only because the Treasury told New GM it could. Through a series of notices, it declared that the statutory limitations on the use of NOLs after a defined ''ownership change'' did not apply if the Treasury owned the stock. The statute itself did not differentiate between government and nongovernment owners. Nonetheless, as we will explain in detail later, Treasury wrote that New GM could continue to claim the NOLs after it sold its stock, and New GM happily deferred.

First, however, we must go into how New GM could possibly acquire the NOLs in the first place. The law is massively opaque,

[7] According to AIG (2009), ''The application of U.S. GAAP requires AIG to evaluate the recoverability of deferred tax assets and establish a valuation allowance, if necessary, to reduce the deferred tax asset to an amount that is more likely than not to be realized (a likelihood of more than 50 percent).''

but that is the point. Precisely because the corporate tax rules are as complex as they are, the administration could successfully deflect attention from what it did.

3.1 Cancellation of Indebtedness and Net Operating Losses

Consider the tax treatment of cancelled debt, relevant here because of the cancellation of Old GM's debt to the Treasury. Suppose a firm has debt outstanding. It negotiates with its creditors and they agree to trade their debt claims for stock. The firm will have cancellation of indebtedness (COD) income equal to the difference between the face amount of the cancelled debt and the market value of the stock distributed (I.R.C. Secs. 61, 108(e)(8); *U.S. v. Kirby Lumber Co.*, 284 U.S. 1 (1931)).

Now suppose the firm is insolvent. If its creditors swap their claims for stock, under general tax principles it will have COD income. In fact, however, the Internal Revenue Code provides that what would otherwise be COD income will not constitute taxable income. Instead, under Sec. 108 of the code, the firm will need to reduce the amount of its other ''tax attributes'' by the amount of the COD income excluded. Most relevant here, it will need to reduce the amount of its NOLs by the amount of the excluded income. Given that $1 of NOL would reduce net taxable income by $1, this obviously leaves the firm (in many cases) in much the same position as if it had included the COD income all along (I.R.C. Sec. 108(a)(1)(B), (b)(2)(A)).

Finally, suppose the firm is solvent but files for reorganization under bankruptcy. If, as part of its bankruptcy reorganization, the creditors swap their claims for stock, the result (for purposes here) is the same as if the firm were insolvent. Under Sec. 108, it can exclude the COD from income, but it must offset the excluded amount against its NOLs (I.R.C. Sec. 108(a)(1)(A), (b)(2)(A)).

3.1.1 Tax Reorganizations

Many reorganizations under the bankruptcy code also constitute ''reorganizations'' under the tax code. If, but only if, a transaction qualifies as a ''reorganization'' under the tax code, a firm that takes the assets of another firm may also take its NOLs. Note that although both the bankruptcy and the tax codes use the term ''reorganization,'' the word refers to different concepts in each. Those concepts are not interchangeable.

In general, reorganizations in bankruptcy are ''G reorganizations'' under the tax code, meaning that they fall under Sec. 368(a)(1)(G) of the Internal Revenue Code:

> [A] transfer by a corporation of all or part of its assets to another corporation in a title 11 or similar case; but only if . . . stock or securities of the corporation to which the assets are transferred are distributed in a transaction which qualifies under section 354, 355, or 356.

Note two points relevant here: First, ''Section 363 sales'' occur in a ''title 11 or similar case.'' ''Title 11'' (not ''Chapter 11'') refers to the Bankruptcy Code, and ''section 363'' refers not to Sec. 363 of the tax code but to Sec. 363 of the Bankruptcy Code. As a result, if a ''debtor in possession'' (a bankruptcy concept) sells its assets under Sec. 363, it sells its assets in a Title 11 case. The court of *In re Motors Liquidation Co.*, 430 B.R. 65 (S.D.N.Y. 2010) explicitly indicated that a Sec. 363 sale (indeed, exactly the GM sale at issue here) could constitute a qualifying G reorganization. This is the position the Treasury has long taken as well (e.g., in Ltr. 8503064 (Oct. 24, 1984); Ltr. 8521083 (Feb. 27, 1985)).

Second, Sec. 354 of the tax code requires merely that *some* security holders (not *only* security holders) of the old firm receive ''stock or securities'' of the new firm. I.R.C. Sec. 354(a) provides:

> No gain or loss shall be recognized if stock or securities in a corporation a party to a reorganization . . . are . . . exchanged solely for stock or securities in . . . another corporation a party to the reorganization.

Suppose the creditors to the old firm include both long-term bond holders and trade creditors. Suppose both receive stock in the new firm. The former held ''securities'' in the old firm, but the latter did not (i.e., bonds are securities, trade credit is not). For at least three decades, the Treasury has taken the position that the transaction qualifies under Sec. 354 even though some of the stock goes to creditors who did not hold securities. Instead, it has argued that a transaction qualifies under Sec. 354 if at least one of the old firm

creditors who received stock in the transaction held a security of the old firm.[8]

3.1.2 Net Operating Losses

Only in a qualifying tax reorganization will a firm that acquires the assets of another also acquire its NOLs. Suppose again that a firm induces its creditors to swap their claims for stock. Suppose further that some NOLs remain after the Sec. 108 adjustments detailed earlier.

Generally, if a debt-for-stock swap occurs as part of a transaction in which a firm sells its assets to another firm, the acquiring firm will not obtain its NOLs, too. After all, the losses are specific to the selling firm. The acquiring firm buys the seller's assets, but it does not—indeed, legally cannot—buy its ''tax losses.'' Conceptually, these tax attributes describe the financial characteristics of a firm; they are not ''things'' that firms can buy and sell.

Under Sec. 381 of the tax code, however, if one firm buys the assets of another firm in a qualifying tax ''reorganization,'' it also acquires its NOLs. More specifically, Sec. 381(a) provides:

> In the case of the acquisition of assets of a corporation by another corporation . . . in a transfer to which section 361 . . . applies, but only if the transfer is in connection with a reorganization described in subparagraph . . . (G) of Section 368(a)(1), the acquiring corporation shall succeed to . . . the items described in subsection (c) of the . . . transferor corporation.

Note two observations. First, if a firm exchanges its assets for stock as part of a G reorganization, Sec. 361 will apply to the exchange. In turn, that section specifies that the two firms recognize no gain or loss on the transaction. Second, Sec. 381(c)(1) lists ''net operating losses.'' Provided the debt-for-stock swap occurs in a G reorganization, an acquirer takes the seller's NOLs along with its assets.

[8] For examples, see ''Bankruptcy Tax Act of 1980: Report of the Committee on Ways and Means, U.S. House of Representatives on H.R. 5043'' (Washington: Government Printing Office, 1980); Ltr. 8503064 (Oct. 24, 1984); Ltr. 8521083 (Feb. 27, 1985); and see generally Pickerill (2009).

3.1.3 The Law Applied to GM

Now turn to the reorganization of GM. Insolvent, GM filed for reorganization in bankruptcy court in the Southern District of New York. It sold its assets to a newly formed corporation (New GM) in a Sec. 363 sale. In exchange, it received stock in the new firm that it distributed to its bond holders and other creditors.

Absent Sec. 108, GM would have had COD income equal to the difference between the amount of its debt and the value of the stock it distributed. We will see next, however, that in bankruptcy the rule may be different.

3.2 Change in Control

A firm that buys another firm's assets in a G reorganization cannot necessarily use the transferor's NOLs immediately. To limit ''trafficking'' in tax losses, Sec. 382 of the tax code limits a firm's ability to use the NOLs of a ''loss corporation'' that it buys (defined at Sec. 382(k)). The limits apply whenever one set of the loss corporation's shareholders sells over 50 percent ownership to another set within a three-year period.[9] And these limits then restrict the amount of the NOLs that the firm can use to a ''section 382 limitation'' amount:

> The section 382 limitation for any post-change year is an amount equal to-
> (A) the value of the old loss corporation, multiplied by
> (B) the long-term tax-exempt rate.

[9] I.R.C. Sec. 382(g)(l). The statute says an ownership change is triggered by an increase of 50 percentage points by a 5-percent shareholder. The statute lumps all small shareholders together as a single fictitious 5 percent shareholder. Thus, if a 100 percent owner sells out entirely to small shareholders, that counts as an increase of over 50 percentage points by a 5 percent shareholder. If, however, the new small owners then trade 60 percent among themselves without anybody reaching 5 percent, that does not count.

The regulations clarify using examples. CFR Sec. 1.382-2T(j)(2)(iii)(B)(2), Example (3) says:

> L is entirely owed by Public L. L commences and completes a public offering of common stock on January 22, 1988, with the result that its outstanding stock increases from 100,000 shares to 300,000 shares. No person owns as much as five percent of L stock following the public offering. . . .
>
> New Public L is a 5-percent shareholder that has increased its ownership interest in L by more than 50 percentage points during the testing period (by 66 2/3 percentage points). Thus, there is an ownership change with respect to L.

Consider how this 382 scheme works. Suppose, first, that a *solvent* firm *not* in bankruptcy convinces its creditors to swap their debt claims for stock. It will recognize COD income. It will apply its NOLs against that income. And if any NOLs remain, then if one set of shareholders sells over 50 percent ownership to another, the firm will be able to use only the product of its earlier value and the long-term tax-exempt rate (I.R.C. Sec. 382(b)(1)).

Suppose, second, that a firm convinces its creditors to swap their claims for stock in a bankruptcy proceeding. As noted earlier, under Sec. 108 it will not recognize its COD as income but will reduce the amount of its NOLs by the amount of that excluded COD. Importantly, under some circumstances Sec. 382 will *not* thereafter limit its ability to use its NOLs *even if* there has been a Sec. 382 change in control. Instead, Sec. 382(l)(5) states that the limits do not apply if

- the transaction occurs in a Title 11 case, and
- ''the shareholders and creditors of the old loss corporation . . . own . . . stock of the new loss corporation'' equal to at least 50 percent (I.R.C. Sec. 382(l)(5)).

Potentially, NOLs could (only ''could''—even under (l)(5) the NOLs do not necessarily live) survive bankruptcy proceedings in full.

Suppose, third, that an insolvent firm does not file for bankruptcy but still induces its creditors to swap their debt claims for stock. Absent more, according to Sec. 382, its NOLs will disappear. They will disappear because the firm can thereafter only use a portion of its earlier value (''the value of the old loss corporation''), and Sec. 382 defines that earlier value as ''the value of the stock'' of the insolvent corporation (I.R.C. Sec. 382(e)(1)). Because the firm was insolvent, its stock was worth nothing (or nearly nothing). The product of the ''value of the old loss corporation'' and the ''long-term tax-exempt rate'' will fall to zero, and the NOLs will disappear.

Finally, suppose an insolvent firm does not meet Sec. 382(l)(5)'s 50 percent test. Provided it negotiates its debt-for-stock swap within a bankruptcy filing, under Sec. 382(l)(6) it may add to the value of the firm used to calculate the amount of annual useable NOLs the value created by canceling the creditors' claims. It can use each year, in other words, a proportional share not just of the value of the pre-reorganization firm but of that value plus any value attributable to the debt cancellation (I.R.C. Sec. 382(l)(6)).

3.2.1 The Law Applied to GM

After its Sec. 363 sale, the creditors of Old GM owned 100 percent of the stock of New GM. Under Sec. 382(l)(5), all of its NOLs may have survived. If the old creditors obtained less than 50 percent of the stock of New GM, then under Sec. 382(l)(6) New GM would have been able to use only an amount of NOLs calculated by adding the value of the canceled debt to the value of Old GM.

3.3 Later Control Shifts

Even for New GM, however, Sec. 382 created a risk. First, suppose that New GM tried to avoid the limits on its NOLs through Sec. 382(1)(5). If within two years of the reorganization, the stock owned by any set of 5 percent shareholders increased by 50 percentage points, then the NOLs disappeared. Subsec. (l)(5) couples its apparent generosity with a draconian penalty: if a firm meets the terms of (l)(5), it potentially enjoys the NOLs without the standard Sec. 382 reduction; but if it then shifts ownership within two years, it loses those NOLs entirely.

Second, even if New GM does not claim the Subsec. (l)(5) benefit, it still jeopardizes much of its NOLs if ownership changes. Suppose New GM claimed the benefit of Subsec. (1)(6) instead. If within three years one set of shareholders sells over 50 percent ownership to another, then the firm will be able to use only the ''section 382 limitation'' amount.

The problem for New GM lay in the fact that it exited its G reorganization with the U.S. government holding 61 percent of its stock. If the government recovers its investment by selling all of that stock within two years (for Subsec. (l)(5)), or three years (for Subsec. (l)(6)), it will probably cause an ownership change under the terms of Sec. 382. We say ''probably'' because we do not know how many other shareholders will trade during the same period. If it does trigger an ''ownership change,'' it will either face the Sec. 382 limits to its NOLs under Subsec. (l)(6) or lose its NOLs entirely under Subsec. (l)(5).

In November 2010, the Treasury did reduce its stake in GM from 61 percent to 33 percent. If Treasury, or any other large shareholder, transfers an additional 22 percent of the stock, GM will face the Sec. 382 limits on its net operating losses.

The cases of AIG and Citigroup are even clearer. Already, the government has triggered an ownership change in both companies. The Treasury acquired a majority of AIG's stock, and it acquired enough of Citigroup's stock that, combined with Citigroup's new

capital issue, it caused a 50 percent ownership change. Thus, by law, both firms should lose their NOLs.

3.3.1 The IRS Notices

If the Treasury lets a firm claim a NOL to which the law does not entitle it, Treasury merely gives the firm a gift. TARP does authorize Treasury to give gifts. As a result, the superficial choice would seem to be, if Treasury wants to enrich a firm, it can either give it money under TARP or let it take an extra NOL. Either way, it transfers funds from the public fisc to the firm.

To give funds under TARP, however, Treasury must follow statutory guidelines. It must give its gifts in amounts and to firms and for purposes described by Congress in the legislation. When it unilaterally authorizes NOLs, by contrast, it escapes all those congressional constraints.

And that is exactly what the Treasury did. From 2008 to 2010, it issued a series of notices exempting firms in specified industries from the statutory restrictions under Sec. 382 on the use of NOLs. The statute establishing TARP authorized Treasury to issue ''regulations and other guidance'' to implement it,[10] and Sec. 382(m) authorized Treasury to issue the regulations necessary to implement Sec. 382.

Treasury issued the first of these notices in mid-2008. Notice 2008-76 exempted from Sec. 382 the acquisition of stock of a loss corporation by the United States under the Housing and Economic Recovery Act of 2008. The notice covered Fannie Mae and Freddie Mac. Notice 2008-83 authorized banks to take certain deductions under 382(h). Commonly called the ''Wells Fargo Ruling,'' it was predicted to cost the government between $105 to $110 billion (Paley 2008). The Jones Day law firm estimated its cost at $140 billion.[11] (As we will see, this notice was terminated, so the actual costs were much smaller.)

[10] *Emergency Economic Stabilization Act of 2008*, P.L. 11-0343, 122 Stat. 3765, Sec. 101(c)(5).

[11] The law firm backtracked some months later to defend the notice strongly and say that it was ''quite modest'' and ''not a significant tax subsidy.'' See *Revisiting Notice 2008-83*, Jones Day, December 2008. Jones Day had estimated the Wells Fargo merger alone to have benefited by some $25 billion. The original Jones Day article was taken down from the web, but it is quoted in Paley (2008). Just one other merger, PNC's acquisition of National City, benefited by an estimated $5.1 billion. See J. Drucker, ''PNC Stands to Gain From Tax Ruling; Acquisition of National City Will Bring Billions in Deductions, Experts Say,'' *Wall Street Journal*, October 30, 2008.

In Notice 2008-84, the Treasury announced that it would not test for ownership changes on days when the United States owned a 50 percent interest in a loss firm.

Notice 2008-100 declared that an acquisition by Treasury of acquired stock in a loss corporation would not trigger the 382 limitations. Since Treasury acquired New GM's stock in a G reorganization qualifying under Sec. 382(l)(5), GM may have escaped the Sec. 382 limitations in its initial reorganization anyway. By contrast, firms like Citigroup and AIG were not G reorganizations.

Notice 2009-14 of February 17, 2009, purported to ''amplify'' 2008-100. In fact, it explicitly covered the auto industry and provided that the Treasury's initial acquisition would not trigger the Sec. 382 limitations (again, given that GM used a G reorganization, ultimately it would not need the assurance 2009-14 offered). Notice 2009-38 continued in much the same vein.

Only in January 2010, half a year after GM's Sec. 363 sale, would the Treasury tackle the firm's real Sec. 382 problem: What happens when Treasury sells its stock? To resolve this question, January 11th's Notice 2010-2 changes the law in two crucial ways.

First:

> For purposes of measuring shifts in ownership by any 5-percent shareholder on any testing date occurring on or after the date on which an issuing corporation redeems stock held by Treasury that had been issued to Treasury pursuant to the Programs. . ., the stock so redeemed shall be treated as if it had never been outstanding.

Picture the problem. Rather than sell its shares to other investors, the Treasury might sell its shares back to the firm. If it did so, the percentage held by the other investors would—necessarily—rise. In Notice 2010-2, the Treasury declared that the increase would not trigger Sec. 382.

Second:

> If Treasury sells stock that was issued to it pursuant to the Programs . . . and the sale creates a public group (''New Public Group''), the New Public Group's ownership in the issuing corporation shall not be considered to have increased solely as a result of such a sale.

Even if the Treasury sells its shares to the public, the sale will not trigger Sec. 382. Thus, in Notice 2010-2, the Treasury finally addressed New GM's Sec. 382 problem.

3.3.2 The Statutory Amendment

But could the Treasury legally issue Notice 2010-2? Could it legally issue any of these Sec. 382 notices?

Congress in its legislation objected to some of what Treasury did, validated some, and left most notices unaddressed. The issues of the Treasury's TARP-related Sec. 382 notices came up in the American Recovery and Reinvestment Tax Act of 2009 (better known as the 2009 stimulus bill).

First, the Conference Committee added a provision to the tax code, Sec. 382(n)(1), to exempt from Sec. 382 advances of TARP funds that had an explicit requirement for a restructuring plan (neither the original House nor the original Senate version had anything like this). From the conference report (U.S. Congress 2009, pp. 560–61):

> The limitation contained in subsection (a) shall not apply in the case of an ownership change which is pursuant to a restructuring plan of a taxpayer which-
>
> (A) is required under a loan agreement or a commitment for a line of credit entered into with the Department of the Treasury under the Emergency Economic Stabilization Act of 2008, and
>
> (B) is intended to result in a rationalization of the costs, capitalization, and capacity with respect to the manufacturing workforce of, and suppliers to, the taxpayer and its subsidiaries.
>
> (2) SUBSEQUENT ACQUISITIONS.-Paragraph (1) shall not apply in the case of any subsequent ownership change unless such ownership change is described in such paragraph.

The same auto-industry Sec. 382 exemption (but explicitly for auto companies) had been proposed in December 2008 in a bailout bill that passed the House and was supported by Republican President George W. Bush, but was killed by Senate Republicans.[12]

[12] See: J. Puzzangherra, ''Auto Bailout Dies in Senate: Big 3 Could Opt for Bankruptcy after a Late Compromise Attempt Fails to Satisfy GOP Opponents.'' *Los Angeles Times*, December 12, 2008; M. Leone, ''Grab Coveted Losses, Buy a Car Company,'' *CFO.com*, December 12, 2008.

Second, the act authorized the Wells Fargo notice as far as bank mergers that happened before January 16, 2010, but not afterward. The drafters explained that Congress did this because it found Treasury's various TARP notices outrageous but thought it should save taxpayers who relied on them anyway. The drafters continued:[13]

> Congress finds as follows:
> (1) The delegation of authority to the Secretary of the Treasury, or his delegate, under section 382(m) does not authorize the Secretary to provide exemptions or special rules that are restricted to particular industries or classes of taxpayers;
> (2) Internal Revenue Service Notice 2008-83 is inconsistent with the congressional intent in enacting such section 382(m);
> (3) the legal authority to prescribe Notice 2008-83 is doubtful;
> (4) however, as taxpayers should generally be able to rely on guidance issued by the Secretary of the Treasury, legislation is necessary to clarify the force and effect of Notice 2008-83.

3.3.3 Notice 2010-2

Now return to Notice 2010-2 and ask the obvious question: Given Sec. 382(n), why did Treasury issue the notice? It did so because Subsec. (n) did not cover a sale by the Treasury to the public. Subsec. (n)(1)(A) may have covered the Treasury's initial stock acquisition. After all, the Treasury took its equity interest as part of its TARP investment, so perhaps it ''required'' the stock ''under a loan agreement.'' Ironically, however, Treasury did not need Sec. 382(n) for GM since GM restructured itself as a tax-free G reorganization. And Sec. 382(n) was not applicable to the purchases of equity in Citigroup and AIG because they were financial firms, not manufacturers.

Subsec. 382(n)(1) did not protect GM from Treasury's re-sale of the stock it acquired. When Treasury lent GM the money, it did not ''require'' its own re-sale under the loan agreement. It would be an odd agreement that required the lender to sell any stock it obtained. And if it did not require the re-sale, then Sec. 382(n)(1) did not

[13] P.L. 111-5, American Recovery and Reinvestment Act of 2009, section 1261 (paragraph indentation added).

exempt Treasury's sale of its shares to the public from the Sec. 382 limitations.

This put Treasury in a bind. Congress claimed not to like the way the Treasury helped the financial institutions. It declared that it had not authorized Treasury to issue the notices it did.[14] But absent a notice, Treasury would trigger the Sec. 382 limitations at GM when it sold its stock.

Apparently Treasury responded, ''Congress won't mind.'' To move $18 billion to New GM, it needed to be able to assure the firm and its investors that GM would continue to have access to the accumulated losses after Treasury sold its stock. Sec. 382(n) did not offer that assurance. Through Notice 2010-2, Treasury offered it anyway.

4. RATIONALE, DEFERENCE, AND RELIANCE

Treasury does not explain why the notices promote the policy behind Sec. 382. Davidson (2011) nicely lays out the case Treasury might have made (without endorsing it; she later gives the counterargument, too):

> Section 382(m) gives the Secretary authority to issue regulations ''necessary or appropriate to carry out the purposes of'' section 382, so one must look to the purpose of section 382.
>
> As a broad matter, section 382 is meant to prevent the trafficking in losses and to preserve ''the integrity of the carryover provisions,'' which perform an ''averaging function by reducing the distortions caused by the annual accounting system.'' More specifically, Congress was concerned with matching items of income and loss.
>
> The TARP Guidance did not violate these principles by trafficking in losses, in the generally understood meaning of the phrase. The government did not acquire shares in these banks in order to use their loss carryforwards; it did so to stabilize the financial sector. Looking beyond the acquirer's motives, because the government does not pay taxes, it is not even

[14] Although Congress spoke sternly in the 2009 stimulus bill of how the Wells Fargo notice infringed on its authority as legislature, it made no comment on the other dubious notices that Treasury had issued by February 2009. A footnote on p. 560 of the stimulus bill conference report (U.S. Congress 2009) mentions the Treasury notices 2008-39, 2008-100, and 2009-14 without commenting on their validity.

> capable of trafficking in losses in the traditional sense. The TARP Guidance also did not violate the integrity of the carryover provisions. Losses created by TARP banks remain with the bank–they will only be used to offset income of that bank. When shares are sold to the public, the guidance was careful to limit its application to buyers in the public group. This prevents another corporation from acquiring the bank to use its NOLs. Losses of a TARP bank will not be able to be used by any other institution by means of a TARP-related acquisition. From the perspective of avoiding the trafficking in losses and maintaining the integrity of the carryover provisions, the TARP Guidance were ''appropriate to carry out the purposes of'' section 382. [Footnotes omitted.]

This is unsatisfactory. The Treasury does not pay taxes, but the other investors in New GM do. For them, the ability to invest in a company that earns its income tax-free for the indefinite future is a major advantage.

What is more, the purpose behind a section does not matter when its language is clear. Sec. 382 routinely covers transactions not motivated by tax avoidance, and the Treasury does not exempt them from the section by appealing to ''purpose.'' Sec. 382 covers nonabusive transactions because it is, at root, a ''prophylactic rule.'' By their very nature, prophylactic rules cover transactions one would not necessarily cover if ''purpose'' were all that mattered.

That the government buys stock does not itself imply that different ownership change rules should apply. The United Kingdom, for example, imposes a rule similar to Sec. 382. It does not make special allowance for government-owned stock. As KPMG explained:

> The UK tax code contains similar provisions preventing the carry forward of losses following a 50 percent or more ownership change, but only when there is a ''major change in the nature or conduct of the trade'' within three years of the change of ownership. But, in contrast to the position in the US, the acquisition of shares by the UK government does count in measuring whether there has been an ownership change. (KPMG 2010)

The U.S. statute does not exempt government-owned stock and neither does the UK's.

Ultimately, tax benefits did play a major role in these transactions. By letting New GM keep NOLs to which it was not legally entitled, Treasury gave the firm (and its owners, including the UAW) $18 billion more in assets. Had the administration tried to give GM $18 billion forthrightly, voters might have complained. By hiding the gift in an obscure tax section, it reduced that electoral scrutiny. But the investors who bought New GM shares noticed. They paid a higher price than they otherwise would have paid.[15] And necessarily, the UAW, the government of Canada, and the former bondholders also noticed.

4.1 Court Deference

The executive branch continually interprets statutes as it issues regulations. Courts do too, and often make interpretations that outsiders such as ourselves consider ridiculous. It is generally accepted that courts should be allowed to have the final word in interpretation nonetheless. Could it be that the executive branch, in interpreting tax law, similarly has the final word? In fact, courts have ruled it does not—a sensible rule. Courts do defer to executive branch interpretations of statutes in many circumstances, but not in those like the TARP notices.

On January 11, 2011, the U.S. Supreme Court made clear in *Mayo Foundation v. U.S.*, 131 S. Ct. 704 (2011), that courts should treat tax regulations just like any other regulations. The case concerned a statute that exempted students from Social Security and Medicare taxes withholding. In 2004, the Treasury promulgated regulations under which medical residents were not students. The Mayo Clinic challenged the regulation, and the Court held it valid. Courts should treat tax regulations like any other, it explained.

Under the well-known ''*Chevron*'' rule by which it sometimes defers to executive agencies (*Chevron U.S.A. Inc. v. Natural Resources Defense Council*, 467 U.S. 837 (1984)), explained the Supreme Court, courts should first ask whether Congress had ''directly addressed the precise question at issue.'' If not, then they should defer to the agency unless the rule was ''arbitrary or capricious in substance, or manifestly contrary to the statute'' (*Mayo* 2011, p. 711). It would not, the Court explained, ''carve out an approach to administrative

[15] Note that this reduces the net cost to the government of the notice, since the Treasury will be able to re-sell its shares at a higher price.

review good for tax law only. . . . The principles underlying our decision in *Chevron* apply with full force in the tax context'' (p. 713).

Nonetheless, this deferential standard applies only when Congress intended to delegate to the agency and the agency followed standard rulemaking procedures. Continued the Court (p. 714):

> We have explained that ''the ultimate question is whether Congress would have intended, and expected, courts to treat [the regulation] as within, or outside, its delegation to the agency of 'gap-filling' authority.'' [*Long Island Care at Home, Ltd. v. Coke*, 551 U.S. 157, 173 (2007)]. In the *Long Island Care* case, we found that *Chevron* provided the appropriate standard of review ''[w]here an agency rule sets forth important individual rights and duties, where the agency focuses fully and directly upon the issue, where the agency uses full notice-and-comment procedures to promulgate a rule, [and] where the resulting rule falls within the statutory grant of authority.''

Notice 2010-2 fails both of those requirements. First, Congress expressly declared that it did not intend to delegate this authority to Treasury. Notice 2010-2 applied only to financial institutions, automobile companies, and other specific TARP recipients. Yet, Congress announced in its committee report, ''section 382(m) does not authorize the Secretary to provide'' ''special rules that are restricted to particular industries or classes of taxpayers.'' As a result, the earlier TARP Notice 2008-83 was ''inconsistent with the congressional intent'' and of only ''doubtful'' ''legal authority.'' Notice 2010-2 is precisely such an industry-specific rule.

Second, Notice 2010-2 is not a regulation. It is a ''notice.'' The *Mayo* Court declared *Chevron* appropriate where an agency uses ''full notice-and-comment procedures to promulgate a rule.''[16] By contrast, the Supreme Court explained in *Christiansen v. Harris County*, 529 U.S. 576 (2000):

[16] The Treasury is notorious for its cavalier attitude toward the Administrative Procedure Act. In *Intermountain Insurance Service of Vail v. Commissioner of Internal Revenue Service*, No. 10-1204 (June 21, 2011), p. 32 (D.C. Circuit, 2011), the Commissioner ''simultaneously issued immediately effective temporary regulations and a notice of proposed rulemaking for identical final regulations and then held a 90-day comment period [receiving just one comment] before finalizing the regulations.'' The opinion goes on to say that this procedure is ''typical of the Commissioner's practice.''

> Interpretations such as those in opinion letters—like interpretations contained in policy statements, agency manuals, and enforcement guidelines, all of which lack the force of law—do not warrant *Chevron*-style deference. Instead, interpretations contained in formats such as opinion letters are [governed by *Skidmore*].

Turning now to *Skidmore v. Swift & Co.*, 323 U.S. 134 (1944), the Supreme Court considered the agency's logic, but made its own decision (p. 140):

> We consider that the rulings, interpretations and opinions of the Administrator . . . constitute a body of experience and informed judgment to which courts and litigants may properly resort for guidance. The weight of such a judgment in a particular case will depend upon the thoroughness evident in its consideration, the validity of its reasoning, its consistency with earlier and later pronouncements, and all those factors which give it power to persuade, if lacking power to control.

In *United States v. Mead Corp.*, 533 U.S. 218 (2001), the Court went further and declared that as a general rule an agency interpretation would have to go through notice and comment to receive *Chevron* deference. In Notice 2010-2, Treasury did not try to reason or persuade. It simply declared the rule so. As Smith (2011, pp. 1260, 1261) puts it:

> *Mayo* benefits taxpayers by clarifying that the *Mead* principles apply in tax. When the *Mead* test is applied to revenue rulings, revenue procedures, and notices, the conclusion is that they are not among the types of agency guidance that receive *Chevron*'s high level of deference.
>
> . . . Any pre-*Mayo* case law on the status of revenue rulings, revenue procedures, and notices should generally be considered obsolete unless the opinion reflects *Mead* analysis. The clear conclusion that those forms of guidance do not qualify for the level of deference described in *Chevron* is another benefit to taxpayers from *Mayo*.

Because the Treasury did not follow notice-and-comment procedures, the GM notices would not qualify for *Chevron* deference, even if the statutes they purport to interpret were indeed ambiguous.[17]

4.2 Taxpayer Reliance

Suppose the TARP notices were invalid. Should taxpayers be able to rely on them anyway, since it is the fault of Treasury and not the taxpayer?[18] Notice 2010-2 provides:

> Taxpayers may rely on the rules described in Section III of this notice. These rules will continue to apply unless and until there is additional guidance.

This is profoundly self-serving, of course. The Treasury cannot change the law by fiat. A bureaucrat cannot give his friend funds illegally and then protect that friend by declaring his friend's reliance protected. If a court held Notice 2010-2 illegal, GM could not cite the notice as authority for deducting $45 billion in NOLs anyway.

The relevant question goes to penalties: May a taxpayer who relies on the notices avoid civil and criminal penalties? As Rogovin and Korb (2008, p. 341) explain:

> As with revenue rulings and revenue procedures, announcements and notices can provide substantial authority sufficient to relieve taxpayers from the negligence and substantial understatement penalties and, consequently, may be relevant to whether certain penalty provisions apply.

[17] We should mention a caveat. In *Intermountain Insurance,* cited earlier, the court gave *Chevron* deference to Treasury regulations in Treasury's appeal, even though those regulations were written after Treasury had already lost in Tax Court. Perhaps Treasury could re-issue the GM regulations with a pretence of notice and comment. The tax provision at issue in *Intermountain*, however, is important and has resulted in split circuits (3–2), and so is likely to go to the Supreme Court. See K. B. Friske and D. Pulliam, ''Circuit Split Deepens on Six-Year Period for Basis Overstatements,'' *Journal of Accountancy*, May 2011.

[18] Before the Treasury and other owners (including the 10 percent given to Old GM) sell enough stock to trigger the 50 percent threshold, use of the NOLs would be legal even without Notice 2010-2. GM is now, however, a publicly traded company and has told the public that the NOLs are part of its assets, though without 2010-2 they will not be if the Treasury sells its stake. Thus, the immediate question would be whether GM has thereby violated federal securities laws.

Sec. 6662 of the code imposes a penalty for any ''substantial understatement of income tax.'' Subsec. (d)(2)(B) protects a taxpayer who relies on ''substantial authority.'' According to the Treasury, its own notices are ''substantial authority'' (Rogovin and Korb 2008, Reg. 1.6662-4(d)(3)(iii)), though it also explains that the ''weight accorded an authority depends on its relevance and persuasiveness'' (Reg. 1.6662-4(d)(3)(ii)).

Consider the weight appropriate to Notice 2010-2. First, the Treasury itself declares it ''substantial authority.'' This is, of course, again self-serving. Acting on behalf of the administration, the Treasury has manipulated tax procedure to route $18 billion to its supporters' car company. In essence, it also argues that its manipulation insulates those favored taxpayers from ''substantial underpayment'' penalties.

Second, Notice 2010-2 does not try to persuade. It simply declares. But if an IRS notice were to announce that Microsoft did not have to pay taxes because Bill Gates paid the Treasury secretary $1 million in bribes, the announcement would hardly give Microsoft substantial authority. Here, the Democratic administration has given a massive tax benefit to one of the party's biggest supporters. Like other labor unions, the UAW provided the Obama campaign with elaborate assistance. Some of the help came in person, and some came as money. From 1989 to 2010, the UAW spent over $27 million on political campaigns, 98 percent of it on behalf of the Democratic Party.[19] In 2008 alone, it spent $2,119,937 on political campaigns, $2,101,187 of that for Democrats.[20]

Suppose that Notice 2010-2 had said:

> The President is grateful to the UAW for the assistance it provided his party. In gratitude for that political support, the Treasury announces that, should it sell the stock that was issued to it pursuant to the Programs . . . and should the sale create a public group (''New Public Group''), the New Public Group's ownership in the issuing corporation shall not be considered to have increased solely as a result of such a sale.

The only difference between this hypothetical notice and the real Notice 2010-2 is the explicit character of the reason for the largesse.

[19] ''Top All-Time Donors, 1989–2010,'' OpenSecrets.org.

[20] ''United Auto Workers,'' OpenSecrets.org.

It is an odd approach to statutory interpretation that would make a notice illegal if it articulates its reason, but legal if it leaves the reason unsaid.

5. LEGISLATIVE RESOLUTION

Return to the problem at stake: the manipulation of the highly arcane minutiae of the corporate tax rules to route huge sums to favored groups. The question is what anyone can do about it.

Political remedies are unlikely to work. Voters do not understand transactions like this well enough to punish a candidate in the next election. Much less will they impeach anyone for a transaction like this. Voters understand politicians who take briefcases stuffed with cash; they do not understand G reorganizations and NOL carryforwards. Congress has complained, asking TARP's inspector general to investigate the validity of the notices and their motivation.[21] Sen. Jim Bunning (R-KY) even introduced a bill with the sole purpose of repealing Notice 2010-2.[22] Unless Congress can override the notices by a veto-proof two-thirds majority, however, it can do little more than badger the administration with its oversight authority and complain to the public.

5.1 The Standing Problem

All this leaves a lacuna in the law. As the GM notices illustrate, it leaves an $18 billion lacuna.

To explore how Congress might try to address the problem, consider the following fantasy IRS notice:

> Internal Revenue Bulletin: 2010-999
> February 24, 2011
> Notice 2011-999
>
> Application of Title 26 to Certain Persons Pursuant to the Emergency Economic Stabilization Act of 2008
>
> I. BACKGROUND
>
> Section 7805(a) of the Internal Revenue Code (''the Code'') provides that except where such authority is expressly given

[21] Office of the Special Inspector General for the Troubled Asset Relief Program, ''Engagement Memo–Review of the Section 382 Limitation Waiver for Financial Instruments Held by Treasury,'' Aug. 10, 2010.

[22] S. 2916 [111th]. The bill was sent to committee and never returned.

> to any person other than an officer or employee of Treasury, the Secretary shall prescribe all needful rules and regulations for the enforcement of Title 26, including all rules and regulations as may be necessary by reason of any alteration of law in relation to internal revenue.
>
> Section 101(c)(5) of EESA provides that the Secretary is authorized to issue such regulations and other guidance as may be necessary or appropriate to carry out the purposes of EESA.
>
> II. GUIDANCE REGARDING CERTAIN PERSONS
>
> Any funds received by J. Mark Ramseyer or Eric B. Rasmusen shall not constitute ''income'' under Sec. 61 of the I.R.C., and shall be entirely exempt from taxation.
>
> DRAFTING INFORMATION
>
> The principal author of this notice is John B. Doe of the Office of Associate Chief Counsel (Individual). For further information regarding this notice, contact Robert B. Roe at (202) 999-9999 (not a toll-free call).

Few readers would dispute the notion that Notice 2011-999 straightforwardly violates the Code. It does not even try to argue that sparing us from the income tax furthers the purposes of the 2008 stimulus bill. If it did, you, our readers, would laugh. But you could not laugh in court. You would not have standing.

Under current law, voters cannot challenge these transactions in court (see Hickman 2008 for discussion). If a rule benefits some people but does not harm others, nobody will have ''standing'' to challenge it. Justice Powell articulated the point most famously:

> I cannot now imagine a case, at least outside the First Amendment area, where a person whose own tax liability was not affected ever could have standing to litigate the federal tax liability of someone else. (*Simon v. E. Ky. Welfare Rights Org.*, 426 U.S. 26, 46 (1975) (Powell, J., concurring))

A more recent example appeared in a Chrysler case in which Justice Roberts held that taxpayers lacked standing to challenge other people's tax benefits. The plaintiffs argued that Chrysler's tax breaks hurt them:

> Plaintiffs principally claim standing by virtue of their status as Ohio taxpayers, alleging that the franchise tax credit

> ''depletes the funds of the State of Ohio to which the Plaintiffs contribute through their tax payments'' and thus ''diminishes the total funds available for lawful uses and imposes disproportionate burdens on them.'' (*DaimlerChrysler Corp. v. Cuno*, 547 U.S. 332,342 (2006))

Justice Roberts said ''No.''

> As an initial matter, it is unclear that tax breaks of the sort at issue here do in fact deplete the treasury: The very point of the tax benefits is to spur economic activity, which in turn increases government revenues.
>
> Plaintiffs' alleged injury is also ''conjectural or hypothetical'' in that it depends on how legislators respond to a reduction in revenue, if that is the consequence of the credit. Establishing injury requires speculating that elected officials will increase a taxpayer-plaintiff's tax bill to make up a deficit; establishing redressability requires speculating that abolishing the challenged credit will redound to the benefit of the taxpayer because legislators will pass along the supposed increased revenue in the form of tax reductions. Neither sort of speculation suffices to support standing. (*DaimlerChrysler*, 547 U.S. at 344)

Various authors have proposed reforms to the standing rules (e.g., Rosenburg 1996). Unfortunately, their proposals simultaneously increase the incidence of frivolous suits, venue shopping, and collusive litigation, as Stearns (1995) points out. In the name of policing frivolous litigation, GM (and Ramseyer and Rasmusen) keep their special deals. Although Treasury cannot get away with arbitrary interpretations of the statutes that increase someone's taxes (since that person would have standing to object in court), it can get away with equally unreasonable interpretations that reduce someone's taxes.[23]

One might also think that giving away tax breaks was criminal. In fact, the Anti-Deficiency Act, 31 U.S.C. Sec. 1341, makes it a criminal offense for a government officer or employee to give away government money that Congress did not appropriate (he may not

[23] For examples of how Treasury gets around Supreme Court decisions using taxpayer-favorable (and hence unreviewable) regulations, see Polsky (2004).

''make or authorize an expenditure or obligation exceeding an amount available in an appropriation or fund for the expenditure or obligation,'' 31 U.S.C. Sec. 1341(a)(1)(A)). If he does, 31 U.S.C. Sec. 1350 provides that he may be fined up to $5,000 or imprisoned for up to two years, and 31 U.S.C. Sec. 3528 requires him to repay the improper expenditure.

Should the Treasury secretary fear the possibility of spending two years in jail and having to repay $12 billion (perhaps splitting the amount with his predecessor, Henry Paulson)? Treasury secretaries have thought about this before; in a November 2008 speech, Paulson said that because of the Anti-Deficiency Act, Treasury could not have bailed out Lehman Brothers.

There are several reasons why the secretary need not fear at the present time. To start, the Anti-Deficiency Act speaks of ''expenditures.'' A ''tax expenditure,'' no matter how big or unlawful, might reasonably be excluded from its scope. Whether it is excluded probably does not really matter, though. Under 31 U.S.C. Sec. 3528(b)(1)(B), the comptroller general may relieve the spendthrift official from liability for repayment if the expenditure was made in good faith or not specifically prohibited by law (see also 31 U.S.C. Sec. 3527). What is more, criminal charges would have to be brought by the attorney general or his subordinates, and they are part of the administration. We do not allow private prosecutions for federal crimes.[24]

Thus, U.S. law must be changed if we are to be able to deal with unlawful tax expenditures in any way other than trying to explain them to voters so as to unseat the offending official at the next election.

5.2 Three Alternatives

There are three possible changes in law that could discourage such tax expenditures in the future. Below, we consider each one.

[24] Another obstacle to unlawful tax rules, in principle, might be the ethical code of the Bar. What would a Good Man IRS attorney do if asked to authorize an unlawful notice? What would a Bad Man IRS attorney do out of fear of the Bar? We do not know what the Good Man would do, but we are sure the Bad Man need not fear disbarment. See Kwon (2010) for a discussion of the ethical obligations of IRS attorneys generally.

5.2.1 The Canadian Rules

In Canada, a taxpayer does have standing. Public-interest standing was extended to taxpayers in *Harris v. Canada (Minister of National Revenue)*, [2001] 4 F.C. 37 (Ct. of App.).

George Harris alleged that the minister of national revenue acted in bad faith and violated his fiduciary duty when he overruled his professional staff and made a favorable tax ruling (an ''advance ruling'') at the request of influential taxpayers, the billionaire Bronfman family.[25] Harris asked the court for a declaration that the minister of national revenue was obliged to try to collect the taxes from a particular transaction.

An appellate court ruled that Harris did have standing, saying:

> In *Borowski*, Martland J. for the majority held that to obtain public interest standing, a plaintiff must (1) demonstrate that there is a serious issue as to the invalidity of legislation, (2) that the plaintiff has a genuine interest, and (3) there is no other reasonable and effective manner to bring the issue before the court. (*Harris v. Canada (Minister of National Revenue)*, [2001] 4 F.C. 37 (Ct. of App.), referring to *Minister of Justice of Canada et al. v. Borowski* [1981] 2 S.C.R. 575))

A few years after *Borowski* gave standing for constitutional law issues, *Finlay v. Canada (Minister of Finance)*, [1986] 2 S.C.R. 607, extended it to standing for statutory issues. Therefore, the *Harris* court gave standing to Harris to contest the application of the tax code. In *Harris v. Canada (Minister of National Revenue)*, 2001 DTC 5322 (Trial Div.), the trial court even granted Harris's application for discovery of internal government documents relating to the Bronfmans' requests for an advance ruling.

Harris did lose his case in the end, but on the merits rather than on standing. In *Harris v. Canada (Minister of National Revenue)*, [2002] 2 F.C. 484 (Trial Div.), the trial court ruled against Harris on the merits, finding no bad faith on the part of the government and no fiduciary duty violation. It even accepted his argument that he was entitled to be paid for out-of-pocket costs because he had benefited the public by arguing the case despite his loss (which in Canada

[25] K. Foss, ''Judge Scolds Tax Officials, but Crusader Loses Case,'' *Globe and Mail*, December 20, 2001.

would ordinarily mean he would pay the other side's costs, though in this case the government waived its claim against him).

English courts have also given people standing to contest tax policy, albeit only if a genuine public interest is at stake. See the 1978 R.S.C., Ord. 53 and *Inland Revenue Comrs v. National Federation of Self-Employed and Small Businesses Ltd*, [1981] 2 All ER 93 (House of Lords). In that case, the Federation challenged a tax amnesty given to casual employees in the printing industry. The Federation lost, but only because the Law Lords all agreed that the government clearly had the discretion to grant an amnesty in this particular case.

Thus, one policy change for the United States would be to adopt the Canadian or English law of standing. We are hesitant to propose this change, however, because of the problems the United States has had with frivolous litigation, forum shopping, and activist judges (on which see, e.g., the forthcoming book edited by F. Buckley).

5.2.2 Congressional Litigants

To limit Treasury's ability to offer special deals to political favorites, we offer two alternatives that might yet constrain frivolous suits. First, Congress could offer standing to members of Congress:

Tax Regulation Enforcement Bill

> Any two members of Congress shall have standing to challenge in court any interpretative or other notices, rules, regulations, or guidelines of the Internal Revenue Service as arbitrary and capricious. The members bringing the action need not be current members of Congress and need not have voted for or against the statute in question. Should they win, they shall each be entitled to liquidated damages of $1,000. The Declaratory Judgement Act (28 U.S.C. sec. 2201) shall not apply to this legal action.[26] As a remedy, the Court may issue injunctions as appropriate, but not temporary restraining orders or preliminary injunctions.

[26] The Tax Anti-Injunction Act of 1867, 26 U.S.C. Sec.7421(a), says, ''no suit for the purpose of restraining the assessment or collection of any tax shall be maintained in any court by any person, whether or not such person is the person against whom such tax was assessed.'' This provision would continue to apply and would restrict our new statute to injunctions to collect more tax, but not less. We suspect that this would help prevent congressmen from filing frivolous suits to demonstrate sympathy with their constituents.

> Any number of these suits may be filed concurrently. They shall be filed in any District Court of the United States.

Limiting challenges to just the grounds of ''arbitrary and capricious'' Treasury rules would narrow the range of suits drastically. The court need only look at the second part of the *Chevron* test ruling for the Treasury unless the statute is unambiguously contrary to the Treasury position. Yet these challenges would narrow the range of unlawful actions Treasury could take even more. There are, at most, dozens of ways the ambiguities in a sentence can be construed, but there is an infinite number of ''interpretations'' that are totally unfounded. A congressman could not successfully challenge an IRS interpretation of ''after several years'' as being anywhere from 2 to 10 years, but he could challenge an interpretation as ''after 200 years'' or ''after the taxpayer has traveled to Kashmir.''[27]

Requiring two congressmen rather than one will help to reduce the number of frivolous suits, though we recognize that we will not eliminate them. We originally thought to require five congressmen rather than two but recalled how in 1940 Vichy, the resolution that France needed a new constitution passed by 395–3 in the Chamber of Deputies and 229–1 in the Senate.[28]

Allowing more than one suit and in different courts will prevent collusive suits that block review. If only the Tax Court had jurisdiction, for example, then a pro-Treasury plaintiff could bring suit there, ''take a dive,'' and refrain from appealing—thereby blocking a real plaintiff.[29]

5.2.3 A Qui Tam Statute

An alternative to allowing congressmen to challenge Treasury notices would be a *qui tam* statute. A short version, worded for

[27] ''After the taxpayer has traveled to Kashmir'' is ridiculous, of course. But we must keep in mind that the Bad Man asks not whether an interpretation is ridiculous but whether he can get away with it.

[28] W. Shirer, *The Collapse of the Third Republic* (New York: Simon and Schuster, 1969), p. 933. Two Socialist and one Radical deputy voted against; the only dissenting senator was the right-wing Marquis de Chambrun.

[29] Congress cannot completely delegate the executive power to enforce the laws. In *Unique Product Solutions, Ltd. v. Hy-Grade Valve, Inc.* (February 23, 2011, N.D. Ohio), the court held that the president could not give a private plaintiff complete authority to pursue a criminal case against someone who labeled a product as patented after the patent expired. To do so was, it explained, an unconstitutional delegation of the president's duty to ''take care'' that the laws be faithfully executed.

contrast to allow many more suits than our previous statute, might go as follows:

Qui Tam Tax Regulation Enforcement Bill

> It shall be illegal for any employee of the Treasury Department to misinterpret a federal statute. Any employee found willfully to have misinterpreted a statute shall pay a civil fine of $500. Any two members of Congress may bring a civil action against such violator in any District Court of the United States.

Conceptually, the *qui tam* statute performs much the same function as the standing rule. Unfortunately, it does present the same nontrivial risk of frivolous litigation. Either version enables two members of Congress to file suit to challenge any action by the Treasury to route funds to politically favored institutions.

It may seem imprudent to enlarge the power of the courts in a notoriously litigious United States already known for accusations that judges abuse their power by imposing their personal political views. The policy area we are opening up to judicial review, however, is not one known for judicial activism. Indeed, it is generally thought that judges dislike deciding tax cases. Even Justice Antonin Scalia, who made his name in administrative law in his academic career, said, ''The constitutional work can be dull, too, but it's not like the tax code. Philosopher-kings do not read the Internal Revenue Code, believe me.''[30] Justice William Douglas, famed for his expertise in business law and his activism, wrote to an ill Justice Black, ''Take good care, lie low, and forget about these dull tax cases—which are now droning on and on'' (Richards 2001). And Judge Learned Hand, known for his common-law decisions in private law, said in 1947:

> In my own case the words of such an act as the Income Tax, for example, merely dance before my eyes in a meaningless procession: cross-reference to cross-reference, exception

[30] ''A Look at the Hidden World of U.S. Associate Justice Antonin Scalia,'' *National Post*, June 12, 1992, as quoted in Richards (2001). Note, too, what former tax lawyer Justice Blackmun said: ''If one's in the doghouse with the Chief, he gets the crud. He gets the tax cases and some of the Indian cases, which I like, but I've had a lot of them.'' (R. Woodward and S. Armstrong, *The Brethren* [New York: Simon and Schuster, 2005].)

> upon exception—couched in abstract terms that offer no handle to seize hold of—leave in my mind only a confused sense of some vitally important, but successfully concealed, purport, which it is my duty to extract, but which is within my power, if at all, only after the most inordinate expenditure of time. I know that these monsters are the result of fabulous industry and ingenuity, plugging up this hole and casting out that net, against all possible evasion; yet at times I cannot help recalling a saying of William James about certain passages of Hegel: that they were no doubt written with a passion of rationality; but that one cannot help wondering whether to the reader they have any significance save that the words are strung together with syntactical correctness.

One can only imagine what the less economics-minded judges must think about tax cases. Yet it is perhaps in tax cases—particularly business tax cases—that even the limited intelligence of the courts most exceeds the intelligence of the voter, just as it is there that we can expect judges to face the least temptation to care enough about policy to impose their own preferences instead of trying to follow the law.[31] Legislatures, in contrast, while also having neutral ideological preferences, can use the opacity of tax law to transfer large sums of money to sophisticated supporters or to conceal extravagance with public funds. Criminal procedure presents the opposite combination of relative expertise and ideological conflict of interest. Judges seem to like deciding this kind of case, if we look at the willingness of the U.S. Supreme Court to accept cert, despite the fact, or perhaps because of the fact, that they involve situations that the average voter can understand and laws that politicians cannot use to transfer money from one interest group to another. (See Stuntz 1997, 2006, for close analysis of the pathological judicialization of the criminal justice process.)

6. CONCLUSIONS

Authority over tax administration is authority easy to abuse. I.R.S. Notice 2010-2 and its predecessors purported to exempt companies

[31] The incentives and expertise of Supreme Court clerks are perhaps just as important, since they customarily do the first cut of cert petitions in deciding which cases are worth consideration by the Court. How many clerks have taken a tax course? We have not found articles on the self-interest of clerks in cert petition triage, but on more measurable considerations in tax cases and cert, see Staudt (2004).

partly owned by the government from taxes they would have had to pay had their owners been entirely private. The case of GM is the clearest in terms of the bailout of a favored constituency because that transaction resulted in a large subsidy to a labor union that had strongly supported the administration's party. Yet all of the notices helped hide the real cost of the TARP bailouts from the public.

It is hard for Congress to overturn executive actions that have no basis in statute, requiring as it does the agreement of two-thirds of both the Senate and the House of Representatives to override a presidential veto. The natural place to check invalid interpretations of statutes is in the courts. Currently no one has standing to challenge tax interpretations that benefit a few at the expense of taxpayers in general. Toward that end, we propose giving standing to members of Congress.

CASES

Chevron U.S.A. Inc. v. Natural Resources Defense Council, 467 U.S. 837 (1984).
Christiansen v. Harris County, 529 U.S. 576 (2000).
DaimlerChrysler Corp. v. Cuno, 547 U.S. 332 (2006).
Finlay v. Canada (Minister of Finance), [1986] 2 S.C.R. 607.
Harris v. Canada (Minister of National Revenue), [2001] 4 F.C. 37 (Ct. of App.).
Harris v. Canada (Minister of National Revenue), [2002] 2 F.C. 484 (Trial Div.).
Harris v. Canada (Minister of National Revenue), [2001] DTC 5322 (Trial Div.).
In re Motors Liquidation Co., 430 B.R. 65 (S.D.N.Y. 2010).
Inland Revenue Comrs v. National Federation of Self-Employed and Small Businesses Ltd, [1981] 2 All ER 93 (House of Lords).
Intermountain Insurance Service of Vail v. Commissioner of Internal Revenue Service, No. 10-1204 (June 21, 2011) (DC Circuit, 2011).
Long Island Care at Home, Ltd. v. Coke, 551 U.S. 157 (2007).
Mayo Foundation v. U.S., 131 S. Ct. 704 (2011).
Minister of Justice of Canada et al. v. Borowski [1981] 2 S.C.R. 575.
Simon v. E. Ky. Welfare Rights Org., 426 U.S. 26, 46 (1975).
Skidmore v. Swift & Co., 323 U.S. 134 (1944).
Twyne's Case, 3 Coke, 80 b. (Star Chamber, 1602).
U.S. v. Kirby Lumber Co., 284 U.S. 1 (1931).
Unique Product Solutions, Ltd. v. Hy-Grade Valve, Inc. (February 23, 2011, N.D. Ohio).
United States v. Mead Corp., 533 U.S. 218 (2001).

REFERENCES

AIG. 2009. *American International Group Inc., 2009 Annual Report*.
Auerbach, A., M. Devereux, and H. Simpson. 2010. "Taxing Corporate Income." In *Dimensions of Tax Design: The Mirrlees Review,* ed. J. Mirrlees et al. Oxford, UK: Oxford University Press.

Buckley, F., ed. Forthcoming. *An American Illness.* New Haven, Conn.: Yale University Press.
Bunkley N. 2011. ''Resurgent G.M. Posts 2010 Profit of $4.7 Billion.'' *New York Times,* February 4.
Ceraso, C. J., R. Moffatt, and S. Pati. 2010. *General Motors Co.* New York: CreditSuisse.
Davidoff, S., and D. Zaring. 2009. ''Regulation by Deal: The Government's Response to the Financial Crisis.'' *Administrative Law Review* 61: 463–535.
Davidson, B. 2011. ''Tax Policy or Treasury Policy? Treasury's Conflicted Position in Its Ownership Stakes under TARP.'' Unpublished manuscript.
General Motors Company. 2010. ''Amendment 2 to Form S-1 Registration Statement under the Securities Act of 1933.''
Hand, L. 1947. ''Thomas Walter Swann.'' *Yale Law Journal* 57: 167–72.
Harberger, A. 2008. ''Corporation Tax Incidence: Reflections on What Is Known, Unknown and Unknowable.'' In *Fundamental Tax Reform: Issues, Choices, and Implications,* ed. J. Diamond and G. Zodrow. Cambridge, MA: MIT Press.
Hickman, K. 2008. ''A Problem of Remedy: Responding to Treasury's (Lack of) Compliance with Administrative Procedure Act Rulemaking Requirements.'' *George Washington Law Review* 76: 1153–215.
Holmes, O. 1897. ''The Path of the Law.'' *Harvard Law Review* 10: 457–78.
J. P. Morgan. 2010. ''General Motors: Reborn, High Octane SAAR and Product Play; Initiative with Overweight.'' December 28.
Kotlikoff, L., and J. Miao. 2010. ''What Does the Corporate Income Tax Tax? A Simple Model without Capital.'' NBER Working Paper 16199, July.
KPMG. 2010. *Frontiers in Tax: People Thinking Beyond Borders in Financial Services,* July.
Kwon, M. 2010. ''Four of the Hardest Ethical Questions for a Government Lawyer.'' Texas Tech University School of Law Working Paper.
Morgan Stanley. 2010. ''General Motors Re-In-Car-Nation.'' December 28.
Murphy, R. 2010. ''Citi's Deferred Tax—An Asset of Dubious Worth.'' *Tax Research UK,* September 7.
PajamasMedia. 2008. ''Detroit's Downturn: It's the Productivity, Stupid: Union Work Rules Make It Almost Impossible for the Big Three to Keep Up with Foreign Competitors.'' December 16.
Paley, A. 2008. ''A Quiet Windfall for U.S. Banks.'' *Washington Post*, November 9.
Pickerill, C. 2009. ''Regarding the Advisability of a Prohibition on the Taxable Asset Sale to Creditors in Bankruptcy.'' *Tax Law* 62: 357.
Polsky, G. 2004. ''Can Treasury Overrule the Supreme Court?'' *Boston University Law Review* 84: 185–246.
Ramseyer, J. 1995. ''Public Choice,'' Chicago Law and Economics Working Paper Series, 34 (second series).
Richards, N. 2001. ''The Supreme Court Justice and 'Boring' Cases.'' *Green Bag*, 2d, 4: 401–7.
Rogovin, M., and D. Korb. 2008. ''The Four R's Revisited: Regulations, Rulings, Reliance, and Retroactivity in the 21st Century: A View from Within.'' *Duquesne Law Review* 46: 223–374.
Rosenburg, J. 1996. ''The Psychology of Taxes: Why They Drive Us Crazy, and How We Can Make Them Sane.'' *Virginia Tax Law Review* 16 (2): 155.
Smith, P. 2011. ''Life after Mayo: Silver Linings.'' *Tax Notes* 131: 1251–64.
Staudt, N. 2004. ''Agenda Setting in Supreme Court Tax Cases: Lessons from the Blackmun Papers.'' *Buffalo Law Review* 52: 889–922.

Stearns, M. 1995. ''Standing Back from the Forest: Justiciability and Social Choice.'' *California Law Review* 83: 1309–414.

Stuntz, W. 1997. ''The Uneasy Relationship between Criminal Procedure and Criminal Justice.'' *Yale Law Journal* 107: 1–76.

———. 2006. ''The Political Constitution of Criminal Justice.'' *Harvard Law Review* 119: 780–851.

Terlap, S. 2011. ''GM Stock Sag Weighs on U.S. Exit.'' *Wall Street Journal*, June 23.

U.S. Congress. 2009. *Making Supplemental Appropriations for Job Preservation and Creation, Infrastructure Investment, Energy Efficiency and Science, Assistance to the Unemployed, and State and Local Fiscal Stabilization, for the Fiscal Year Ending September 30, 2009, and for Other Purposes: Conference Report to Accompany H.R. 1*. February 12.

Warburton, A. 2010. ''Understanding the Bankruptcies of Chrysler and General Motors: A Primer.'' *Syracuse Law Review* 60: 531–82.

Comment

Efraim Benmelech

This provocative paper by Mark Ramseyer and Eric Rasmusen provides a useful overview of the restructuring of General Motors, and in particular highlights the political economy of the GM deal in which the U.S. Treasury wore two hats, being both an equity holder and a regulator. They focus on one of the main assets GM had on its balance sheets: its net operating losses (NOLs) valued at $45 billion. The reorganization of ''Old GM'' into ''New GM'' enabled New GM to retain the NOLs. Owning the NOLs increased the value of New GM and facilitated a restructuring deal that was favorable to the United Auto Workers (UAW) pension and health plans. However, as Ramseyer and Rasmusen argue, because of the 1986 Tax Reform Act, once the Treasury sells its holdings in New GM, the NOLs should be canceled and the value of New GM should decline dramatically.

THE GM BANKRUPTCY

Ramseyer and Rasmusen do an excellent job describing the details of the GM case, and the reader should refer to their article for the fine details. In my discussion, I provide only a brief summary of the facts.

GM filed for bankruptcy under Chapter 11 of the Bankruptcy Code. Under this reorganization, Old GM was sold under Section 363 of the Bankruptcy Code to a new company, New GM. Typically, when one company acquires another company's assets, it does not acquire its tax losses, but in this specific case, New GM attained the NOLs of Old GM.

However, given that the Treasury plans to sell the shares it acquired in New GM, a problem may arise in the future: Under the

Efraim Benmelech is the Frederick S. Danziger Associate Professor of Economics at Harvard University and an NBER faculty research fellow.

1986 Tax Reform Act, a corporation's ability to carry forward NOLs (and other tax credits) is limited when more than 50 percent of the stock changes hands over a three-year period (Ross, Westerfield, and Jaffe 2006). To solve this problem, the Treasury issued a series of notices declaring that Section 382 of the tax code does not apply to the Treasury. According to these notes, when the Treasury sells its shares in New GM, Section 382 will not be triggered even if more than 50 percent of ownership will change hands.

RAMSEYER AND RASMUSEN'S CRITIQUE

Ramseyer and Rasmusen make two points: First, Treasury had no *legal* justification to exempt GM NOLs from Section 382, hence the Treasury gave GM an illegal tax break. Second, the Treasury had no *economic* justification to exempt the NOLs from Section 382. In fact, Ramseyer and Rasmusen argue, there is a political economy explanation in which the exemption from Section 382 led to overvaluation of GM, which in turn made the government's position in GM look better and resulted in a transfer from the Treasury to other stakeholders—most notably the UAW, which held unsecured claims of $21 billion in GM.

THE ECONOMIC RATIONALE

In my discussion, I will focus on the second point, according to which the Treasury had no economic justification to exempt GM's NOLs from Section 382. In order to assess the economic rationale behind the decision to exempt the NOLs from taxes, we need to evaluate the cost to the Treasury if the NOLs were not allowed to be carried forward to New GM. Ramseyer and Rasmusen argue that the UAW, as a junior creditor, got a very good deal in the restructuring of GM and that crafting such a deal was possible because of the ''overvaluation'' of GM stemming from the exemptions of the NOLs from Section 382. However, what would have been the cost to the Treasury if it failed to reach an agreement with the UAW?

Consider, for example, the case of GM retirees' medical benefits. As part of the restructuring, GM's Voluntary Employees' Beneficiary Association (VEBA) received from GM $2.5 billion of new notes, $6.5 billion in preferred stock with a 9 percent cash dividend, 17.5 percent of New GM common stock, as well as warrants for an additional 2.5 percent of the common stock of New GM. Ramseyer

and Rasmusen argue that the Treasury actions led to a transfer to the UAW VEBA, which in turn is responsible for providing medical benefits to retirees.

Yet, had the restructuring of GM failed, VEBA's assets would have been depleted, and it would have been unable to pay benefits in 2009.[1] As a result, it is likely that many more of GM's retirees would have had to rely on federal health insurance programs such as Medicare, imposing additional costs on the Treasury.

What about GM's pension plans? The restructuring agreements of GM provided that New GM take over the responsibility for the GM UAW pension plan. However, had the restructuring of GM failed, those pension liabilities would not have been assumed by New GM but would have rather been reneged. Moreover, had GM dumped its pension, it could have triggered other companies with underfunded pension plans to make a similar play. For example, other automakers could have tried to rid themselves of their defined benefit plans.[2]

The wrinkle is, however, that GM's UAW pensions are insured by the Pension Benefit Guaranty Corporation (PBGC), which is a U.S. government agency. Had GM's pension plans collapsed, the PBGC would have picked up a large part of the tab. As Brown (2008) argues, since the PBGC receives no tax revenues, and given that it relies on premiums that are set by Congress, the PBGC's financial position has deteriorated, having in 2006 an $18.9 billion deficit. This is another example in which the Treasury could have ended up paying more had the restructuring of GM failed—and it is likely that GM would have failed to emerge from bankruptcy if its NOLs were not allowed to be carried forward.

SUMMARY

One can think about additional implications of a failure to restructure GM. Those include—but are not limited to—failures of auto-parts makers and suppliers, further increases in unemployment, and other forms of local economic activity, resulting in even higher costs for the federal government.

[1] See ''A Message to UAW GM Retirees'' available at http://bankrupt.com/misc/gm_uawretireeletter.pdf.

[2] See, for example, Benmelech, Bergman, and Enriquez (2011) for an analysis of pension dumping in the airline industry.

There is some rationale in having the Treasury structure a deal that leads to higher recovery by the UAW. An analysis of the transfer from Treasury to the UAW needs to take into account the different hats and pockets of the government. It is not clear that, on economic grounds, Treasury was not making the correct calculations.

REFERENCES

Benmelech, E., N. K. Bergman, and R. Enriquez. 2011. ''Negotiating with Labor under Financial Distress.'' NBER Working Paper 17192.

Brown, J. R. 2008. ''Guaranteed Trouble: The Economic Effects of the Pension Benefit Guarantee Corporation.'' *Journal of Economic Perspectives* 22 (11): 177–98.

Ross, S. A., R. W. Westerfield, and J. Jaffe. 2006. *Corporate Finance,* 8th edition. Columbus, Ohio: McGraw-Hill/Irwin.

Comment

F. H. Buckley

Mark Ramseyer and Eric Rasmusen ask three questions in their paper. First, was the Treasury notice that allowed the reorganized ''New'' General Motors to take the benefit of ''Old'' GM's past operating losses inconsistent with American tax law? Second, if it was inconsistent, might this have been an abuse of executive power? Third, if it was an abuse, is there a remedy for this? What the answers to the first two questions might be, I do not know. The third question I think I can answer.

I shall assume that the Treasury notice was inconsistent with general principles of American law. If so, the Treasury Department's decision to waive compliance with the tax laws amounted to a gift to all of the debt- and equity-holders of New GM other than the United States, including the United Auto Workers ($18.5 billion) and the Canadian and Ontario governments ($8.2 billion). Ramseyer and Rasmusen suggest that this amounted to a sweetheart deal for a labor union that was a prominent political supporter of the Obama administration. The gift, moreover, was not easily detected, and this makes it all the more suspicious.

This is not to say that the Treasury notice was corrupt and devoid of reason. It is true that much of GM's trouble had resulted from an overly generous contract with the UAW; that the sale to New GM gave an unsecured creditor, the UAW, more than it would have received under the priority rules of a Chapter 11 bankruptcy; that the claims of equally senior creditors were disregarded; that rescue bids from third parties were not accepted unless they offered the same sweetheart deal for the UAW; that, since firms in Chapter 11 have the ability to reject union contracts, GM might have ripped up

F. H. Buckley is Foundation Professor of Law at the George Mason University School of Law.

the UAW contract; that, given unemployment rates, one might have thought that an employer would be in the driver's seat; and that investors must now ask how strongly America is committed to the rule of law (see Skeel 2011). First, there was the GM bailout, then there was the UAW bailout. However, the propriety of the administration's decision is a deeply partisan issue, like every administration decision today, and it is not without its defenders who argue that it is prudent for a firm in reorganization to make a special accommodation for its employees, on whose loyalty the success of the firm depends. And so I suspend judgment on the second question.

One thing I do know: the gift to Canada was a splendid method of reaffirming the traditional friendship of the American and Canadian peoples.

A JUDICIAL REMEDY?

That leaves the third question. Assuming that the tax break was inconsistent with U.S. tax law and that this might have been an abuse of executive power, what is the remedy for it? Ramseyer and Rasmusen argue that political solutions, in which a misbehaving government is held accountable by voters, are not feasible. The separation of powers under the Constitution immunizes the executive, and in any event, voters are too ignorant to deal with matters as convoluted as this. In place of a political remedy, they propose a judicial one: let the matter be litigated before the courts.

If the courts are to confront this issue, two questions arise: First, should the executive *ever* have the discretion to waiver compliance with a law for the benefit of a single person or group? Second, if the executive does have such power, is it impracticable for a court to distinguish between a proper and improper exercise of that discretion? If the answer to both questions is yes, then the Ramseyer-Rasmusen proposal is a nonstarter.

At first glance, it might seem odd that the executive should ever have the power to dispense with a law of general application on behalf of anyone, for good reason or bad. The dispensing power would seem to invite abuse, and indeed was the subject of the first two articles of the 1689 English Bill of Rights:

> The Lords Spiritual and Temporal and Commons . . . do . . . (as their ancestors in like case have usually done) for the

> vindicating and asserting their ancient rights and liberties declare:
>
> > That the pretended power of suspending the laws or the execution of laws by regal authority without consent of Parliament is illegal;
> >
> > That the pretended power of dispensing with laws or the execution of laws by regal authority, as it hath been assumed of late, is illegal.

This would make the Ramseyer-Rasmusen proposal an easy matter for any judge. Every executive waiver would be null unless Parliament or Congress had specifically authorized it in the legislation in question. And there is indeed something to be said for a prophylactic measure of this sort. When the executive has the power to waive compliance with a law, Congress can be expected to take less care in drafting it, with the result that more bad laws are enacted. Moreover, the need to repeal the law is lessened, with the result that bad laws will stay on the books. There is, further, a concern that the executive will cut special deals for its friends, imposing the whole cost of a bad law on its enemies. For example, that concern has been voiced in the waivers for Obamacare that have been granted to labor unions (Hemingway 2011; are we seeing a pattern here?). Finally, giving the executive the power to dispense with compliance with a law might be thought to weaken the separation of powers by strengthening an already oversized executive branch (Posner and Vermeule 2011).

However, a flat prohibition of executive waivers would undoubtedly go too far. Think of waivers granted to states to come up with alternatives to federal welfare or educational mandates. Even Locke saw a value in the dispensing power. The legislature will inevitably enact overbroad laws, he said, which only the executive can easily remedy:

> The good of the society requires, that several things should be left to the discretion of him that has the executive power: for the legislators not being able to foresee, and provide by laws, for all that may be useful to the community, the executor of the laws having the power in his hands, has by the common law of nature a right to make use of it for the good of the society, in many cases, where the municipal law has

> given no direction, till the legislative can conveniently be assembled to provide for it. . . . Nay, it is fit that the laws themselves should in some cases give way to the executive power. (Locke 1689, at XIV)

For that matter, the dispensing power asserted by King James II, to which Parliament so strenuously objected, would have freed Catholic priests from the most sanguinary of punishments for the exercise of the religion they shared with their monarch. Even that good Whig, T. B. Macaulay, could find no fault with this exercise of the king's prerogative:

> For to place a Papist on the throne, and then to insist on his persecuting to the death the teachers of that faith in which alone, on his principles, salvation could be found, was monstrous. In mitigating by a lenient administration the severity of the bloody laws of Elizabeth, the King violated no constitutional principle. He only exerted a power which has always belonged to the crown. Nay, he only did what was afterwards done by a succession of sovereigns zealous for Protestantism, by William, by Anne, and by the princes of the House of Brunswick. (Macaulay 1849, Ch. 4)

Assume therefore that the executive has a dispensing power. Assume further that some waivers are benign and some corrupt, that (as Ramseyer and Rasmusen put it) the executive might be a Good Man or a Bad Man. The role of the courts, then, would be to distinguish between the two kinds of executives, between a proper and improper exercise of discretion in granting waivers.

The quite obvious problem here is that making such a distinction would necessarily involve political questions that courts wisely decline to answer. Would we want to turn over to unelected judges the question whether the bailout was needed and whether it amounted to a sweetheart deal to a loyal supporter of the Democratic Party? If we could do so, why would we need a legislature or an executive? This explains why, in a case cited by Ramseyer and Rasmusen, the Supreme Court wisely refrained from granting standing to taxpayers who claimed they had been prejudiced because another taxpayer benefited from a waiver.[1] In similar circumstances,

[1] *DaimlerChrysler Corp. v. Cuno*, 547 U.S. 332 (2006).

Canadian courts granted standing to a complaining taxpayer, but then rejected his claim because he had failed to show that the tax authorities had acted in bad faith. Same result; the American courts just got there faster.[2]

A POLITICAL SOLUTION?

If political problems should be kept from the courts, should we then look to the political process for a remedy and leave politics for the politicians? But Ramseyer and Rasmusen argue that this would not cure the UAW bailout. I think the authors are right, but for the wrong reason.

Ramseyer and Rasmusen first note that the separation of powers in the Constitution immunizes executive decisions, such as the GM reorganization, unless Congress is able to muster a supermajority to override a likely presidential veto. From this they conclude that corruption of this kind cannot be policed through the political process. There is something to this, but the paper nevertheless fails to account for the fact that parliamentary governments have been defeated for the same kinds of sweetheart deals, notwithstanding the dominance of the prime ministers in parliamentary systems.

In Canada, Prime Minister Pierre Trudeau famously described his backbenchers as ''nobodies,'' and their lack of power was recently underlined by another Liberal prime minister, Paul Martin:

> Over the last forty years or so, Canadians have seen the influence of individual members of parliament eroded as the power of the prime minister and the executive branch of government grew. . . . They vote according to the dictates of their party, and too often, when their party is in power, no one in the government cares particularly what they have to say. (Martin 2008, pp. 244–245)

The party, in turn, is dominated by the prime minister's office, which has no parallel in American politics.

Like American presidents, then, Canadian prime ministers are largely immunized from legislative control. There is always a possibility of a backbencher revolt in Parliament, but these happen very rarely. Instead, a prime minister takes his government to the people

[2] *Harris v. Canada (Minister of National Revenue)*, [2002] 2 F.C. 484.

in an election, and, if defeated, steps down and is replaced as prime minister by the opposition and as party leader by his party. And that is just what happened after a scandal that in some ways resembles the UAW sweetheart deal. The ''sponsorship scandal,'' in which a Liberal government directed revenues to favored advertising firms from 1996 to 2004 to promote the image of Canada in Quebec, was a prominent reason for the government's defeat in the 2006 general election. The government gave out $2 million in no-bid contracts to its friends, and $1.5 million was awarded for work that was never done. Small potatoes compared to the New GM reorganization, but enough to topple a government.

The Canadian example shows the weakness of another Ramseyer-Rasmusen argument against political solutions to government misbehavior. They argue that voters are irredeemably ignorant about anything so convoluted as the GM reorganization (and, having read their paper, I see the force of this objection). If so, they ask, how could we expect voters to discipline their bad executive?

And yet, in 2006, Canadian voters turned out a government that engaged in an equally questionable and obscure payoff. What Ramseyer and Rasmusen forget is the role that informational intermediaries can play in reducing a complicated set of facts to a simple message: a government of rogues is giving away your money to its friends.[3] These intermediaries include political parties (which would not exist if they failed to cure an informational asymmetry), the media (new and old), and (in Canada at least) government watchdogs. The sponsorship scandal came to light because Auditor-General Shelia Fraser had a nose for corruption and a taste for digging up government shenanigans. She became a media figure in her own right, and in a CBC poll was ranked as 66th on a list of the ''Greatest Canadians'' (behind Pamela Anderson but ahead of Joni Mitchell).

That couldn't happen here. It's hard to imagine an American comptroller general becoming a media figure. In fact, when President Obama fired Inspector General Gerald Walpin after the latter had suspended an Obama supporter for financial misdealings (*Wall Street Journal* 2009), there was barely a ripple of protest. This sort of thing helps to explain

[3] Voters are regrettably ignorant about economics, as noted in Caplan (2007). However, they seem more than able to discipline a government that has been tarnished by scandal, as the Canadian example shows.

Table 1
Transparency International's Perception of Corruption Index

	Rank	Score
Denmark	1	9.3
Sweden	4	9.2
Canada	6	8.9
Australia	8	8.7
Switzerland	8	8.7
Hong Kong	13	8.4
Germany	15	7.9
Japan	17	7.8
United Kingdom	20	7.6
United States	**22**	**7.1**

Source: Transparency International Corruption Perception Index 2010, http://www.transparency.org/policy_research/surveys_indices/cpi/2010/results.

why the United States does not come out particularly well on cross-country measures of corruption. Transparency International conducts surveys of business leaders on their perceptions about bribery, kickbacks, and public-sector anti-corruption efforts, and it ranks the United States behind many of its first-world competitors.[4]

This likely understates America's corruption problem, if corruption is understood to embrace wasteful congressional earmarks. One doesn't see legislative earmarks in Trudeau's Parliament of nobodies. Take Ruth Ellen Brosseau M.P., for example. In the 2011 Canadian election, the voters of Berthier-Maskinongé in Quebec elected the comely Brosseau, a 27-year-old barmaid. Brosseau did not visit the riding during the election campaign because she did not speak the language, and instead holidayed in Las Vegas. Her party's website notes that ''one of her passions is rescuing and rehabilitating injured animals. For many years Ruth Ellen has committed her time and energy to finding homes for stray animals in her community.'' Did I mention she is comely?

When members of Parliament are ''nobodies,'' voters don't expect them to bring any pork back to the riding. Instead, any pork comes from the national party, which has broader incentives than, say, a

[4] The Transparency International corruption rankings are quite similar to those of the World Bank. See ''Worldwide Governance Indicators,'' www.worldbank.org.

John Murtha does. Brosseau might not possess Murtha's legislative skills, but a parliament of Brosseaus more closely resembles the idealized assembly described by Edmund Burke in his Address to the Electors of Bristol, an assembly ''of *one* nation, with *one* interest, that of the whole; where, not local purposes, not local prejudices, ought to guide.''[5]

In sum, the Ramseyer-Rasmusen conclusion that the political process will not afford a remedy for the UAW bailout is likely correct. But it's not because the executive is too strong; and it's not because voters are too stupid to understand political corruption when it is pointed out to them. Rather, it's because the bailout is business as usual here.

REFERENCES

Caplan, B. 2007. *The Myth of the Rational Voter.* Princeton, N.J.: Princeton University Press.

Hemingway, M. 2011. ''Over Half of All Obamacare Waivers Given to Union Members.'' *WeeklyStandard.com*, May 16.

Locke, J. 1689. *The Second Treatise on Government.* London: Awnsham Churchill.

Macauley, T. B. 1849. *The History of England.* Philadelphia: Porter and Coates.

Martin, P. 2008. *Hell or High Water: My Life In and Out of Politics.* Toronto: McClelland and Stewart.

Milligan, K., and M. Smart. 2005. ''Regional Grants as Pork Barrel Politics.'' Center for Economic Studies and Ifo Institute for Economic Research.

Posner, E. A., and A. Vermeule. 2011. *The Executive Unbound: After the Madisonian Republic.* Oxford, UK: Oxford University Press.

Skeel, D. 2011. ''The Real Cost of the Auto Bailouts.'' *Wall Street Journal*, June 6.

Wall Street Journal. 2009. ''The White House Fires a Watchdog.'' June 17.

[5] This is not to say that pork barrel spending is unknown in parliamentary systems. On average, government spending in Canada is higher in constituencies represented by the party in power. See Milligan and Smart (2005).

Free to Punish? The American Dream and the Harsh Treatment of Criminals

Rafael Di Tella
Juan Dubra

ABSTRACT

We describe the evolution of selective aspects of punishment in the United States over the period 1980–2004. We note that imprisonment increased around 1980, a period that coincides with the ''Reagan revolution'' in economic matters. We build an economic model where beliefs about economic opportunities and beliefs about punishment are correlated. We present three pieces of evidence (across countries, within the United States, and an experimental exercise) that are consistent with the model.

Rafael Di Tella is the Joseph C. Wilson Professor of Business Administration at Harvard Business School. Juan Dubra is a professor of economics at the Universidad de Montevideo, Uruguay.

We thank Ricardo Perez Truglia for extremely helpful suggestions, as well as Anthony Doob for generous help understanding Canadian data. We also thank Jeff Miron (the editor), our commentators (Glenn Loury and Justin McCrary), as well as Sallie James, Marc Mauer, Eric Rasmusen, and participants at the Cato Papers on Public Policy *conference for extremely helpful suggestions. For ideas and exceptional research assistance, we thank Javier Donna, Ramiro Galvez, Irene Mussio, Ricardo Perez Truglia, and James Zeitler.*

Free to Punish? The American Dream and the Harsh Treatment of Criminals

1. INTRODUCTION

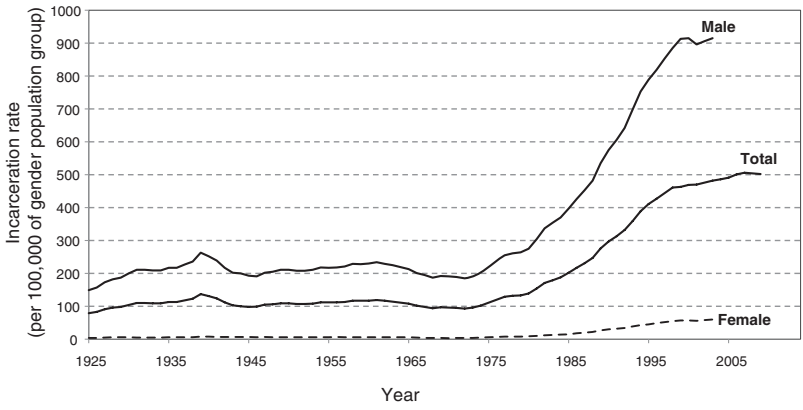
Several pieces of data suggest that contemporary policies concerning criminal punishment in America are harsh, both relative to other rich countries and relative to the country's own history. For example, the incarceration rate in the United States in the early 1970s was around 100 per 100,000 of total population, whereas it is now over 700 per 100,000. Figure 1 illustrates. It is also the highest in the world. In comparison, the average incarceration rate for European countries is somewhat over 100 (see, e.g., Walmsley 2007). Other aspects of America's penal policy also appear harsh when compared with other countries at similar levels of development, such as the use of the death penalty.[1] An important question, and one we take up in this paper, concerns the causes of harsh punishment in America.

The answer proposed in this paper is that beliefs concerning economic opportunities cause desired punishment levels in society. Although the explanation we present is relatively narrow, it is connected to the more ambitious notion that Americans punish criminals at this unprecedented scale because it is considered legitimate to do so. This stands in contrast to commonly discussed alternatives such as deterrence or the political economy of the ''prison-industrial complex.'' To emphasize (and at the risk of exaggerating), we are claiming that, even if there were well-estimated deterrent effects of imprisonment widely accepted by criminologists, this would not explain the observed increase in U.S. imprisonment, because somebody

[1] According to Amnesty International, in 2008, the United States was one of only eight countries with more than 500 prisoners on death row. With 3,263, it was second behind Pakistan. The other six countries included China, Thailand, Kenya, Bangladesh, Nigeria, and Uganda.

Many countries have designed reforms based on what they see as best practices in the United States (see, e.g., the contributions in Di Tella, Edwards, and Schargrodsky 2010).

Figure 1
Combined U.S. Incarceration Rate (Federal and State Jurisdiction) 1925–2009



Source: Bureau of Justice Statistics and University of Albany, *Sourcebook of Criminal Justice Statistics, 2003*.

would need to produce evidence that voters agree that this is a good idea.[2] Part of the difficulty is to include explanations for policies that are in all likelihood counterproductive from a recidivism standpoint (such as charging inmates telephone rates that are significantly higher than those for the general population; see, e.g., Dannenberg 2011). And of course, it would be hard to write down a deterrence model that fits the magnitude of the incarceration changes without dramatic (and implausible) changes in the other variables of the model.[3] A similar difficulty affects many explanations based on the political economy of the prison-industrial complex. If the expansion is driven by corruption or lobbying by interest groups, why do so

[2] Experimental evidence by Carlsmith, Darley, and Robinson (2002) suggests that individuals are motivated by retribution concerns (over deterrence) when choosing punishment. They study individuals who are given a short vignette describing a theft and are asked for a sentence recommendation. They show that when the probability of catching the thief in the vignette changes, the sentencing recommendation does not change, contrary to what deterrence suggests. On the other hand, sentences were harsher when the thief's motivation changed (in one case he wanted money to redistribute to the poor and in another he needed it for canceling betting debts).

[3] For example, it would require a large increase in the income of the lowest decile (the legal alternative for many criminals) in the United States relative to France.

many Americans support these policies? To qualify as an answer to the question of why such harsh punishment in America, we think, there has to be an explanation for why so many Americans are happy to support harshness levels that in other countries would be considered completely out of all proportion.

We organize our paper around a model and several pieces of evidence that are consistent with this hypothesis. In the model, we focus on agents that differ in the expected rewards for work and hence in their preferred economic system (as in Piketty 1995). Differences in the power of incentive schemes used (or in tax rates) induce further differences in effort and, a posteriori, differences in the propensity to commit crime. Inferences about the characteristics of criminals (for example, those formed by judges) differ across economic systems. This provides an economic explanation for why some ideological beliefs go together. Specifically, we show that people whose values and beliefs simultaneously include the harsh treatment of criminals and the virtues of free markets (and support for low taxes) hold a coherent model of how the world works. Put differently, criminals are ''meaner'' in systems where there are more economic opportunities, so the belief that there are more economic opportunities (for example, in America relative to Europe, or within the United States after 1980 and the Reagan presidency) is the driver of the demand for harsh punishment.[4] Our explanation is thus connected to work on the expressive content of the law, where policymakers ''send a message'' about society's values by setting harsh sentences (see Sunstein 1996 and Benabou and Tirole 2011).

We provide different pieces of evidence that are consistent with the idea that economic beliefs cause punishment. First, we compare the United States to other countries and show that the desire to punish criminals and certain economic beliefs (such as that effort,

[4] Merton (1938) argued that high crime rates in America were a result of the psychological stress created by the gap between a reality of limited opportunities and a generalized belief in the ''American dream.'' See also Messner and Rosenfeld (2001) and Cullen and Agnew (2003). They do not explain, however, why such harsh punishment might be associated with these beliefs, particularly if such mitigating circumstances are present. Closer in spirit to our approach is the fascinating comparative historical study by Whitman (2003). He argues that American rejection of status-oriented European societies based on a strong state led to the adoption of egalitarian harsh punishment.

rather than luck, matters in the determination of income) are relatively more widespread in the United States. We also note that there is a positive correlation between these two variables in a small cross section of countries. This section reveals two key limitations of our paper. The first is causality: obviously this correlation does not establish a causal link, and even if it did (later in the paper we have some causal evidence), it does not show that the link originates in the particular mechanisms outlined in our model. The second is measurement error: Any study dealing with people's beliefs and with punitiveness (either people's desire to punish or as expressed in the classifications of the legal system) has to deal with imprecise measures, particularly when it involves people living in different time periods or geographical jurisdictions. This makes it difficult to design convincing tests to distinguish between alternative hypotheses.

Our second piece of evidence reveals that, within the United States, beliefs about the economic system have moved toward the right end of the ideological spectrum over time, particularly for African Americans. We also show that the proportion of people who support the death penalty and the average belief in ''effort pays'' are positively correlated across U.S. states. The data also show that there is a correlation between beliefs and punitiveness at the individual level: people who believe effort pays also support the death penalty.

Finally, we conduct an experiment to provide at least suggestive evidence on one aspect of the causal link between beliefs and punitiveness. Students are randomly exposed to hypothetical situations involving criminals from neighborhoods with different economic opportunities. Students who were exposed to a criminal who grew up in a neighborhood with good educational prospects that were associated with economic progress supported tougher punishment (for the same crime) than those exposed to a criminal who did not have those opportunities. Although the causal link we develop in the model is more complex and there is obviously a question of the external validity of this empirical exercise, the evidence suggests that beliefs in economic opportunities cause punitiveness.

Our paper is related to a large literature on the structure of ideology. Several authors have studied the nature of political beliefs, many of them observing the fact that ideological beliefs often come in bundles (see, e.g., Lipset 1979, *inter alia*; see also the discussions

in Rokeach 1973, Feldman 1988, Inglehart 1990, and Zaller 1991).[5] Two important questions are why beliefs about one issue differ across people within the same society who presumably observe the same reality, and why beliefs about different processes (e.g., whether firms pollute too much and whether effort pays) are often correlated. One interesting approach puts emphasis on explaining the structure of beliefs as a coherent outcome when individuals organize information using metaphors (see, e.g., Lakoff 1996). An alternative approach is taken by psychologists who study belief bundling as originating in personality traits and goes back to the work on fascism and authoritarian personality by Adorno et al. (1950). Views about motivated social cognition emphasize that belief systems are adopted largely to satisfy some psychological need (see Jost et al. 2003 for a recent example and discussion of the relevant literature).[6]

An alternative approach, which we emphasize, focuses on how the economic structure might connect beliefs across issues through political and economic choices (see, e.g., Hall and Soskice 2001). A classic example in economics is Piketty (1995), who shows that people who believe effort pays are more likely to believe that low taxes are best—a connection that might be reinforced when people choose compensation schemes. In this paper, we take this approach by emphasizing that people who believe effort pays will vote for (and choose privately) high-powered incentive schemes, which will have a consequence on desired sentences because the type of people committing crimes in such settings will differ from those choosing to be criminals in places with low-powered incentives.

In Section 2, we present a simple model to illustrate how beliefs may cause differences in the way societies organize their economic

[5] Some of this work emphasizes how left/right political choices reflect the basic cleavages in society (see, e.g., Lipset and Rokkan 1967, who emphasize the importance of religion and class). For descriptions of American's beliefs and attitudes, see Hochschild (1981), Inglehart (1990), and Ladd and Bowman (1998).

[6] The specific connection between meritocratic beliefs (sometimes approximated as ''free will'') and punitiveness has also been explored in experimental settings. On the one hand, subjects manipulated to believe less intensely in free will have been shown to be more likely to lie, cheat, steal, and become aggressive (Vohs and Schooler 2008, Baumeister et al. 2008). On the other hand, when people can place blame for an offense on someone, even if undeservedly, they become more punitive (Sanfey et al. 2003). Shariff et al. (2011) show that an eroded belief in free will can soften retributive impulses toward violent criminals.

systems, the types of criminals, and the desired punishments. In Section 3, we present our three pieces of evidence: some cross-country evidence, evidence for the United States, and finally evidence from our experimental exercise. Section 4 concludes.

2. A MODEL WHERE BELIEFS ABOUT THE ECONOMIC SYSTEM CAUSE PUNISHMENT

In this section, we present a variation of the model in Di Tella and Dubra (2008) that incorporates several improvements. First, in order to analyze the increase in punitiveness and in the belief that effort pays in the United States in the past 30 years, we have incorporated income changes in order to study the role of GDP growth, an element that seems important in a model where beliefs matter.[7] Second, we provide a better (more precise) approach to modeling whether exerting effort is profitable. Finally, the model is more flexible because the source of variation across individuals is a ''type,'' which can now be interpreted in several ways. For example, types can include ''laziness'' (in accordance with the World Values Survey question concerning whether poverty is due to laziness or because of bad luck) or, more generally, any innate or ''environmental'' factor that makes effort by the individual more costly (for example, if the individual has erroneous perceptions about the ''profitability'' of exerting effort or if the individual's education was not conducive to good work habits). This allows naturally for discussions of several topics that others have argued are important in the decision to commit crime (like segregation in particularly ''bad'' neighborhoods, or identity).[8]

The basic model has three agents: firms, workers, and the government who must simultaneously choose their actions. Firms must choose whether they want a market technology, M, where effort and training by workers matters, or a bureaucracy, B, where output is independent of effort. Workers must choose whether they will be criminals, work with low effort $e_L = 0$, or work with high effort

[7] More specifically, one of the potential advantages of an economic system where belief in ''effort pays'' prevails is that individuals end up putting forth more effort and there are material gains.

[8] See Sampson and Loeffler (2010) for fascinating evidence on the concentration of prisoners.

$e_H = 1$. The government must choose a punishment level, time in jail, t for criminals.

For a parameter g representing technological progress, the wealth level $gw(s, e)$ of the individual when facing a technology $s = M, B$ and exerting effort $e = e_H, e_L$ is given by $w_h = w(M, e_H)$, $w_l = w(M, e_L)$, and $w_m = w(B, e_H) = w(B, e_L)$. In this paper, w is exogenous, but it can easily be made the endogenous result of a competitive model.

Workers are of one of two effort types: low θ_L or high θ_H, and let ρ denote the probability of a type θ_H. For a wealth level gw and effort e, an individual of type θ has a utility $u(gw, e; \theta) = gw - (1 - \theta)e$ if he chooses to work. As will be clear shortly, low types will be more likely to become criminals.

From the form of the utility function, at least three interpretations arise: First, one can interpret θ_H as a hard-working type, since the cost of effort is lower than for the ''lazy'' type θ_L: for the hard-working individual, the cost of effort is $1 - \theta_H$, while for the lazy one, it is $1 - \theta_L$. A second interpretation is that a type θ_L was raised in an environment with ''low-quality'' work habits, so that a greater effort level is required to obtain the same results as somebody who was raised in an environment conducive to ''high-quality'' work habits. In this case, the effort level e is not measured in ''hours'' but rather in effective units of effort. Finally, a somewhat related interpretation is that a type θ_L is one who believes that effort is not very useful (say, has a low productivity), and so a lot of hours of effort would be needed to obtain a certain objective; meanwhile, a type θ_H thinks that effort is highly productive and that a small number of hours would suffice to obtain the given objective. In this interpretation, for example, e_H could be ''obtain a university degree,'' while e_L could be ''be a high school graduate.'' Types θ_H may then think that obtaining a degree would involve 20 hours of study per week, while types θ_L could believe that it would require 40 hours.

The payoffs for the worker and the per-worker profit of the firm are presented in the matrices shown in Figure 2, where the matrix on the right is simplified using $e_H = 1$ and $e_L = 0$.

As we explain below, this economic structure gives rise to two different equilibria: the ''American equilibrium'' and what can be called the ''French equilibrium.'' In the American equilibrium, most workers choose a high level of effort (or training) because they believe that effort pays; in this equilibrium, given that workers are

Figure 2
Firm's Technology Choice

Worker	Market	Bureaucracy		M	B
High Effort	$gw_h - (1-\theta)\, e_{H_1}\, \pi_H$	$gw_m - (1-\theta)\, e_{H_1}\, \pi_M$	H	$gw_h + \theta - 1_{H_1}\, \pi_H$	$gw_m + \theta - 1_{1}\, \pi_M$
Low Effort	$gw_l - (1-\theta)\, e_{L_1}\, \pi_H$	$gw_m - (1-\theta)\, e_{L_1}\, \pi_M$	L	$gw_l \quad {}_1\, \pi_L$	$gw_m \quad {}_1\, \pi_M$

exerting effort, it is profitable for firms to choose a market technology, and this ''confirms'' that workers' beliefs that effort pays are correct in equilibrium. In the French equilibrium, workers anticipate that effort does not pay and choose low effort; firms correctly forecast this behavior and choose a bureaucracy, ensuring that workers' beliefs that effort doesn't pay are correct in equilibrium.[9]

If the individual decides to participate in the labor market, he collects his lifetime wealth and pays his effort cost. If he opts for crime, his payoff is $c + \mu$, where c is the expected utility of crime and μ is a taste shock for criminal activity. In order to link the punishment rate with the utility of the individuals, we assume that the direct utility from crime is a continuous function $c(t)$, where the variable t is time in jail. Of course, other dimensions (like the probability of apprehension) matter, but in order to simplify our analysis, we take them as given. We also assume that c is decreasing in t. The taste shock μ is the meanness of the individual (a large μ is an unkind individual), which is drawn from a density f that is positive in its support $[\mu_{\min}, \mu_{\max}]$ and has a cumulative distribution function F. In Di Tella and Dubra (2008), we discuss alternative interpretations of μ.

Given a pair of presumed strategies (s,t) for the firm and the government, the individual decides whether to enter the crime market by comparing his utility $gw - (1 - \theta)e$ with $c + \mu$. He commits a crime if and only if $gw - (1 - \theta)e < c + \mu$.

The government must choose the time in jail for a criminal, which in turn determines c. In other words, once the government has proved that the individual has committed a crime, it must decide

[9] For a model where stigma and self-fulfilling expectations of criminality lead to multiple equilibria, see Rasmusen (1996).

the time t he must spend in jail. We assume that for some increasing function q, the government has a utility $1(q(\mu) - t)^2$ of punishing with t years a type μ; if the government knew that the individual was of a certain meanness μ, it would choose a punishment level $t = q(\mu)$. Since q is increasing, it means that the government wants to punish ''worse'' individuals more. More generally, and denoted by E_h the expected value with respect to a belief h about μ, the government must choose t to maximize $v(t, \mu) = -E_h[(q(\mu) - t)^2]$. This yields a desired punishment of $t = E_h[q(\mu)]$.

To see why we obtain our basic results (higher punishment in America than in ''France'' and higher punishment in America today than 30 years ago), note that the government's beliefs about the types of apprehended criminals, h in the formulation above, depends on the economic system. For example, if criminals in a certain environment are ''meaner,'' on average, than in another environment because economic opportunities are better (and hence only really mean individuals commit crimes), then the government will choose a harsher punishment.

2.1 Two Worked-out Examples

We now present two worked-out examples in order to illustrate how the model operates.

Set $w_h = 2$, $w_m = \frac{3}{2}$, $w_l = \frac{5}{4}$, and $g = 1$. Let $\theta_H = \frac{3}{4}$ and $\theta_L = 0$, and let $\rho = \frac{3}{4}$ be the probability of type θ_H. Let $\pi_H = 4$, $\pi_M = 2$, and $\pi_L = 1$. Let us also assume that $c(t) = \frac{3}{2} - t$, that μ is uniformly distributed in $[-2, 2]$, and finally that $q(\mu) = (104\mu - 99) \div 92$.

With these parameters, two equilibria arise:

- The American Dream equilibrium, where the firm chooses a market technology, high-effort types exert effort while low-effort types don't (a portion of each type commits crimes), and the government chooses a high punishment level.
- The French equilibrium, where the firm chooses a bureaucracy technology, all types exert low effort (and again, some individuals of each type commit crimes), and the government chooses a low punishment level.

It is easy to check that these are the unique equilibria in pure strategies. We first analyze the French equilibrium, which is easier.

Since workers are choosing low effort, it is a best response for the firm to choose a bureaucracy. Assume now that the desired punishment by the government is $t^F = \frac{1}{8}$, and we will then check that this is indeed optimal. Given a bureaucracy, neither θ_H nor θ_L would choose to exert high effort, so the only choice is between low effort, which yields $w_m = \frac{3}{8}$, or crime that gives $c(t^F) + \mu = \frac{3}{2} - t^F + \mu = \frac{11}{8} + \mu$. Individuals in this equilibrium commit crimes if and only if $\frac{11}{8} + \mu > \frac{3}{2}$, or $\mu > \frac{1}{8}$. Hence, the expected value of μ if a crime has been committed (in a bureaucracy) is the midpoint between 2 and $\frac{1}{8}$: $E[\mu|C, B] = \frac{17}{16}$. Then, the optimal strategy of the government is to choose $t^F = E[q(\mu)] = (104E[\mu] - 99) \div 92 = \frac{1}{8}$, as was to be shown.

The American equilibrium is somewhat more involved, since different θs behave differently.[10] In order to analyze the equilibrium, assume that the desired punishment by the government in this case is $t^A = \frac{1}{4}$, and we will then check that this is indeed the optimal thing to do. A type θ_H with meanness μ has to choose among high effort, which yields $gw_h - (1 - \theta)e = 2 - \frac{1}{4} = \frac{7}{4}$; low effort, which gives utility $\frac{5}{4}$; and crime, which nets him $c(t_A) + \mu = \frac{3}{2} - t^A + \mu = \frac{5}{4} + \mu$. Therefore, he commits a crime if and only if $\mu > \frac{1}{2}$. Similarly, low types θ_L commit crimes if and only if $\mu > 0$. Since all types μ greater than $\frac{1}{2}$ commit crimes and only θ_L individuals with types $0 < \mu < \frac{1}{2}$ become criminals, the probability that an individual becomes a criminal (in a market technology) is

$$P(C; M) = P\left(\frac{1}{2} < \mu < 2\right) + P(\theta_L) \times P\left(0 < \mu < \frac{1}{2}\right)$$
$$= \frac{2 - \frac{1}{2}}{4} + \frac{1}{4} \times \frac{\frac{1}{2} - 0}{4} = \frac{13}{32}$$

[10] Specifically, in the French equilibrium, both μ_H and μ_L are associated with the same cutoff in μ for which meaner types commit crimes, whereas in the American equilibrium they have different cutoffs, with the one associated with μ_H higher than the one for μ_L.

Therefore, the posterior belief that a criminal has a type $0 < \mu < \frac{1}{2}$ is the probability of a type in that range, times the probability that it is a θ_L, divided by the probability of a crime being committed:

$$P\left(0 < \mu < \frac{1}{2} \middle| C, M\right) = \frac{P(\theta_L)P\left(0 < \mu < \frac{1}{2}\right)}{P(C; M)}$$

$$= \frac{1}{4} \times \frac{\frac{1}{2} - 0}{4} \times \frac{32}{13} = \frac{1}{13}$$

Hence the expected value of μ in the American equilibrium is the probability that μ is in $[0, \frac{1}{2}]$ multiplied by the expected value conditional on μ in that interval (which is just the midpoint of the interval), plus the probability that μ is greater than $\frac{1}{2}$ multiplied by the expected value conditional on that interval (again, the midpoint between $\frac{1}{2}$ and 2). That is,

$$E(\mu; C, M) = \frac{1}{13} \times \frac{1}{4} + \frac{12}{13} \times \frac{5}{4} = \frac{61}{52}$$

Then, the optimal strategy of the government is $t^A = E[q(\mu)] = (104E[\mu] - 99) \div 92 = \frac{1}{4}$, as was to be shown. Given the strategies of workers, where most exert high effort (a proportion larger than $\rho = \frac{3}{4}$), it is optimal for firms to choose a market technology. This completes the analysis of the first example, where punishment in the American equilibrium is larger than in the French equilibrium.

The above example concerns a cross section of punitiveness levels. Our second worked-out example concerns the analysis of a ''time series'' of what happens when the economy grows. In order to analyze this case, we leave all parameter values as before but increase g from 1 to $g = \frac{6}{5}$. This has the effect of raising wages in both the French and American equilibria. Following the same steps as before, it is easy to check that $t^F = \frac{103}{720}$ while t^A is approximately $\frac{16}{25}$. The

desired punishment increased by 15 percent in the French equilibrium, while it increased by 160 percent in the American equilibrium. If we interpret growth in g as the increase in incomes during the 1980s and 1990s, this example illustrates two stylized facts from the imprisonment literature: a small increase in severity in the French equilibrium (and more generally around the world, see Walmsley 2007) and an even larger increase in the desired punishment in the American equilibrium.

The appendix discusses possible ways to extend the model, connecting it to issues that others have claimed to be relevant to the crime-punishment discussion (such as biased sampling in segregated neighborhoods).

3. THREE PIECES OF EVIDENCE

We now focus on three pieces of evidence connecting beliefs and punitiveness. As emphasized above, the evidence is only suggestive of the relationship outlined in the model, as establishing tight causal links is beyond the scope of this paper. Note also that there are many peculiarities in the U.S. penal system (and several of them contribute to increases in punitiveness, such as ''truth-in-sentencing'' laws), but we do not review them here (see, e.g., Austin et al. 2000).[11] Instead, we selectively include pieces of evidence that we see as relevant to a theory connecting imprisonment to beliefs.

Before presenting the evidence, we note some selected observations related to the evolution of the U.S. data. First and most basic is that punitiveness in the United States is higher now than it was historically. See Figure 1 above. Several legal initiatives gradually loosened restrictions on the activities of law enforcement officials in the 1970s. Later on, the Comprehensive Crime Control and Sentencing Reform Acts of 1984 introduced stricter sentencing (mandatory minimums for many categories of drug- and gun-related offenses) and new search and seizure powers. Over time, truth-in-sentencing laws have been introduced federally and in several states. These require prisoners to serve 85 percent of their sentence before being

[11] The case of Maricopa County (Ariz.) Jail offers several such peculiarities, with poor conditions including ''chain gangs for men and women,'' inmates that are ''forced to wear old-fashioned prison stripes and pink underwear,'' and that ''prohibited items include cigarettes, adult magazines, hot lunches and television.'' See *CNN.com*, ''Arizona Criminals Find Jail Too—in Tents,'' July 27, 1999.

eligible for parole. In 1994, a popular ballot initiative brought in California's controversial ''Three Strikes Law,'' with lengthy and mandatory prison terms for repeat offenders. Simultaneously, it is possible to observe reductions in prison alternatives (electronic monitoring) and re-entry programs (including parole, probation, psychiatric care, and rehabilitation). The increase in imprisonment was not steady, with a clear break around 1980—a time when ideological changes associated with the Reagan revolution took place (some are documented in Section 3.2. below). The rate of incarceration in the United States hovered around 100 per 100,000 population from the 1920s (when we first have readily available data) to 1980, when it began an upward trend. During the early 2000s, it stabilized somewhat. Indeed, formal estimates (as in Perron 2005) indicate structural breaks in 1978 and 2001.

A substantial part of this increase has taken place in minimum security prisons.[12] Between 1979 and 2005, the percentage of inmates held in maximum security was halved, from 40 percent to 20 percent; the percentage of inmates in minimum security nearly doubled, from 18 percent to 34 percent.[13] In 1979, state prisons held less than one minimum security prisoner for every maximum security prisoner; in 2005, state prisons held nearly three minimum security prisoners for every one maximum security prisoner. It appears that a lot of the changes in incarceration rates involve offenders who are judged to be less dangerous.

[12] One factor is longer sentences for less severe crimes. The ''war on drugs'' has played a role, as there has been a substantial increase in people incarcerated for drug offenses. Austin et al. (2000, p. ii) write:

> [I]n 1980 the number of prisoners convicted for a drug offense was six percent of the state prison population, which numbered less than 300,000. By 1998 the numbers had increased by 237,000, or 21 percent of the state prison population. Furthermore, the average sentence for drug offenses had increased from 13 months in 1985 to 30 months by 1994.

At the federal level the increase was 10-fold. Mauer (2008) reports that burglars in the United States serve an average of 16 months in prison, whereas in Canada they serve 5 months on average (and 7 months in England).

[13] The estimates in this paragraph use population estimates from the U.S. Census Bureau (various years) and inmate statistics from the U.S. Department of Justice, Bureau of Justice Statistics, *Census of State and Federal Adult Correctional Facilities*, 1979, 1984, 1990, 1995, 2000, and 2005.

As is well known, some minorities are imprisoned at disproportionate rates. For example, the black incarceration rate (relative to the black population) is substantially higher than the white incarceration rate, in some states by a factor of almost 10 (see, e.g., Mauer and Ryan 2007). See Table 1. Convincing evidence of racism is provided by Alesina and La Ferrara (2011), who study all death penalty appeals that became final between 1973 and 1995 and show that the probability of judicial error is up to 9 percentage points higher for minority defendants who killed white victims than for those who killed minority victims.[14] There is a large body of work on racism and the mass incarceration of so many black (and Hispanic) young men, which we do not review here (for a recent example, see Alexander 2010). Even though these accounts make many valid points, they fail to account for the simple fact that few people who support punishment see themselves as racist. Interestingly, a first look at the evidence suggests that the increase in the overall incarceration rate has approximately preserved the 1980 differences in incarceration rates by race. Given that there was a large difference in 1980, the increase in imprisonment has affected blacks disproportionately. As a percent of the total population in each group, the incarceration rates of blacks was over six times that of whites both in 1980 and in 2009.[15]

It is interesting to compare the United States with other countries. In 2007, the incarceration rate in the United States was 756 per 100,000 population, whereas it was significantly lower for Europe (average of 125). Although in some countries there certainly was an increase in imprisonment, the dynamics were nowhere as extreme as in the United States. Canada, which in many ways is a good

[14] Other work in criminology has studied bias in the legal system using the assumption that racial differences in arrests indicate differences in criminal involvement. One study concluded that close to 76 percent of the racial bias in imprisonment can be attributed to differences in criminal involvement of racial groups (see Blumstein 1993; in 1978 this proportion was 80 percent). Although the assumption of unbiased arrest rates seems contrary to anecdotal evidence (we were unable to find convincing studies that could be generalized), Hindelang (1978) found that the racial differences in arrests mirrored racial identities of offenders as reported by the victims in the National Crime Victimization Survey, although this evidence refers to the period prior to the escalation in imprisonment.

[15] Calculations based on ''Correctional Populations in the United States and National Prisoner Statistics,'' U.S. Department of Justice. Population figures taken from U.S. Census estimates.

Table 1
Racial Disparity in Incarceration Rates

Year	Percentage of White Population Incarcerated	Percentage of Black Population Incarcerated	Ratio of Black Percentage to White Percentage
1980	0.18	1.11	6.3
1990	0.36	2.36	6.6
2000	0.41	3.41	8.3
2009	0.43	2.99	6.9

Source: Authors' calculations using data from U.S. Census and U.S. Department of Justice.

counterfactual, is characterized by its stability: since the 1950s, it has imprisoned approximately 100 per 100,000 population (Webster and Doob 2007).[16]

Crime rates in the United States are not generally much higher than those prevailing in Europe. See Table 2 for data for a U.S.–Europe comparison across crime categories during the late 1990s. Homicide is the one possible exception: Figure 3 provides a graph of incarceration rates and a measure of homicides (from the World Health Organization). Incarceration and homicide rates have a mildly positive correlation (see Bushway and Paternoster 2009 and Durlauf and Nagin 2010 for clear discussions, including the difficulties in making causal interpretations given the possible presence of deterrent and political economy effects, as well as references to previous work). The United States is still an outlier, with extremely high levels of incarceration. This is the same conclusion emerging from the study by Raphael and Stoll (2009), who decompose the changes in incarceration and find that only a small proportion of the increase is attributable to increases in criminal behavior (at most

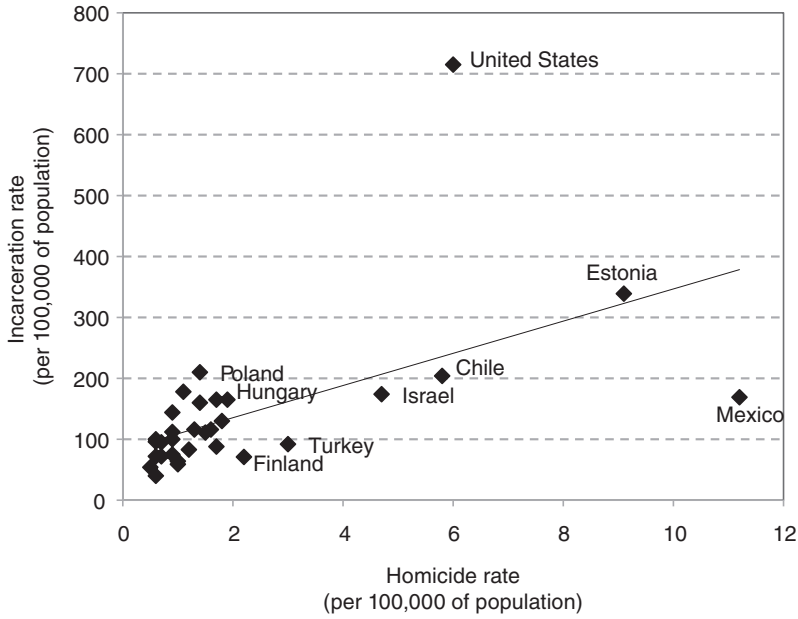
[16] Unfortunately, it is not possible to see if Quebec has different punitiveness than the rest of Canada, looking at imprisonment rates, due to several features of the Canadian legal system (for example, there is a lot of shifting of prisoners across provinces). See Doob and Webster (2006). Furthermore, the Canadian federal system (which holds prisoners sentenced to more than two years) is quite substantial and it is not easy to obtain data on the province in which prisoners were sentenced. For 1995 and 2003, the three largest provinces (British Columbia, Ontario, and Quebec) are quite similar in their *overall* imprisonment rates. We are extremely grateful to Anthony Doob for providing us with these data. Kensey and Tournier (1999) describe prison inflation in France in detail. Walmsley (2007) writes that prison populations have risen in 69 percent of European countries.

Table 2
Crime and Punishment in the United States and Europe
(Late 1990s)

	Incarceration Rate	Victim Rate	Crime				
			Total	Car	Property	Sex	Person
United States	645	24.2	5,375	19.6	10.8	2.5	5.7
Europe	88	25.2	7,984	19	9.6	2.9	4
Austria	85	18.8	6,285	11.7	6.6	3.8	2.1
Canada	115	25.2	9,979	17.3	13.1	2.7	4
England and Wales	125	30.9	n/a	24.7	12.8	2	5.9
Finland	55	18.9	7,650	12.9	5	2.5	4.1
France	90	25.3	6,765	20.7	9.5	0.9	3.9
Netherlands	85	31.5	7,422	25.9	13.3	3.6	4
Sweden	60	24	12,670	20	7.5	2.9	4.5
Switzerland	90	26.7	5,116	18.6	9	4.6	3.1

Notes: The figures for Europe correspond to the unweighted average across the European countries in the table. Data on incarceration rates come from the first edition of the ''World Prison Population List.'' *Incarceration Rates* are prison populations per 100,000 of national population. The incarceration rates are from 1997, except for England and Wales and France, where the data are from 1998. *Crime* is the total recorded crime per 100,000 population from the United Nations Surveys on Crime Trends and the Operations of Criminal Justice Systems. All data are from 1994 except for the Netherlands, which are for 1986. *Victim Rate* is the victimization rate (the proportion of the population victimized in one year), for 1995 (the latest year available) in P. Mayhew and J. J. M. van Dijk *Criminal Victimization in 11 Industrialized Countries: Key Findings from the 1996 International Crime Victims Survey*, The Hague: Ministry of Justice, WODC, 1997. *Car* is victimization rates for car theft, theft from car, car damage, motorcycle theft, and bicycle theft from the same surveys. *Property* is victimization for burglary, attempt at burglary, robbery, and theft of personal property from the same surveys. *Sex* is sexual offenses victimization from the same surveys. *Person* is assault and threat victimization from the same surveys.

Figure 3
Prison Population and Homicides in OECD Countries, 2004



Notes: The variable on the y-axis (*Prison Population*) is the number of prisoners in the country's national prison system (including pretrial detainees/remand prisoners) per 100,000 of the country's national population. The source is the *World Prison Brief*, International Centre for Prison Studies, 2003. The variable on the x-axis (*Homicides*) is the number of homicides (defined as unlawful death purposefully inflicted on a person by another person) per 100,000 of the country's national population in 2004. The data were obtained from the United Nations Office on Drugs and Crime's homicide statistics, which are based on public health sources. The sample covers 31 OECD countries.

17 percent of total growth). These authors attribute the bulk of the increase to longer time served and to an increase in the likelihood of being sent to prison (conditional on committing a crime).[17]

[17] These authors note that average time served in the aggregate has not increased even though we now have longer sentences (conditional on type of crime.). The reason is that prison admissions have shifted toward less serious offenses, consistent with the increase in minimum security prisons we document.

3.1 Punishment and Economic Beliefs in the United States and Other Developed Countries

Given the difficulties in interpreting data involving legal definitions across countries, and that incarceration confounds the amount of crime, enforcement efforts, and other factors with desired punitiveness, it is useful to study alternative measures. We derive a measure of desired punishment from the 2004–2005 International Crime Victims Survey (ICVS). This is a comprehensive survey developed to monitor crime, perception of crime, and attitudes toward the criminal justice system in a comparative international perspective, financed largely by the United Nations and the European Union.[18] The main question for our purposes is:

> People have different ideas about the sentences which should be given to offenders. Take for instance the case of a man of 20 years old who is found guilty of burglary for the second time. This time, he has stolen a colour TV. Which of the following sentences do you consider the most appropriate for such a case: (1) Fine, (2) Prison, (3) Community service, (4) Suspended sentence, (5) Any other sentence.

A simple way to summarize the data is through the percentage of respondents opting for imprisonment as punishment for the recidivist burglar. The percentage of the public opting for imprisonment as punishment for a recidivist burglar in the United States was 47, while the average for 22 European countries included in the sample was 25.4 (s.e. 2.4).

Data on beliefs about the economic system come from the fifth wave of the World Values Survey (2005–2008). The first belief that we use is based on the standard question on self-placement on the ideological spectrum:

> In political matters, people talk of ''the left'' and ''the right.'' How would you place your views on this scale, generally speaking?

[18] Standardized questionnaires and other aspects of data collection provide some reassurance regarding data quality. The biggest apparent drawback is that it is telephone-based, although it appears that experimental work in the Netherlands comparing answers to the ICVS survey using telephone (CATI) interviews with face-to-face interviews produce similar results (see Scherpenzeel 2001, cited in van Dijk et al. 2008).

The response takes values from 1 to 10, where 1 is Left and 10 is Right.

The second belief is constructed based on the following question:

> Now I'd like you to tell me your views on various issues. How would you place your views on this scale? 1 means you agree completely with the statement on the left [side of the page]; 10 means you agree completely with the statement on the right [side of the page]; and if your views fall somewhere in between, you can choose any number in between.
>
> [Left-side statement:] In the long run, hard work usually brings a better life.
>
> [Right-side statement:] Hard work doesn't generally bring success—it's more a matter of luck and connections.

We inverted the scale so that 1 means ''Hard work doesn't generally bring success—it's more a matter of luck and connections,'' and 10 means ''In the long run, hard work usually brings a better life.''

The last belief is constructed based on the following question:

> Many things may be desirable, but not all of them are essential characteristics of democracy. Please tell me for each of the following things how essential you think it is as a characteristic of democracy:
>
> Governments tax the rich and subsidize the poor.
>
> [Left-side statement:] Not an essential characteristic of democracy
>
> [Right-side statement:] An essential characteristic of democracy.

We inverted the scale so 1 means ''An essential characteristic of democracy,'' and 10 means ''Not an essential characteristic of democracy.'' Note that these beliefs are coded so that higher numbers indicate the respondent is closer to what is typically interpreted as the right end of the ideological spectrum. As revealed by several prior papers (see, e.g., Alesina et al. 2001), beliefs in the United States are more toward the right end of the ideological spectrum than in Europe.

More interestingly, Figure 4 reveals that there is a positive relationship between right-wing answers (using the three measures of beliefs) and the percentage of people recommending prison in the ICVS question. In Di Tella and Dubra (2008), similar results are presented using somewhat different samples.

3.2 The Punishment–Economic Beliefs Correlation in the United States

We now turn to evidence within the United States. We divide the evidence into movements in the aggregate data in the United States and correlations across states; we then turn to individual-level correlations.

Data on beliefs come from the U. S. General Social Survey (GSS), a repeated cross section of randomly sampled Americans (for a description see Davis and Smith 2005). Each survey is an independently drawn sample of English-speaking persons 18 years of age or over living in the United States. One of the basic purposes of the GSS is to gather data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. We focus on two concerning the role of effort (vs. luck) in the income-generating process, which might loosely be called ''self-reliance.''

The first question in the GSS that can serve such purpose is:

> Some people say that people get ahead by their own hard work; others say that lucky breaks or help from other people are more important. Which do you think is most important?''

The options were ''Hard work most important,'' ''Hard work, luck equally important,'' and ''Luck most important.'' We created the variable *Effort Pays*, which takes the value 1 if the individual responded *''Luck most important*,'' 2 if the individual responded ''*Hard work, luck equally important*,'' and 3 if the individual responded ''*Hard work most important*.'' (We treat ''*Don't Know*'' as a missing value.) Thus, higher values of *Effort Pays* can be interpreted as an individual that is more likely to believe that effort pays.

The second alternative measure of self-reliance can be created exploiting the answers to the following question:

> Some people think that the government in Washington should do everything possible to improve the standard of

Figure 4
Beliefs and Demand for Punishment across Countries

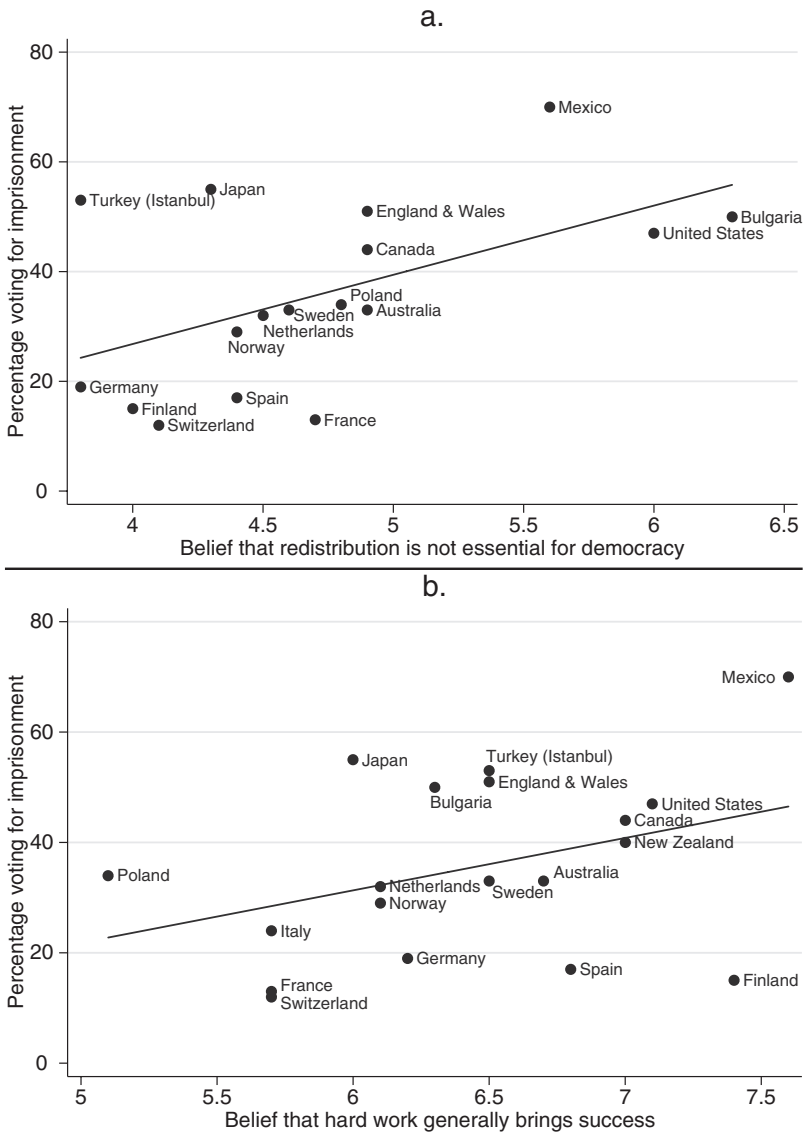
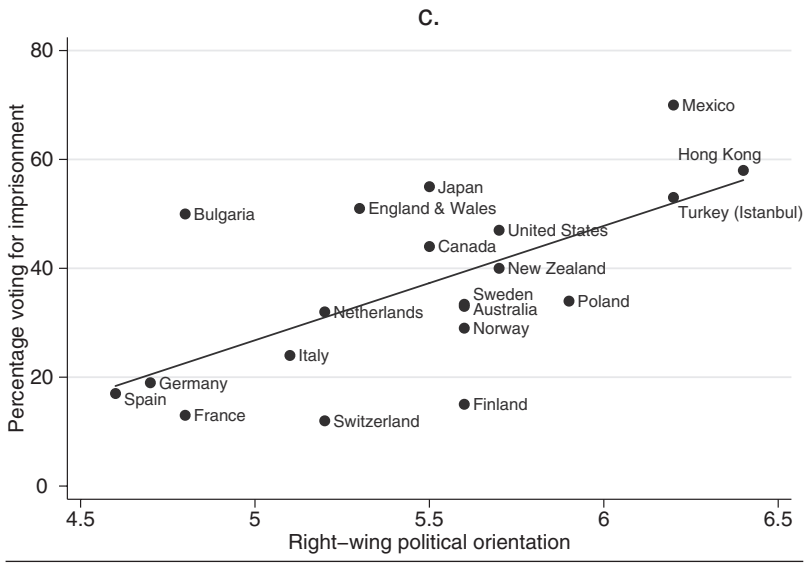


Figure 4
(continued)



Notes: The variable on the y-axis is derived from the question: ''*People have different ideas about the sentences which should be given to offenders. Take for instance the case of a man of 20 years old who is found guilty of burglary for the second time. This time, he has stolen a colour TV. Which of the following sentences do you consider the most appropriate for such a case: (1) Fine, (2) Prison, (3) Community service, (4) Suspended sentence, (5) Any other sentence.*'' (ICVS, 2004–2005) The y-axis variable is the percentage of the public opting for imprisonment as punishment for the recidivist burglar. The x-axes use data on beliefs about the economic system from the fifth wave of the World Values Survey (2005–2008). In panel (a) the belief comes from the question: ''*Many things may be desirable, but not all of them are essential characteristics of democracy. Please tell me for each of the following things how essential you think it is as a characteristic of democracy: Governments tax the rich and subsidize the poor.*'' We inverted the scale used in the survey so that 1 means ''*It definitely is an essential characteristic of democracy*,'' and 10 means ''*Not at all an essential characteristic of democracy*.'' In panel (b) the question used is: ''*Now I'd like you to tell me your views on various issues. How would you place your views on this scale? 1 means you agree completely with the statement on the left [side of the page]; 10 means you agree completely with the statement on the right [side of the page]; and if your views fall somewhere in between, you can choose any number in between. Agreement: Hard work brings success.*'' We inverted the scale such that 1 means ''*Hard work doesn't generally bring success—it's more a matter of luck and connections*,'' and 10 means ''*In the long run, hard work usually brings a better life*.'' Panel (c) uses self-placement: ''*In political matters, people talk of ''the left'' and ''the right.'' How would you place your views on this scale, generally speaking?''* The response takes values from 1 to 10, where 1 is Left and 10 is Right.

> living of all poor Americans; they are at Point 1. Other people think it is not the government's responsibility, and that each person should take care of himself; they are at Point 5. Where would you place yourself on this scale, or haven't you made up your mind on this?

We created the variable *Not-Washington*, which is simply the answer to the question, so that higher values mean that the respondent is more ''individualist'' in the sense that he believes that each person should take care of himself. (We treat ''*Don't Know*'' as a missing value.)

The questions discussed above are not present in all of the years in the GSS, although they are present in most years after 1983. As a consequence, we will use 1984–2008 as our sample frame.[19] Data definitions appear in Table 3 and descriptive statistics in Table 4.

3.2.1 Aggregate Data

Figure 5 shows the co-evolution of imprisonment rates and two measures of self-reliance beliefs (*Effort Pays* and *Not-Washington*) over the sample period.[20] The incarceration rate increased sharply during the period 1984–98 and has stabilized since then. Both *Effort Pays* and *Not-Washington* increased during the same period 1984–98, and they have decreased somewhat since then. Figure 6 splits the GSS sample into white respondents and those self-identifying as black. The black sub-sample is considerably smaller, so Figure 6 is only suggestive of any real differences across whites and blacks. Still, it is interesting to note that

- blacks tend to believe less in the prevalence of the ''American Dream'': values for both *Effort Pays* and *Not-Washington* are lower for blacks than for whites; and
- there seems to be some increase in the percentage of blacks reporting the typical right-wing answers.

[19] In this period, the sample is reasonably continuous over time. None of the variables is missing for two consecutive years, and all the holes except *Not-Washington* in 1985 and *Effort Pays* in 1986 correspond to the years for which there are no GSS data at all: 1992, 1995, 1997, 1999, 2001, 2003, and 2005.

[20] We measure punishment as the number of prisoners in state correctional facilities per 100,000 of state population (also includes prisoners sentenced in federal courts, but serving in state prisons) compiled by the U.S. Federal Bureau of Investigation using data from the U.S. Census Bureau.

Table 3
Data Definitions

Desire to Punish	
Death Penalty	''Do you favor or oppose the death penalty for persons convicted of murder?'' 1 if the individual answered ''Favor'' and 0 if the individual answered either ''Oppose'' or ''Don't Know.''
Courts	''In general, do you think the courts in this area deal too harshly or not harshly enough with criminals?'' (1) Too harsh; (2) About right; (3) Not harsh enough.
Beliefs about Self-Reliance	
Effort Pays	''Some people say that people get ahead by their own hard work; others say that lucky breaks or help from other people are more important. Which do you think is most important? Hard work most important (3); Hard work, luck equally important (2); or Luck most important (1).''
Not-Washington	''Some people think that the government in Washington should do everything possible to improve the standard of living of all poor Americans; they are at Point 1. Other people think it is not the government's responsibility and that each person should take care of himself; they are at Point 5. Where would you place yourself on this scale, or haven't you made up your mind on this?'' Scale is inverted.

Note: Data from the General Social Survey. See Table 4 for descriptive statistics.

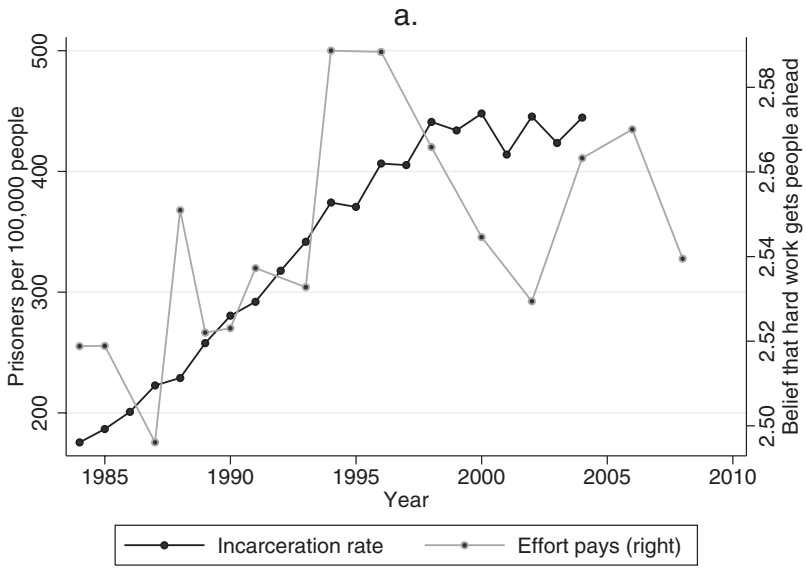
Table 4
Descriptive Statistics

	Observations	Mean	Standard Deviation	Minimum	Maximum
Death Penalty	27,915	0.7	0.46	0	1
Courts	31,056	2.75	0.55	1	3
Effort Pays	19,092	2.54	0.7	1	3
Not-Washington	18,667	2.91	1.16	1	5

Notes: Data from the General Social Survey over the years 1984–2008. See Table 3 for data definitions.

Even though the data are far too noisy for definite conclusions, this last observation points to the intriguing possibility that what enables harsher punishment in America is the increase in the belief

Figure 5
Self-Reliance Beliefs and Incarceration Rates over Time in the United States

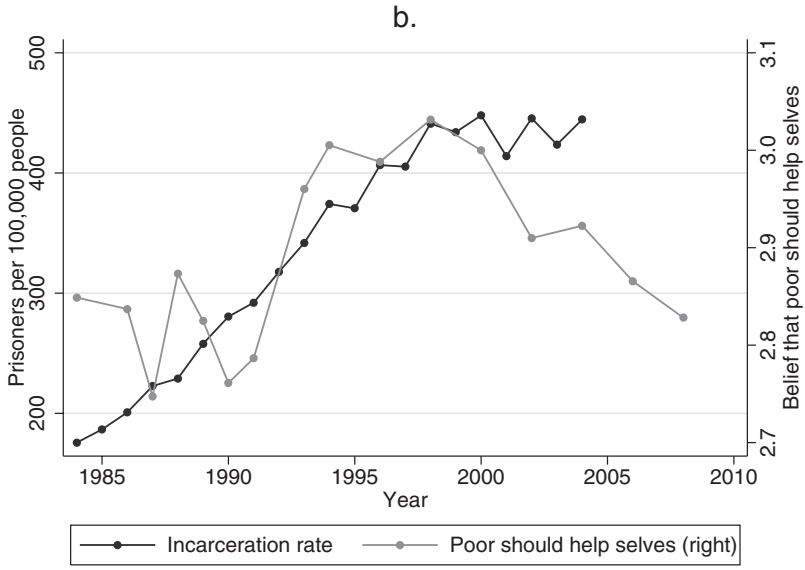


in economic opportunity among blacks, which is the group most affected by the policy. In other words, the legitimacy of punishment also appears to have increased during this period.

The GSS also reports data that can be interpreted as ''desired punishment'': the answers to the question: *''Do you favor or oppose the death penalty for persons convicted of murder?''* We create the variable *Death Penalty*, which takes the value 1 if the individual answered ''Favor'' and 0 if the individual answered either ''Oppose'' or ''Don't Know.''[21] Note that *Death Penalty* measures a particularly extreme form of punishment, which may differ from desires to punish using jails and prisons. Our model does not distinguish between these two forms of punishment, but in richer psychological models, these two desires may differ. For example, a person that is religious may

[21] The share of individuals answering ''Don't Know'' is very stable over time and close to 5 percent. All the results are practically the same if we treat those observations as missing values instead.

Figure 5
(continued)



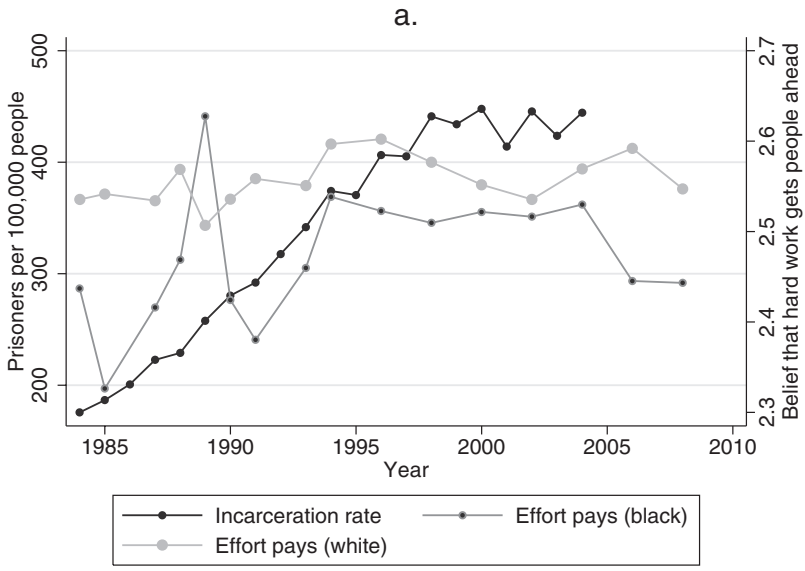
Source: U.S. Federal Bureau of Investigation.

Notes: Data on beliefs are year averages from the General Social Survey. See *Not-Washington* and *Effort Pays* in Table 3 for the data definitions. The incarceration rate is the number of sentenced inmates incarcerated under state and federal jurisdiction per 100,000 population.

cherish all forms of life and refuse to kill convicted criminals but may certainly favor long sentences for criminals. Another question available from the GSS on desired punishment is *Courts*, namely "*In general, do you think the courts in this area deal too harshly or not harshly enough with criminals?*" and gives very similar results.

These data (on punitiveness from the GSS) are useful for many reasons. First, while the incarceration rate gives a measure of actual punishment, *Death Penalty* is a measure of desired punishment. The relationship between individual beliefs and incarceration rates is indirect and involves a time lag: beliefs affect political choices, which in turn may affect aspects of the economic system (such as tax rates), and then there would be an effect on future incarceration rates. On the contrary, the relationship between individual beliefs and *Death*

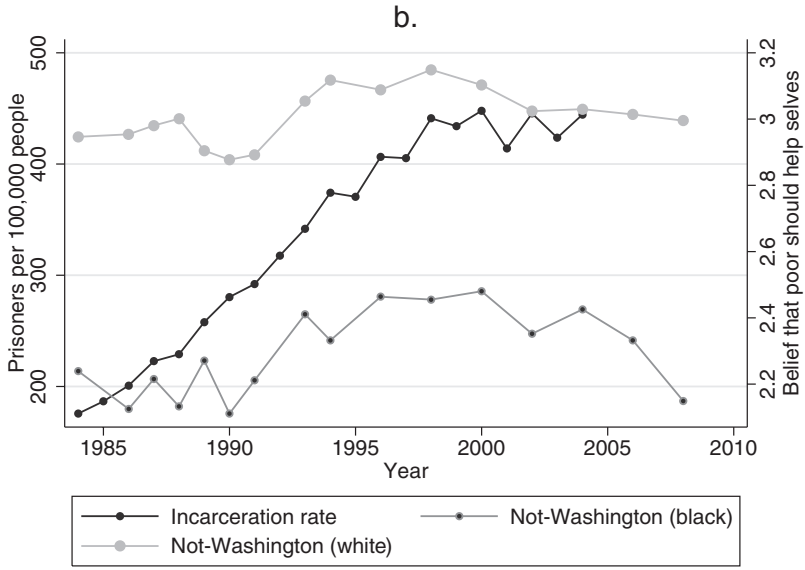
Figure 6
Self-Reliance Beliefs and Incarceration Rates over Time in the United States



Penalty is both direct and contemporaneous. Second, the data on *Death Penalty* vary at the individual level, while the data on incarceration rates vary at the state level only. Third, the GSS data are only representative at the national level, not at the state level. This implies a noisy relationship between GSS state-average beliefs and state-average incarceration rates. This can be avoided by using all data from the GSS (e.g., *Death Penalty* or *Courts* to capture punitiveness).

Figure 7 looks at the raw correlation of beliefs on self-reliance and desired punishment for the cross section of U.S. states, with *both* measures originating in the GSS sample (so the lack of representative sample is not as serious). The data correspond to the state averages for the period 1984–2008. States where individuals have more self-reliance beliefs display a higher share of the population in favor of the death penalty. This is consistent with the cross-country evidence presented before.

Figure 6
(continued)



Source: U.S. Federal Bureau of Investigation.

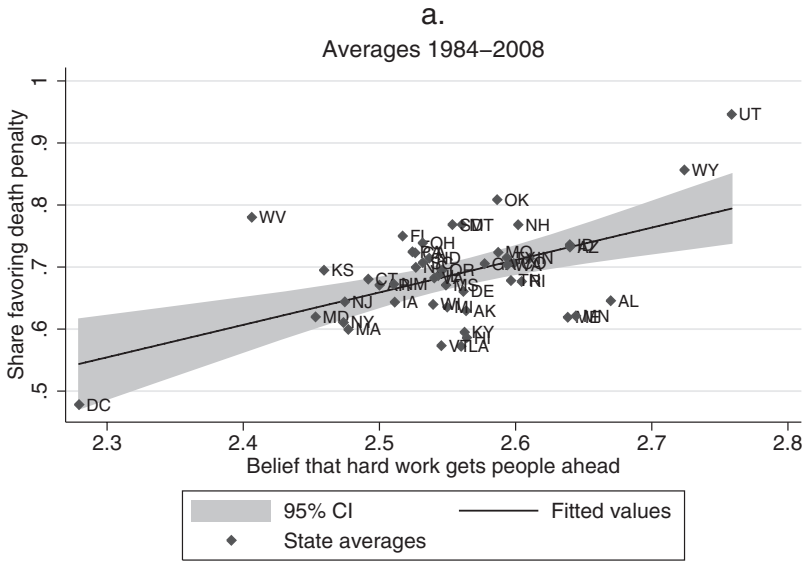
Notes: Data on beliefs are within-race year averages from the General Social Survey. See *Not-Washingon* and *Effort Pays* in Table 3 for the data definitions. The incarceration rate is the number of sentenced inmates incarcerated under state and federal jurisdiction per 100,000 population.

3.2.2 Individual Data

The GSS data allow us to study further the aggregate correlation between self-reliance and punishment in more detail, for example conditioning on other observable information (e.g., individual controls, state income inequality, state crime rates, etc.). Table 5 studies the individual-level relationship between self-reliance and punishment using a regression framework. The variable on the left-hand side of the estimating equation is *Death Penalty* (columns 1–4) and *Courts* (columns 5–8). As right-hand side variable, we use *Effort Pays* (and in separate regressions, *Not-Washington)*.[22] The regressions are

[22] In Table 12 of their paper, Alesina et al. (2001) present a regression that connects *Death Penalty* with *Not-Washington* with a different set of controls (and also obtain a strong correlation) and provide a broader discussion of the possible reasons for the differences in welfare policy across Europe and America.

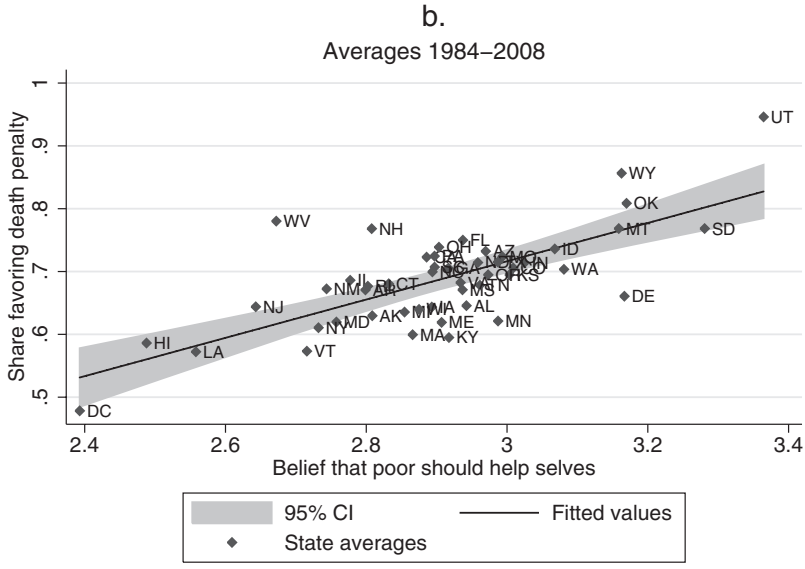
Figure 7
Self-Reliance Beliefs and Desired Punishment across U.S. States



OLS, and the results are qualitatively the same if, instead, we use a logit/probit model. All regressions use heteroskedasticity-robust standard errors clustered at the state level. All regressions include time effects and state fixed effects. The individual-level control variables are age of respondent, gender, a dummy for African-American race, a dummy for whites, a set of three dummies for marital status, income, a set of five dummies for employment status, education, number of adults, and number of children in household. The state-level control variables are crime rate for homicides, property crime rate, current real GDP per capita, GDP growth, income inequality (Gini coefficient), share of African-American population, and the share of white population.[23] In order to control semi-parametrically for other macro variables, we also include a set of state-specific time trends.

[23] There is a large literature on inequality and incarceration (see, e.g., Western 2002, 2006).

Figure 7
(continued)



Source: Data on beliefs are state averages from the General Social Survey.

Notes: See *Not-Washington*, *Effort Pays*, and *Death Penalty* in Table 3 for the data definitions.

As seen in Table 5, the correlation between *Effort Pays* and *Death Penalty* is positive and statistically significant, even after accounting for many parametric and semi-parametric controls. That is to say, self-reliance beliefs are associated with harsher desired punishment. Nevertheless, the magnitude of the coefficient does decrease when including individual controls. The same findings are true for the correlation between *Not-Washington* and *Death Penalty*, or when we use *Courts* (columns 5–8) as the left-hand side variable.

The coefficient on *Effort Pays* is significant from an economic point of view. For example, in a causal interpretation, a one-standard-deviation increase in *Effort Pays* (0.7) would increase *Death Penalty* by almost 8 percentage points (0.7×0.11). The coefficient on *Not-Washington* is also economically significant: a standard-deviation increase in *Not-Washington* (1.16) would increase *Death Penalty* by almost 4.5 percentage points (1.16×0.038). Similar conclusions are obtained when *Courts* is used.

Table 5
Punitive Regression Results

Dependent Variable:	*Death Penalty*				*Courts*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Effort Pays	0.022***	0.016***			0.044***	0.037***		
	(−0.004)	(−0.004)			(−0.005)	(−0.005)		
Not-Washington			0.057***	0.039***			0.046***	0.037***
			(−0.003)	(−0.003)			(−0.004)	(−0.004)
Time Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	No	Yes	No	Yes	No	Yes	No	Yes
State Controls	No	Yes	No	Yes	No	Yes	No	Yes
State-Specific Time Trends	No	Yes	No	Yes	No	Yes	No	Yes
Observations	22,309	22,093	21,843	21,620	21,074	20,871	20,580	20,378
R-squared	0.01	0.08	0.03	0.09	0.04	0.07	0.05	0.08
Number of States	49	49	49	49	49	49	49	49

Sources: Data from the General Social Survey (GSS) for years: 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1993, 1994, 1996, 1998, 2000, 2002, 2004, 2006 and 2008. See Table 3 for data definitions, and Table 4 for descriptive statistics.

Notes: Heteroskedastic-robust standard errors clustered at the state level in parentheses. *** significant at 1%. Left-hand side variable: columns (1) to (4) is *Death Penalty*; Columns (5) to (8) is *Courts*. The individual-level control variables are: age of respondent, gender, a dummy for African-American race, a dummy for white, a dummy for American citizen, a set of three dummies for marital status, income, a set of five dummies for employment status, education, number of adults, and number of children in household. The state-level control variables are: crime rate for homicides, property crime rate, current real GDP per capita, GDP growth, income inequality (Gini coefficient), share of African-American population, and the share of white population.

3.3 An Experiment Where Some Students Face Criminals Who Had Opportunities Growing Up

The experiment took place at a large business school of an Ivy League university in April 2011. The participants were potentially a highly selected sample. All 180 students from a second-year class on macroeconomics were sent invitations to participate in the online survey. The name of the survey was ''Survey of Attitudes,'' and students had to prepare a class on Jamaica and were given material on the macroeconomic performance of Jamaica and the country's relationship with the International Monetary Fund. They were told that the survey was anonymous and that it would only take five minutes to complete it. Most students were second-year MBA students, although there were a handful of exceptions (e.g., Ph.D. students). Students were not offered any money or course credit for participating in the survey. However, the professor in charge of the class sent the invitation with the link to the online survey from his own email address; this probably contributed to the high response rate: 128 out of 180 students logged in and completed the entire survey. The survey was posted on a Monday morning and students were given until Tuesday midnight to participate. Some 115 out of the 128 respondents (90 percent) completed the survey on Monday.

The online survey consisted of four consecutive screens. The first screen was exactly the same for all respondents and included a series of nine demographic questions about the respondent (e.g., gender, age, relative income). Once they finished answering those questions, respondents were shown a second screen with some brief information about education and crime in Jamaica.

Participants were randomized into two groups. The *Cherry Gardens* group saw the following description:

> Jamaica's development has been extremely uneven. In some regions of the country, economic growth was significant and there was substantial progress in areas like health and education. For example, in the neighborhoods around Cherry Gardens in Kingston, public schools (which are free, government run) had very attractive teacher/pupil ratios (on average 24:1), with a large proportion of students graduating high school (on average 81%), and most of them obtaining jobs, many of them very well paid. The statistics reveal that crime is a serious problem, both in rich and in poor neighborhoods.

> Recently, there has been an intense debate regarding the sentences that should be given to offenders. We would like to know your opinion about this issue. Take for instance the case of a 21-year-old man from the Cherry Gardens area who was found guilty of burglary for the second time. This time, he has stolen a TV.

The *Jones Town* group saw the following description:

> Jamaica's development has been extremely uneven. In some regions of the country, economic growth was non-existent and there was no progress in areas like health or education. For example, in the neighborhoods around Jones Town in Kingston, public schools (which are free, government run) had very unattractive teacher/pupil ratios (on average 41:1), with a low proportion of students graduating high school (on average 31%), and only a minority of them obtaining jobs, few of which were well paid. The statistics reveal that crime is a serious problem, both in rich and in poor neighborhoods.
>
> Recently, there has been an intense debate regarding the sentences that should be given to offenders. We would like to know your opinion about this issue. Take for instance the case of a 21-year-old man from the Jones Town area who was found guilty of burglary for the second time. This time, he has stolen a TV.

Relative to the *Jones Town* treatment, the *Cherry Gardens* treatment depicts a more positive image of Jamaica, where most people can get a job if they put their minds to it.

Right after the randomized treatment, respondents were asked their opinion about what the government should do with the individual in the example:

> Which of the following sentences do you consider the most appropriate for such a case? A. Fine; B. Prison; C. Community service; D. Suspended sentence.

This question closely resembles the question in the ICVS discussed in Section 3.1.

After answering this question, the respondent was presented with another question about the example:

> The judge decided to send him to prison. For how long do you think he should go to prison? A. 1 month or less; B. 2–6

> months; C. 6 months–12 months; D. 1 year; E. 2 years; . . .; N. Life Sentence.

This is another question included in the ICVS.

After answering that question, the respondent was asked a final question about the example:

> The government is considering a proposal whereby prisoners would be offered reductions in their sentences if they complete their education (primary and secondary courses would be expanded and made available in all Jamaican prisons). Do you agree with this proposal? A. Strongly Disagree; B. Disagree; C. Neither Agree Nor Disagree; D. Agree; E. Strongly Agree.

The third and fourth screens had all the information about the treatment introduced in the second screen, in case the respondent needed a refresher.

The data definitions of the variables used appear in Table 6, and their corresponding descriptive statistics appear in Table 7. A total of 65 respondents were in the *Cherry Gardens* group and 63 in the *Jones Town* group. As a routine check that the treatment was balanced, Table 8 shows the differences by treatment group in responses to pre-treatment questions.

The hypothesis is that respondents in the *Cherry Gardens* group will want to punish criminal behavior more severely because they perceive that the individual in the example had better opportunities not to become a criminal. As expected, the three measures of desired punishment suggest that people in the *Cherry Gardens* group desired more severe punishments compared to the *Jones Town* group. The first measure of punishment is the type of sentence. Figure 8 shows the distribution of responses for both groups. There are no major differences in the proportion of people choosing fine and suspended sentence, but there are major differences in the percentage of people choosing prison and community service. The simplest way to compare the answers is to look at what percentage of respondents chose prison, the most severe option. Some 45 percent of respondents in the *Cherry Gardens* group chose prison, compared to 32 percent in the *Jones Town* group. The p-value of the two-sided mean difference

Table 6
Data Definitions

Cherry Gardens	Dummy variable that takes the value 1 if the individual in the example belonged to the *Cherry Gardens* group and 0 if belonged to the *Jones Town* group.

Respondents were then given the following text: ''Take for instance the case of a 21-year-old man from the [Cherry Gardens/Jones Town] area who was found guilty of burglary for the second time. This time, he has stolen a TV.''

Post-Treatment Questions	
Punitiveness	Which of the following sentences do you consider the most appropriate for such a case? (1) Fine; (2) Prison; (3) Community service; (4) Suspended sentence.
Prison	Dummy variable that takes the value 1 if the individual chose the option ''Prison'' for the above question.
Months of Incarceration	''The judge decided to send him to prison. For how long do you think he should go to prison?'' The options were given in bins, as in the original ICVS question. To construct months of incarceration, we compute the mean (in months) of each bin (noted in parentheses): 1 month or less (0.5); 2–6 months (4); 6 months–12 months (8.5); 1 year (18); 2 years (30). Since there were no responses above 2 years, we will not care about those cases.
Rehabilitation	The government is considering a proposal whereby prisoners would be offered reductions in their sentences if they complete their education (primary and secondary courses would be expanded and made available in all Jamaican prisons). Do you agree with this proposal? (1) Strongly Disagree; (2) Disagree; (3) Neither Agree Nor Disagree; (4) Agree; (5) Strongly Agree.

(continued)

test is 0.155 (one-sided yields 0.078). Although the difference is (marginally) not significant at the 10 percent level, it is statistically significant once we include a set of control variables in order to improve precision, as shown later.

Figure 9 shows the distribution of responses to the second post-treatment question by treatment group. Relative to the *Jones Town* group, people in the *Cherry Gardens* group are less likely to choose prison sentences between 2 and 12 months and more likely to choose

Table 6
(continued)

Pre-Treatment	
Female	Dummy variable that takes the value 1 if the individual is female.
Age	Age of respondent in years.
American	Dummy variable that takes the value 1 if the individual grew up in the United States.
Never Married	Dummy variable that takes the value 1 if the individual has never been married.
Number of Children	Number of children that the respondent has.
Relative Income	Compared to your classmates. . . , would you say that your family income while you were growing up was below average, about average, or above average? (1) Well below average; (2) Somewhat below average; (3) About average; (4) Somewhat above average; (5) Well above average.
Stay in America	Are you planning on staying in the United States after graduation, for at least 5 years? (1) No, unlikely; (2) Undecided; (3) Yes, likely.
Monday	Dummy variable that takes the value 1 if the individual answered the survey on Monday, and 0 if answered on the next day.

Source: Data from the online experiment survey.

prison sentences between 1 and 2 years. In order to be able to compare the responses cardinally, we constructed the variable *Months of Incarceration*, which takes the value of the mean number of months in the corresponding option (e.g., 4.5 months for the category ''3–6 months''). The difference between the *Cherry Gardens* and *Jones Town* groups is statistically significant at conventional levels: the p-value of the two-sided test of mean difference is 0.097.

Finally, Figure 10 presents the distribution of answers for the question on the support for the rehabilitation program. In both *Cherry Gardens* and *Jones Town* groups, most people responded either ''Partially Agree'' or ''Strongly Agree'' (89 percent of respondents). However, relative to the *Jones Town* group, respondents in the *Cherry Gardens* group were much more likely to agree partially rather than strongly. The difference is not statistically significant at conventional

Table 7
Descriptive Statistics

	Observations	Mean	Standard Deviation	Minimum	Maximum
Cherry Gardens	128	0.51	0.5	0	1
Punitiveness	127	2.57	0.7	1	4
Prison	127	0.39	0.49	0	1
Months of Incarceration	127	6.32	6.31	0.5	30
Rehabilitation	128	4.13	0.9	1	5
Female	128	0.34	0.47	0	1
Age	128	27.24	1.6	24	33
American	128	0.46	0.5	0	1
Never Married	128	0.83	0.38	0	1
Number of Children	128	0.05	0.32	0	3
Relative Income	128	2.52	1.16	1	5
Stay in America	128	2.2	0.92	1	3
Monday	128	0.9	0.3	0	1

Source: Data from the online experiment survey.

Table 8
Mean Difference Tests for Pre-Treatment Variables

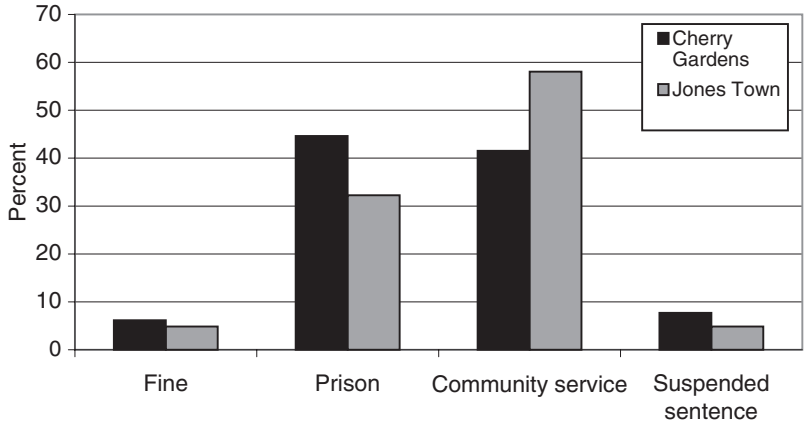
	Cherry Gardens	Jones Town	Difference[1]	|t| [2]
Female	0.385	0.286	0.099	1.182
Age	27.015	27.476	−0.461	1.644
American	0.4	0.524	−0.124	1.405
Never Married	0.846	0.81	0.037	0.546
Number of Children	0.046	0.063	−0.017	0.31
Relative Income	2.569	2.476	0.093	0.454
Stay in America	1.923	1.683	0.241	1.482
Monday	0.923	0.873	0.05	0.933
Observations	65	63		

Source: Data from the online experiment survey.

Notes: The first two columns display the mean of the variables within each group. [1]Cherry Gardens – Jones Town. [2]The t-statistic from the mean-difference test whose null hypothesis is that the means are equal between the Cherry Gardens and Jones Town groups.

levels: the p-value of the two-sided difference test is 0.228 (the p-value of the two-sample Wilcoxon rank-sum test is 0.056). Nevertheless, the results are statistically significant when introducing control variables as a means of increasing precision, as shown below.

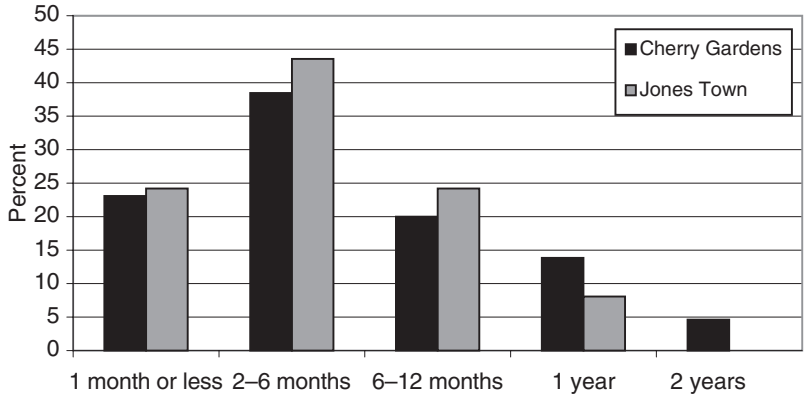
Figure 8
Differences in Desired Sentences by Treatment Groups



Source: Data from the online experiment survey.

Notes: Number of observations: 127. See *Punitiveness* in Table 6 for data definition.

Figure 9
Differences in Desired Incarcerations by Treatment Groups

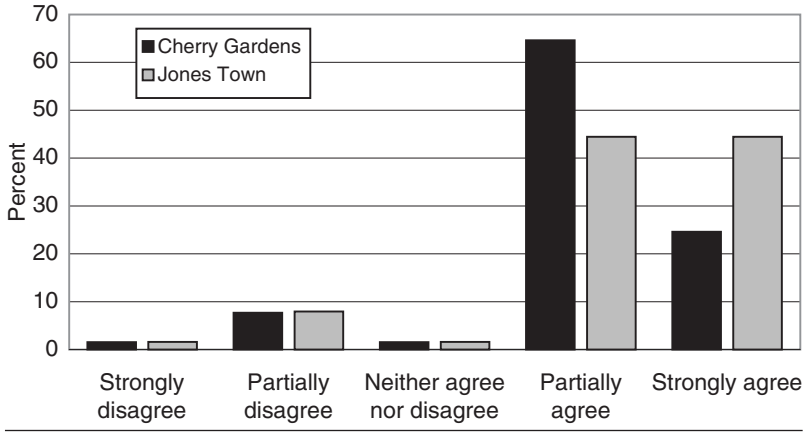


Source: Data from the online experiment survey.

Notes: Number of observations: 128. See *Incarceration* in Table 6 for data definition.

In order to increase the precision of the estimates, we will perform the same mean-difference test, but in a regression fashion. The

Figure 10
Differences in Support for Rehabilitation Programs by Treatment Groups



Source: Data from the online experiment survey.

Notes: Number of observations: 128. See *Rehabilitation* in Table 6 for data definition.

dependent variables are going to be three different measures of desired punitiveness. The first is *Prison*, a dummy variable that takes the value 1 if the individual recommended a prison sentence instead of a fine, community service, or suspended sentence. The second dependent variable is the recommended prison sentence in months, *Months of Incarceration*. The third dependent variable is *Rehabilitation*, a categorical value that represents how much the individual agrees or disagrees with the proposal of a rehabilitation program for prisoners. The first set of control variables includes just gender, age, and a dummy for growing up in America. The extended set of controls includes all the rest of the pre-treatment questions available in the online survey, including a dummy for completion of the survey on the same day it was released (perhaps those who were more conscientious might also be more conservative in their attitudes toward crime, but see Table 8 for mean difference tests across pre-treatment variables). For the dependent variable *Prison*, we estimate a logit model and then report the marginal effects at the mean of the control variables. For *Months of Incarceration,* we report OLS coefficients. For *Rehabilitation,* we use the ordered logit model, and

we present the marginal effect on the probability of the highest outcome (''Strongly Agree'') at the mean of control variables. We always report heteroskedasticity-robust standard errors.

The regression results are presented in Table 9. The treatment had an economically and statistically significant effect on all the dependent variables: relative to *Jones Town*, being in the group *Cherry Gardens* increases the probability of recommending a prison sentence by more than 15 percentage points, it increases the desired incarceration rate by approximately two months, and it decreases the probability of strongly agreeing with the implementation of the rehabilitation program by 15 percentage points.

We interpret this experimental evidence as supportive of the hypothesis that beliefs cause punishment. We should note that it has limited value, however, as a means to identify the particular channels highlighted in our model, as these are much more specific. And, of course, strong causal inferences are not feasible with such a small-scale exercise (for example, although the two scenarios involve burglars with similar criminal history who are caught stealing the same thing—a TV—a longer survey might be able to develop a better control for income).

4. CONCLUSIONS

Incarceration in the United States is the highest it has ever been, and it is the highest in the world. What is the reason for this? Given that incarceration affects minorities disproportionately, it is easy to see racism as the basic cause (see, for example, Bonczar and Beck 1997). One problem with racism as a cause is that few people who support increases in punitiveness consider themselves racist. Thus, one restriction that we impose on candidate answers is that people accept their own theories (explaining their support for increases in punitiveness). Our answer is based on beliefs: we argue that the increase in punitiveness is associated with widespread belief in economic opportunities for those willing to put in the effort. Our explanation connects incarceration and differences in pay (and inequality), as argued by Western (2006), although in our model, both are caused by beliefs about economic opportunities. In brief, we argue that harsh punishment is caused by the American dream.

The paper describes selective facts related to the evolution of punishment in the United States over the period 1980–2004. We note that

Table 9
Regression Results

Dependent Variable:	*Prison*			*Months of Incarceration*			*Rehabilitation*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Cherry Gardens	0.124	0.154*	0.190**	1.860*	1.940*	2.135*	−0.128*	−0.154**	−0.157**
	(−0.086)	(−0.087)	(−0.092)	(−1.101)	(−1.111)	(−1.16)	(−0.075)	(−0.075)	(−0.076)
Basic Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Extended Controls	No	No	Yes	No	No	Yes	No	No	Yes
Observations	127	127	127	127	127	127	128	128	128

Source: Data from the online experiment survey.

Notes: Heteroskedastic-robust standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. The coefficients for *Prison* are the marginal effects at the mean of the independent variables for a logit model. The coefficients for *Months of Incarceration* are OLS coefficients. The coefficients for *Rehabilitation* are the marginal effects on the probability of reporting the highest category, ''Strongly agree,'' at the mean of the independent variables using an ordered logit model. The basic set of controls includes gender, age, and a dummy for growing up in America. The extended set of controls includes *Never Married*, *Number of Children*, *Relative Income*, *Stay in America*, and *Monday*. See Table 6 for data definitions and Table 7 for descriptive statistics.

imprisonment started increasing around 1980, a period that coincides with the ''Reagan revolution.'' A large part of the increase involves expansion of the use of minimum-security prisons. While minorities are imprisoned at a disproportionate rate, the ratio of the incarceration rates for blacks versus whites has not changed even as these rates increased substantially. The contrast with the European experience, where imprisonment rates are much lower, suggests that differences in beliefs and ideologies could play a big role, as suggested by Tonry (1998).

We then build an economic model where beliefs about economic opportunities and beliefs about punishment are correlated. There is a ''French equilibrium,'' where workers believe effort does not pay, firms set up bureaucratic systems (low-powered incentive schemes), and criminals on average are ''kinder.'' There is an ''American equilibrium'' where workers believe effort pays (and exert effort), firms set up a market technology (high-powered incentive schemes), and the proportion of mean types who become criminals is larger than in the French equilibrium. With increases in income, one can observe that there is a demand for harsher punishment. We present three pieces of evidence (across countries, across states in the United States, and an experimental exercise) that are consistent with the model.

5. APPENDIX: EXTENSIONS OF THE MODEL

In this section, we present an extension of the model of Section 2. The idea is to sketch how several interesting questions concerning crime can be incorporated into the model and therefore illustrate what the beliefs-punitiveness connection has to say about those questions.

The first simple extension of the model we consider is incorporating ''opportunities available to criminals'' in the utility function of the government. In this modification, the government's utility of a strategy $s = M, B$ by the firm and t by the government, when beliefs about μ are given by h, is for a parameter x,

$$v(s,t,\mu) = -E_h[(q(\mu) + xO(s) - t)^2]$$

In the above equation, $O(s)$ represents the opportunities available to individuals when the firm chooses s. A natural definition of opportunities is $O(M) = g(w_h - w_l)$, the difference in income between choosing high and low effort. Similarly, $O(B) = 0$.

In order to understand why incorporating opportunities in the utility function of the government is interesting, consider the following situation: The firm had chosen a market technology, a criminal was caught,

and we know both his type μ, and that he is a θ_L who chose e_L. It seems natural that individuals (or the government) would want to set a harsher punishment if they knew that w_k was high so that, by exerting effort, he could have avoided becoming a criminal. One possible reason for this harsher desired punishment is ''identity'': people want to believe that they are not the kind of people who can be ''fooled'' or ''taken advantage of''; they are probably willing to forgive a theft by somebody who had no opportunities, but they wouldn't forgive a thief who could have made an honest living but is taking advantage of their forgiveness. The formulation above captures the idea that when there are more opportunities—a larger $O(s)$—the government chooses a harsher punishment: the desired punishment is given by $t = E_h[q(m) + xO(s)]$.

In the above formulation, we have postulated that $O(s)$ is calculated from the model. But one can also interpret $O(s)$ as the ''ideology'' of the government, make it an exogenous parameter, and calibrate it from external data (say, opinion polls of officials) and note that increases in $O(s)$ lead to increases in punishment.

A second addition one can make to the model is

- make returns to effort depend on the individual's type;
- include a choice of one of two neighborhoods; at the time of choosing effort, the individual picks a rule that specifies a choice of neighborhood conditional on realized income.

In this variation, the types would be irrelevant regarding the cost of effort (say, setting $\theta_L = \theta_H$) but making the return to each effort random and dependent on the type. Also, and just to simplify, the choice of neighborhood would be made so as to minimize the distance between one's expected income and the neighborhood's average income.

This extension can be used to address the important issue of the criminal behavior of African Americans and the harsh punishment they face. At least two different explanations for the harsh punishment arise in this model. The first (which does not use the neighborhood choice feature) is that the government holds a belief that criminals (regardless of their race) face very good opportunities in the legal market and therefore should be punished harshly.[24]

[24] This prediction runs in the opposite direction of models based on deterrence: the better the opportunities in the legal market, the less one has to punish criminals to deter them.

A second explanation is rooted on the intriguing observation that acquiring skills (e.g., obtaining a university degree) may be *more* profitable for poor African Americans than for whites, but that they are less prone to doing so than their white counterparts. That is, some data suggest that although African Americans earn less than whites in either category, the wage increase of obtaining a degree is larger for African Americans. In this model, the government punishes harshly because opportunities are in fact large. And one could obtain the result that African Americans are less prone to acquiring a degree, and therefore more likely to engage in criminal behavior, through one of two methods.

In one variation of the model, individuals have a belief about the return to each effort level and choose an effort level and a neighborhood rule (what neighborhood to choose, depending on the income); the beliefs about the return to each effort level must be consistent with the distribution of effort levels and incomes in the neighborhood he settles on. The story told by this version of the model is that individuals living in poor neighborhoods incorrectly estimate the returns to schooling because they only get to meet the lower tail of college-graduate wage earners (those who returned to the poor neighborhood).

In order to sketch the second variation, imagine that the distribution of wages for college graduates is either $1 for sure, or $1 with probability 95 percent and $100 with probability 5 percent. The prior belief of the individual is that each distribution has the same probability. In order to estimate the returns to schooling, the individual samples a few people, but since sampling a graduate who earns $1 is so likely, the individual is likely to finish his sampling with a (downwardly) biased estimate of the returns to schooling. In fact, if the individual samples only once, with a probability of 97.5 percent, he will estimate that the distribution ''degenerates at $1'' is the most likely. One can incorporate this idea (developed in Benoît and Dubra 2011) into this model (without the need of biased sampling or neighborhood choices) to obtain the same predictions as in the previous paragraph.

REFERENCES

Adorno, T. W., et al. 1950. *The Authoritarian Personality.* Harper.

Alesina, A., E. Glaeser, and B. Sacerdote. 2001. ''Why Doesn't the United States Have a European-Style Welfare State?'' *Brookings Papers on Economics Activity*, 2001(2): 187–278.

Alesina, A., and E. La Ferrara. 2011. ''A Test of Racial Bias in Capital Sentencing.'' Harvard Institute of Economic Research Discussion Paper No. 2192. May.
Alexander, M. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New York Press.
Austin, J., et al. 2000. ''The Use of Incarceration in the United States.'' National Policy White Paper. American Society of Criminology, National Policy Committee.
Baumeister, R. F., et al. 2008. ''Free Will in Consumer Behavior: Self-Control, Ego Depletion, and Choice.'' *Journal of Consumer Psychology* 95: 1367–82.
Benabou, R., and J. Tirole. 2011. ''Laws and Norms.'' Mimeo. Princeton University and University of Toulouse.
Benoît, J.-P., and J. Dubra. 2011. ''Apparent Overconfidence.'' *Econometrica*. Forthcoming.
Bonczar, T. P., and A. J. Beck. 1997. ''Lifetime Likelihood of Going to State or Federal Prison.'' U. S. Department of Justice, Bureau of Justice Statistics. March.
Blumstein, A. 1993. ''Racial Disproportionality of U. S. Prison Populations Revisited.'' *Colorado Law Review*, 64: 743–60.
Blumstein, A., J. Cohen, and D. Nagin. 1978. ''Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates.'' National Academies Press.
Bushway, S., and R. Paternoster. 2009. ''The Impact of Prison on Crime.'' In *Do Prisons Make Us Safer? The Benefits and Costs of the Prison Boom*, ed. S. Raphael and M. A. Stoll, pp. 119–50. New York: Russell Sage Foundation.
Carlsmith, K. M., J. M. Darley, and P. H. Robinson. 2002. ''Why Do We Punish? Deterrence and Just Desserts as Motives for Punishment.'' *Journal of Personality and Social Psychology* 83: 284–99.
Cullen, F., and R. Agnew. 2003. ''Anomie/strain Theories of Crime, Part V.'' In *Criminological Theory: Essential Readings,* ed. F. Cullen and R. Agnew. Los Angeles: Roxbury Publishing Company.
Davis, J. A., and T. W. Smith. 2005. *General Social Surveys, 1972–2004.* Machine-readable data file. Roper Center for Public Opinion Research, University of Connecticut.
Dannenberg, J. 2011. ''Nationwide PLN Survey Examines Prison Phone Contracts, Kickbacks.'' *Prison Legal News* 22 (4): 1–16.
Di Tella, R., and J. Dubra. 2008. ''Crime and Punishment in the American Dream.'' *Journal of Public Economics* 92: 1564–84.
Di Tella, R., S. Edwards, and E. Schargrodsky. 2010. *The Economics of Crime: Lessons for and from Latin America.* Chicago: University of Chicago Press.
Doob, A. N., and C. M. Webster. 2006. ''Countering Punitiveness: Understanding Stability in Canada's Imprisonment Rate.'' *Law and Society Review* 40 (2): 325–67.
Durlauf, S. N., and D. S. Nagin. 2010. ''The Deterrent Effect of Imprisonment.'' Mimeo. Carnegie Mellon University.
Feldman, S. 1988. ''Structure and Consistency in Public Opinion: The Role of Core Beliefs and Values.'' *American Journal of Political Science* 32 (2): 416–40.
Flanagan, S. 1987. ''Value Changes in Industrial Societies.'' *American Political Science Review* 81: 1303–19.
Hall, P., and D. Soskice. 2001. *Varieties of Capitalism: The Institutional Foundations of Comparative Advantage.* Oxford, UK: Oxford University Press.
Hindelang, M. 1978. ''Variations in Sex-Age-Race-Specific Incidence Rates of Offending.'' *American Sociological Review* 43(1): 93–109.
Hochschild, J. 1981. *What's Fair? American Beliefs about Distributive Justice.* Cambridge, MA.: Harvard University Press.

Inglehart, R. 1990. *Culture Shift in Advanced Societies.* Chicago: Chicago University Press.

International Crime Victims Survey 2004/2005. Netherlands: Tilburg University, 2006.

Jost, J. T., and M. R. Banaji. 1994. ''The Role of Stereotyping in System-Justification and the Production of False Consciousness.'' *British Journal of Social Psychology* 33 (1): 1–27.

Jost, J. T., et al. 2003. ''Political Conservatism as Motivated Social Cognition.'' *Psychological Bulletin* 129 (3): 339–75.

Kensey, A., and P. Tournier. 1999. ''Prison Population Inflation, Overcrowding and Recidivism: The Situation in France.'' *European Journal on Criminal Policy and Research* 7: 97–119.

Ladd, E. C., and K. Bowman. 1998. *Attitudes towards Economic Inequality.* Washington: AEI Press.

Lakoff, G. 1996. *Moral Politics: How Liberals and Conservatives Think.* Chicago: University of Chicago Press.

Lipset, S. M. 1979. *The First New Nation.* New York: W. W. Norton.

Lipset, S. M., and S. Rokkan. 1967. ''Cleavage Structures, Party Systems, and Voter Alignments: An Introduction.'' In *Party Systems and Voter Alignments,* ed. S. M. Lipset and S. Rokkan. New York: Free Press.

Liptak, A. 2008. ''Inmate Count in U.S. Dwarfs Other Nations.'' *New York Times*, April 23.

Mauer, M. 2008. ''Testimony of Marc Mauer before the Maryland Commission on Capital Punishment on Race and the Criminal Justice System.'' The Sentencing Project. August 19.

Mauer, M., and R. S. King. 2007. ''Uneven Justice: State Rates of Incarceration by Race and Ethnicity: A Report by the Sentencing Project.'' Mimeo. The Sentencing Project.

Merton, R. K. 1938. ''Social Structure and Anomie.'' *American Sociological Review* 3 (October): 672–82.

Messner, S., and R. Rosenfeld. 2001. ''Crime and the American Dream.'' In *Criminological Theory: Essential Readings,* ed. F. Cullen and R. Agnew. Los Angeles: Roxbury.

Perron, P. 2005. ''Dealing with Structural Breaks.'' Mimeo. Boston University.

Putterman, L., J. Roemer, and J. Sylvestre. 1998. ''Does Egalitarianism Have a Future?'' *Journal of Economic Literature* 36: 861–902.

Piketty, T. 1995. ''Social Mobility and Redistributive Politics.'' *Quarterly Journal of Economics* 110 (3): 551–84.

Rapahel, S., and M. Stoll. 2009. ''Why Are So Many Americans in Prison?'' In *Do Prisons Make Us Safer? The Benefits and Costs of the Prison Boom,* ed. S. Raphael and M. A. Stoll, pp. 27–72. Washington: Russell Sage Foundation.

Rasmusen, E. 1996. ''Stigma and Self-Fulfilling Expectations of Criminality.'' *Journal of Law and Economics* 39: 519–44.

Rokeach, M. 1973. *The Nature of Human Values.* New York: Free Press.

Sampson, R. J., and C. Loeffler. 2010. ''Punishment's Place: The Local Concentration of Mass Incarceration.'' *Daedalus* 139 (Summer): 20.

Sanfey, A. G., et al. 2003. ''The Neural Basis of Economic Decision-Making in the Ultimatum Game.'' *Science* 300: 1755–8.

Shariff, A., et al. 2011. ''For They Know Not What They Do: Diminishing Free Will Beliefs Reduces Retribution and Increases Forgiveness.'' Mimeo. University of Minnesota.

Scherpenzeel, A. 2001. *Mode Effects in Panel Surveys: A Comparison of CAPI and CATI.* Neuchâtel, Switzerland: Bundesamt fur Statistik.

Sunstein, C. 1996. ''On the Expressive Function of Law.'' *University of Pennsylvania Law Review* 144 (5): 2021–53.

Tonry, M. 1998. ''Introduction: Crime and Punishment in America.'' In *The Handbook of Crime and Punishment*, ed. M. Tonry. Oxford, UK: Oxford University Press.

Van Dijk, J. J. M., J. N. van Kesteren, and P. Smit. 2008. *Criminal Victimisation in International Perspective, Key Findings from the 2004–2005 ICVS and EU ICS.* Annandale, Australia: Boom Legal Publishers.

Vohs, K. D., and J. R. Schooler. 2008. ''The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating.'' *Psychological Science* 19: 49–54.

Walmsley, R. 2007. *World Prison Population List,* 7th ed. International Centre for Prison Studies, School of Law, King's College London.

Webster, C. M., and A. N. Doob. 2007. ''Punitive Trends and Stable Imprisonment Rates in Canada.'' *Crime and Justice: A Review of Research* 36: 297–369.

Western, B. 2002. ''The Impact of Incarceration on Wage Mobility and Inequality.'' *American Sociological Review* 67 (4): 526–46.

———. 2006. *Punishment and Inequality in America.* Washington: Russell Sage.

Whitman, J. Q. 2003. *Harsh Justice: Criminal Punishment and the Widening Divide between America and Europe.* Oxford, UK: Oxford University Press.

World Values Survey, 5th Wave. 2005–2008. World Values Survey Association.

Zaller, J. 1991. ''Information, Values and Opinions.'' *American Political Science Review* 85 (4): 1215–37.

Comment

Justin McCrary

The starting point for ''Free to Punish?'' is the simple empirical observation that the current incarceration rate in the United States is high relative to that of other rich countries, as well as that seen in the United States 40 years ago. This growth in incarceration has been well-documented in the academic literature and is rediscovered anew in the popular press several times annually.

Overlooked aspects of the growth in incarceration include its homogeneity and its accretive nature. The growth in incarceration is seen in all 50 state systems as well as in the federal system. Within each of these jurisdictions, no single policy shift has led to the increased use of incarceration. Rather, there has been an accretion of small policy reforms that together have had very large consequences for the use of incarceration. A great many sentence-enhancement bills have been passed over the last 40 years. These were passed contemporaneously with a decreasing tolerance for judicial discretion in sentencing, a decreasing tolerance for the use of parole, the rise of ''supermax'' facilities, and an increase in the rate at which we all transport our children to school in the safety of vehicles rather than letting them walk. Overall, it is hard to read the increased use of incarceration as anything other than a general shift in ideas regarding what kind and scale of punishment is appropriate for criminality, or perhaps simply a broad increase in the taste for safety.

From this fundamentally empirical starting point, ''Free to Punish?'' develops a simple theoretical model seeking to connect the evolution of income distribution to shifting attitudes regarding crime. As is common in economics, the details of the model may be hard to follow if one is reading quickly, but the intuition is rather straightforward. There is a latent variable, meanness, that is a shock

Justin McCrary is professor of law at the University of California, Berkeley.

inducing a marginal offender to commit crime. Government aspires to punish more severely those with more meanness. If there is economic opportunity, then only the truly mean commit crime; if there is little economic opportunity, then criminals are a broader cross section. Government infers likely meanness. That is, it is public knowledge whether opportunities are prevalent or rare, and government reasons that punishment is more desirable in a good economy than in a bad economy, by virtue of statistical discrimination.

As evidence in support of this proposition, the authors note that ideology typically comes bundled, with many different beliefs comprising a belief system. Perhaps more persuasively, the authors emphasize the empirical correlation between an individual's belief in the importance of self-reliance as a means of getting ahead and an individual's support for punitive treatment of criminals. This correlation is quite robust in the American data. The authors also note that, relative to Europeans, Americans dislike redistribution and like punishment.

The paper is self-aware that these correlations may not have much to do with causal mechanisms. Additional evidence from a small classroom experiment suggests that randomization to better economic conditions makes subjects more likely to support punishment for individuals convicted of crime.

I agree that the evidence presented is consistent with the theoretical model put forward. However, I do not think that this prediction is unique to this particular theory. Moreover, while the conceptual connection between opportunities and support for punishment may well be operative, I suspect it is second-order. First, real wages in the United States are roughly flat during the entire time period, and possibly even declining at the bottom of the wage distribution. The model could be interpreted as being about the variance rather than the mean, but as shown in the literature on skill-biased technological change, much of the increase in the standard deviation of log wages occurs during punctuated periods. Both of these observations are inconsistent with the smooth, nearly linear shift in the use of incarceration as a sanction.

In terms of first-order issues pertaining to the criminal justice system that deserve study, I would have instead emphasized the spectacular increase in prisoners per capita (roughly 500 percent) as compared to the tiny increase in police per capita (roughly 20 percent) from 1970 to today. I would emphasize the decreasing public

tolerance for minor infractions such a graffiti, drug use while on parole, and the more general decrease in public tolerance for risk and insistence on safety, even when safety comes at the price of civil liberties. I would emphasize that the current organization of the criminal justice system is slow, unpredictable, and draconian, whereas if crime reduction were the goal, it would be better to organize the system to be swift, predictable, yet moderate. I would emphasize the phenomenon of passage of sentence-enhancement bills that are so esoteric that offenders are not aware of them until after apprehension.

On the plus side, the question of why the United States has found it optimal to use incarceration as much as we do is a great one. To paraphrase Sen. Jim Webb (D-VA), we are either the most evil nation in the world or our government is doing something very, very wrong. The sense that the question deserves thinking about is only increased once one recognizes that there is no literature on the topic.

Competition and Innovation

Michele Boldrin
Juan Correa Allamand
David K. Levine
Carmine Ornaghi

ABSTRACT

Which kind of intellectual property regime is more favorable to innovation: one that enforces patents or one that does not? Economic theory is unable to answer this question, as valid arguments can be made both for and against patents; hence we must turn to empirical evidence. In this paper, we review empirical evidence gathered by other researchers and add new evidence of our own. We conclude that the evidence suggests that patents do not promote innovation, but instead retard it.

Michele Boldrin is the Joseph Gibson Hoyt Distinguished Professor of Economics at Washington University in St. Louis and a research fellow at the Federal Reserve Bank of St. Louis. Juan Correa Allamand is a researcher at the Ministry of Finance, Chile. David K. Levine is the John H. Biggs Distinguished Professor of Economics at Washington University in St. Louis and a research fellow at the Federal Reserve Bank of St. Louis. Carmine Ornaghi is a lecturer in economics at the University of Southampton.

Competition and Innovation

1. INTRODUCTION

Our concern is with a straightforward policy question: Which kind of intellectual property regime is more favorable to innovation, hence more conducive to technological progress, increasing factor productivity, and—eventually—economic growth? More precisely, what is the role that patents play, or do not play, in fostering technological innovation? A practical policy question such as this rarely has a straightforward answer; this may be one of those rare cases when it does. Our final answer, in fact, will be that there is no evidence whatsoever that patents and copyright foster innovation and creation at large, while there is abundant evidence that they damage it in particular instances, apart from inducing substantial efficiency, rent-seeking, and distributional costs.

To allow for a focused discussion, we will pretend that a starker difference between legal regimes exists (and is the object of contention) than is probably feasible in reality, where legislation must act on a multidimensional continuum of choices. On the one hand, we consider a regime in which patents are assigned easily and then strictly and forcefully enforced for a long time—as has been the case in the United States for more than a century and increasingly so since the middle 1980s. On the other hand, we consider a counterfactual regime in which patents are hard to obtain, loosely enforced, and last for brief periods of time only, that is, a system in which imitation, and the free-entry competition it brings about, rules. Subject to the qualifications discussed in the paper, we call the latter arrangement ''competition'' and the former one ''monopoly.'' We ask: Which regime is more desirable from a social point of view?

At a theoretical level, this question receives two straightforward—but completely opposite—answers. Within a certain class of models, ''monopoly'' is the necessary evil without which innovation is impossible. Within a second class of models, ''competition'' nicely provides innovation and, in some cases, even reaches first-best.

Depending on the kind of assumptions one is willing to make about the set of available technologies, the cost of imitation, and the speed at which productive capacity can be built, economic theory may go either way. In one case—when innovations are dramatic and well identifiable, the cost of technological discontinuities is very high, productive capacity is easily built, and imitation is costless—intellectual property is likely to be a growth-enhancing arrangement. This is recognized today as the ''received wisdom.'' In the opposite case—when innovations are a distributed and incremental phenomenon the cost of which, while sometimes high in the aggregate, is relatively low at each single step, productive capacity cannot be rapidly adjusted, and imitation is costly—patent protection is likely to be a growth-reducing arrangement. This latter view is considered, nowadays, a form of ''economic heresy.'' It happens to be a view of the innovation process that, among others, two of us have been suggesting as appropriate for quite a while (see, e.g., Boldrin and Levine 2008b).

The question of which regime actually induces more innovation and is more socially desirable has been hardly considered at the empirical level, where most researchers have been taking it for granted that patents and copyright are the best incentives to innovation, hence to economic growth. Because, at the end of the theoretical day, the issue is an empirical one, it is to the small amount of available empirical evidence—to which we add a bit of fresh material—that we dedicate the bulk of this paper. Our examination leans strongly toward the heretic option: all things considered, patents do not have a favorable impact on technological innovation and tend to hurt economic welfare. Nevertheless, we will not even try to claim the issue can be settled on the basis of the available evidence, if anything because the latter is still scarce and unsystematically organized. As mentioned, maybe because so little relevant data have been collected or maybe because the received wisdom has been so dominant for so long that empirical verification appeared redundant, the number of empirical papers addressing the policy question we have asked turns out to be very small. Better data and more systematic analysis are therefore needed before any conclusive policy prescription may be drawn.

We proceed through the following steps: In Section 2, we start with a brief summary of the main theoretical features of the received wisdom. That wisdom has two sharp and testable predictions:

(a) The stronger the monopoly power granted by patents, the higher the incentive to innovate, hence the faster the growth rate of productivity.
(b) The innovation process consists of a sequence of discrete jumps; at each step a new innovator ''jumps over'' the incumbent monopolist, acquires a patent to protect her new innovation, and becomes a temporary monopolist, soon to be overtaken by an even better newcomer.

The empirical support for these predictions is considered in Section 3, and we find it to be scant. In fact, we show that neither (a) nor (b) receives any support in the statistical or historical literature. More cogently, empirical research has focused on two lines of investigation:

- The ''patent puzzle''—that is, the fact that, in recent decades, as patent protection was progressively strengthened, we did not observe an innovation explosion.
- The ''inverted-U curve''—that is, the claim that, in the data, while it is apparent that strong monopoly power does not favor innovation, free competition may not be all that good either, and innovation is maximized somewhere in the middle.

While we have little to add to the interesting findings collected in the patent-puzzle literature, we believe that something more can be said about the inverted-U claim. Here we make two points, one general and the other specific. The general point is that using the number of patents (or their citations count) to measure the actual number of innovations (or, better, their impact on productivity) across sectors and firms may lead to somewhat biased results. While many innovations lead to a patent, at least equally many others do not; while many patents are associated with an actual innovation, at least equally many others are obtained for legal or rent-seeking purposes that have little or nothing to do with actual innovations. As we show, while patent citations may be vaguely correlated with productivity growth, they are far from being a reliable measure of it. This suggests treating with the greatest care those studies that assume a close identity between the two: showing that, in the data, X may increase patents is a long shot from showing that X also increases innovations and labor productivity. The specific point—developed in great detail in the Ph.D. dissertation of one of us

(Correa 2010)—is that the inverted-U finding does not actually withstand a careful statistical examination and disappears rather quickly. When we reexamine the very same dataset in which the inverted-U relation was claimed to exist, we find instead a robust monotone relation between measures of competition and the number of patent citations. As the heretic view predicts, this relationship is increasing.

Section 4 summarizes the heretic line of thought according to which imitation does not hamper, and in fact stimulates, innovation. The heretic view predicts that we should observe a higher rate of innovation in those sectors in which competition is stronger and patent protection weaker. Because, as we argue, the heretic line of thought rests on the assumption that most technological progress is embodied in either human or physical capital, we also briefly discuss the cumulated empirical evidence supporting this claim.

In Section 5, we test our theory using recent high-quality data. By merging three different micro datasets on firms' patents, patent citations, and productivity growth at the sectorial level, we show there is a clear, monotone, increasing relationship between measures of sectorial competition, the innovative activity of firms, and—most importantly—labor productivity and total factor productivity (TFP) growth. The implications of our findings for future research and actual policymaking are collected at the end, in Section 6.

2. RECEIVED WISDOM

The intellectual roots of the currently dominant consensus about the role that patents and monopoly power play in the innovation process were planted long ago: Schumpeter (1942) ought to be seen as its starting point, while Scherer (1990) is a leading current rendition among industrial organization scholars.

A famous paper by Kenneth Arrow (1962) provided debatable but authoritative information-theoretical support for the view that ''innovation is a public good.'' That view, in turn, gave rise to two completely opposite lines of research and related policy implications. According to one line of thought, if innovation is a public good, then it cannot be provided by the markets, be they competitive or monopolized; innovation needs, therefore, to be either produced directly by government agencies or, at least, subsidized by means of fiscal revenues and tax incentives. This was the line of thought Arrow's initial argument was meant to support and that still appears

as a logical consequence of the (in our view patently falsified) view that innovation is a public good. According to the second line of research stemming from Arrow's assumption, innovation becomes an undersupplied public good because the property right system is poorly designed: those contracts through which innovation can be made excludable are not available in a competitive market system as long as the latter allows for imitation. Patents can solve this particular version of the ''tragedy of the commons'': they make innovations excludable, thereby insuring at least a second-best outcome—Gilbert and Shapiro (1990) contains a relatively recent, and more elaborated, exposition of this point of view. This approach, which had come to dominate theoretical and applied work in the area of industrial organization by the late 1960s, was incorporated into dynamic models of economic growth and capital accumulation with the arrival of the so-called ''New Growth Theory'' (NGT) in the late 1980s and early 1990s.

From a theoretical perspective, these two renditions of the view that patents and legal monopoly are essential for innovation are what we have in mind when we speak of the currently received wisdom. Purely because of its very simplified analytical structure, we will use here the Aghion and Howitt (1992) model as our representative stand-in for this otherwise vast literature. Almost everything we say applies equally well to all other NGT models produced since the middle 1980s (in particular to those of Gene Grossman and Elhanan Helpman, and of Paul Romer, which were both prior to that of Aghion and Howitt) as well as to the classical contributions in the field of industrial organization.

This literature posits a tradeoff between ''static efficiency,'' achievable under free competition insofar as this leads to an efficient allocation of resources for a given technology, and ''dynamic efficiency,'' which is due to technological progress driven by patented innovations aimed at acquiring a monopoly. While competition may yield static efficiency, it fails to deliver the dynamic one; because the latter is, arguably, more important than the former, free competition is not good for society. Patents are therefore a ''necessary evil'': even if they sustain a technologically artificial monopoly power, thereby reducing consumer surplus while they last, they are a necessary tool for fostering dynamic efficiency over time. Innovative efforts would find no reward if the monopoly power that patents create were absent—no patents, no party.

There are two key technological assumptions underlying the prediction that, under competition and free entry, dynamic inefficiency (i.e., suboptimal innovative efforts) would be the outcome:

- The private fixed costs of innovation are ''large,'' at least relative to the size of the market and the evaluation (willingness to pay) that the marginal consumer would give of the new good.
- Imitation is both simple and cheap, so much so that, within a small span of time, so much productive capacity can be built up that every competitor would be led to play a Bertrand game when it comes to pricing.

It is obvious that, in a world where these two assumptions are simultaneously realized, one would observe no innovation absent patent's protection. Innovators would know that, as soon as they had sunk their fixed cost and introduced the new good or the new production method, imitators would flock into the market at zero cost, building up excessive capacity and, because of Bertrand competition, driving prices to variable marginal costs. Because the latter makes it impossible to cover the initial fixed cost, entrepreneurs would not even bother trying to innovate, thereby leading the economy to technological stagnation. From this reasoning there follows, on the one hand, the prediction that industries where imitation is easy will not see much technological progress and, on the other, the prediction that, in the presence of strong patents, innovation will instead thrive. This is statement (a) in the introduction.

It is worth stressing that statement (b) is neither marginal nor dispensable in this framework. When (b) fails, the model predicts that (a) also will fail and innovation will come to a halt even in a world where patents guarantee innovators a strong monopoly power. To see this, assume the current incumbent monopolist can—either through a sequence of very small innovations or by purchasing the key factors its competitors need in carrying out their research and development efforts—maintain its lead on the pack, thereby progressively reducing their chances to ever be able to ''jump over'' the incumbent with a breakthrough innovation. In other words, assume the incumbent monopolist can make valuable innovations extremely costly for its potential competitors; in this case, prediction (b) fails together with (a). In the Schumpeterian model, when the monopolist is so far ahead of its competitors to feel safe enough,

the innovation machine stops operating and the economy becomes stagnant (see Piazza 2010 for an insightful formalization of this argument in a dynamic general equilibrium context). The monopoly-based theory of creative destruction does require monopolists to be ''destroyed'' every so often: if the monopolist never gets overtaken, then according to the theory, there cannot be any innovation.

3. CONCERNS ABOUT THE RECEIVED WISDOM: FACTS

Why should one doubt the sharp and reassuring predictions of such an elegant theoretical apparatus? In the Schumpeterian-NGT world, we can have our cake and eat it too: patented entrepreneurs are safely protected by the legal monopoly bestowed upon them, making higher and safer profits, while innovative activity (hence, technological progress and labor productivity growth) proceeds at the highest possible pace. This distinct characteristic of the Schumpeterian-NGT paradigm may well be the main reason behind its wide academic and public success. By telling us that monopoly power is socially benign because it generates ''dynamic efficiency,'' hence economic growth, at the very minor cost of some short-run losses of consumer surplus, it reconciles the desire of protecting what is already established with the urge to foster progress.

A world of low risk and high returns, though, seems too good to be true. At a minimum, it contradicts the basic economic intuition according to which there are no free lunches and, in a world of tradeoffs, rewards always entail some costs. More to the point, economic theory has been arguing for centuries that competition is conducive to good economic performances and good management practices, and all the available evidence (Nickell 1996; Blundell, Griffith, and Van Reenen 1999; Carlin, Schaffer, and Seabright 2004; and Okada 2005 are some of the main references within a very wide literature; for a decently recent survey, see Van Reenen 2010) suggests that, indeed, this is the case. If, in general, more competition has been shown to foster better allocation of resources, lower production costs, better management practices, higher workers' and entrepreneurs' effort, adoption of technological best practices, and so on, then it would be hard to believe on purely theoretical grounds that technological change and innovation should make such an extraordinary exception to this general and repeatedly confirmed rule.

This simple observation, after all, was at the root of an unfortunately little-known but remarkably prescient paper by George Stigler (1956), who—somewhat lonely—tried to argue that, to the best of his knowledge, most of the great innovations of any century had been created under conditions of competition and in the absence of patents and legally enforced monopoly power. That same paper contains a simple, somewhat artisanal empirical test in which Stigler uses pre–World War II data to regress a raw estimate of sectorial labor productivity growth against an index of sectorial concentration, finding that, indeed, the less concentrated an industry was, the faster labor productivity grew, at least back then. As far as we can tell, that was the first empirical refutation ever published of what is now (and was becoming then) the established wisdom. Times may have changed since (and they have indeed changed), hence the need for replicating Stigler's test using more recent data and more advanced statistical techniques, which we do in Section 5.

Before getting to it, let us consider a few other facts casting serious doubts on the realism of the Schumpeterian-NGT theory.

3.1 Who Innovates?

A key prediction of the Schumpeterian view of innovation is that the latter is the special product of large business organizations—their unique contribution to economic growth and to the increased social welfare it brings about. As a matter of fact, Schumpeter's argument rests more on the deep pockets and planning capabilities of large conglomerates than on patents and intellectual property per se. After convincing himself (do not ask us how) that socialist planning was the best way to achieve economic growth, he was intent on convincing his readers that through the actions of the protected monopolies, we could have our cake and eat it too: socialist planning for workers and consumers, and private property for the few lucky monopolists. Unfortunately, the historical and statistical evidence points to the exact opposite of what Schumpeter predicted. While large corporations are often (not always, maybe not even most of the time, but certainly often) the outcome of major innovations—sometimes patented and sometimes not; think of Ford and Edison, Sears and Amazon, Microsoft and Google, AT&T and Wal-Mart—major innovations are seldom if ever the product of large and

monopolistic corporations. On the contrary, breakthrough innovations are more often than not the product of small firms—competitive and creative outsiders that must compete with established incumbents and are able to do so because of their creative superiority. This fact, which has been remembered and then forgotten innumerable times in the applied economics literature, is documented once again in the recent work of William Baumol, one of the most articulate supporters of the central role played by the ''creative destruction'' mechanism in generating economic growth. So, for example, in Baumol (2010, chap. 2) we find an excellent summary of a series of studies, produced between 1995 and 2004 by the U.S. Small Business Administration, documenting that, in spite of the fact that the bulk of R&D activity is carried out by large corporations, the lion's share of technological breakthroughs come from small firms. One may argue that these findings are biased by the fact that the SBA's task is to argue exactly that. This is a correct observation that, nevertheless, cannot erase the fact that the data reported in those studies have withstood various years of attempted criticism and that the author quoting them with approval is one of the strongest supporters of the Schumpeterian view of technological progress. A second, critical observation may point out that a number of large corporations (IBM, Xerox, Bell Labs, or Microsoft) have been innovating a lot. Leave aside the fact that one would have a very hard time making the list longer than the one above; the fact is that the first three—IBM, Xerox, and Bell/AT&T—came up with lots of good ideas when they were the dominant monopolists in their industries but left those ideas in their drawers, not turning them into actual innovations until the breakup of their monopoly power forced them to compete with new entrants. Only then did their labs' good ideas turn into actual innovations, which is exactly the opposite of what the Schumpeterian's viewpoint implies. As for Microsoft, the answer is even simpler, if historically symmetrical to the previous one. Microsoft *was* a great innovative company until it had to fight to stay ahead of its competitors, but it stopped innovating once it acquired its monopoly power. As a matter of fact, Microsoft would already be history if it were not for the monopolistic grip it maintains on the market through its Windows operating system.

Establishing that the bulk of innovations does not come from large corporations but from small ones may or may not matter for those

subscribing to a more ''elastic'' view of the creative destruction mechanism. But it should at least serve as a reminder that, if patents are socially helpful, it is not because they allow the big conglomerates to innovate (as in Schumpeter's initial theory) but—maybe—because they allow the newcomers to enter a market. Hence, as far as evidence is concerned, we are restricted to considering the hypothesis that patents are good mostly for small firms and are instrumental to the inception of new industries. Is this the case?

This leads to a more sharply defined empirical question. Consider the 10 or 20 most economically relevant ''new industries'' of the last century or two: from electrical machines and appliances to chemical and then pharmaceutical products, from the car to the aviation industries, from software to hardware to mobile devices, from retail distribution to steel, from the movie and television industry to banking and finance. Most of these sectors are now relatively mature and dominated by a few large companies, either country-by-country or even worldwide, though to different extents. Nevertheless, when we look back at their inception, be it more than a century or half-century ago, not only do we not see any big conglomerate ''inventing'' (say) the personal computer or the motor-car and protecting it with a patent, but—practically always—we see groups of small and unprotected entrepreneurs introducing the new goods or services without much patent protection, at least in the very initial stages of the industry. The fact is that—even in the few cases when the innovation that revolutionized an industry comes from within pre-existing and relatively large companies in the same industry (think of the role that Wal-Mart has played, for example, in dramatically increasing the productivity of the retail and distribution sector, or that the now-all-defunct investment banks played in the 1970s and 1980s in creating the modern financial industry)—patents played either a minor role or no role at all in the gestation of the initial innovation!

In the light of these and plenty of similar observations, we wonder why in the applied industrial organization literature the following question seems to have never attracted the attention of researchers: what does actually happen over the life cycle of a new industry? The question is, at the same time, historical and statistical—purely descriptive in some sense, but of great theoretical relevance. How did the new industries that have actually emerged since the inception of the industrial revolution about two centuries ago come about?

How many were fostered by a preexisting large conglomerate and/or protected by an array of patents? Along the life cycle of a new good/technology, when does intellectual property enter the scene and start playing a (positive) role in fostering innovation? Is intellectual property there from the very beginning? Is it really the essential obstetrician without whom the healthy infant would never see the light of day, let alone grow to become the powerful athlete it eventually will? Or is it something that comes much later in the life of innovative industries, when the rate of innovation slows down, the technology tends to mature, and the classic ''shake out'' stage is under way or has just passed? Our guess, maybe anecdotal but not completely uneducated, is that the ''stylized fact'' emerging from such a careful historical investigation would be exactly the one we are suggesting here through our rhetorical questions. As far as we can tell, the main role intellectual property has played in the initial development of the great industries of the last two centuries has been either none or to delay some of them—see, for example, Boldrin and Levine (2008a, chaps. 1 and 8) for the cases of the steam engine and the airplane, respectively. In the few cases in which patents helped some among the initial innovators, such help was either damaging to the growth of the industry overall (as in the case of radio through the controversial Marconi patents), or it played an ambiguous role by first enabling either ''theft'' or something pretty close to it and then leading to the creation of long-lasting and scarcely innovative monopolies (as in the cases of Alexander Graham Bell and the telephone, and of RCA and the television). The only important exceptions to these stylized facts we can think of are associated with the name of Thomas Edison—the light bulb and related electrical equipment.

Most certainly we may be wrong, which is why we hope someone better equipped than we are will find the time and energy to look carefully into this issue. For the time being, though, all the modern industrial history available in print and online says that the Schumpeterian-NGT prediction is much more a fable than history.

3.2 Revolving-door Monopolists?

Next is the observation that, while in the established model of creative destruction progress comes from a new (patent-protected)

monopolist overcoming the incumbent one with a superior innovation rendering the latter's old patent irrelevant, this seldom happens in practice.

This observation readily applies to mature industries, where we observe two kinds of long-run market structures emerging. Most common is the case in which the efficient technology is so widely accessible that the sector has remained competitive and still experiences a large amount of entry and exit even many decades after having reached its maturity—for example, textile and apparel, furniture and house appliances, food and drinks, retail distribution, restoration and hospitality, and so on and so forth for a long array of ''traditional'' goods and services. These mature industries still experience a fairly large amount of creative destruction, with local or temporary leaders emerging and disappearing at a relatively high frequency and, more importantly, a sustained growth of labor productivity. Still, all this is achieved with little or no use of patents as a tool for enforcing monopoly power, thereby contradicting at least one of the two basic predictions of the Schumpeterian-NGT model (innovations keep taking place in these industries without the help of patents). The same model is also contradicted for the complementary reason by those mature industries that evolved toward a long-run oligopolistic market structure and in which patents are frequently used to enforce monopoly rights over innovations. In these cases—for example, cars and trucks, trains and airplanes, personal computers, proprietary software applications, shipbuilding, medicines, durable goods in general—market leaders tend to remain the very same few for very long periods of time (decades and decades, in fact) with very little, often literally zero, entry of creative outsiders. In all these industries, patents are certainly used and a more-or-less high level of innovative activity does take place, but the creative destruction mechanism predicted by the Schumpeterian-NGT model never materializes. There may be ''creation,'' but there is very little destruction: after a couple of decades, Microsoft Word—no matter how cumbersome and little different from 20 years ago it happens to be—still is the market leader in word processing, not to speak of Excel. The last time the market for word processing (or for spreadsheets) experienced any kind of ''creative destruction,'' patents were *verboten* in the software industry. This is not an exceptional case, as a quick examination of the industries listed above (or

of your own consumption basket, for that matter) will easily confirm; there is a lot of healthy creative destruction in free-market economies, but little of it comes from the new creative patentee overtaking the old, and no longer creative, patentee.

3.3 More Patents = Higher Productivity?

Let us move on to a more delicate issue: Is there a sharp and well-defined one-to-one relationship between patents and actual technological innovations? No matter how you turn it around, the Schumpeterian-NGT theory predicts that the stronger the monopoly power and the higher the (impact-weighted) number of patents issued, the stronger productivity growth should be—without ''if'' and without ''but.'' Coherent with such a prediction, patents, their numbers, and the number of times they are cited by subsequent patents are by far the most commonly used measures of innovation in most empirical industrial organization studies. Apart from the predictions of the received wisdom, this common academic habit has another reasonable justification: plenty of high-quality econometric work (starting with Pakes 1986) has shown that citation indices are a good measure of patents' quality and market impact. They are, for example, important predictors of the stock market valuation of the firm owning the patents. Frequently cited patents are economically very valuable, while scarcely cited patents tend to be associated with valueless innovations. In going from this finding to the conclusion that patent citations are an overall good measure of innovative activity, though, there is a far-from-obvious logical step, which we consider next.

The problem is twofold: It seems reasonable to believe that if a given innovation is both valuable and patentable, the innovator will have an incentive to seek a patent. While a number of surveys (Levin et al. 1987; Cohen, Nelson, and Walsh 2000) show that innovative firms very often rely on tools other than patents (e.g., secrecy, lead time, and so on) to achieve and maintain their competitive advantage, there is no doubt that, when possible, firms do patent their valuable innovations. It follows, therefore, that some innovations are contained in patents and that some patents contain actual innovations. The problem arises once we try replacing the two occurrences of the word ''some'' in the last statement with the words ''almost all.'' Safely performing such substitutions is necessary for treating ''patents and patent citations'' as adequate empirical measures of ''technological change.''

In very many industries, patenting was not even an option until a short while ago (e.g., the 1970s for plants, the 1990s for software, the 2000s for business practices, etc.) and it is still not a truly usable option for many other ones (e.g., retail and distribution, financial products, fashion and design, telecommunications, and so on). What this means is that lots of valuable innovations are not captured either by patents or by patent citations, making the latter an incomplete measure of the former. As the data below show, a very large number of productivity-enhancing innovations are most certainly not captured by patents. The other side of the problem is subtler: how many patents are issued (consequently, how often a patent is cited by subsequent ones) is determined by the specific legal and industrial organization characteristics of each different industry. Tens of thousands of patents are awarded each year in the software industry, while those in the car industry are counted in the hundreds. Some of these patents have innovative contents, but many others are purely ''defensive'' patents taken out for purely legal purposes and with little or no innovation inside them. The cumulative research work of James Bessen and his coauthors (e.g., Bessen and Hunt 2003) has made this fact plain in the case of the software industry.

As patenting became legally possible and spread widely across the industry, the actual amount of innovation carried out certainly did not increase and probably diminished, thereby making patents (and citations) a very poor measure of what we often take them to be. Our own somewhat artisanal examination of the agricultural sector (e.g., Boldrin and Levine 2008a, chap. 3) has led us to the same conclusion. Since they were legally allowed in the early 1970s, the number of agricultural patents has exploded—especially in recent decades thanks to the introduction of bio-engineered products—but there is no evidence of a structural break whatsoever in the time series of agricultural TFP either in the United States or in Europe. Hence, whatever it is that the dramatic increase in agricultural patents may be measuring, it cannot be the movement in agricultural productivity. Because it is the latter we care about, one should conclude that neither almost all innovations are contained in patents nor almost all patents have innovative content.

Case studies aside, how serious is this problem, statistically speaking, across a wide array of sectors and firms? In other words, for the overall economy, how poor a measure of productivity growth

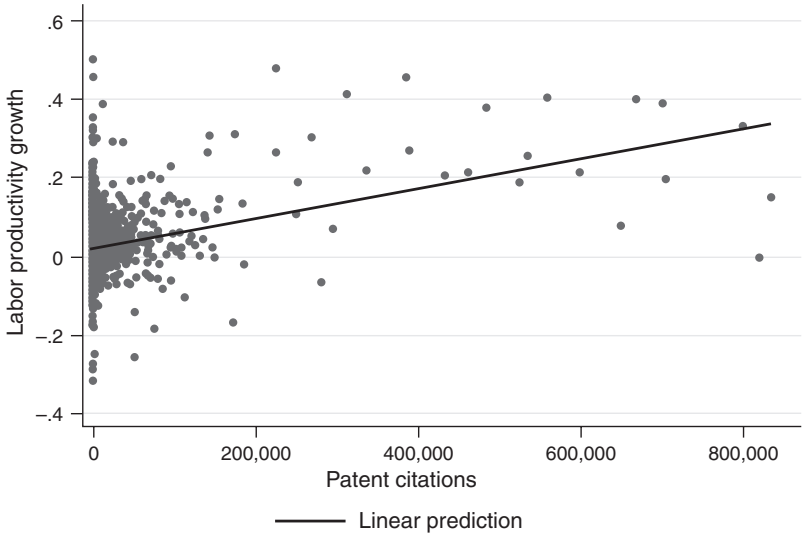
are patents and their citations? We try shedding some light on this question by computing conditional and unconditional correlations between patents and productivity measures from the same datasets we use later in Section 5 to carry out our main tests. In this instance, we have computed measures of labor productivity and TFP, sector by sector, together with patent numbers and citations for the firms in each sector. Next, we have regressed both their levels and growth rates on each other in a variety of reasonable combinations, both unconditionally and conditioning for investment activity in previous years and a number of common-sense control variables.

While not perfectly homogenous, our findings are nearly so. Their basic message is that, except for a couple of specifications reported below, there is not even a significant positive correlation between patents and productivity. This is a more surprising result than one would predict on the basis of purely theoretical considerations. In fact, one would expect patents to be at least a decent predictor of productivity growth across sectors—certainly for the last couple of decades, during which their use was extended to more and more sectors. Instead we find a weak correlation in only one of the many possible specifications, and no correlation at all otherwise! The data suggest, in other words, that the use of patents either as a defensive or as a rent-seeking tool (Boldrin and Levine 2004, 2008b) is actually more widespread than we had estimated it to be.

This specification yields a statistically significant, but very small, positive coefficient when we apply it to U.S. Bureau of Labor Statistics (BLS) output data, and a coefficient not statistically different from zero when National Bureau of Economic Research (NBER) output data are used. A quick look at Figure 1 suggests that a few data points may be driving the result, hence Figure 2 reports the same data without the top 5 percent of observations for citations. In Figure 2, the correlation is gone.

The role that the three statistical outlier sectors play in satisfying the received wisdom's predictions is revealed by computing the 20-year mean, at the North American Industry Classification System 4-digit level, for both variables in each sector. This is depicted in Figure 3. Also, for sectorial averages, once the top 5 percent values of the ''citations'' variable are dropped, the statistical correlation weakens substantially and, also in this subsample, appears to be determined by a small number of high-citation sectors. This is depicted in Figure 4.

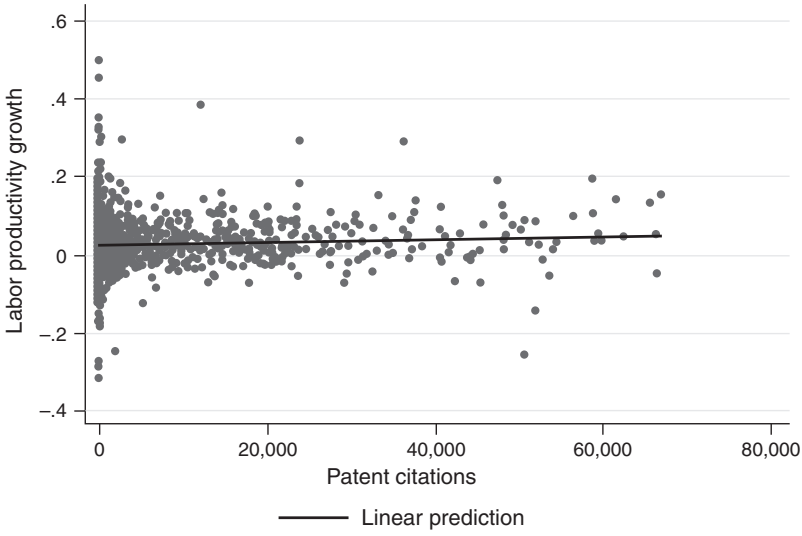
Figure 1
Patent Citations vs. Labor Productivity Growth
(1987–2008)



Very similar (in fact, weaker) results obtain when measures of TFP are used. In the Appendix, we report a number of additional cloud-like plots that are representative of the kind of results one obtains with different specifications. As the various graphs vividly illustrate, even in the very long run, patents and their citations measure only vaguely those innovations that actually increase either labor productivity or TFP, and do so only for a handful of sectors. Most observations are concentrated near the origin of the horizontal axis, corresponding to very few or no patents and patent citations at all; they display as much variance in the measured growth rates of productivity as the overall sample does, signaling that the larger fraction of productivity-enhancing innovations are not captured by patents.

This suggests that patents should not be used to assess which firm innovates more and which firm innovates less, unless one is capable of controlling for industry-specific effects and the various strategic and legal considerations different firms face at different points in time. From the point of view of economic theory and social

Figure 2
Patent Citations vs. Labor Productivity Growth
(Top 5 percent dropped)

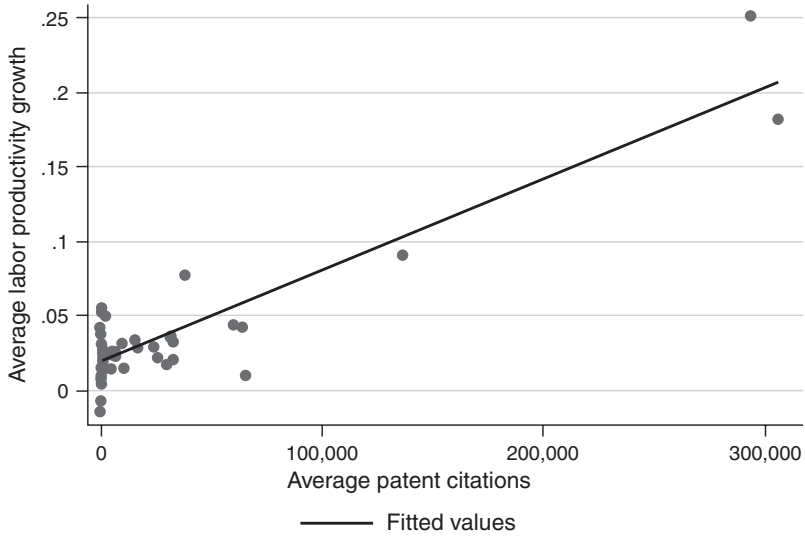


welfare, what matters is the growth of factors' productivity and not of the number of patents per se. Contrary to what it has been doing for decades under the influence of received wisdom, empirical research in this area should focus on the former as the proper measure of socially desirable outcomes, treating the latter as a tool—sometimes just a legal tool—that may or may not foster desirable innovations and productivity growth. Whether patents do or do not improve productivity should be the question we ask the data, not the preset answer from which we start our investigation.

3.4 Inverted-U Relation?

In recent years, an empirical ''middle of the road'' position has emerged within the Schumpeterian-NGT framework of analysis. This position is widely seen as a way of reconciling the theoretical predictions of that line of thought with the growing evidence that increasing monopoly power generates less, not more, innovation. Among others, the paper by Aghion et al. (2005) contains the main empirical findings giving support for this position, while the book

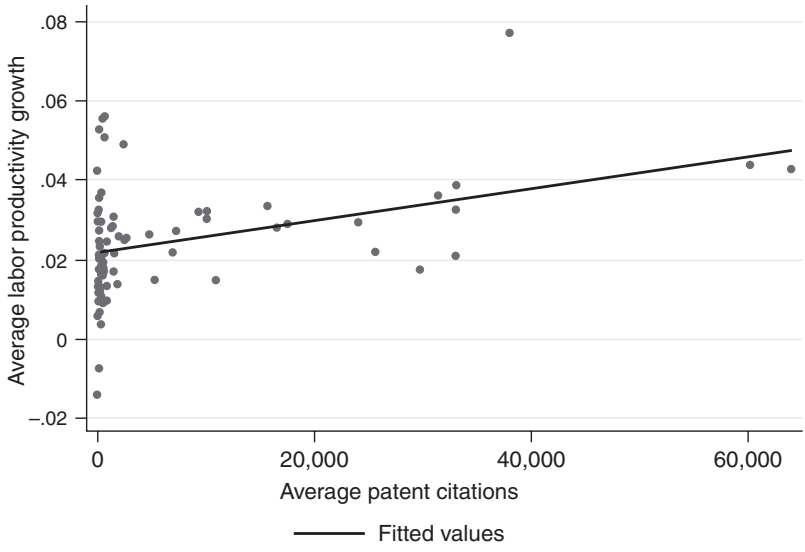
Figure 3
Average Patent Citations and Average Labor Productivity Growth



by Aghion and Griffith (2005) provides a broader overview of the whole line of analysis.

Aghion et al. (2005) develop a ''variety'' model in which competition is measured by the elasticity of substitution within pairs of goods, each pair produced by a duopolistic industry. In this setting, the higher the elasticity of substitution between the two goods is, the higher the return from innovation for either of the two duopolists will be. Assuming that sometimes one of the duopolists finds itself to be a technological leader relative to the other, a high elasticity of substitution between the two goods reduces the incentives to innovate for the laggard when its distance from the leader becomes particularly large. The authors interpret this kind of model as predicting that the maximum innovative effort will obtain at some ''intermediate'' position between perfect substitutability (competition) and perfect complementarity (monopoly). They compare these predictions with the patenting activity in the United States for a panel of United Kingdom manufacturing firms, claiming that their

Figure 4
Average Patent Citations and Average Labor Productivity Growth
(Top 5 percent dropped)



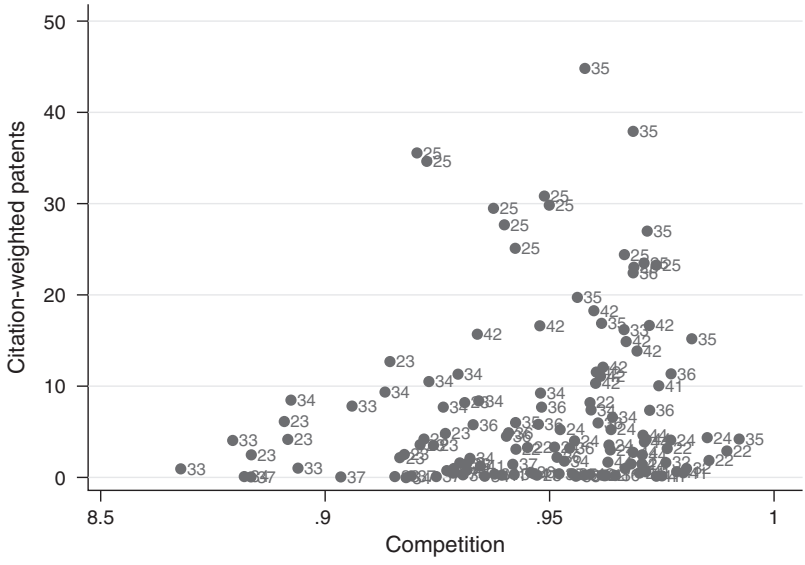
model's predictions are supported by the data. This is interpreted as a vindication of the basic intuition behind the Schumpeterian-NGT theories, according to which at least some degree of patent protection and monopoly power is good for technological progress. *In medio stat virtus.*

The robustness and even the very same existence of such empirical findings are questioned in the works of Correa (2010), first, and then of Hashmi (2011). The latter re-examines the inverted-U relationship between competition and innovation by using a different dataset covering publicly traded manufacturing firms in the United States between 1976 and 2001. Apart from a minor statistical detail, Hashmi mimics the estimation strategy of Aghion et al. (2005). He replicates their results for the UK dataset but, when using the U.S. dataset, he finds a robust positive relationship between competition (as measured by the inverse of markups) and innovation (as measured by citation-weighted patents). Hashmi conjectures that the different findings may be due to the underlying characteristics of the two

different economies. In particular, he claims that a possible explanation of the sharply different results may come from the fact that (in his view) the U.S. manufacturing industries are technologically more neck-and-neck than their counterparts in the UK. In other words, there is a lot more competition among firms in the United States than in the UK. In this case, because no competitor is ever ''far behind'' the leader, the incentives to innovate that competition induces are stronger; because there are not many laggards giving up the race, there are also very few leaders that are able to keep the lead without further innovations. The argument is more than a bit lopsided because ''competition,'' in these regressions, means a low markup, and the argument we just summarized does not explain why, in the United States, firms with very low markups should innovate a lot while in the UK the same kind of firms innovate very little. Further, as Hashmi (2011) remarks, this conjecture is not really supported by the data, and there is no independent evidence in the literature that this should be the case. Finally, we note that, even if the conjecture were supported by the data, it would simply mean that, indeed, the more competition there is, the better competition works. This tautology, at the very end, does not account for the different empirical results between the U.S. and UK firms, which remains a puzzle.

The investigation carried out by Correa (2010) shows, instead, that there is no puzzle and that this particular conjecture is not necessary to account for the different correlations observed in the two datasets. This is because, most likely, the inverted-U pattern that Aghion et al. (2005) claim to have discovered in the UK data is either not robust to very reasonable perturbations or it is not there at all. Guided by the well-documented recent history of the U.S. patent system, Correa begins his investigation by carrying out a simple Chow test on the Aghion et al. (2005) data to find evidence that a structural break took place in the early 1980s. This structural break coincides with the establishment of the United States Court of Appeals for the Federal Circuit (CAFC) in 1982. Scholars of intellectual property have amply documented that the establishment of the CAFC had a dramatic impact on the enforceability of patents and greatly strengthened the position of patent holders, see, for example, Jaffe and Lerner (2004), Kortum and Lerner (1998). It therefore increased the incentive to apply for patents, and those applications started to

Figure 5
Competition and Patent Citations Prior to the CAFC
(1973–1982)

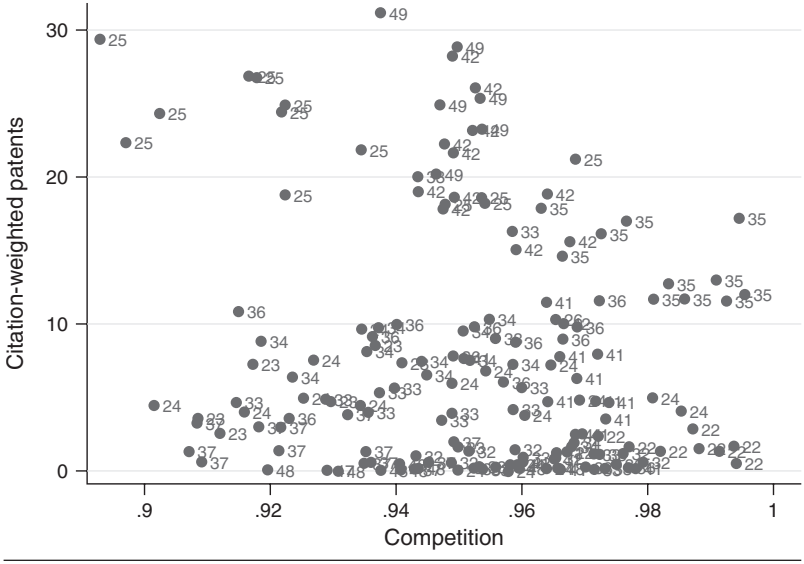


increase at about that time. By taking into account this break, Correa finds a positive relationship between innovation and competition during the period 1973–1982 and no statistically significant innovation-competition relationship during the period 1983–1994.

Figure 5 displays the sample observations of innovation and competition levels before the establishment of the CAFC (period 1973–1982). We can see that some particular industries, such as Motor Vehicles (Standard Industrial Classification 2-digit code 35) and Chemicals (SIC-2 25), have higher levels of innovation than the rest of the manufacturing sector. However, as Figure 6 shows, after the establishment of the CAFC, both the Motor Vehicles and the Chemicals industries decreased their innovation intensity, while industries such as Other Manufacturing (SIC-2 49), Sugar Beverage and Tobacco (SIC-2 42), and Office and Computing Machinery (SIC-2 33) increased their number of citation-weighted patents.

The Chow test looks for only one structural break without using additional information about its location in time. Clearly, the break

Figure 6
Competition and Patent Citation after the Creation of the CAFC
(1983–1994)



could have taken place in another year or there could be more than one break, hence a sup-Wald test is performed, which makes no assumptions about either the number or the location of the structural breaks. This test finds just one structural break and places it in 1981, while the CAFC was established only in 1982. However, it is worth noting that in the dataset, a patent is assigned to the year in which the application was filed rather than the one in which it was granted. While it is now greatly increased, reaching almost three years, a substantial lag between applications and grants existed already in the early 1980s. Furthermore, the establishment of the CAFC did not come as a surprise to anyone but was the consequence of a long and public debate about the opportunity of introducing such a court to strengthen the U.S. patent system. It seems therefore reasonable to expect patent applications to react somewhat earlier on, in anticipation of the October 1982 establishment.

We leave it to the reader to decide which way the final jury should lean. We are satisfied with having shown that, indeed, once one takes into account a little tiny bit of well-known history, the inverted-U relationship is gone, and a clear upward-sloping relation connects the degree of competition with the strength of innovation also in the particular dataset used by Aghion et al. (2005). That is, we like to recall, the one and only dataset discovered so far in which a modified version of the Schumpeterian-NGT theory appears to be vindicated.

4. A DIFFERENT POINT OF VIEW

In the data, one can hardly find any support either for the claim that stronger intellectual property regimes favor technological change or for the general theory according to which monopoly power fosters innovation. (There is, in fact, good evidence for the opposite.) There is scant evidence even for the claim that patents are a reliable measure of productivity growth. Thus it is important to understand how markets for innovation and technology adoption might function in the absence of intellectual property. Innovative activity has taken place historically and does take place currently, as plenty of evidence shows, in markets where intellectual property restrictions are either absent or even not allowed. Hence, in order to explain the existence of something that standard theory predicts should not even exist (i.e., competitive innovation), one needs to develop an alternative theoretical framework, which we briefly outline next.

In the paper we quoted earlier, Stigler (1956, p. 274) argues that monopoly is completely unnecessary to provide incentives for innovation.

> There can be rewards—and great ones—to the successful competitive innovator. For example, [consider] the mail-order business.... The innovators ... were Aaron Montgomery Ward, who opened the first general merchandise establishment in 1872, and Richard Sears.... Sears soon lifted his company to a dominant position by his magnificent merchandising talents, and he obtained a modest fortune, and his partner Rosenwald an immodest one. At no time were there any conventional monopolistic practices, and at all times there were rivals within the industry and other industries making near-perfect substitutes.

In more recent times, Liebowitz (1985), Hellwig and Irmen (2001), Quah (2002), Legros (2005), and Boldrin and Levine (1997, 2005, 2006, 2008a,b) have all examined the competitive rents that accrue to innovators due to ''limited capacity''—the fact that, in a competitive market, the owners of a fixed factor (first copy of an idea) are the recipients of all downstream rents originating from it, and that an infinite number of copies cannot be made instantaneously. The conclusion is that innovation will take place even without intellectual property, as it often has in the past—see, for example, the cases mentioned by Moser (2005). While some of this work shows that there may be too little innovation under competition due to the indivisible nature of the initial copy of ideas, it also suggests that the appropriate remedy is unlikely to be a government-granted monopoly.

On the one hand, it seems transparent that providing a reward for innovation in the form of a monopoly can only increase innovation, at least on impact. Indeed, this idea is so obvious it seems to have an intellectual grip on the economics profession that a half-century of empirical evidence to the contrary has been unable to break. There is, however, a flip side to a patent system as opposed to an individual patent: a patent system not only rewards innovators, it also makes it more costly to innovate in the face of the many licenses that need to be acquired in order to bring a new product to market. Hence, from a purely theoretical point of view, the impact of patenting on innovation is ambiguous: it both encourages and discourages innovation. Once this fact is taken into due account, it is perhaps not such a surprise that no evidence has emerged that there is a net positive effect. Scotchmer (1991) is among the first to point out the problem that occurs when innovations build on existing ideas. Formal models, first in the form of a simple example and later in the form of a detailed model of innovation, were provided in Boldrin and Levine (2005).

On the positive side, Boldrin and Levine (2006) examine innovation that is driven by the ''need'' for new technologies when old technologies have exhausted their growth potential. In such a setting, perfect competition delivers the first-best, as competitive rents are exactly what is needed to provide the socially optimal level of invention. Hence, any intervention in the form of a government-granted monopoly strictly reduces welfare because a monopolist would accumulate each technology more slowly than under competition,

thereby increasing the time it takes to move from one technology to the next, more productive one, thereby reducing the long-run growth rate of consumption.

A further aspect of the theoretical problem, as Stigler briefly noted in his 1956 article, is that the formal definition of competition is capable only of capturing the final, stylized outcome and fails to describe the process. In a (perfectly) competitive industry, the participating firms are small relative to the size of the market, produce the same identical goods with the same identical technology, and act as price-takers because there is nothing they can do that can alter anything of relevance in the market in which they participate. Such a description of ''competitive outcomes'' may be useful for the study of general equilibrium arrangements in certain settings, but it hardly seems capable of describing the underlying intuition according to which ''competing entrepreneurs'' try to outsmart each other by both imitating the best practices and improving upon them, reducing production costs as much as possible (thereby de facto implementing technological change) to finally bring about some form of cost/price equalization among the surviving firms. It is the latter, though, and not the former that one has in mind when saying, intuitively, that ''competition fosters innovation.''

Is ''monopolistic or imperfect'' competition, then, the proper analytical answer to this unpleasant state of affairs? While one could be tempted to say ''yes,'' at least instinctively, and then follow along the lines of Aghion et al. (2005) and identify the extent of competition with the degree to which one good is a substitute for another, there are good reasons not to go that way. The extent to which a new good is, or is not, a substitute for an already existing one is an endogenous equilibrium choice, not an institutional or technological parameter, as innovators elect which niche to enter on the basis of market incentives. This is, in fact, one of the channels through which competition fosters innovation: the free entry of imitators/innovators choosing either to expand total productive capacity or to build up on previous innovations by introducing close substitutes of the existing goods. Until we are able to build workable models of free entry in which this dynamic aspect of competition is captured over time, we must face the unpleasant choice of either going to the extreme of ''perfect competition,'' in which everyone always does the same thing, or to that of ''monopolistic competition,'' in which

every firm is actually a monopolist acting in parallel to other, similar monopolists.

In any case, the key conceptual problem with the ''monopolistic competition'' approach to modeling competition is that it is grounded in the idea that imitation is costless, hence competitive rents are irrelevant and competition would always bring about Bertrand pricing. When imitation is costly but competitive rents are not immediately equal to zero—therefore, imitation is profitable—monopolistic competition evaporates, and the dynamic process of competition with free entry sets in. In practice, imitation *is* costly and, as noted above, a patent system serves only to discourage downstream imitation. The latter point is important because in practice—as shown by the managerial surveys cited earlier—innovations tend to be protected more by the difficulty of imitation than through government intervention in the form of patents. This means that, de facto, the role of patents is to discourage competition, free entry, and indeed innovation; old incumbent firms acquire large patent portfolios covering tens of thousands of known ideas with the aim of using them as barriers against any entrant that does not have a similar portfolio.

Finally, let us consider the following often-heard criticism of the theoretical argument according to which competition favors innovation more than (legal) monopoly does:

> The Boldrin & Levine model has a built-in first mover advantage. Of course, then, if the first mover advantage is big enough you do not need patents. I never understood why we should pay attention to such a trivial point (theoretically, the empirical question of whether first mover advantages would be enough is of course relevant). Is there something that I have been missing all of these years?

It is true that the essence of the Boldrin and Levine model is a first-mover advantage generating competitive rents. Boldin and Levine repeatedly mention it in their original paper on perfectly competitive innovation (Boldrin and Levine 1997), which was in fact meant to show how the traditional Edgeworth-Marshall framework of a competitive industry with free entry and minimum plant size could easily be adapted to model repeated innovation in a growth-theoretical setting. But there are two aspects in Boldrin and Levine's

paper that are not trivial at all, at least not for the modern innovation and growth literature. First, they make the argument that there is always a first-mover advantage and that one can predict how large it is on the basis of observables. In other words, they give some structure to what the above criticism calls an advantage ''big enough'' and show under which conditions it may or may not obtain in practice. For instance, given the elasticity of demand, they show that the first-mover advantage (and the incentive to undertake an innovation) depends on two things: the initial cost of getting the first unit of output (the prototype), and the rate at which the invention can be copied. Second (and related to the previous point), even in markets where it is very easy to reproduce the innovation, the rent of the first mover can potentially increase if consumptions between periods are substitutes.

Actually, we think there is a deeper point that the expression ''first-mover advantage'' misses. Also, in a regime of monopolistic competition, there is a first-mover advantage—a short-run monopoly. But whether the first-mover advantage is a legally induced monopoly power (i.e., a market distortion) or competitive rent (fully efficient) is quite relevant for both positive and normative analysis. It is relevant for positive empirical analysis because one can use statistical data to quantify the two uses of ''enough'' in the statement, ''[W]hen the initial fixed cost is small enough relative to the size of the market and the cost of copying is high enough, we will observe sustained innovation under conditions of competition.'' It is relevant for normative analysis because, to the extent it allows us to explain the thousands of episodes of competitive innovation observed in reality, it provides us with guidance for setting policies in this area. Second, the expression ''first-mover advantage'' seems to imply that, once there is a ''second mover,'' the first-mover advantage goes away. It does not, and the framework mentioned earlier captures this important fact well, while the one of monopolistic competition misses it completely. This is because in the Edgeworth-Marshall framework, one assumes that capacity is bounded at every point in time and costly to accumulate, which is not the case in the monopolistic competition case. Because of this assumption, in a competitive industry, rents are generated more or less for the entire life of the industry; in the baseline case, without any external effects, all the rents do go to the first mover, but in more general cases, they also

accrue to imitators and downstream innovators. We do not think that is what has been historically meant by ''first-mover advantage,'' and we would be pretty surprised if the above criticism meant that. As far as we can tell from the strong resistance this set of propositions still faces more than a decade after they were first published, these statements are not widely accepted. The propositions may be wrong, but they are definitely not ''trivial.''

There is another point in which the reported criticism is perfectly correct. Ultimately, it is an empirical question whether absence or presence of patents distorts markets more. The ''this is trivial'' criticism seems unaware of a very long literature in and outside of economics asserting that this is not the case and that, as a matter of theory, absence of patents distorts markets more than their presence. As we have documented, there is an ever-growing, long list of empirical economists surprised to learn that, in the data, patents lowered rather than raised innovation, and they have no idea why that might be true. The theory of competitive innovation provides a testable answer to this long-standing puzzle.

4.1 The Embodiment Issue

The ''embodiment hypothesis'' is a crucial step in the theoretical argument claiming that competition, in the sense of free entry and the right to imitate, fosters innovation and economic progress. Before moving forward to consider recent microeconomic evidence supporting our claim, we should briefly explain why the embodiment/disembodiment controversy is relevant in the context of our research.

First off, what is the embodiment/disembodiment controversy? It centers on the fact that technological advances may or may not be obtained without embodying them in something material and expensive to either produce or acquire—that is, in some ''capital,'' be it physical or human or organizational. Traditional treatments of the TFP measure describe the latter as completely disembodied and unrelated to investment expenditure. Most likely, this did not correspond to the intuition many researchers had of the nature and causes of TFP—certainly, it was not what Robert Solow had in mind. Nevertheless, it was the formalism adopted in writing $Y = AF(K, L)$, where Y is output, A is TFP, K is the stock of capital, and L is the flow of labor entering the neoclassical production function F. The same is true for the literature we have classified as Schumpeterian-NGT:

while some of those arguments would, formally, go through even if the new technology were embodied in some kind of capital good, most models use a disembodied representation of technological progress to formalize their argument. There are purely technical reasons for this choice (i.e., deriving a balanced growth path), but there are also intuitive and conceptual ones, which should be understood, as they are relevant for the issue concerning us here.

The fundamental link between disembodiment and the Schumpeterian-NGT view of the world is related to the assumption of unbounded productive capacity that the Schumpeterian-NGT view needs to make in order to reach its specific conclusions. In fact, for the ''imitation will destroy innovator's rents'' argument to work, one must assume that imitators will force Bertrand pricing almost immediately. For Bertrand pricing to make sense, one needs a situation in which each of the competitors is capable of satisfying total demand all alone, which requires productive capacity to be instantaneously expandable. While one may pretend that the available quantities of physical and human capital of a certain kind (i.e., the kind embodying the new technology) can be increased at an infinite speed at no cost, this seems in clear violation of the basic economic intuition according to which there is no free lunch. In such context, the hypothesis that the new idea or technology is actually disembodied and that, therefore, one can make a very large number of copies of it at no cost and in no time seems necessary to move the argument forward. More generally, the view according to which ideas, once discovered, can be immediately copied by anyone at no cost requires, conceptually, a form of general disembodiment of the idea itself that, like a light in the sky, appears to, and is immediately usable by, everyone else, were it not for the legal restriction that patents imply.

At the opposite extreme, when a new idea is fully embodied either in an object owned by the creator or in her human capital, it is up to her to decide what to do with it, and unauthorized imitation becomes, if not impossible, certainly neither easy nor legal. This is the key fact allowing innovators to be rewarded for their inventions even in the absence of the specific privileges that intellectual property introduces. As a matter of fact, the theory of competitive innovation rests on the twin hypotheses that

- inventors have control of their inventions and will require appropriate payment to make them available to others, and

- imitation is always and everywhere costly because it requires either producing or acquiring the material object or the human capital embodying the innovation.

These two assumptions, which we consider most natural, imply that productive capacity embodying the new technology will build up only slowly, the price of the new good will be determined by demand for quite a while, and the position in which price equals average cost will be reached only in the long-run equilibrium of the competitive industry. As a consequence, competitive rents will be attained by both the innovator and her imitators. Also in this case, one can certainly conceive of circumstances in which such rents could obtain even when the innovation is completely disembodied, but those cases appear to be either exceptional or far-fetched.

In conclusion, the theory of competitive innovation rests on the fact that innovations are embodied while, symmetrically, the Schumpeterian-NGT view of innovation rests on the fact that innovations are disembodied. The latter is a factual issue, and it should be resolved empirically. Common-sense inspection of actual innovations fails to deliver (as far as we can tell) convincing examples of actual innovations that were completely disembodied and could be instantaneously reproduced in an unlimited amount with zero costs of imitation. Even the commonly used examples of ''digital goods'' (whose historical relevance is clearly limited, given that they were not around until a couple of decades ago) fail the test. Unless the original creator voluntarily releases the master copy of the digital good in question, making a large number of copies of it available to consumers is technically impossible, even assuming the machines through which reproduction and distribution of digital goods take place are costless, which they are not. In a world of free competition, creators will charge a price to release the master copy to imitators and, in equilibrium with free entry, it is easy to show that such a price will be equal to the net present value of all future rents (Boldrin and Levine 2008b).

The applied statistical literature on this argument is very large, and our reading is that—since Zvi Griliches' path-breaking 1957 Ph.D. dissertation ''Hybrid Corn: An Exploration in the Economics of Technological Change''—it leans almost completely in favor of the embodiment hypothesis. It should be pointed out here that Robert Solow himself, with whom many people tend to associate the

idea of ''disembodied TFP,'' always squarely supported the view that the quantitatively relevant type of technological progress is, in fact, embodied in one form of capital or another.

In any case, this is not the appropriate place to engage in a full survey of the embodiment controversy, even if it may be the place to point out that an updated and careful summing-up of the half-century-old debate would be most welcome in light of its fundamental relevance when it comes to modeling innovation. We will therefore limit ourselves to summarizing the findings of a recent micro-investigation of this issue, which uses an original and very powerful dataset and that is one of the few to address squarely the questions that concern us, that is, the empirical relevance of ''disembodied external effects'' from technological innovation across competing firms.

Castiglionesi and Ornaghi (2011) make use of an unbalanced micro-panel dataset of Spanish manufacturing firms observed with annual frequency during the period 1990–2006. This dataset proves to be particularly suitable for disentangling the impact of specific individual sources of productivity growth, as it includes detailed observations on firms' outputs, inputs, proportion of skilled employees, types of capital investment undertaken, and modifications of the production processes. Moreover, a unique feature of this dataset is that it provides growth rates of firm-specific prices for outputs and intermediary inputs, thus allowing for the construction of a more reliable measure of firms' productivity change.

Their estimation builds up progressively from a simple regression that reveals a large and unexplained residual representing the (unweighted) average TFP growth across firms. Traditionally, this is taken as a measure of ''disembodied technological progress''; their goal is to demonstrate that, using the micro-observations listed above, one can show it is actually accounted for by very embodied investments in some kind of capital. To this end, they start by analyzing the contributions of traditional disembodied variables as sources of average TFP growth. They consider firm-specific learning-by-doing (LBD) and unpriced externalities such as those acting through aggregate human capital and the spillovers of R&D from firm to firm. The quantitative role played by such variables in accounting for TFP growth at the firm level is often argued to be strong evidence, possibly the strongest, in favor of the Schumpeterian-NGT theory

of innovation and growth. Because, this theory says, technological progress is mostly disembodied and generates large externalities at no explicit cost to the recipients, patents are necessary to allow inventors to make their products excludable, therefore appropriating at least part of their social returns. Without patents, private appropriation of the returns from innovative investments would become impossible, and the value of the innovation for the innovator would dissipate through the external effects induced by free imitation. Castiglionesi and Ornaghi replicate in their dataset the well known earlier results showing that certain aggregate variables—interpreted as the source of disembodied technological progress and unpriced externalities—are correlated to average TFP growth at the industry level.

Next, they take into account the relevance of (human and physical) capital-embodied technological progress as an engine of TFP growth. They measure the impact of new capital goods on TFP by means of two variables: the average vintage of the physical capital and an index of new technology usage. They account for differences in human capital using two variables: firm wages and the percentage of R&D employees at the firm level. To deal with classical endogeneity issues, they estimate a specification with the ratio of skilled workers at the firm level instead of firm wages. Once the measures of embodied technological progress are considered, the variables that capture firm-specific LBD, human capital externalities, and R&D spillovers no longer show any relevance in affecting average TFP growth either at the firm or industry level. In other words, once they account for variations in physical and human capital that are measurable at the firm level, the external effects completely evaporate and are no longer relevant in ''explaining,'' even statistically, the movements in TFP. In fact, Castiglionesi and Ornaghi find that embodied variables alone can fully explain average TFP growth across firms and industries. Last but not least, in all specifications, constant returns to scale cannot be rejected.

Finally, in order to better assess firm-specific LBD and to look more carefully into the presence of potentially external effects, they consider two alternative measures that, arguably, are closer in spirit to the theoretical idea behind LBD: cumulative output since the introduction of a process-innovation, and time after the introduction of a process-innovation. These two variables should capture the idea

that a change in the methods and techniques used for production must trigger a new learning cycle of the workforce at the firm level. When considered together with the embodied variables, these alternative measures of firm-specific LBD retain explanatory power. This is coherent with the classical definition of LBD: internal to the firm, short-lived, and due to the adoption of new processes. In other words, there is no evidence of unpriced spillovers in this dataset. These results, taken together with those relative to the embodiment of TFP, cast a long shadow on the idea that spillover effects external to the firm play a major role in technological progress and in the increase of TFP along the lines assumed in the Schumpeterian-NGT paradigm.

Summing up: Castiglionesi and Ornaghi's findings prove that things such as ''free imitation'' and ''disembodied technological progress''—generated via a chain of external spillovers from one firm to another—find scant support in actual microeconomic data. Innovations and technological change appear to originate at the firm level and are mostly embodied in investment decisions, which are therefore both costly and internal. Average TFP growth is fully explained by the kind of technical progress that is embodied in actual physical and human capital; economy-wide neutral (or disembodied) technical change plays virtually no role.

5. NEW FINDINGS

We have already seen in Section 3 that, contrary to what received wisdom keeps repeating, the available empirical evidence about the statistical relation between measures of competition and patents point to a positive link: more competition, more (highly cited) patents. Still, because we have also seen that patents and their citations are poor predictors of productivity growth, these results imply relatively little with respect to the policy issue at stake: Does competition foster productivity growth more or less than monopoly?

With the aim of replicating Stigler's test on more recent micro data, we have constructed a dataset that includes, among other measures, patent counts and patent citations for 220 4-digit SIC code industries over the period 1990–2001 and productivity growth for 85 4-digit NAICS code industries over the period 1987–2008. The raw data used to construct this dataset come from different sources. Firm balance-sheet and financial data available in Compustat are

matched with firm-level data on patents retrieved from the NBER Patent Data described by Hall, Jaffe, and Trajtenberg (2001). Total output data for SIC-4 manufacturing industries are taken from the NBER and U.S. Census Bureau's Center for Economic Studies Manufacturing Industry Database described by Bartelsman and Gray (1996). Information on output, inputs, and productivity for NAICS-4 manufacturing industries are obtained from the BLS. Finally, we retrieve information on U.S. imports at SIC-4 and NAICS-4 industry level from the U.S. Department of Commerce and the U.S. International Trade Commission. The accounting data retrieved from Compustat include sales *S* (item 12), gross capital *K* (item 7), operating profits *OP* (item 13), and advertising expenditure *A* (item 45) for the period 1990–2006. The data we use refer to all firms in the manufacturing sector; this includes 7,432 firms divided among 220 industries according to the 4-digit SIC code.

This large dataset allows us to construct more accurate measures of innovation and competition and to test the robustness of our findings when different outcomes are used to capture technological change and productivity growth. Specifically, innovation is measured with two different sets of variables. The first set consists of number of patents and number of citations received by those patents. As opposed to a simple patent count, citations can capture not only the quantity of ideas produced but also the quality of those ideas. The main advantages of patents compared to other R&D indicators are that they provide a measure of successful research output and they are objective in so far as they are not influenced by accounting practices. At the same time, there are a number of limitations in measuring innovation through patents, which we have already discussed and documented in Section 3 above. As we showed there, patents and citations measure only a relatively tiny fraction of the actual output of innovative activity. Moreover, patents cannot account for efficiency gains due to the adoption of the most efficient technologies and best managerial practices as long as these are not patented, nor can they account for the increased productivity that follows the introduction of new goods and services not covered by patents.

Coherent with the discussion carried out in Sections 3 and 4, and in order to overcome these limitations and to provide a check of robustness of our findings, the second variable we use to capture

technological advances refers to firms' productivity growth computed either as TFP or as labor productivity for 85 different NAICS-4 manufacturing industries over the period 1987–2008, as calculated on the basis of either NBER or BLS output data. Competition is an index based on the profitability of the industry and takes values from 0 (low competition/high profits) to 1 (high competition/low profits).

On the basis of the data so organized, first we have replicated and extended the empirical model discussed in Section 3 (see Correa 2010), with innovation (patents or patent citations) on the left-hand side and competition (inverse of profitability) on the right-hand side. In line with the theoretical arguments developed in Section 4 as well as with the empirical results discussed in Section 3, we find a positive relationship that is remarkably robust to changes in industry classification (SIC-4 vs. NAICS-4), time sample period (1990–2001 vs. 1975–2001), and set of sampled industries (manufacturing vs. all industries).

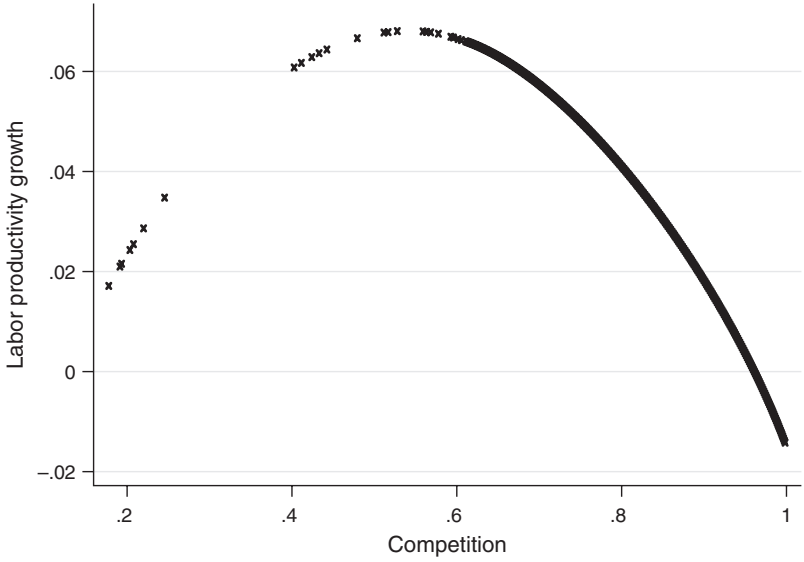
Granted, then, that the traditional measures of innovations are—in our dataset very much like those considered in Section 3—positively correlated with the most natural index of competitive pressure, we moved next to the issue that is most important for us, that is, the correlation between the latter and objective measures of productivity growth. The next two figures (Figures 7 and 8) are based on NAICS-4 industries for the period 1987–2007. Figure 7 seems to suggest that, when one regresses productivity growth on the inverse of profitability, the inverted-U relationship that had been pushed out the door in the case of patents strikes back with a vengeance when studying labor productivity growth.

Nevertheless, as amply discussed in Correa and Ornaghi (2011) and Hashmi (2011), the (inverse of) profitability is clearly an endogenous outcome that depends, among other things, on the ability of firms (in each sector) to reduce costs and increase prices through innovations that increase labor productivity.

To deal with this problem, Figure 8 is obtained by regressing labor productivity growth on lagged inverse profitability, that is, on measured sectorial competition during the previous period. This ''simple'' change has a dramatic effect: from an inverted-U we move to a clear positive relationship—and a very robust one as the tests reported in Correa and Ornaghi (2011) show.

This result is confirmed when the observations are aggregated sector by sector and the regression is performed using the sectorial

Figure 7
Labor Productivity Growth and Competition



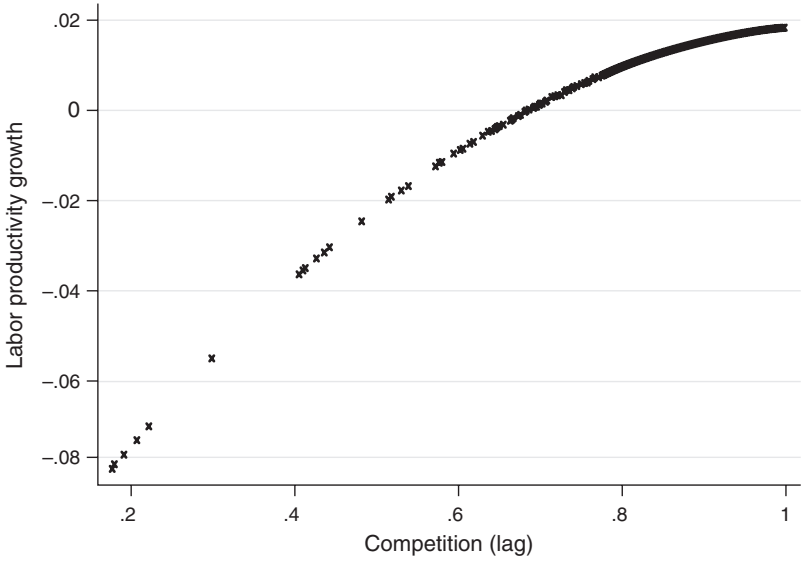
averages on both sides of the equality sign. If possible, the positive relationship displayed in Figure 9 is more pronounced than that in Figure 8.

It is worth stressing that, because of re-scaling, the interval of variation of productivity growth, as reported in Figure 9, may incorrectly appear as exceedingly small. This is not the case; when expressed in percentage points per year, the effect of increased competition is actually quite strong. Once the few very monopolized sectors in which our index of competition is lower than 0.7 are dropped, the actual lower bound is a growth rate of 1.5 percent per year (recall that these are sectorial averages), and the upper bound is 3.5 percent. In other words, the average annual growth of productivity in the sectors with the highest level of competition is up to 2 percent bigger than in the sectors with the lowest level of competition. These are strikingly large differences when cumulated over various decades, as is the case in our dataset.

6. OPEN ISSUES AND CONCLUSIONS

Economic theory is ambiguous when it comes to assessing if either competition and free entry or patent protection and the monopoly

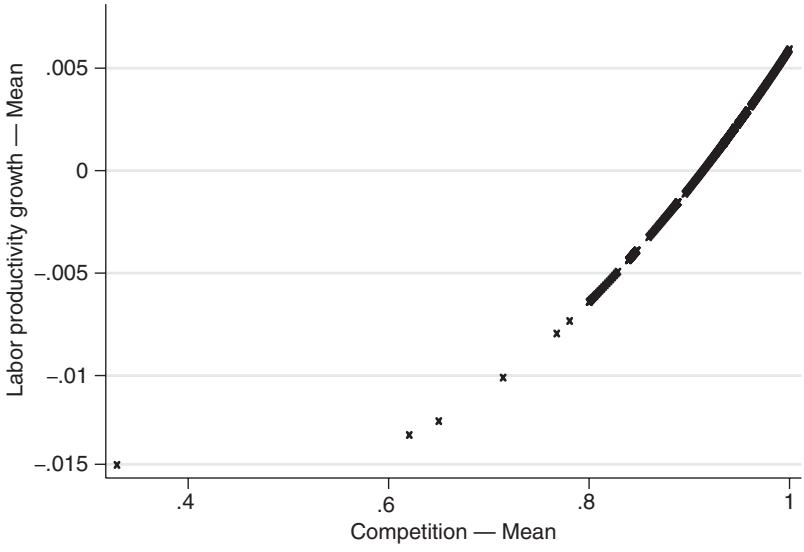
Figure 8
Labor Productivity Growth and Lagged Competition



power it induces define the best institutional environment to foster innovation and technological progress. The empirical evidence, though, is overwhelming: far from fostering and stimulating innovation and productivity growth, patents are likely to hamper them, while free entry and competition appear to be at least correlated with (if not the cause of) labor productivity and TFP growth. This has an obvious implication: There are no objective reasons whatsoever to strengthen patents any further than we have already done. In fact, it seems urgent to begin undoing some of the damage we have been doing to ourselves over the last 30 years and work to reform—slowly but surely—the overall system of intellectual property.

Reforming the system is neither politically easy, nor is it obvious, at least to us, which steps are appropriate and in which order they should be undertaken. As a matter of fact, the number of empirical studies that have carried out well-designed counterfactual exercises capable of assessing how an alternative legal system would affect innovative activity and which parts of the current one are the most damaging to technological progress is exceedingly small. What this

Figure 9
Labor Productivity Growth and Lagged Competition
(Sectorial averages)



means is that the very first policy goal may just be that of advocating for, supporting, and carrying out well-designed empirical studies of how innovation and technological progress fare across sectors, countries, and time periods, under different legal and incentive systems. Our basic intuition, grounded on the research either presented or referred to here, is that the current systems of intellectual property are tantamount to the trade restrictions in existence until a few decades ago, and their dismantlement should be approached in the very same fashion, piece by piece and quite patiently.

We do not have, therefore, any grand strategy to propose, just a short list of general goals plus a few somewhat more specific ones in those particular areas where, we believe, economic analysis has managed to dig deeper. Because policy proposals are better digested and metabolized when served in the form of small pills, here is our list:

- Stop the still-rising tide that, since the early 1980s, is both extending the set of ''things'' that can be patented and shifting

the legal and judicial balance more and more in favor of patent holders.

- Because competition fosters productivity growth, antitrust and competition policies should be seen as key tools to foster innovation. This is of particular relevance for high-tech sectors, from software, to bioengineering, to medical products and pharmaceuticals.
- Free trade is a key part of competition policies, hence the role that the World Trade Organization, World Intellectual Property Organization, and Trade Related Aspects of Intellectual Property Rights agreement play should be redefined to move away from the current neomercantilist approach to free trade in goods and ideas. The aim here should be that of stopping the policy of exporting our intellectual policy laws to other countries while adopting a policy of exporting free trade and competition in innovation. This seems to us an urgent goal because, within a couple of decades, the ''balance of trade in ideas'' between the United States and Europe on one hand and Asia on the other may easily reverse. At that point, the temptation to engage in ''mercantilism of ideas'' may well affect the now-developing Asian countries, leading to a general increase in intellectual property protection worldwide.
- Cross-industry variation in the importance of patents suggests we may want to start tailoring patent length and breadth to different sectorial needs. Substantial empirical work needs to be done to implement this properly, even if there already exists a vast legal literature pointing in this direction.
- Reverse the burden of proof. Patents should be allowed only when monopoly power is justified by evidence about fixed costs and actual lack of appropriability. The operational model should be that of ''regulated utilities'': patents should be awarded only when strictly needed on economic grounds.
- Use prizes and competition to nurture innovation. An interesting approach is to change the role that the National Science Foundation and the National Institutes of Health play in fostering innovation. The basic goal, in this case, is that of reversing the principle according to which federally financed investigation can lead to private patents. As a first step, we would advocate going back to the old rule according to which the

results of federally subsidized research cannot lead to the creation of new private monopolies but should be available to all market participants. This goal is particularly important in the pharmaceutical industry.

- With regard to the pharmaceutical industry, we advocate reforming pharmaceutical regulation to either treat Stage II and III clinical trials as public goods (to be financed by NIH on a competitive basis) or by allowing the commercialization (at regulated prices equal to the economic costs) of drugs that satisfy the Food and Drug Administration requirements for safety even if they do not yet satisfy the current, over-demanding requisites for proving efficacy. It is ensuring the efficacy—not the safety—of drugs that is most expensive, time-consuming, and difficult. All the usual mechanisms of ensuring the safety of drugs would remain firmly in place. While pharmaceutical companies would be requested to sell new drugs at ''economic cost'' until efficacy is proved, they could start selling at market prices after that. In this way, companies would face strong incentives to conduct or fund appropriate efficacy studies where they deem the potential market for such drugs to be large enough to bear the additional costs. At the same time, this ''progressive'' approval system would give cures for rare diseases the fighting chance they currently do not have. This solution would substantially reduce the risks and cost of developing new drugs.
- If this progressive approval approach works for rare diseases, there is no reason it should not be adopted across the board. The current system favors a small number of blockbuster drugs that can be sold to millions of patients. The coming revolution in medicine will rely on carefully targeting hundreds or even thousands of drugs to the correct patients. But lawmakers must first usher in a new system that makes developing these precision treatments possible. The regulation reform we are suggesting here would be a first important step to achieve such a goal.
- Finally, software patents are a particularly egregious and bad form of intellectual property for a sector where we also have very detailed micro evidence about the role of patents in (not) promoting innovation (see, e.g., Bessen and Meurer

2008 and references therein). The same arguments are likely to apply to bio-engineering and genetic research at large. The goal of policy, in these cases, should be just that of slowly but surely decreasing the strength of intellectual property interventions.

7. APPENDIX

Figures A1 and A2 reproduce the above Figures 1–2 and 3–4, respectively, with each industry's NAICS-4 number appended to its data point. This should allow the interested reader to make up his mind about the extent to which, sector by sector, patent citations counts are, or are not, a good proxy for those factors that, in practice, do increase productivity. We remind readers that these are averages over the 20-year period 1987–2007.

Figures A3 and A4 are from the same dataset. They have labor and TFP growth on the vertical axes, respectively, and citations growth on the horizontal axes. There is no correlation in either figure.

Figures A5–A8 should allow for a better understanding of the findings reported in Section 5. A5 and A6 show how the inverted-U relationship and the positive relationship fit the actual observations. Both curves are, overall, quite flat. A7 and A8 replicate those displayed in Section 5. They are a ''zoom'' of the fitted line reported above. The graphs reveal the ''fragility'' of the empirical analysis based on these specifications and underline the need for additional research.

Figure A1
Patent Citations and Labor Productivity Growth
(Right panel excludes the top 5 percent of performers)

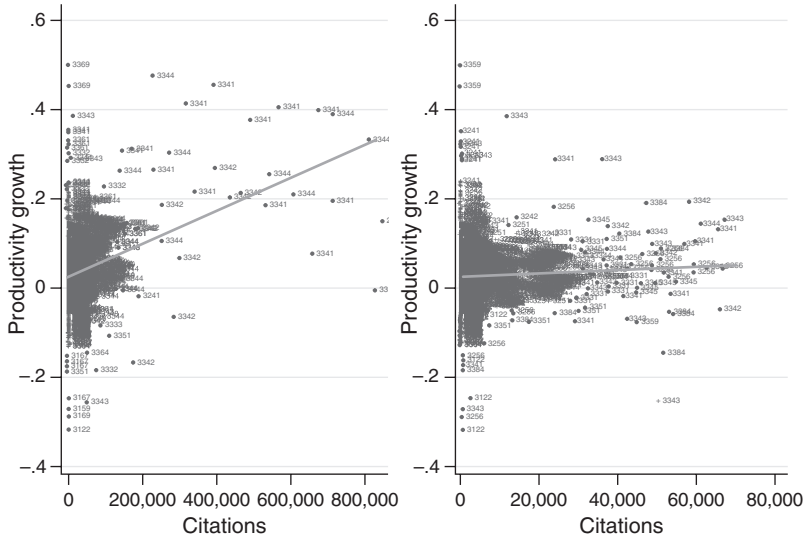


Figure A2
Average Citations and Average Labor Productivity Growth
(Right panel excludes the top 5 percent of performers)

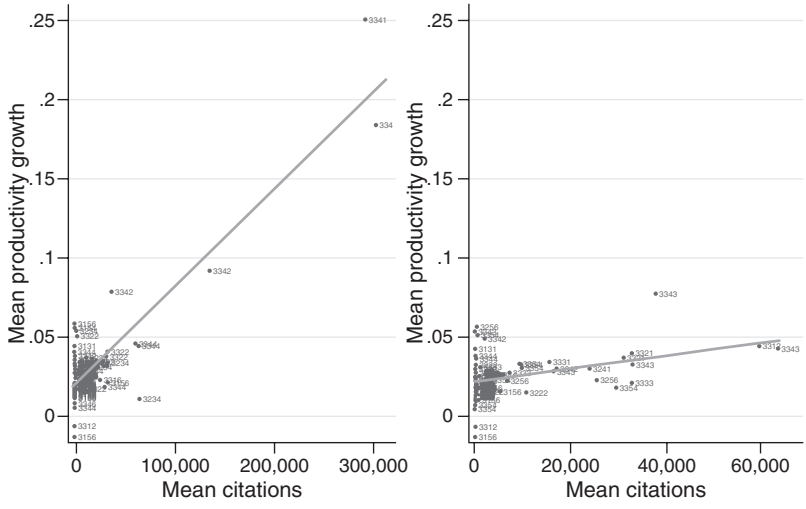


Figure A3
Citation Growth and TFP Growth

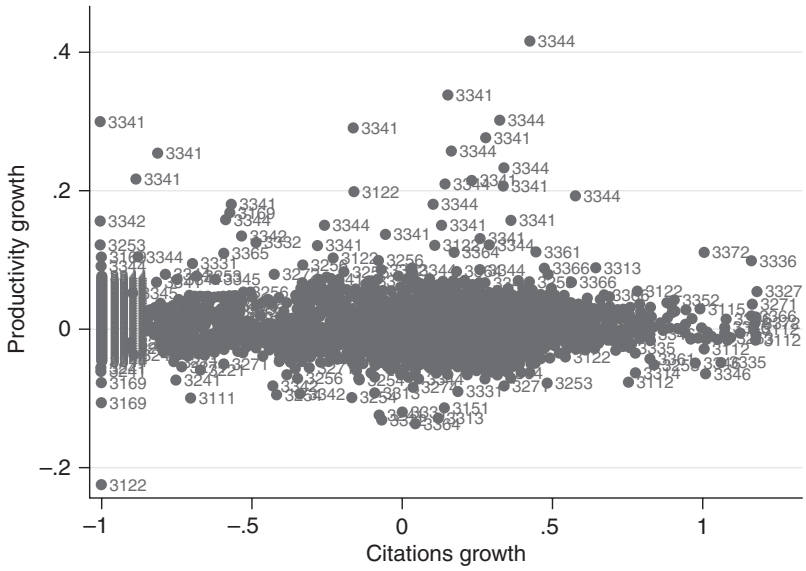


Figure A4
Citation Growth and Labor Productivity Growth

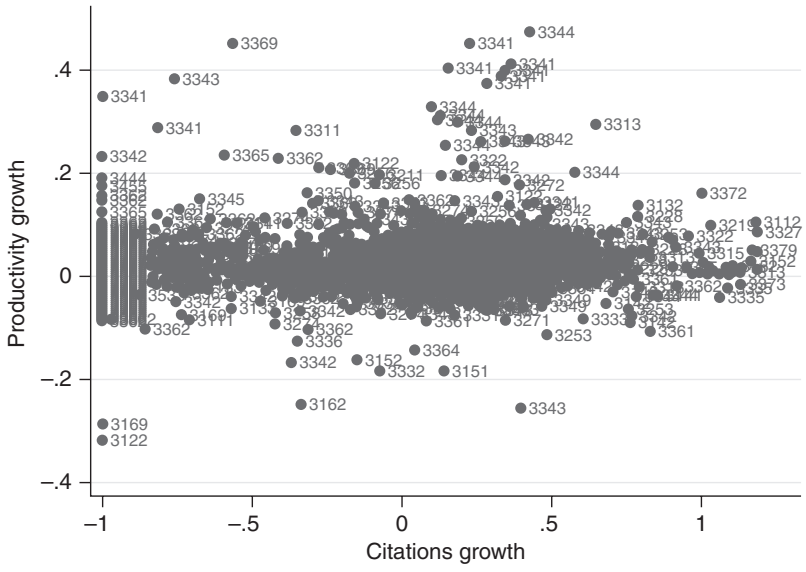


Figure A5
Relationship between Competition and Labor Productivity Growth

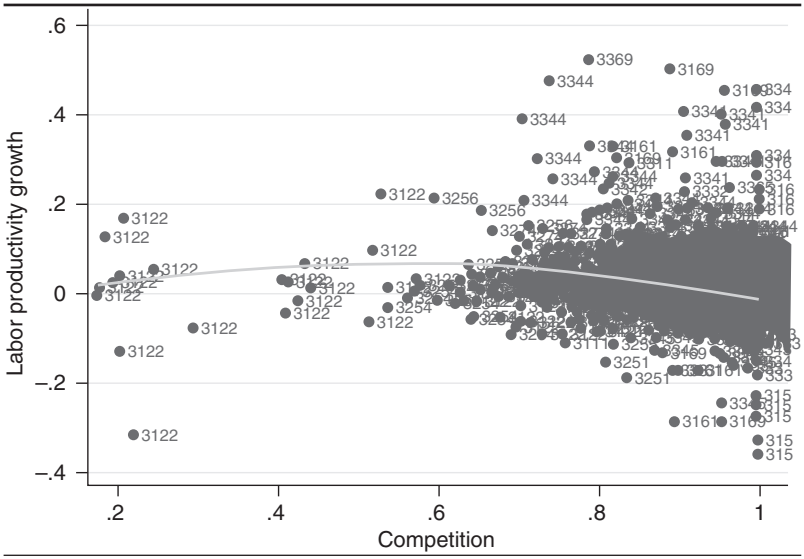


Figure A6
Relationship between Lagged Competition and Labor Productivity Growth

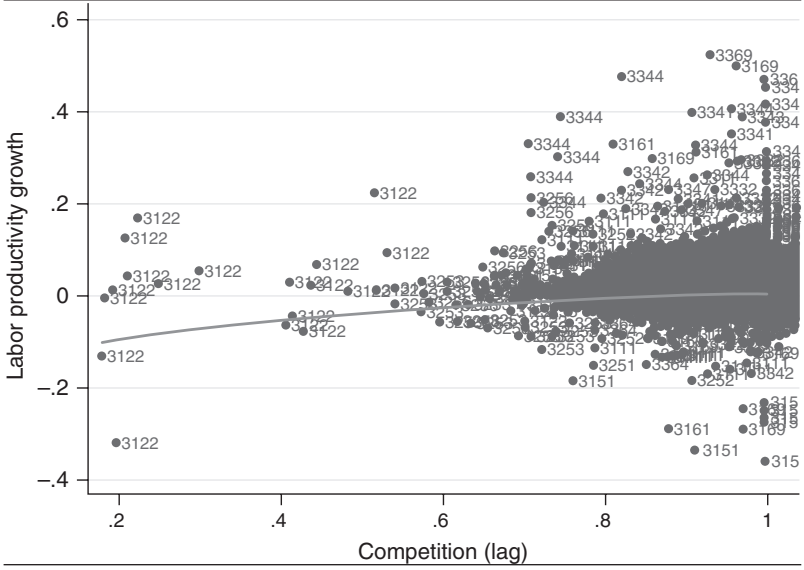


Figure A7
Labor Productivity Growth and Competition

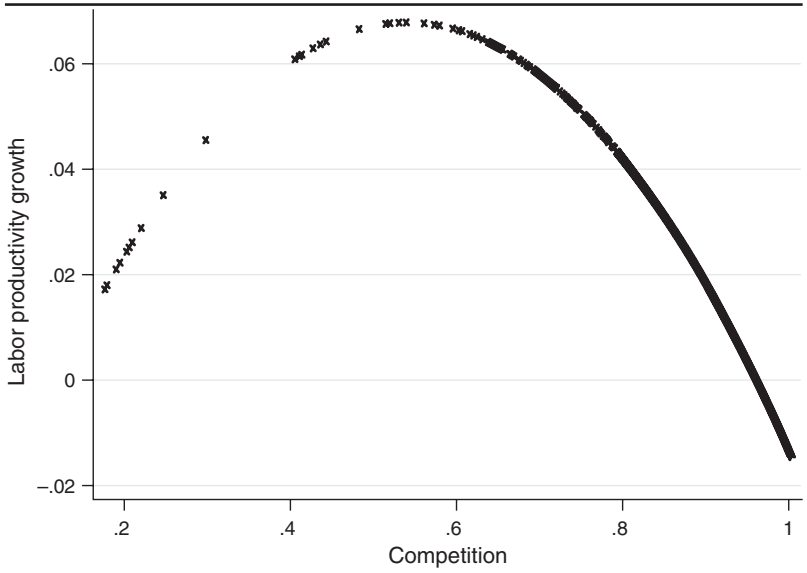
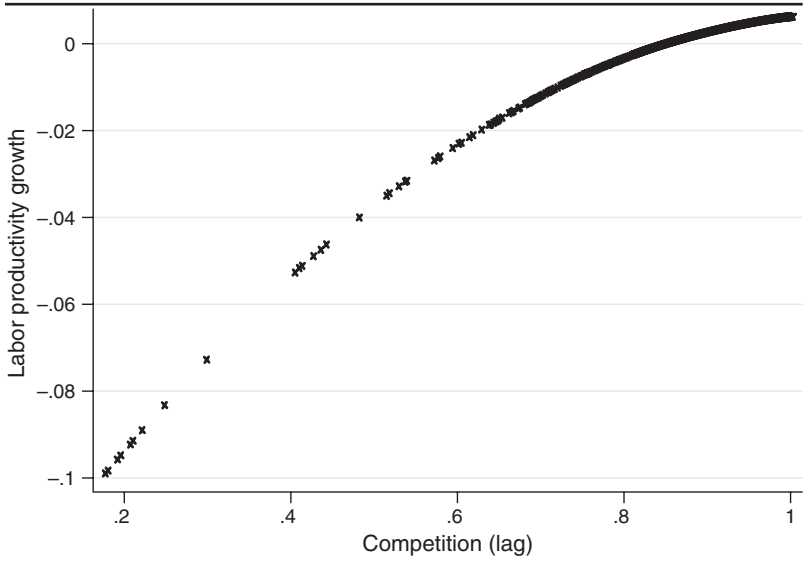


Figure A8
Labor Productivity Growth and Lagged Competition



REFERENCES

Adams, J., and A. Jaffe. 1996. ''Bounding the Effects of R & D: An Investigation Using Matched Establishment–Firm Data.'' *RAND Journal of Economics* 27 (4): 700–21.

Aghion P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt. 2005. ''Competition and Innovation: An Inverted U Relationship.'' *Quarterly Journal of Economics* 120: 701–28.

Aghion, P., and T. Griffith. 2005. *Competition and Growth: Reconciling Theory and Evidence*. Cambridge, MA: MIT Press.

Aghion, P., and P. Howitt. 1992. ''A Model of Growth through Creative Destruction.'' *Econometrica* 60 (2): 323–51.

Arrow, K. 1962. ''Economic Welfare and the Allocation of Resources for Invention.'' In *The Rate and Direction of Innovative Activity*, ed. R. Nelson. Princeton, NJ: Princeton University Press.

Bartelsman, E. J., and W. Gray. 1996. ''The NBER Manufacturing Productivity Database.'' National Bureau of Economic Research Technical Working Paper No. 205.

Baumol, W. J. 2010. *The Microtheory of Innovative Entrepreneurship*. Princeton, NJ: Princeton University Press.

Baumol, W. J., R. E. Litan, and C. J. Schramm. 2007. *Good Capitalism, Bad Capitalism, and the Economics of Growth and Prosperity*. New Haven, CT: Yale University Press.

Bessen, J., and R. M. Hunt. 2003. ''An Empirical Look at Software Patents.'' Mimeo. Abridged version in *Journal of Economics and Management Strategy* 16: 157–89.

Bessen, J., and M. J. Meurer. 2008. ''Of Patents and Property.'' *Regulation* 32 (4): 18–26.

Blundell, R., R. Griffith, and J. Van Reenen. 1999. ''Market Share, Market Value and Innovation in a Panel of British Manufacturing Firms.'' *Review of Economic Studies* 66: 529–54.

Boldrin, M., and D. K. Levine. 1997. ''Competitive Equilibrium Growth.'' Mimeo. Universidad Carlos III and University of California, Los Angeles. October.

———. 2004. ''Rent Seeking and Innovation.'' *Journal of Monetary Economics* 51: 127–60.

———. 2005. ''The Economics of Ideas and Intellectual Property.'' *Proceedings of the National Academy of Sciences* 102: 1252–6.

———. 2006. ''Quality Ladder, Competition and Endogenous Growth.'' Mimeo. Washington University in St Louis.

———. 2008a. *Against Intellectual Monopoly*. Cambridge, UK: Cambridge University Press.

———. 2008b. ''Perfectly Competitive Innovation.'' *Journal of Monetary Economics* 55: 435–53.

———. 2009a. ''IP and Market Size.'' *International Economic Review* 50: 855–81.

———. 2009b. ''A Model of Discovery.'' *American Economic Review: Papers and Proceedings* 99: 337–42.

Carlin, W., M. Schaffer, and P. Seabright. 2004. ''A Minimum of Rivalry: Evidence from Transition Economies on the Importance of Competition for Innovation and Growth.'' *Berkeley Electronic Journal of Economic Analysis and Policy: Contributions* 3: 1–43.

Castiglionesi, F., and C. Ornaghi. 2011. ''On the Determinants of TFP Growth: Evidence from Spanish Manufacturing Firms.'' Mimeo. Tilburg University and University of Southampton, 2008. Forthcoming in *Macroeconomic Dynamics*.

Cohen, W. M., R. R. Nelson, and J. P. Walsh. 2000. ''Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not).'' NBER Working Paper No. 7552. February.

Correa, J. A. 2010. ''Innovation and Competition Relationship in a Memory Process.'' Ph.D. dissertation. UK: University of Southampton. Forthcoming in *Journal of Applied Econometrics* under the title ''Innovation and Competition: An Unstable Relationship.''

Correa, J., and Ornaghi, C. 2011. ''Competition and Innovation: New Evidence from U.S. Patent and Productivity Data.'' Unpublished manuscript. UK: University of Southampton.

Dean, E. R., and M. J. Harper. 1998. ''The BLS Productivity Measurement Program.'' U.S. Bureau of Labor Statistics.

Gilbert, R., and C. Shapiro. 1990. ''Optimal Patent Length and Breadth.'' *Rand Journal of Economics* 21: 106–12.

Griliches, Z. 1957. ''Hybrid Corn: An Exploration of the Economics of Technological Change.'' Ph.D. dissertation. Chicago: University of Chicago.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg. 2001. ''The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools.'' NBER Working Paper No. 8498.

Hashmi, A. R. 2011. ''Competition and Innovation: The Inverted-U Relationship Revisited.'' Mimeo. National University of Singapore. February.

Hellwig, M., and A. Irmen. 2001. ''Endogenous Technical Change in a Competitive Economy.'' *Journal of Economic Theory* 101: 1–39.

Jaffe, A. B., and J. Lerner. 2004. *Innovation and Its Discontents*. Princeton, NJ: Princeton University Press.

Klette, T. J., and S. Kortum. 2004. ''Innovating Firms and Aggregate Innovation.'' *Journal of Political Economy* 112: 986–1018.

Kortum, S., and J. Lerner. 1998. ''Stronger Protection or Technological Revolution: What Is Behind the Recent Surge in Patenting?'' *Carnegie-Rochester Conference Series on Public Policy* 48: 247–304.

Legros, P. 2005. ''Art and Internet: Blessing the Curse?'' Mimeo. Belgium: ECARES and the Université Libre de Bruxelles.

Lerner, J. 2009. ''The Empirical Impact of Intellectual Property Rights on Innovation: Puzzles and Clues.'' *American Economic Review: Papers and Proceedings* 99 (2): 343–8.

Levin, R. C., A. K. Klevorick, R. R. Nelson, and S. G. Winter. 1987. ''Appropriating the Returns from Industrial Research and Development.'' *Brookings Papers on Economic Activity* 3: 783–820.

Liebowitz, S. J. 1985. ''Copying and Indirect Appropriability: Photocopying of Journals.'' *Journal of Political Economy* 93: 945–57.

Moser, P. 2005. ''How Do Patent Laws Influence Innovation?'' *American Economic Review* 95 (4): 1214–36.

Nickell, S. J. 1996. ''Competition and Corporate Performance.'' *Journal of Political Economy* 104: 724–46.

Okada, Y. 2005. ''Competition and Productivity in Japanese Manufacturing Industries.'' NBER Working Paper No. 11540.

Pakes, A. S. 1986. ''Patents as Options: Some Estimates of the Value of Holding European Patent Stocks.'' *Econometrica* 54: 755–84.

Piazza, R. 2010. ''Leadership Contestability, Monopolistic Rents and Growth.'' Mimeo. International Monetary Fund. December.

Quah, D. 2002. ''24/7 Competitive Innovation.'' Mimeo. London School of Economics.

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

Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. New York: Harper and Brothers.

Scotchmer, S. 1991. ''Standing on the Shoulders of Giants: Cumulative Research and the Patent Law.'' *Journal of Economic Perspectives* 5: 29–41.

Sheshinski, E., R. J. Strom, and W. J. Baumol, eds. 2007. *Entrepreneurship, Innovation, and the Growth Mechanism of the Free-Enterprise Economies*. Princeton, NJ: Princeton University Press.

Solow, R. M. 1956. ''A Contribution to the Theory of Economic Growth.'' *Quarterly Journal of Economics* 70 (1): 65–94.

———. 1957. ''Technical Change and the Aggregate Production Function.'' *Review of Economics and Statistics* 3 (3): 312–20.

Stigler, G. J. 1956. ''Industrial Organization and Economic Progress.'' In *The State of the Social Science*, ed. L. D. White, pp. 269–82. Chicago: University of Chicago Press.

Van Reenen, J. 2010. ''Does Competition Raise Productivity through Improving Management Quality?'' CEP Discussion Paper No. 1036. December.

Comment

Samuel Kortum

As someone who got the ''received wisdom'' about the economics of innovation as a part of his schooling, I approached the ''economic heresy'' of Michele Boldrin and David Levine with some skepticism. I was also intrigued. This paper, with coauthors Juan Correa and Carmine Ornaghi, (hereafter BCLO) discusses those theoretical arguments, presents empirical evidence, and draws out the policy implications. While I don't think they make a convincing case for dismantling the patent system, I do find parts of the argument compelling.

COMPETING THEORIES OF TECHNOLOGICAL CHANGE

In the introduction to his 1990 article in the *Journal of Political Economy*, Paul Romer articulates what I think of as the received wisdom. It starts with a careful account of the production function. Economic output is produced with factors of production, unskilled labor, skilled labor, machinery, etc. Technology is the recipe book for how to combine these ingredients (the factors of production) to produce goods that we like. In principle, you could always make twice as much of any good you're currently making by using twice the quantity of each ingredient. You don't need a new recipe to do that. Similarly, with the same recipe, anyone else could produce more of what you are currently making. A recipe is said to be *nonrival*, as any number of people can use it simultaneously. That's why, in a competitive equilibrium, all the revenue from production is paid to the factors of production, with nothing left to pay for the recipes.

Numerous studies have shown that the growth of factors of production leaves much to be explained in accounting for economic growth. It seems that the recipes must be getting better, and casual

Samuel Kortum is professor of economics at the University of Chicago.

observation, for example using a new smart phone, supports that view. Since technological change appears to be a crucial source of rising living standards, yet a competitive market doesn't provide incentives for inventive activity, we are fortunate to have institutions that fill the gap. The patent system is one of them, allowing the owner of a patented recipe to restrict its use by others. The patent holder gets some monopoly power. The expectation of monopoly profit is an incentive to invent new recipes.

What could be wrong with this received wisdom? The starting point of the ''economic heresy'' is the observation that it can be difficult to separate the recipe from the chef. New techniques are often embodied in people or machines. It takes time and effort to train people or build machines. While a component of the technique may be nonrival in principle, it may not be easy to replicate in practice. The good techniques end up being used by some and not by others. A consequence is that there can be rents to the inventor of a new technique, even in the absence of patent protection.

This starting point is quite compelling and is, in fact, a common view of technology. A recent paper by Robert Lucas (2008), while not modeling any payments, envisions the diffusion of techniques in this way. A technique only spreads from one individual to another via a chance meeting and an ensuing transfer of information (from one brain to the other). Note that techniques are nonrival goods in this economy. But, because they are embodied in individuals, the good techniques aren't instantaneously replicated. In this economy, one could imagine that an individual with knowledge of a good technique might put a price on the service of teaching it to someone else.

These arguments become less abstract with a specific example in mind. Consider a new technique for folding metal into various shapes without weakening it. The idea behind it is nonrival. A good engineer could look at it, get the idea, and decide how it could be put to use in his manufacturing process. Without patent protection, the inventor of this technique would seem to have no means of extracting any return from it. But actually adopting the new process of folding metal requires educating and training engineers. The inventor knows how to make it work, but the potential users do not. In principle, even in the absence of a patent, the inventor could profit from manufacturers who hire him as a consultant to get the process up and running.

THE CHALLENGE FOR EMPIRICAL WORK

Faced with these two scenarios and their starkly different policy implications for the patent system, how can we use data to resolve the issue? In principle, we'd like to build an economic model that encompasses these two scenarios. The model would also incorporate features of the current patent system, as well as other mechanisms inventors use to appropriate returns, such as trade secrecy. We would then confront this model with microeconomic data of various sorts in order to estimate the relevant parameters. Finally, using these estimates of the parameters, we would perform various counterfactual simulations of the model to examine the outcomes for economic growth and economic welfare under various possible reforms of the patent system. We could then make policy recommendations with some confidence, based on such a systematic quantitative analysis.

Back to reality: While a worthy goal, we have a long way to go before we can carry out such a research agenda in a satisfactory manner. In the meantime, we're left looking at correlations between measures of productivity, patents, firm size, R&D, and competition (measured as the inverse of profitability). We have to be pretty cautious about what we can conclude from such correlations about the worthiness of either scenario. The implications for patent policy are even more tenuous.

In some cases, the mapping from the empirical findings to the different models of innovation is not very clear. The received wisdom is not that large firms are the most inventive. As formalized in economic models, all the innovation is done by new entrants. Similarly, since the models associated with the received wisdom typically assume free entry into research activity, it is not clear that the industries with more research activity will appear less competitive (more profitable). In these models, profits get used up paying for research.

The paper is quite right, on the other hand, to point to the weak connection between productivity growth and patenting. It is a genuine puzzle on many levels that the industries with more inventive activity do not experience much more rapid productivity growth. Furthermore, this puzzle persists if you replace patents with R&D spending. In his presidential address to the American Economic Association, Zvi Griliches (1994) dwelled on this issue. Figure 2 in his paper gives roughly the same message as Figures 1–4 in BCLO.

Why are these relationships so weak? Griliches pointed to the difficulty of measuring productivity. There are also conceptual issues. An industry experiencing technological improvements will typically expand into less productive activities, so that the expansion itself may dampen measured productivity growth. This point is carefully worked out in a recent paper by Costinot, Donaldson, and Komunjer (2011).

The empirical work in BCLO is a useful reminder that we ought to be humble given how little we know in this field. The authors suggest that the weakness of the evidence in favor of the received wisdom means that there is little justification for our current patent system. Given our limited knowledge, I think the appropriate response is to be very cautious in recommending dramatic changes in that institution.

OTHER PROMISING APPROACHES

My critique of the empirical work in the paper may appear unconstructive. What *is* a good way of determining which view of innovation and competition is more reasonable? Short of the rather heroic approach I laid out in the beginning, how else might we use data to answer the important question of how we can best reform the patent system? I see at least three useful lines of attack:

- examining the key premises of the economic heresy view,
- exploiting macro evidence on patent-system reform, and
- conducting case studies on how actual inventors and users of new technology interact with the patent system.

I consider each of these in turn.

Key Premises

The economic heresy view is built on the premise that technology, while nonrival, is not easy to replicate. That premise appears very solid but ripe for empirical exploration. At the macro level, it relates to why many countries are not exploiting state-of-the-art technology. At the micro level, it relates to why individual technologies don't spread faster. For example, even Wal-Mart took a long time to spread across the United States, as graphically illustrated in Holmes (2011). But what is it that holds things back? Is it that technology is inherently costly to replicate, generating rents to inventors? Or is it that

users, either consumers or producers, are resistant to new technology, thus reducing the discounted flow of profits to inventors? A better understanding of this phenomenon of slow replication seems central to the issues of this paper and economics more generally.

Macro Evidence

Many countries have recently strengthened patent enforcement. Sometimes these changes have been driven by internal pressure from their own inventors and sometimes by external pressure from developing countries who want their intellectual property protected. In either case, these changes provide a laboratory within which to study the consequences of changing the patent system. While Frederic Scherer is mentioned in BCLO as buying into the received wisdom, I have seen him point to evidence that when Italy adopted stronger pharmaceutical patents, its pharmaceutical industry suffered.

Case Studies

Finally, in this area, case studies can be extremely valuable. In preparing for this discussion, I had a long conversation with my cousin Max W. Durney, the inventor of the technology for folding metal that I mentioned above. On the basis of this invention, which is protected with numerous patents, he founded the company Industrial Origami. He argues that his extensive use of patents was crucial in attracting financing. Up-front financing was itself necessary to invest in developing the technology and in marketing it to manufacturers who were initially resistant to adopting a new process. The current patent system, while not without problems, can enable creative individuals to make a career inventing new technologies.

REFERENCES

Costinot, A., D. Donaldson, and I. Komunjer. 2011. ''What Goods Do Countries Trade? A Quantitative Exploration of Ricardo's Ideas.'' *Review of Economic Studies*. Forthcoming.

Griliches, Z. 1994. ''Productivity, R&D, and the Data Constraint.'' *American Economic Review* 84: 1–23.

Holmes, T. J. 2011. ''The Diffusion of Wal-Mart and Economies of Density.'' *Econometrica* 79: 253–302.

Lucas, R. E. 2008. ''Ideas and Growth.'' *Economica* 76: 1–19.

Romer, P. M. 1990. ''Endogenous Technical Change.'' *Journal of Political Economy* 98 (5): 71–102.

Comment

Andrew Atkeson

Michele Boldrin and David Levine, in previous work and now in this paper with Juan Correa and Carmine Ornaghi, have issued a powerful challenge to the idea that patent protection plays an essential role in fostering technological innovation and improved welfare for consumers. In fact, these authors go further at times and argue that patent protection may in fact hinder rather than help technological innovation.

The authors start this paper with the premise that it is a straightforward task to construct economic models such that patent protection improves welfare when the models' parameters are in one region and reduces welfare when the models' parameters are in another region. They argue that, as a result, the question of whether patent policy has a beneficial or harmful effect on innovation and consumer welfare must be resolved on the basis of empirical evidence rather than theory. In this paper, the authors aim to shed some light on the empirical link between patents, competition, and technological progress.

I anticipate that in his discussion, commenter Sam Kortum will assess the contribution of this paper to the broader empirical literature on the relationship between patenting, innovation, and technological progress. Given the dictates of comparative advantage, I will specialize in my discussion on the theoretical basis for arguing that patent protection may be harmful rather than helpful to innovation and welfare. I agree with the basic premise of the paper's authors that the idea that patents are essential to support innovation in equilibrium is deeply ingrained in many economists' and policymakers' minds. While many practitioners would argue that there are important flaws in patent policy as currently implemented in the

Andrew Atkeson is the Stanley M. Zimmerman Professor of Economics and Finance at the University of California, Los Angeles.

United States (see, e.g., Shapiro 2007), the authors of this paper look to go further and argue that the property rights conferred by patent protection, even if executed well, may be harmful to innovation and welfare. This view is the ''economic heresy'' the authors refer to in their introduction. Given the strength of the authors' theoretical arguments, I see my best chance of adding value to this debate is to sketch a simple theoretical model of innovation with and without patent protection to see how straightforward it is to have the welfare implications of patents go either way.

My goal in sketching this model is to capture some of the ideas the authors discuss in their paper about the difference between *innovation* and *imitation*, and the role of a *first-mover advantage* for innovators of some kind in supporting innovation without patent protection in a way that complements the previous theoretical work by these authors. The authors are certainly correct that the theoretical work assessing the impact of patent policy in general equilibrium in the ''New Growth Theory'' literature is too special and narrow and that much more theoretical and empirical work needs to be done. What follows is a small contribution in that direction.

The central idea I would like to illustrate with this model is that, in an economy with imperfect competition, incumbent firms built on a successful innovation already have strategies available to preempt entry by imitating firms even in the absence of patent protection. As a theoretical matter, it is not at all clear that welfare is improved if we enhance the strategic position of these incumbents by granting them patent protection as an additional tool to deter entry of competing firms.

The model I use to illustrate this point starts with a demand structure in which we can say that some products are closer substitutes than others, so that we can think about oligopoly in a particular market nested in a larger general equilibrium economy as a whole. To do so, posit a nested demand system with a continuum of potential *sectors* and many (but countable) potential *goods* within each sector. A firm *innovates* by being the first to introduce a good into a particular sector. To innovate, a firm must pay a high fixed cost to introduce this first good in the sector. A firm *imitates* by being the second or later to introduce a good into a particular sector. An imitator benefits from following an innovator into a sector in that the imitator pays a lower fixed cost to introduce this additional good in the sector.

To be more specific, let aggregate consumption be a constant elasticity of substitution aggregate (CES) across sectors with

$$C = \left[\int_0^N y_j^{1-\frac{1}{\eta}} dj \right]^{\frac{\eta}{\eta-1}}$$

Here y_j is the output of sector j and N is the measure of sectors that have active firms producing goods in those sectors. The parameter η is the elasticity of substitution across sector inputs in producing final consumption. Standard arguments give that the induced demand by final consumption producers for the output of sector j is given by the CES (inverse) demand function

$$\frac{P_j}{P} = \left(\frac{y_j}{C} \right)^{-\frac{1}{\eta}}$$

where P_j is the price index for output of sector j and P is the price index for final consumption C.

Output in sector j is a second CES aggregate across the output of the K_j firms active in that sector

$$y_j = \left[\sum_{k=1}^{K_j} q_{jk}^{1-\frac{1}{\rho}} \right]^{\frac{\rho}{\rho-1}}$$

where q_{jk} is the output of the kth firm in sector j and ρ is the elasticity of substitution across goods within the sector. To capture the idea that goods within a sector are closer substitutes than goods in different sectors, assume that $\rho > \eta$. With this ranking of elasticities, we have the inverse demand curve for a particular good given by

$$\frac{P_{jk}}{P_j} = \left(\frac{q_{jk}}{y_j} \right)^{-\frac{1}{\rho}}$$

where P_{jk} is the price of the kth good in sector j and the price index for the sector P_k is constructed in the standard manner.

Now consider some simple economics of firms' decisions to innovate and imitate when facing this demand structure. The returns to either of these decisions will depend on the number of firms that introduce goods in a particular sector. Assume that when it comes

time to produce, firms active in a sector engage in price (Bertrand) competition. With a finite number of goods in a sector, the perceived elasticity of residual demand for each firm in a sector—and hence the markup over marginal cost charged by that firm—depends on that firm's share of the market in that sector.[1] If there are no imitators that follow an innovating firm into a particular sector, then that innovating firm enjoys a high profit from its innovation not only because it commands the entire market but also because it charges a high markup corresponding to the low elasticity of substitution η across sectors. If many imitators follow an innovating firm into a sector, each firm has a small market share in the sector, and markups are small, corresponding to the high perceived elasticity of residual demand for each firm as determined by ρ. A firm that innovates pays an entry cost c_1, while a firm that imitates pays an entry costs c_2, with $c_1 \geq c_2$.

Are patents *necessary* to support innovation in this environment? The answer to this question depends on parameters.

If the cost of imitation is zero and there are potential profits for an imitator (i.e., $\rho < \infty$), then the answer is *yes*. We cannot have an equilibrium in which an incumbent innovator earns the positive profits necessary to recoup the innovation cost $c_1 > 0$ without attracting entry from imitators. This particular parameter configuration corresponds to the standard assumption that imitation is truly costless, and in this special case, it is likely impossible to support innovation in equilibrium without some regulation of imitation such as patent protection. As the authors of this paper correctly point out, this special case of zero imitation costs likely lies at the heart of most thinking by policymakers about intellectual property.

How general is this argument for patent protection? Not very. Consider, for example, the case in which goods within a sector are perfect substitutes ($\rho = \infty$) and imitation costs are positive ($c_2 > 0$). In this case, with goods within a sector being perfect substitutes,

[1] This dependence of elasticities on market shares arises as a result of the assumption that there are only a finite number of goods K_j being produced in any given sector. As a result, firms must take into account that the price that they choose affects the sectoral price index P_j and thus overall demand for the sectoral output y_j in addition to the share of spending in the sector on their specific good. In contrast, with a continuum of sectors, individual firms take the overall consumption price index P and level of consumption C as given.

entry by a second firm as an imitator in a sector eliminates all profits for both the innovator and the imitator by driving prices down to marginal cost through Bertrand competition. Here, there are no returns to imitation. All firms prefer to innovate in a new sector and enjoy monopoly profits in that sector rather than imitate. In this extreme case, innovators here are protected from imitation as long as there is some cost to imitation.[2] Similar logic holds that patent protection is also unnecessary if imitation costs are equal to innovation costs $c_2 = c_1$ regardless of the elasticity ρ between goods in a sector. In these cases, adding patent protection to the economy has no impact on equilibrium innovation and welfare.

In these special cases, the impact of patent protection on innovation and welfare is easy to see. What happens, however, for more general values of imitation costs relative to innovation costs ($0 < c_2 < c_1$) and elasticities across sectors and goods ($\eta < \rho < \infty$)? Are the impacts of patent policy on innovation, imitation, and consumer welfare purely a problem of parameter values in this more general case? Or are there some more general lessons we might draw even in this simple environment?

It is impossible to give a fully worked-out answer to these questions without specifying the details of the model of entry and the post-entry competition between innovators and imitators, but I conjecture that older ideas about how innovators might preempt entry by imitators in the absence of patent protection may well be useful for drawing more general conclusions about the impact of patents on welfare. The key economic idea to note is that, even in the absence of patent protection, an innovating firm in a particular sector has a *first-mover advantage* in that it can use *product proliferation* through imitation of its own products as a strategy to simultaneously maintain monopoly profits with high markups in equilibrium and to effectively deter entry by imitating firms that might seek to compete with it in the sector.[3]

[2] This idea that firms in a market with imperfect competition will choose to avoid competing with each other in equilibrium is a long-standing one in economics. See, for example, Prescott and Visscher (1977) and Vogel (2008).

[3] See Ellison and Ellison (2011) for recent work documenting how pharmaceutical companies use product proliferation as a strategic device to deter entry by generics as the expiration of their patent protection approaches.

The basic idea is as follows: A single firm that has paid both the innovation cost c_1 for the first good in a sector and the imitation cost c_2 for additional goods in that same sector can charge a high markup for all of those goods by setting high prices (and low quantities) in a coordinated fashion across its product line if there are no other imitators in the sector and, at the same time, credibly convince any potential imitators that the competitive outcome that would emerge if that firm were actually to enter the sector would be very unfavorable to that firm because it would have such a small market share. Essentially, an innovator has the strategic incentive to build on his own innovation through imitation to fill up the sector with his own products. In contrast, if a new firm considers entry into the sector through imitation, it faces the prospect of competition with the original innovator as an incumbent in a powerful position.

What can we conjecture about equilibrium in an economy of this kind in which an innovator can obtain patent protection for his innovation that is broad enough in scope to prevent imitation by other firms looking to introduce goods into the sector? Here patents may well serve the imitation preemption role for innovators that product proliferation serves in the economy without patents. What are the implications of this comparison between an economy without patents and with patents for product creation and welfare? With patents, innovators will clearly charge high prices and enjoy monopoly profits, just as in the equilibrium without patents, so in this respect the two economies might look similar. With patents, however, innovating firms will not feel the competitive pressure from the threat of imitation by other firms to introduce additional products into the sector that they have innovated in, and hence consumers will not benefit from the additional (and relatively cheap) product creation that occurs in the economy without patents when innovating firms fill up their sector with products to deter entry. Hence, by introducing patent protection for innovators, policymakers will in fact have reduced one incentive for product creation without necessarily increasing other incentives to innovate sufficiently to deliver a compensating increase of entry into new sectors.

The overall impact of patent protection on welfare would depend on the general equilibrium reallocation of products across sectors and would require a much fuller analysis of a fully specified model, but the possibilities that patent protection may be a drag on the

creation of new products should be clear from the arguments laid out here. What is even clearer is that the economic thinking on this topic has not been well fleshed out in the academic literature. The authors of this paper have set a bold agenda that should provoke considerable further development of our thinking on this important topic.

REFERENCES

Ellison, G., and S. F. Ellison. 2011. ''Strategic Entry Deterrence and the Behavior of Pharmaceutical Incumbents Prior to Patent Expiration.'' *American Economic Journal: Microeconomics* 3 (1): 1–36.

Prescott, E., and M. Visscher. 1977. ''Sequential Location among Firms with Perfect Foresight.'' *Bell Journal of Economics* 8 (2): 378–93.

Shapiro, C. 2007. ''Patent Reform: Aligning Reward and Contribution.'' *Innovation Policy and the Economy* 8: 111–56.

Vogel, J. 2008. ''Spatial Competition with Heterogeneous Firms.'' *Journal of Political Economy* 116 (3): 423–66.

Labor Market Dysfunction during the Great Recession

Kyle F. Herkenhoff
Lee E. Ohanian

ABSTRACT

This paper documents the abnormally slow recovery in the labor market during the Great Recession and analyzes how mortgage modification policies contributed to delayed recovery. By making modifications means-tested by reducing mortgage payments based on a borrower's current income, these programs change the incentive for households to relocate from a relatively poor labor market to a better labor market. We find that modifications raise the unemployment rate by about 0.5 percentage points and reduce output by about 1 percent, reflecting both lower employment and lower productivity, which is the result of individuals losing skills as unemployment duration is longer.

Kyle F. Herkenhoff is a doctoral student in economics at the University of California, Los Angeles. Lee E. Ohanian is professor of economics at UCLA.

Labor Market Dysfunction during the Great Recession

1. INTRODUCTION

The Great Recession, which began in December of 2007, differs considerably from most other significant U.S. economic declines, as the recovery—particularly recovery in the labor market—has been remarkably slow. In fact, the Great Recession and the Great Depression are the only severe U.S. downturns in which job loss persisted so long following respective business cycle troughs. This paper documents the very weak labor market recovery during the Great Recession and evaluates mortgage modification policies as one channel for understanding why high unemployment has continued for so long during the Great Recession. We study mortgage modifications because some economists have presented evidence linking housing market weakness to labor market weakness (Ohanian and Raffo 2011) and because mortgage modification programs are means-tested and thus change the incentives for home borrowers to relocate to labor markets with more favorable job prospects.

Means-tested mortgage modifications reduce the cost of staying in a home by reducing mortgage payments, with the payment reduction based on the household's current earnings. This includes cases in which the borrower's income is limited to unemployment benefits and the borrower's current debt-to-income ratio is well above standard levels, so that a modified mortgage payment can be much lower than current payments. Mulligan (2009 and 2010b), among others, has suggested that this policy may be significantly contributing to high unemployment by distorting incentives. In addition to concerns about incentives, modification programs are controversial because redefault rates, which are the percentage of defaults on the modified mortgage, are high, ranging between 30 percent and 50 percent within one year of modifying.

This paper evaluates the impact of mortgage modification programs on unemployment and other macroeconomic variables by constructing a very simple model that integrates a model of search unemployment along the lines of Ljungqvist and Sargent (1998, 2004) with a model of homeownership, including mortgages, mortgage modifications, and location choice. By reducing mortgage payments based on current income, mortgage modification changes the incentives to accept job offers and to relocate to labor markets with more favorable job prospects.

In the model, households are located in a particular local labor market (island) in which they receive stochastic job offers and their skills evolve over time, as in Ljungqvist and Sargent (1998, 2004). Households can accept the job offer and remain in their local labor market, reject the offer and receive unemployment benefits, or relocate to another labor market. If the household has a mortgage, they choose whether to continue with an existing mortgage, which preserves their current flow of housing services, whether to walk away from the mortgage and rent either in their current labor market or in an alternative labor market, or whether to seek a one-time modification of their mortgage that reduces mortgage payments. Relocating is costly but offers a job-finding probability that stochastically dominates the job-finding probability on a household's current island. While employed, households accumulate skill in expectation, and, while unemployed, households decumulate skill in expectation. By changing the cost to a borrower of maintaining an existing mortgage, modifications increase the incentives for a household to remain in their current location and forgo more favorable job prospects in another location.

Our model of the modification process is motivated by various modification programs that have been in place since 2007, in which modifications reduce mortgage payments to a debt-service-to-income ratio (DTI) that depends on current income. Thus, households with low income, including those whose income is limited to unemployment benefits, can receive substantial reductions in their payments that increase the opportunity cost of relocating in the model economy.

We first analyze the implications of modifications by examining steady states of two economies that are identical, except one has modifications. We next conduct an economic turbulence experiment

along the lines of Ljungqvist and Sargent (1998) to assess how modifications impact the economy during a major recession. This experiment consists of two shocks:

- the layoff rate in the model is doubled for one period, and
- for those who are laid off, their human capital stock is reduced by one level.

The main finding from this experiment is that the unemployment rate rises by about 0.5 percentage points and real GDP declines by about 1 percent for several years after the modification policies are in place. A number of empirical features in the model economy correspond to data, including the rate of modifications, the redefault rate on modifications, and mobility. We also estimate that a 10 percent reduction in payments reduces the chance of a person walking away from a home by 11.3 percent, which is comparable to Haughwout, Okah, and Tracy (2009). We are then able to estimate that mortgagors who have a job at the date they request a modification are 48 percent less likely to default again.

The paper is organized as follows: Section 2 compares the Great Recession to other U.S. downturns. Section 3 summarizes mortgage modification programs during the Great Recession with a focus on the labor market impact of these policies. Section 4 presents the model economy. Section 5 presents quantitative experiments that assess the impact of today's modification programs. Section 6 concludes.

2. THE GREAT RECESSION COMPARED TO OTHER U.S. ECONOMIC DECLINES

This section compares the Great Recession to other economic declines. Figures 1 and 2 compare the Great Recession labor market to other postwar recessions and highlight a number of patterns that contrast sharply with those in other downturns. Figure 1 shows employment following each recession and Figure 2 shows employment during the Great Recession compared to the average of all other postwar recessions. Even abstracting from the size of employment loss during the recession, these two figures clearly suggest labor market dysfunction in which employment during the Great Recession is not recovering at a normal rate.

Figure 1
Change in U.S. Employment: Recoveries

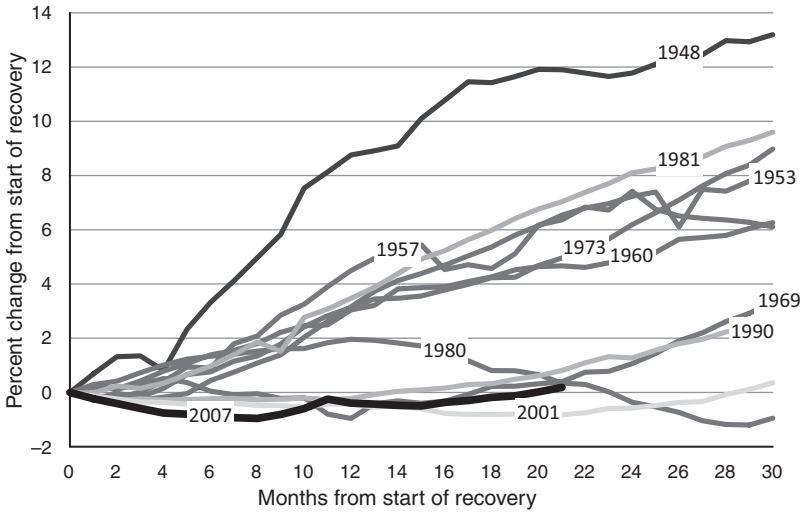
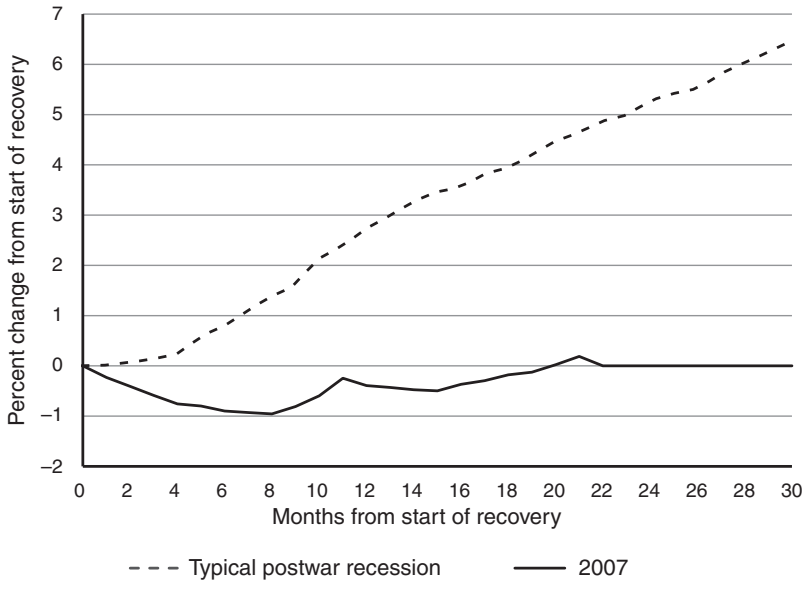


Figure 3 combines information from both the recession and the recovery by showing employment change in all post–World War II recessions 36 months after the start of the recession. There are only two postwar recessions that do not feature employment recovery three years after the start of a recession—the 2000–2001 recession and the Great Recession. Note, however, that the 2000–2001 recession was relatively mild so that, despite the slow recovery, employment was only about 1 percent lower three years after the start of that recession. In contrast, employment during the Great Recession is nearly 6 percent lower three years after it started. In terms of the most recent severe recessions, both the 1973 and 1981 recessions featured rapid labor market recoveries, with employment rising 3 percent above previous business cycle peak values.

Tables 1 and 2 provide a more comprehensive comparison between the Great Recession and other recessions and present additional evidence that the recovery from the Great Recession has been very slow. Table 1 shows the average recovery for detrended (2 percent annual growth) per capita output and its components and per capita employment through six quarters for all postwar recessions except the Great Recession. In the average recovery from a

Figure 2
Change in U.S. Employment: Recoveries



postwar recession, the economy is quite close to returning to trend. Table 2 shows the same variables for the Great Recession, which shows virtually no recovery relative to trend for any variable, with the exception of investment. This pattern is qualitatively very similar to that in the Great Depression.

Table 3 shows the same variables for the 1981–82 recession, which is the last severe recession in the United States. The recovery from the 1981–82 recession was quite fast, with all variables, including employment, either very close to trend or even above trend. This rapid recovery following the 1981–82 recession is consistent with standard neoclassical growth theory, which predicts that recoveries should be relatively fast following severe recessions, reflecting transition dynamics associated with diminishing marginal product of capital and low investment during the recession.

This evidence indicates that the recovery from the Great Recession has been comparatively very slow, with the restoration of jobs and output proceeding much more slowly than their postwar averages.

Figure 3
Total Change in U.S. Employment
(36 months after trough)

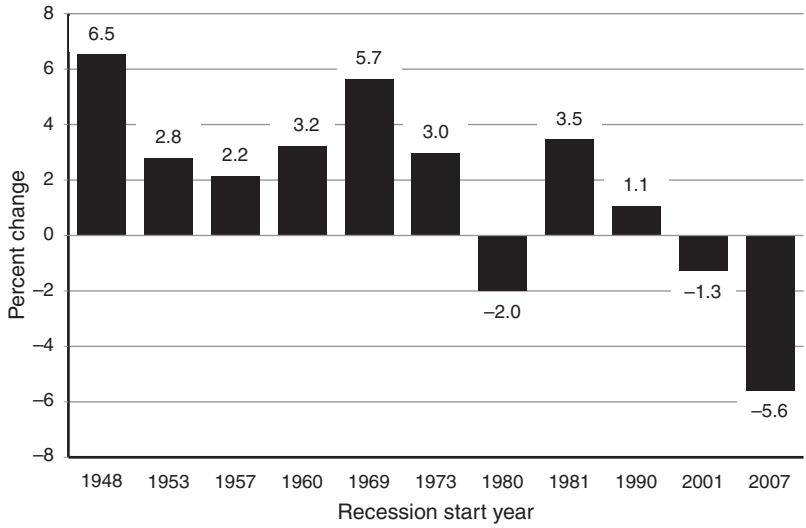


Table 1
Detrended Levels of Output and Its Components in a Typical Postwar Recovery (Excludes 2007 Recession)
(Measured quarterly from trough, Peak = 100)

Quarters from Trough	Output	Consumption	Investment	Government Purchases	Employment	Compensation to Employees
0	95.4	97.8	82.0	98.9	95.8	99.2
1	96.3	98.3	86.9	97.8	95.5	100.0
2	97.0	98.7	91.6	97.9	95.9	100.1
3	97.9	99.5	95.0	97.4	96.6	100.0
4	98.6	99.7	99.5	97.5	97.4	100.3
5	98.7	99.9	99.3	98.1	97.8	100.2
6	99.0	99.8	100.9	99.5	98.1	100.4

Sources: Output and components, Bureau of Economic Analysis; Employment, Bureau of Labor Statistics.

The only comparable episode in which a severe downturn was followed by such a slow recovery is the Great Depression. Table 4

Table 2
Detrended Levels of Output and Its Components in Great Recession

(Measured quarterly from trough, Peak = 100)

Quarters from Trough	Output	Consumption	Investment	Government Purchases	Employment	Compensation to Employees
0	92.1	93.7	65.6	99.7	92.8	100.0
1	91.7	93.5	67.0	99.3	91.7	99.1
2	92.1	93.0	70.5	98.2	91.0	98.9
3	92.4	92.8	74.7	97.3	90.9	98.5
4	92.2	92.7	78.6	97.5	91.0	99.2
5	92.1	92.5	80.8	97.7	91.0	98.9
6	92.1	92.7	76.2	96.6	91.1	98.2

Sources: Output and components, Bureau of Economic Analysis; Employment, Bureau of Labor Statistics.

Table 3
Detrended Levels of Output and Its Components in 1981-II to 1982-IV Recession

(Measured quarterly from trough, Peak = 100)

Quarters from Trough	Output	Consumption	Investment	Government Purchases	Employment	Compensation to Employees
0	97.6	98.8	74.4	99.8	95.1	100.6
1	98.2	99.0	76.5	99.8	95.0	101.3
2	99.5	100.2	83.2	99.9	95.7	100.6
3	100.9	101.2	88.1	100.8	96.6	100.0
4	101.9	102.0	96.7	98.3	98.0	99.4
5	102.3	102.0	105.8	98.4	99.1	99.2
6	102.8	102.6	108.6	99.8	100.1	98.8

Sources: Output and components, Bureau of Economic Analysis; Employment, Bureau of Labor Statistics.

shows the severe depth and duration of the Depression, with relatively little recovery after its 1933 trough. Specifically, relative to 1929, per capita output is about 39 percent below trend, consumption is about 28 percent below trend, investment is about 75 percent below trend, and employment is about 25 percent below trend. And both the Great Depression and the Great Recession feature virtually no recovery in consumption, indicating that factors that are considered to be permanent are contributing to the slow recovery.

Table 4
Consumption, Investment, and Other Components of GNP, 1930–1939
(1929 = 100)

Year	Real GNP	Consumption		Investment (Nonresidential)	Government Purchases	Foreign Trade		Employment
		Durables	Nondurables			Exports	Imports	
1930	87.4	76.2	90.9	79.2	105.1	85.3	84.9	93.8
1931	78.1	63.4	85.4	49.4	105.4	70.6	72.4	86.7
1932	65.2	46.7	76.0	27.9	97.3	54.5	58.1	78.9
1933	61.9	44.4	72.2	24.6	91.7	52.8	60.8	78.6
1934	64.6	49.0	72.1	28.4	101.1	52.8	58.3	83.7
1935	68.1	58.9	73.1	34.4	100.1	53.8	69.3	85.4
1936	74.9	70.8	77.0	45.9	113.9	55.1	71.9	89.8
1937	76.0	72.2	77.2	53.6	106.3	64.3	78.3	90.8
1938	70.6	56.3	74.3	37.8	112.0	62.8	58.3	86.1
1939	73.5	64.3	75.0	40.5	112.9	61.7	61.6	87.5

Source: Cole and Ohanian (2001).

What accounts for such slow recoveries, particularly in the labor market, during these episodes? Ohanian (2009) and Cole and Ohanian (2004) present theoretic and empirical evidence that the severity and continuation of the Great Depression significantly reflected industrial and labor policies that increased industrial cartelization and increased labor bargaining power that, in turn, substantially increased relative prices and real wages. Real manufacturing wages (relative to trend) rose approximately 17 percent from 1929 to 1939, which is abnormal from the perspective of the normal forces of supply and demand. Specifically, low consumption and high unemployment during the decade should have reduced real wages and expanded employment relative to its low level. Ohanian (2009) presents evidence that high real wages during the early phases of the Depression were the result of Hoover's nominal wage maintenance program. Hoover promised firms protection from labor unions provided that industry maintain nominal wage levels and preserve jobs through work sharing. Cole and Ohanian (2004) present evidence that New Deal policies that promoted monopoly and union formation, including the National Industrial Recovery Act and the National Labor Relations Act, fostered higher real wages.

This interpretation of the Great Depression places economic policy, in particular policies that distorted competition and prevented some markets from clearing, at the center of the Great Depression and its labor market dysfunction. Some economists have also suggested that economic policies are contributing to the persistence of high unemployment today. Specifically, the federal minimum wage increased by about 24 percent, rising from $5.85 in 2007 to $6.55 in 2008 and then to $7.25 in July 2009, which may have priced a number of lower-skill workers out of the job market. Some economists also point to a number of executive orders signed by President Obama designed to promote the use of union contractors on large-scale federal construction projects.

This paper analyzes an alternative policy channel that can connect the coincidence between the severity of housing sector depression and the failure of the labor market to recover. In particular, Ohanian and Raffo (2011) document that the only OECD countries to experience severe labor market dysfunction during the Great Recession were also the countries with the most severe housing market downturns: Ireland, Spain, and the United States. The other OECD countries had much less employment loss and much less housing sector

contraction. We therefore analyze the impact of mortgage modifications on persistently high levels of U.S. unemployment. Mulligan (2009 and 2010b) has suggested that these policies distort individual behavior by changing the incentives to take jobs. We pursue this idea by considering the impact of mortgage modifications on location choice. In particular, by reducing the cost of servicing a mortgage, modifications change the incentives to relocate to labor markets with better job opportunities.

3. MORTGAGE MODIFICATIONS DURING THE GREAT RECESSION

This section summarizes mortgage modifications since 2007. Table 5 provides a national perspective on mortgage accounts (90 million accounts by 2010), broad modifications as defined below (11.4 million since 2007), and foreclosure starts (5.8 million since 2007). Following Adelino, Gerardi, and Willen (2009), we define a mortgage modification as a change in the interest rate, principal, or term of the mortgage, or more broadly, as any change to a mortgage that increases or decreases the present value of the loan (many modifications merely tack the current ''forgiven'' portion of the debt onto the end of the loan as a balloon payment). This may include an immediate payment in exchange for forgiveness of principal, such as a short sale (a pre-foreclosure sale) and a deed-in-lieu, in which a mortgagor hands over collateral property in exchange for a release from all obligations.

HOPE NOW, the alliance of mortgage industry entities that was initiated by the federal government, estimates that there have been about 14.2 million modifications that satisfy one of these modification definitions. Adelino, Gerardi, and Willen (2009), in their sample of mortgages, which covers roughly 60 percent of all mortgages through the fourth quarter of 2008, find that roughly 3 percent of borrowers who were 90 days or more in arrears received a narrowly defined modification while 8 percent received a broader form of modification.

Modifications fall into one of two categories:

- government modification programs developed by, or associated with, the Federal Deposit Insurance Corporation, the Federal Housing Finance Agency, the Troubled Asset Relief Program, or the Home Affordable Modification Program, and
- other modification programs.

Table 5
Modifications, Foreclosures, and Mortgage Accounts

	2007-III & IV	2008	2009	2010-I, II, III
Mortgage Accounts (Stock Var.)	97,205,000	97,705,000	93,760,000	90,893,333
Modifications Started (Broad Definition, Flow Var.)	912,671.00	2,258,603.00	4,253,364.00	3,691,320.00
Subset: HAMP Modifications (Broad Definition, Flow Var.)			1,023,224	891,967
Modifications Completed (Flow Var.)	206,240	961,355	1,239,428	1,413,271
Percent of Mortgage Accounts Modified (Broad Definition)	0.9	2.3	4.5	4.1
Percent of Mortgage Accounts Modified (Narrow Definition)	0.21	0.98	1.32	1.55
Foreclosure Starts (Flow Var.)	673,960	1,755,860	2,037,940	1,397,580
Percent of Foreclosure Starts	0.69	1.80	2.17	1.54

Sources: N.Y. Fed Consumer Credit Report; HOPE NOW.

Almost all modifications change the mortgagor's current payment by changing the mortgage interest rate, and/or the time profile of payments, and/or changing the term of the mortgage, and/or deferring payment of principal or forbearance. There have been about 1.9 million HAMP modifications, which account for about 17 percent of all permanent modifications, including the broad definition of modifications, and all other programs account for about 83 percent of modifications.

The three main government programs are HAMP, the FDIC Loan Modification Program, and the Federal Home Finance Agency Streamlined Modification Program. All the government programs feature modifications that reduce payments to either 31 percent or 38 percent of current income (DTI). This is accomplished by reducing the mortgage payment, subject to a minimum interest rate. This minimum interest rate is operative for a fixed period, after which the interest rate rises over time. If the initial interest rate adjustment does not satisfy the DTI requirement, then the term of the mortgage is increased, up to a maximum of 40 years. If these modifications together do not generate the required DTI, then principal is deferred to the end of the mortgage as a balloon payment in order to satisfy the DTI requirement of the program, and this deferred principal does not accumulate interest. Some programs also provide benefits to borrowers by paying off principal, waiving late fees, and recapitalizing arrears in principal, interest, and taxes. The median DTI for those receiving HAMP modifications is about 45 percent of current income when debt only includes mortgage principal, interest, homeowners insurance, and property taxes (''front-end'' DTI). The median ''back-end'' DTI is nearly 80 percent for HAMP modifiers, as this broader measure of debt includes other mandated payments, including credit card, auto, and other debt, and spousal and child support.

We focus on HAMP modifications since they represent the most frequently used modification among the government programs. Other government programs and nongovernment programs are similar along several dimensions. The main exception is that some programs reduce DTI for eligible applicants to 38 percent, rather than HAMP's 31 percent. Descriptions of some of the other programs are in Herkenhoff and Ohanian (2011b).

3.1 Home Affordable Modification Program

This section summarizes the HAMP program and how it evolved over time.

The Making Home Affordable Program was announced in February of 2009 and was operative by March of 2009. The program touted $75 billion for mortgage modifications. However, according to the Treasury's expense report, only $1 billion was spent through 2010. There were several changes to the program on June 1, 2010, but for the sake of space, we will only describe the pre–June 1, 2010 version of HAMP:

Eligibility: HAMP eligibility criteria are listed below, loosely quoted from Fannie Mae's *HAMP Servicing Guide* (2009):

- The mortgage loan is a first lien mortgage loan originated on or before January 1, 2009.
- The mortgage loan has not been previously modified under HAMP.
- The mortgage loan is delinquent or default is reasonably foreseeable.
- The borrower documents a financial hardship by completing a Home Affordable Modification Program Hardship Affidavit and provides the required income documentation. The documentation supporting income may not be more than 90 days old.
- The borrower has a monthly mortgage payment ratio of greater than 31 percent (mortgage payment over gross income).
- A borrower actively involved in a bankruptcy proceeding is eligible for HAMP at the servicer's discretion.
- The current unpaid principal balance is no greater than $729,750.
- The loan must pass a standardized net present value (NPV) test that compares the NPV result for a modification to the NPV result for no modification. If the NPV result for the modification scenario is greater than the NPV result for no modification, the servicer *must* offer the modification; otherwise, the servicer has the option of performing the modification at its discretion.

Unemployment Eligibility: Unemployed persons are eligible, and unemployment benefits count as qualified income. The February

2010 report includes statistics on the main hardship reasons. Roughly 57 percent of the permanent modifications were for people with employment problems, including outright unemployment.

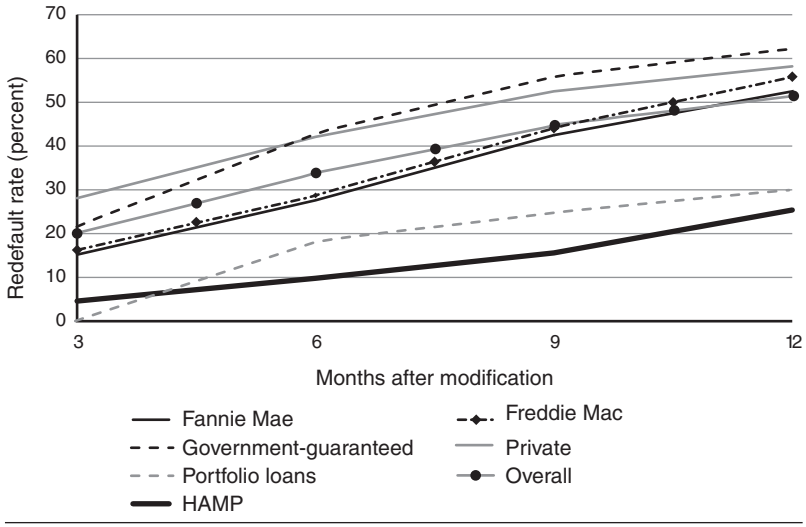
Terms of Modification:

- A borrower may be asked to complete a trial period. The trial period typically lasts three months. During this period, the bank verifies income and assesses whether the borrower can make the new payments.
- If a borrower has an adjustable-rate mortgage or interest-only mortgage, the existing interest rate will convert to a fixed–interest rate, fully amortizing loan.
- The following steps outline the process for determining the 31 percent monthly mortgage payment ratio:
    - Capitalize accrued interest, out-of-pocket escrow advances to third parties, and any other third-party fees that are reasonable and necessary.
    - Reduce the interest rate. The interest rate floor in all cases is 2.0 percent. The reduced rate will be in effect for the first five years, followed by annual increases of 1 percent per year (or such lesser amount as may be needed) until the interest rate reaches the Interest Rate Cap[1], at which time it will be fixed for the remaining loan term.
    - If necessary, extend the term and re-amortize the mortgage loan by up to 480 months to achieve the target monthly mortgage payment ratio.
    - If necessary, the servicer must provide for principal forbearance to achieve the target monthly mortgage payment ratio. The principal forbearance amount is non–interest bearing and nonamortizing. The amount of principal forbearance will result in a balloon payment fully due and payable upon the earliest of the borrower's transfer of the property, payoff of the interest-bearing unpaid principal balance, or maturity of the mortgage loan.

Performance: By June 1, 2010, there were 398,021 permanent modifications and 1,300,526 trials that had been started. The January 2011

[1] The ''Interest Rate Cap'' is the Freddie Mac Weekly Primary Mortgage Market Survey Rate for 30-year fixed-rate conforming loans as of the modification date.

Figure 4
Redefault Rates
(At least 60 days late)



Source: MHA and OTS.

MHA report claims that the pre–June 1, 2010 conversion rate (from trial to permanent) was roughly 41 percent. For HAMP modifications started in the fourth quarter of 2009, 81.6 percent had payment reductions of 20 percent or more (OTS Report 2010-III).

Figure 4 illustrates the redefault rates (which means the loan is again at least 60 days delinquent) for a cohort of HAMP permanent modifications. Annual redefault rates reach 30 percent even with sizeable reductions. Figure 4 also illustrates the redefault rates for government-guaranteed loans (e.g., Federal Deposit Insurance Corporation–insured), government-sponsored enterprise loans (e.g., Fannie Mae and Freddie Mac, see below), private loans with no government affiliation, and the loan portfolio of the participating institutions (i.e., the loans that the banks do not service for someone else).

Failed Trials and Rejections: Of those who had their trial period canceled, 44.2 percent received an alternative modification, 5.9 percent were somewhere in the foreclosure process, and only 6 percent

were current (MHA 2011). A person who fails a trial is no longer eligible for HAMP; however, a rejected person may re-apply. Pre–June 1, 2010, the lack of paperwork necessary led to many unsuccessful modifications (Norris 2009).

3.2 Alternative Modifications

HOPE NOW reports that of the approximately 4.2 million permanent modifications, 3.6 million were completed independently of HAMP guidelines. Pre–March 4, 2009, there were no HAMP criteria, and the streamlined modifications were used scarcely. The pre-HAMP modification performance is relatively poor. Very few loans are current—a mere 24 percent—and even fewer have actually been foreclosed upon—11 percent completed, 16 percent in process (OTS Report 2010-IV). Table 6 details a post-HAMP comparison of the modifications. The HAMP reductions hover around 35–40 percent, and the alternative reductions are also considerably generous, around 15–20 percent. The alternative modifications were similar in several respects to HAMP, though often focused on reducing payments to 38 percent DTI for eligible applicants instead of 31 percent as in HAMP.

For those who did not satisfy one or more eligibility criteria, modifications were still performed outside of HAMP. But in these cases, payments were reduced less than under HAMP and, as we will see below, led to higher redefault rates. Specifically, about 44 percent of canceled HAMP trials obtained an alternative modification, and about 30 percent of denied applicants obtained an alternative modification.

We now describe the performance of these modifications in terms of redefault rates, which are defaults on modified mortgages. Adelino, Gerardi, and Willen (2009) find that the redefault rate for pre-HAMP loans lies between 30 percent and 50 percent. Furthermore, there is a large fraction of loans (30 percent) that have a larger NPV after the modification. They also report that ''fewer than 5% of all of our troubled [90+ days past due] borrowers repaid their mortgages.''

In terms of reallocation, the government is allowing people to maintain their mismatch with the local labor market by providing loan modifications. Clearly, those who could not pay their mortgages initially are redefaulting. The process of modifying and then redefaulting is precisely the delay that we focus on in the model below.

Table 6
Comparison of HAMP Modifications and Alternative Modifications
(Percent)

	2009-IV	2010-I	2010-II	2010-III	2010-IV
Average HAMP Reduction	–39.4	–37.2	–37.8	–35.6	–35.9
Average Alternative Modification Reduction	–14.7	–15.1	–18.5	–20.7	–21.6
HAMP Modifications, Redefault Rates since 2009-I	11	13	17		
Alternative Modifications, Redefault Rates since 2009-I	12	24	32		

Source: OTS Report, 2010-IV.

3.3 Analyses of Current Modification Programs

We are unaware of studies that quantify the impact of modification programs on unemployment within an optimizing model framework. There are several related papers, including Mulligan (2010b), who considers the implicit marginal tax rates generated by actual guidelines for FDIC and HAMP modifications. Mulligan (2009) also discusses the incentive effects of mortgage modifications on employment and suggests that modifications have significantly increased unemployment, but he does not provide a quantitative assessment. Chatterjee and Eyigungor (2009) consider the HAMP program in a single-location dynamic stochastic equilibrium model with endogenous prices, but their model lacks an employment margin and therefore does not allow for employment incentive effects or relocation. Adelino, Gerardi, and Willen (2009) argue that there are few modifications because lenders expect to make more money foreclosing than modifying. Their conclusion is that preventable foreclosures are rarer than most people believe. The probability of redefault in their sample ranges between 30 percent and 50 percent depending on the quality of the mortgage and the type of modification. If a modification uses resources, it may be socially optimal to have few modifications. Gerardi and Li (2010) provide a useful summary and timeline of the policies that were enacted to save homes.

In terms of the link between housing and unemployment, Oswald (1996) hypothesizes that areas with high homeownership rates have higher unemployment rates. Green and Hendershott (2001) use a 1988–1992 Panel Study of Income Dynamics panel to track 9,000 U.S. household outcomes over time. Among their findings, they present evidence that supports the Oswald hypothesis and find significant heterogeneity in the effect of homeownership on unemployment outcomes. There are also studies that have presented evidence against the Oswald hypothesis. These studies primarily are based on European data. Vuuren (2009) studies panel data from the Netherlands and rejects a number of predictions of the Oswald hypothesis.

DiPasquale and Glaeser (1998) consider homeownership and mobility and argue that homeowners are less mobile. Studies based on U.S. data from the recent recession include Ferreira, Gyourko, and Tracy (2010), who use a panel from the American Housing Survey to document that negative equity (which is much more prevalent than modifications) greatly reduces mobility. Schulhofer-Wohl

(2010) disputes these mobility claims, suggesting that the empirical methodology of Ferreira, Gyourko, and Tracy is flawed. Winkler (2011) analyzes homeownership and homeowner mobility, and finds that homeownership reduces mobility by 40 percent and that homeownership also negatively impacts income.[2] While Winkler does not consider mortgage modifications, his economic environment is perhaps the closest to ours in that he also uses an optimizing model that includes locational choice.

4. MODEL ECONOMY

This section presents the model economy we use to assess the impact of mortgage modification programs on economic activity. We blend a search model of unemployment with housing and with the choice of relocating from one local labor market (island) to another island. To focus on the relocation effect and keep the model tractable, we model only the consumer side of the economy and treat prices exogenously. Herkenhoff and Ohanian (2011a) consider the employment incentive effect along with the large eviction delays in a related paper.

Households face a constant probability of death (δ) and maximize the sum of expected utility discounted using a fixed interest rate (r_h). With a one-month period, this means households discount the future using a discount factor $\beta = \frac{(1-\delta)}{(1+r_h)^{\frac{1}{12}}}$. They have preferences over sequences of a nondurable consumption good (c) and a flow of housing services, which is higher if a household owns a home (z_m) rather than rents (z_r). Mortgages are treated as a perpetuity. Thus, owning a home requires making a fixed mortgage payment each period ($\bar{m}$). The mortgage payment is tax-deductible. Renters pay a rental payment (r) each period, but this payment is not tax-deductible.

[2] He also finds that, after a labor shock, the homeowner subgroup has an unemployment rate that is 6 percent higher one or more years after the shock as compared to before the shock. However, renters show no significant difference in unemployment rates. The estimated job offer equation implies that the probability of receiving a job offer from another location is increasing in education (skill in our model). He also estimates the following offer rates: 16 percent of renters receive offers per period, 13 percent of homeowners receive offers per period.

Households are located on one of two symmetric islands, *A* or *B*. They are either employed (*W*) or unemployed (*U*). Each period, each household receives a wage offer drawn from a stationary Markov chain. Households can either accept the offer, or reject the offer and receive unemployment benefits. The household can also choose to relocate to the alternative island, which offers a job-finding distribution that stochastically dominates the distribution in their current island. If they relocate, they exit from their home permanently and incur a onetime utility cost (*MC*).

Mortgage modifications are challenging to model, as these programs may include many changes to the mortgage contract, several of which are difficult to represent recursively. Because mortgages are a perpetuity, all modifications are a temporary reduction in payments, which is exactly what HAMP and other modification programs do. This temporary reduction can lower payments enough so that people who would otherwise move choose to remain in their current location. We call this change in incentives to relocate the *relocation effect* of mortgage modifications.[3]

In our model, households may request a modification, but one time only, as is the case with many modification programs. A modification in the model works as follows: As long as an agent has a DTI between 31 percent and 75 percent (in the model notation described below, $\overline{DTI} > \frac{i\bar{m}}{w\pi_A} > \underline{DTI}$), the agent is eligible for a temporary modification that reduces mortgage payments as a fraction of his current gross income to 31 percent. However, to keep the state space tractable, the modification term ends with probability η. Since the modification depends on the mortgagor's current income, this means-testing of modifications means that the incentives to relocate in order to sample a better wage distribution are distorted, as the opportunity cost of relocating is higher, reflecting the fact that relocating means losing the modification.

Each agent's skills evolve over time, and skill evolution depends on employment status, which is motivated by Ljungqvist and Sargent (1998, 2004). Specifically, while employed, an agent's skills

[3] This model of modifications does not include another important element of the HAMP modification, which is a trial period for a modification. Specifically, in a trial period, the borrower may decline some offers because this would result in a more expensive modification later.

can increase or decrease, and the probability of increasing his skill level exceeds the probability of receiving a lower skill. For unemployed agents, their skills on average depreciate. Let π_A denote the skill of an agent on island A. The skill transitions are governed by a Markov chain, which is described in Section 4.3. We assume that the probability of finding a job increases monotonically with skill. To capture this, $f(\pi_A)$, which is the job-finding probability for an agent with skill π_A, is monotonically increasing. Layoffs occur with probability $1 - p_e$, with p_e denoting the probability of job continuation.

The period budget constraint for an employed mortgage borrower with taxable income I is given by:

$$c + \bar{m}(1 - \tau_l(I)) + T_m = w\pi_A(1 - \tau_l(I))$$

Employed agents earn income $w\pi_A$ where w scales the skill level π_A. Agents face a progressive income tax schedule that is summarized by the function $\tau_l(\cdot)$. We use a progressive income tax, as it allows the model to help match observed homeownership rates by increasing the incentive to own a home as income rises. This allows the model to generate regions in the state space in which consumers prefer renting and regions in which consumers prefer a mortgage. The tax function is piecewise linear, which we describe in detail in Section 4.2.

After-tax income finances nondurable consumption (c), the after-tax mortgage payment, $\bar{m}(1 - \tau_l(I))$, and T_m, which is obligated payments corresponding to other homeownership costs, including property insurance and homeowner association fees, and in addition, includes other mandated payments such as revolving debt service, child support, and spousal support. These other obligated payments are important to include in the model since their level affects the incentive to request a mortgage modification. Renters also face obligated costs, (T_r), where $T_r < T_m$.

If an agent cannot finance a mortgage, which means that the level of other obligated payments is greater than or equal to income, then the mortgagor is forced to leave the home and rent. While unemployed, agents are provided with unemployment benefits $b(\pi_A)$. Benefits are weakly monotone in the skill level with a 50 percent replacement rate and a benefit cap of $\bar{b}$ (see Section 4.2). As skills decumulate, benefits expire. This declining path of benefits is

adopted because it allows us to formulate the problem recursively while maintaining computational tractability. Moreover, this declining time path of benefits in the model reflects the fact that benefits do indeed decline over time.[4] Specifically, extended benefits or emergency unemployment compensation, both of which apply in many states after 26 weeks, can fall to only 24 percent of the original benefit level. Moreover, it is likely that other sources of financial support that unemployed individuals receive—including support from family, friends, unions, and charities—also decline.

An agent that searches on another island finds a job with probability $f(\pi'_B)$, where π'_B is stochastically drawn and depends on the previous island's skill, π_A.

4.1 Value functions

Let $S \in \{W, U, W^M, U^M, W^R, U^R\}$ represent the status of an agent. $W(\pi_A)$ is the value function of an agent with an offer and skill level π_A. $U(\pi_A)$ is the value function of an agent without an offer and skill level π_A. $W^M(\pi_A, \kappa)$ is the value function of an agent with an offer and a mortgage payment that has been reduced by $100 \times (1 - \kappa)$ percent. In other words, $\kappa = 0.75$ indicates a 25 percent reduction in payments. $U^M(\pi_A, \kappa)$ is defined similarly for an agent without an offer. $W^R(\pi_A)$ is the value function of a renter that has an offer and skill level π_A. $U^R(\pi_A)$ is defined similarly for an agent without an offer. In general, the superscript M indicates that the agent currently has a modification and the agent is no longer eligible to have a modification in the future. The M superscript will stay with an agent even when $\kappa = 1$, which means that the temporary modification period is over and the agent pays $1 \times \bar{m}$. The superscript R indicates that the agent is a renter.

As indicated above, mortgages are perpetuities with fixed payments. Once an agent defaults on a mortgage, the agent is a renter for the remainder of his lifetime. Agents are only allowed one modification in a lifetime, and the modification is structured to reduce payments to 31 percent of gross income. The modification term is

[4] Emergency Unemployment Compensation has different stages called ''tiers.'' With each tier there is a duration and a replacement rate. Tier 1 lasts 20 weeks and pays 80 percent of the maximum benefit amount. Tier 2 lasts 14 weeks and pays 54 percent of the maximum benefit amount, and so on. The last tier pays 24 percent of the maximum benefit amount.

stochastic: with probability η a modification ends. When a modification ends, the mortgage payment returns to its original level $\bar{m}$.

Gross income is $w\pi_A$ for employed people and $b(\pi_A)$ for unemployed people. At the time of the modification request, κ, which denotes the mortgage payment reduction, is set so that the new payment $\kappa \times \bar{m}$ is 31 percent of income: $\frac{\kappa(w\pi_A)\bar{m}}{w\pi_A} = 0.31$. After this initial period, κ becomes a fixed-state variable and will only change once the modification ends. When the modification period is over, $\kappa \rightarrow 1$, reflecting the fact that payments return back to $\bar{m}$. Mortgagors who decide to default or redefault are subject to a one-time moving cost $-MC$, which reflects the costs of leaving the home.

There are two important states for a mortgagor with no previous modification activity:

- the skill level π_A, and
- the employment status, which is summarized by S.

For this type of agent, $\mathbb{Q}_S(\pi_A)$ describes the choice set. This choice set will reflect eligibility restrictions for modifications. For instance, if the agent has an offer ($S = W$) and the payment ratio falls between the cap and the eligibility cutoff, $\overline{DTI} > \frac{\bar{m}}{w\pi_A} > \underline{DTI}$, then the choice set includes a modification option,

$$\mathbb{Q}_W(\pi_A) = \{\text{Accept Offer and Pay, Accept Offer and Modify,}$$
$$\text{Reject Offer and Pay, Reject Offer and Default}\}$$

If $\frac{\bar{m}}{w\pi_A} < \underline{DTI}$ or $\frac{\bar{m}}{w\pi_A} > \overline{DTI}$, then no modification is allowed and the choice set is now restricted,

$$\mathbb{Q}_W(\pi_A) = \{\text{Accept Offer and Pay, Reject Offer and Pay,}$$
$$\text{Reject Offer and Default}\}$$

There are three key states for a modified mortgagor:

- the skill level π_A,
- the modification payment reduction κ, and
- the employment status summarized by S.

For this type of agent, $\mathbb{Q}_S(\pi_A, \kappa)$ summarizes the choice set of the agent. Consider an unemployed modified agent ($S = U^M$),

$$\mathbb{Q}_{U^M}(\pi_A, \kappa) = \{\text{Search for Job and Pay, Redefault and Move}\}$$

In the value functions below, we drop the state π_A, which is already summarized in the value function, and we refer to $\mathbb{Q}_S$ as the choice set that implicitly summarizes the qualification criteria.

An agent that begins with an offer and has not previously modified starts the period with a value function $W(\pi_A)$. Recall that taxable income is given by (I). The agent has several choices:

- pay the mortgage, receive a utility flow $u([w\pi_A - \bar{m}](1 - \tau_l(I)) - T_m, z_m)$, accumulate on-the-job skills and face some probability of being fired $(1 - p_e)$;
- skip a payment and request a modification (so long as the payment ratio lies between $\overline{DTI}$ and $\underline{DTI}$);
- reject the offer (*notice that there is no lag between certain states*); or
- default and rent, which gives the agent the option to search on the other island.

For the model,

$$\begin{aligned} W(\pi_A) = \max_{\mathbb{Q}_W}\{&u([w\pi_A - \bar{m}](1 - \tau_l(I)) - T_m, z_m) \\ &+ \beta E_{\pi'_A|\pi_A,W}[p_e W(\pi'_A) + (1 - p_e)U(\pi'_A)], \\ &u(w\pi_A(1 - \tau_1(I)) - T_m, z_m) \\ &+ \beta E_{\pi'_A|\pi_A,W}[p_e W^M(\pi'_A, \kappa(w\pi_A)) + (1 - p_e)U^M(\pi'_A, \kappa(w\pi_A))], \\ &U(\pi_A), -MC + W^R(\pi_A)\} \end{aligned}$$

An agent that begins with an offer and a modified mortgage (either currently modified or modified in the past) starts the period with a value function $W^M(\pi_A, \kappa)$. With probability η, payments step back up:

$$\begin{aligned} W^M(\pi_A, \kappa) = \max_{\mathbb{Q}_{W^M}}\{&u([w\pi_A - \kappa\bar{m}](1 - \tau_l(I)) - T_m, z_m) \\ &+ \eta\beta E_{\pi'_A|\pi_A,W}[p_e W^M(\pi'_A, 1) + (1 - p_e)U^M(\pi'_A, 1)], \\ &+ (1 - \eta)\beta E_{\pi'_A|\pi_A,W}[p_e W^M(\pi'_A, \kappa) + (1 - p_e)U^M(\pi'_A, \kappa)], \\ &U^M(\pi_A, \kappa), -MC + W^R(\pi_A)\} \end{aligned}$$

An agent that begins the period without an offer and has not previously modified starts the period with a value function $U(\pi_A)$. The agent has several choices:

- pay the mortgage, receive a utility flow $u([b(\pi_A) - \bar{m}](1 - \tau_l(I)) - T_m, z_m)$;
- decumulate skills while unemployed and search locally, which results in a job with probability $f(\pi_A)$;
- skip a payment and ask for a modification (so long as the payment ratio lies between $\overline{DTI}$ and $\underline{DTI}$); or
- default and rent, which allows the agent the option to also search on the other island.

For the model,

$$\begin{aligned} U(\pi_A) = \max_{\mathbb{Q}_U} \{ & u([b(\pi_A) - \bar{m}](1 - \tau_l(I)) - T_m, z_m) \\ & + \beta E_{\pi'_A|\pi_A,U}[f(\pi'_A)W(\pi'_A) + (1 - f(\pi'_A))U(\pi'_A)], \\ & u(b(\pi_A)(1 - \tau_l(I)) - T_m, z_m) \\ & + \beta E_{\pi'_A|\pi_A,U}[f(\pi'_A)W^M(\pi'_A, \kappa(b(\pi_A))) \\ & + (1 - f(\pi'_A))U^M(\pi'_A, \kappa(b(\pi_A)))], \\ & -MC + U^R(\pi_A)\} \end{aligned}$$

An agent that has no offer and a modified mortgage (either currently modified or modified sometime in the past) starts the period with a value function $U^M(\pi_A, \kappa)$. With probability η, payments increase to their original level ($\kappa = 1$):

$$\begin{aligned} U^M(\pi_A, \kappa) = \max_{\mathbb{Q}_{U^M}} \{ & u([b(\pi_A) - \kappa\bar{m}](1 - \tau_l(I)) - T_m, z_m) \\ & + \eta\beta E_{\pi'_A|\pi_A,U}[f(\pi'_A)W^M(\pi'_A, 1) + (1 - f(\pi'_A))U^M(\pi'_A, 1)], \\ & + (1 - \eta)\beta E_{\pi'_A|\pi_A,U}[f(\pi'_A)W^M(\pi'_A, \kappa) \\ & + (1 - f(\pi'_A))U^M(\pi'_A, \kappa)], \\ & -MC + U^R(\pi_A)\} \end{aligned}$$

An agent that begins the period renting with an offer has a value function $W^R(\pi_A)$. This agent has two choices:

- continue to work on the same island, or
- quit and pick an island to search for a new job.

For the model,

$$\begin{aligned} W^R(\pi_A) = \max_{\mathbb{Q}_{W^R}} \{ & u(w\pi_A(1 - \tau_l(I)) - r - T_r, z_r) \\ & + \beta E_{\pi'_A|\pi_A,W}[p_e W^R(\pi'_A) + (1 - p_e)U^R(\pi'_A)], U^R(\pi_A)\} \end{aligned}$$

An agent that begins the period renting without an offer has a value function $U^R(\pi_A)$:

$$U^R(\pi_A) = u(b(\pi_A)(1 - \tau_l(I)) - r - T_r, z_r) + \beta \max\{E_{\pi'_A|\pi_A,U}[f(\pi'_A)W^R(\pi'_A) + (1 - f(\pi'_A))U^R(\pi'_A)], E_{\pi'_B|\pi_A,U}[f(\pi'_B)W^R(\pi'_B) + (1 - f(\pi'_B))U^R(\pi'_B)]\}$$

4.2 Functional Forms, Parameters, and Results

The utility function is given by:

$$u(c, z) = \log(c) + z$$

The job-finding probability is strongly monotone in the interior and weakly monotone in the tails. This functional form captures the intuition that it is easier for persons with high skills to find jobs.[5]

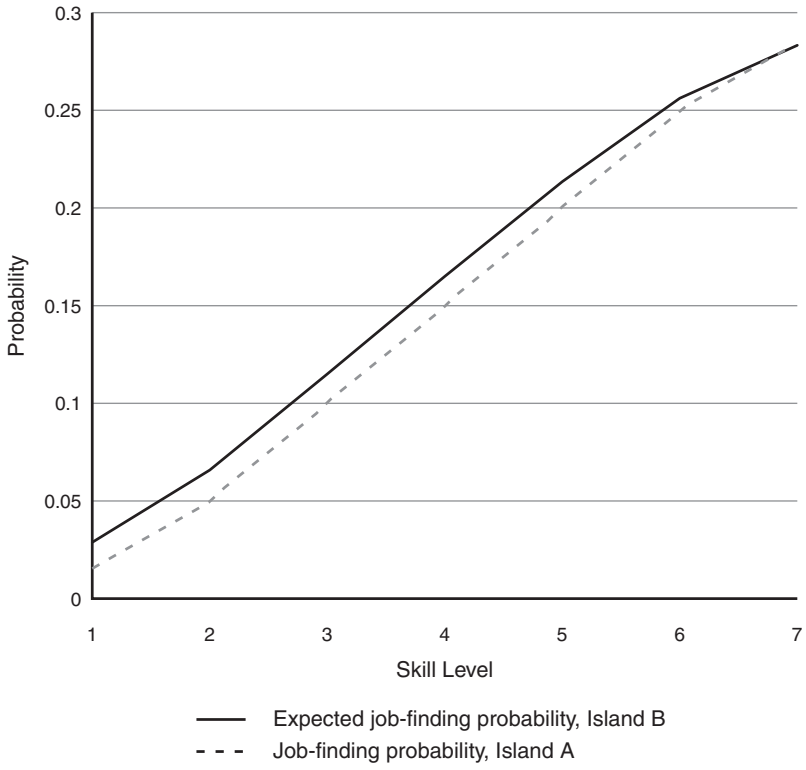
$$f(\pi_A) = f_c \cdot \max\left\{ \underline{f}, \min\left\{ \bar{f}, \frac{\pi_A - \underline{\pi}_A}{\overline{\pi}_A - \underline{\pi}_A} \right\} \right\}$$

This job-finding function is graphed in Figure 5 for $f_c = \frac{3}{10}$, which is used in the simulations. The expected job-finding rate on the alternative island is also graphed and shows the expectation of an island A agent with skill π_A finding a job on island B. This functional form is in line with Shimer's (2008) estimate that the probability of being reemployed in the next month, on average, for all workers in the Current Population Survey dataset is 28.6 percent. We match on average that renters have 3 percent more job offers, as estimated in Winkler (2011).

Unemployment benefits are monotone in the lower half of the support with a 50 percent replacement rate, but benefits are capped by $\bar{b}$. For the simulations that follow, $\bar{b}$ is set at approximately one-half the mean observed wage. Given the skill process described below, benefits last on average for two years:

[5] Mincer (1991) presents evidence supporting this choice. We make use of the following notation: $\underline{x}$ is the lower bound of x and $\bar{x}$ is the upper bound of x.

Figure 5
Job-Finding Probability for Different Skill Levels
(7 is highest skill level)



$$b(\pi_A) = \frac{1}{2} \cdot w \cdot \min\{\pi_A - \underline{\pi}_A, \bar{b}\}$$

The income tax function is described below:

$$\tau_l(I) = \tau_{1/5}\mathbb{I}(I < I_{1/5}) + \cdots + \tau_{4/5}\mathbb{I}(I_{3/5} < I < I_{4/5}) + \tau_{5/5}\mathbb{I}(I_{4/5} < I)$$

This describes an average rate that is applied to all labor income, and the cutoffs are income quintiles, which are defined by $I_{x/5}$ for $x \in \{1, \cdots, 5\}$.

4.3 Parameter Values

The period length is one month. Given the period length, there are several parameter values: the wage (w), the interest rate (r_h), the death probability (δ), the probability that a modification ends (η), the probability of continued employment (p_e), the modification cutoffs for debt to income ($\overline{DTI}$ and $\underline{DTI}$), the mortgage payment ($\bar{m}$), the rental payment (r), the housing utility flow (z_m), the renter utility flow (z_r), non-mortgage debt payments for a homeowner (T_m), non-mortgage debt payments for a renter (T_r), the costs of foreclosure and leaving a house, (MC), the tax schedule $\tau_l(\cdot)$, the grid for skill levels, the transition probabilities, and the initial draws.

The wage rate is set to unity. The annual household discount is set to 6 percent, which is in line with Livshits, MacGee, and Tertilt (2007). The death probability δ is set such that the average lifetime is 42 years. η follows from the HAMP modification program, which reduces payments for five years. The probability of remaining employed, p_e, is set to match the average job duration of 4.6 years (Ljungqvist and Sargent 1998). Modifications reduce mortgage payments to 31 percent of current gross income. The upper limit on debt to income, $\overline{DTI} = 0.75$, is set to match the modification rate (note that by picking the cap to match this moment, the initial front-end DTI of a mortgagor is larger than in the data since the option value of modifying skews this decision) and DTI = 0.31 is taken from the HAMP servicer manual. The fixed mortgage payment $\bar{m}$ is set such that, in the absence of modifications, the average mortgage-payment-to-income ratio is 20 percent, as in Corbae and Quintin (2010). The rental payment r is set to 90 percent of the mortgage payment (U.S. Census Bureau 2011, Table 2-13). The flow from housing z_m is picked to be the log of the average mortgage payment $z_m = \log(\bar{m})$. z_r is scaled in proportion to payments: $z_r = \frac{r}{\bar{m}} z_m$. The fixed cost for a mortgagor, T_m, is set to match the difference between the back-end and front-end DTI of a modifier, which is roughly 30 percent in the HAMP data. T_r is set to match the fraction of people renting, which is about 35 percent.[6] The cost of foreclosure and leaving a house to become a renter, MC, is set to one year's worth of the median wage in order to match the annual migration rate of

[6] The 2000 Census shows a 66 percent homeownership rate.

6.3 percent as reported in Davis, Fisher, and Veracierto (2010).[7] Once an agent becomes a renter, the agent is free to move between locations without any additional cost. The tax schedule $\tau_l(\cdot)$ matches the average effective tax rates by income quintile as published by the Congressional Budget Office (2007). By quintile, the tax rates are 4.3 percent, 9.9 percent, 14.1 percent, 17.3 percent, and 25.2 percent.

The grid for π_A has seven nodes that are evenly spaced between $[\frac{1}{2}, 14\frac{1}{2}]$. Ljungqvist and Sargent (2004) use a process calibrated to two weeks with 11 nodes. In their model, agents lose one node with a 10 percent chance. We follow their setup and have agents move down twice as fast in the unemployed state (10 percent chance of moving down one level every four weeks while unemployed; 5 percent chance of moving down one level every four weeks while employed). Agents keep their original skill level 80 percent of the time while employed, 85 percent of the time while unemployed, and 70 percent of the time when searching on another island. Employed agents move up one slot with a probability of 15 percent, and unemployed agents move up one slot with a probability of 5 percent. (In our model, unlike Ljungqvist and Sargent, the unemployed can increase their skill level.) A person who searches on another island moves up one slot 15 percent of the time and moves up two slots 10 percent of the time. While this matches the wage gains in Kennan and Walker (2011) and renter offer rate in Winkler (2011), this generous process for movers generates an upper bound on the effects of modifications on unemployment.

This human capital process captures much of the dispersion and volatility of monthly income. According to monthly Survey of Income and Participation Program (SIPP) data from 2001, the coefficient of variation (σ/μ) for monthly income ranges from 0.78 to 0.26 depending on where an individual falls relative to the poverty line; for example, those at least 150 percent above the poverty line have an average coefficient of variation of 0.31 for monthly earnings and 0.28 for monthly income (Bania and Leete 2009).

Table 7 illustrates the moments that we try to match by picking appropriate parameters. Several other references are included in the tables for completeness.

[7] While this may seem high, this is conservative in lieu of Kennan and Walker (2011), who find an average moving cost of $312,000.

Table 7
Reference Moments

Description	Data	Model	Parameter	Source
Demographics and Unemployment				
Average Time until Layoff	4.3 years	4.3 years	Firing rate	Ljungqvist and Sargent (1998)
Expected Duration of Working Life	42.7 years	42.7 years	Death rate	Ljungqvist and Sargent (1998)
Homeownership Rate (HR)	66.20%	61.00%	Renter's fixed payment	U.S. Census Bureau (2000)
Adjustment to HR for Negative Equity	−5.60%			Haughwout, Peach, Tracy (2009)
Mean Duration of Unemployment in Weeks (Great Recession)	18.5 weeks	18.1 weeks	Job-finding rate	Bureau of Labor Statistics (Table A-12)
Renting and Default				
Ratio of Renter Payment to Mortgage Payment	0.9	0.9	Ratio of payments and ratio of housing flows	American Housing Survey (2003)
Modifications				
1-Year HAMP Redefault Rate (HAMP Data Summary 2010-III)	25.40%	29.82%	Transition probabilities and DTI cap	OTS 2010-IV (Table 3)
Duration of Modification	5 years	5 years	Stochastic term probability	HAMP Servicing Guide (2009)
Median of Back-End DTI Minus Front-End DTI	33.9%	30.0%	Mortgagor fixed costs	HAMP Data Summary (p. 3, January 2011)

Table 7
Continued

Description	Data	Model	Parameter	Source
Median Front-End DTI before Modification	45.3%	59.0%	DTI cap	HAMP Data Summary (p. 3, January 2011)
Median Back-End DTI before Modification	79.2%	88.0%	DTI cap	HAMP Data Summary (p. 3, January 2011)
Median Front-End DTI after Modification	31.0%	31.0%	Modification scalar	HAMP Data Summary (p. 3, January 2011)
Median Back-End DTI after Modification	62.4%	60.0%	Modification scalar	HAMP Data Summary (p. 3, January 2011)
Modification Rate Per Annum (Broad)	1.4% to 4.5%	1.76%	Transition probabilities and DTI cap	(FRBNY Quarterly Report on Household Credit and Debt, Nov. and HOPE NOW)
Average Quarterly Foreclosure Rate until 2006	0.25%	0.38%	Transition probabilities	Corbae and Quintin (2010)
Migration and Moving				
1-Year MSA Migration Rate, Tax Data	6.36%	8.52%	Moving costs	Davis, Fisher, and Veracierto (2010)
Income Process				
Coefficient of Variation $\times$ 100, Monthly Income, 150+% of Poverty Income	24.8	21.9	Transition probabilities	Bania and Leete (2009), Table 2

5. THE IMPACT OF MORTGAGE MODIFICATIONS IN THE MODEL ECONOMY

The following section presents analyses to evaluate the quantitative impact of modifications on unemployment levels, unemployment duration, and skill levels (productivity). We consider two experiments:

- a comparison of steady states between an economy with no modifications and one with modifications, and
- a one-time economic turbulence analysis along the lines of Ljungqvist and Sargent (1998 and 2004), in which we follow the economy over time after there is a one-time, unanticipated, large, exogenous destruction of jobs.

5.1 Steady State Comparison

We solved for a stationary mass of agents using the techniques outlined in Hopenhayn (1992). We use value function iteration on the grids described above to solve for the policy functions, and we proxy the unit mass of agents on each island with a large number of simulated agents. The stationary mass of 300,000 agents is symmetric across islands, with island *A* movers exactly offset by replica island *B* movers. The results in Table 8 are for an economy that gives modifications in the same proportion as observed in HAMP data. To capture current conditions, the newly born agents are born with mortgages; they are randomly endowed with skills over skill slots 2 to 6; and 9 percent of them start unemployed.[8]

5.2 Steady State Discussion

The duration of unemployment increases in the modification economy, which is the consequence of the lower incentive to relocate. Specifically, low-skill workers and unemployed mortgagors receive, on average, large mortgage payment discounts to reduce payments to 31 percent of their current income. The modification program thus subsidizes unemployment/low skills by reducing the opportunity cost of staying in the local labor market. In the no-modification world, there is no such subsidy, and as a result, the incentive to relocate to the labor market with better job prospects is higher. As a consequence, the modification policies generate 30 basis points of

[8] Recall, there are seven possible skill slots, so agents are not started in the extremes.

Table 8
Steady State Comparison

Modification Policy in Place?	Yes	No	Ratio
Unemployment Rate	7.71%	7.40%	1.084
	(0.015085)	(0.014023)	-
Average Unemployment Duration	18.1081 Weeks	17.3597 Weeks	1.043
	(0.024106)	(0.021994)	-
Average Renter Unemployment Duration	17.695 Weeks	17.0168 Weeks	1.040
	(0.033485)	(0.028446)	-
Average Skill of Employed	12.8805	12.9526	0.995
	(0.0024433)	(0.0025297)	-
Average Skill of Unemployed	11.0629	11.5008	0.995
	(0.0068475)	(0.0068613)	-
Annual Migration Rate	8.53%	10.68%	-
	(0.051711)	(0.05259)	-
Quarterly Foreclosure Rate	0.38%	0.56%	0.685
	(0.0029566)	(0.0041526)	-
Modification Rate Per Quarter	0.44%	NaN%	-
	(0.0016293)	(NaN)	-
Redefault Rate within 12 Months	29.82%	NaN%	-
	(0.22319)	(NaN)	-
Mean Pay Reduction	0.45126%	NaN%	-
	(0.00039674)	(NaN)	-
Fraction of Modifiers with Offer	0.40316%	NaN%	-
	(0.0022109)	(NaN)	-
Average Mortgagor DTI	0.13713%	0.1864%	0.736
	(0.00018456)	(0.00014419)	-
Percentage Renting	0.38397%	0.4807%	0.799
	(0.002414)	(0.0024185)	-

Note: Standard errors in parentheses. NaN = not a number.

higher unemployment. Unemployment is about 50 basis points higher in the turbulence experiment, which is described below.

In addition to the 30–basis point steady state difference in unemployment, the average duration of unemployment in the modification economy is about one week longer, which is about a 10 percent increase. Duration increases because there are more households staying in local (poor job prospect) labor markets. By moving, households expect to move up one skill level, and this is proportional to their chance of finding a job. This implication of higher unemployment duration in the modification economy is consistent with a key fact reported by Winkler (2011), which is that homeowners have a lower hazard rate out of unemployment after an adverse labor shock.

Mobility falls in the modification economy to a migration rate of about 8 percent per year, compared to a 10 percent migration rate in the economy without modifications. This impact on mobility is moderate compared to that estimated by Kennan and Walker (2011), who find that a $10,000 subsidy for moving results in a 2 percentage point rise in mobility. In our model, agents receive a 45 percent reduction in payments, on average, for five years. Mapping this into the $6,000 median annual reduction in payments observed in HAMP data, there is an undiscounted subsidy to modifiers of about $30,000, which is about one-third as large as that estimated by Kennan and Walker.

We also find that the quarterly foreclosure rate is about 20 basis points lower in the modification economy, and the fraction of renters is much lower. The foreclosure rate is higher in the no-modification economy for the same reason that unemployment is lower, which is because more households leave the local labor market when modifications are unavailable. While the difference in the foreclosure rates between the two economies is fairly small, there are large differences in the number of renters. Because modifications are always available, this quarterly 20–basis point difference generates a much higher steady state mass of renters. Specifically, about 48 percent are renters in the non-modification steady state, while about 38 percent are renters in the modification steady state.

To compare our results to those in the literature, we estimated the following equation:

$$D_{i,t=12} = \beta_0 + \beta_1 \pi_{i,t=1} + \beta_2 \kappa_{i,t=1} + \beta_3 JO_{i,t=1} + \epsilon_i$$

The variable i indexes the individual. We estimate this in the cross section where $D_{i,t=12}$ is an indicator of default within 12 months after modification, $\pi_{i,t=1}$ is the skill level at the date of modification, $\kappa_{i,t=1}$ is the payment reduction expressed as a fraction, $JO_{i,t=1}$ is a job offer indicator at the date of modification, and ϵ_i is the error term.

The estimated results, which are in Table 9, are very similar to the empirical results presented in Haughwout, Okah, and Tracy (2009), who estimate a proportional hazard model of the form $D_t = \exp(\alpha(t)) \exp(X_\gamma)$. Since their coefficients are in a nonseparable exponential form, it is difficult to directly compare the results. However, they report that ''the data indicate that a 10 percent reduction in the required monthly mortgage payment is associated with around a

Table 9
Hazard for Redefaulting within 12 Months
(Dependent variable: 12-month redefault indicator)

Constant	0.85509
	(0.17707)
$\pi_{(i,t=1)}$ (Skill)	−0.14873
	(0.008327)
$\kappa_{(i,t=1)}$ (New Payment)	1.1396
	(0.52317)
$JO_{(i,t=1)}$ (Job Offer Dummy)	−0.47614
	(0.070691)

Note: Standard errors in parentheses.

13 percent reduction in the re-default hazard.'' In our model, if payments are reduced by 10 percent (i.e., κ is reduced by 0.1), then the redefault probability falls by 11.3 percent $= 0.1 \times 1.13$. This is roughly the same as their empirical results. Haughwout et al. did not have data on individual employment status, but our model provides results about this effect on redefault. Our estimated equation predicts that a person who has a job at the date of requesting a modification is 48 percent less likely to default one year later as compared to an unemployed person. Likewise, an increase in skill, which is a proxy for income, also reduces the probability of default.

The median modified mortgage payment declines by about 45 percent, compared to a median decline of about 40 percent reported by HAMP. Without the cap on the qualifying DTI, the reduction in the model would be much larger, as households would tend to wait to take the modification until their income is very low. Given that payments are reduced to 31 percent of current income and a modification can be taken one time only, there is an option value to wait to modify.

5.3 Economic Turbulence Experiment

Ljungqvist and Sargent (1998, 2004) analyze the impact of labor market policies by conducting what they refer to as ''turbulence experiments.'' Specifically, they analyze policies in an economy that has a one-time large exogenous destruction of jobs, and in which the skill level of the unemployed declines. We pursue a similar turbulence experiment to analyze the consequences of mortgage modifications in the model with a one-time large job destruction and a reduction in skills of those who are unemployed.

Specifically, our turbulence experiment is as follows:

- we double the layoff rate ($2 \times (1 - p_e)$) for one period; and
- for those who are laid off, their skill level is cut by one notch from their initial condition skill level.

The layoff shock and skill shock are unanticipated. We simulate a unit mass (approximated by 300,000 individuals) and follow them for two years after the shock, and compare the following variables over time between the modification economy and the nonmodification economy:

- the unemployment rate,
- average skill level,
- modification rate,
- the redefault rate, and
- the unemployment survival function.

Modifications are allowed one time only and can be applied for from the initial date of the shock until the end of the second year. The results of this experiment are illustrated in Figures 6 through 9.

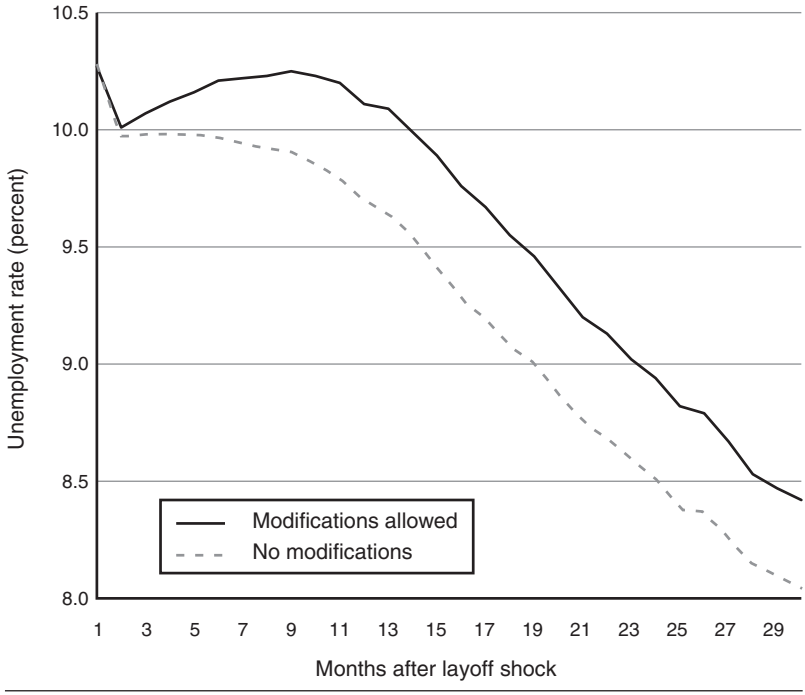
The initial conditions, which are identical across the modification and nonmodification economies, are as follows: 35 percent begin as renters, 65 percent begin as mortgagors, and the initial mean front-end DTI is 14 percent with a 2 percent standard deviation. The initial mean back-end DTI is 21 percent with a 4 percent standard deviation. The initial skills are distributed uniformly over skill slots 3 to 6.[9] Figures 6 through 9 compare the turbulence experiments for the modification and nonmodification economies.

5.4 Turbulence Discussion

The main findings are that modifications raise unemployment, increase the duration of unemployment, reduce the average skill level, reduce worker mobility, and reduce foreclosures. Specifically, the unemployment rate in the modification economy is about one-half percentage point higher than in the nonmodification economy. The 0.5 percent difference in unemployment is reached after about 10 months and continues at about that level for the 30-month horizon that we have examined. While this program does not account for

[9] There are seven slots in total, and no one begins in the extremes.

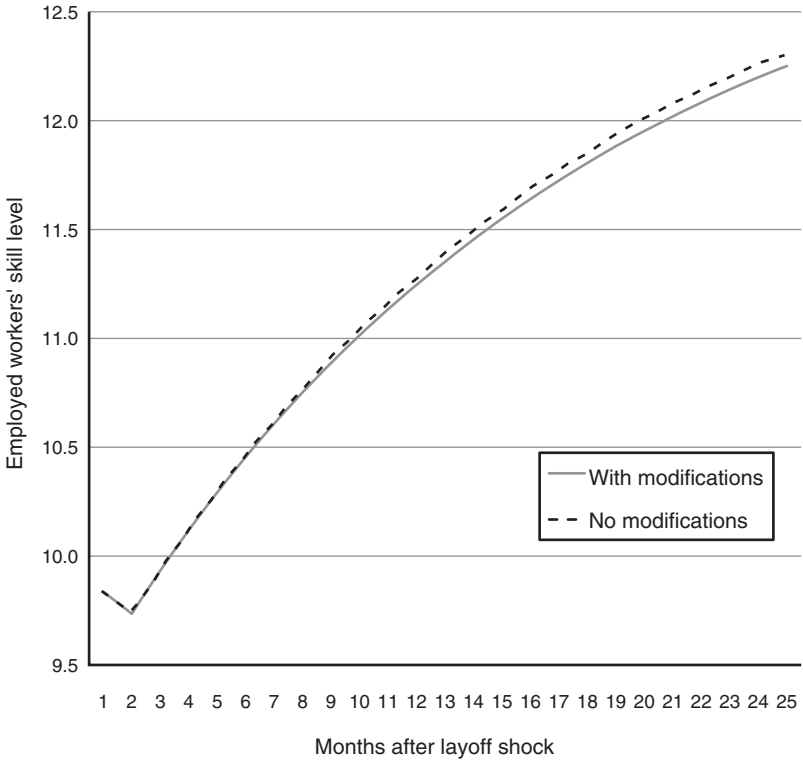
Figure 6
Turbulence Experiment: Unemployment Rate



the bulk of the increase in current unemployment, it does generate a persistent increase in unemployment, corresponding to about 730,000 unemployed individuals given the current size of the U.S. labor force. Unemployment duration in the modification economy is about 18.1 weeks, compared to 17.3 weeks in the economy without modifications.

The average skill level of the employed in the modification economy is about 0.5 percent lower than in the nonmodification economy. Figure 7 illustrates that this difference grows over time as the low-skilled unemployed modifiers eventually reintegrate back into the workforce and drag down the average. It is interesting to note that the modification rate in the first year following the job destruction shock is about 4 percent, which is close to the peak rate of modifications of 4.5 percent in 2009. The median modifier in this economy

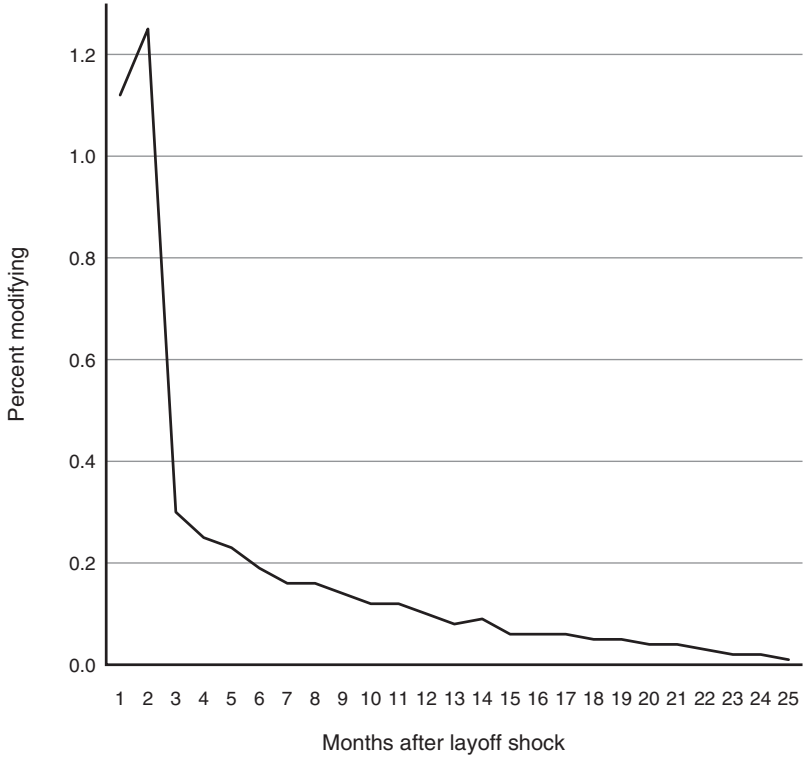
Figure 7
Turbulence Experiment: Employed Workers' Skill



has a back-end DTI of 0.88 as compared to a back-end DTI of 0.79 in the data.

Despite relatively generous modifications, there are a number of redefaults in the modification economy. Many of those who lose their job in the turbulence experiment choose to modify immediately. Of those who modify, many redefault shortly afterward. Specifically, 41 percent of modifiers redefault within one year, which is very similar to the 48 percent rate reported by the Office of Thrift Supervision for overall mortgages (OTS 2010-IV, Table 3). As in actual experience, many modifications are unsuccessful from the perspective of keeping the mortgagor in his home. Of those who successfully modify, it is precisely those with low skill who pay their modified mortgages that create the difference between the two economies. These

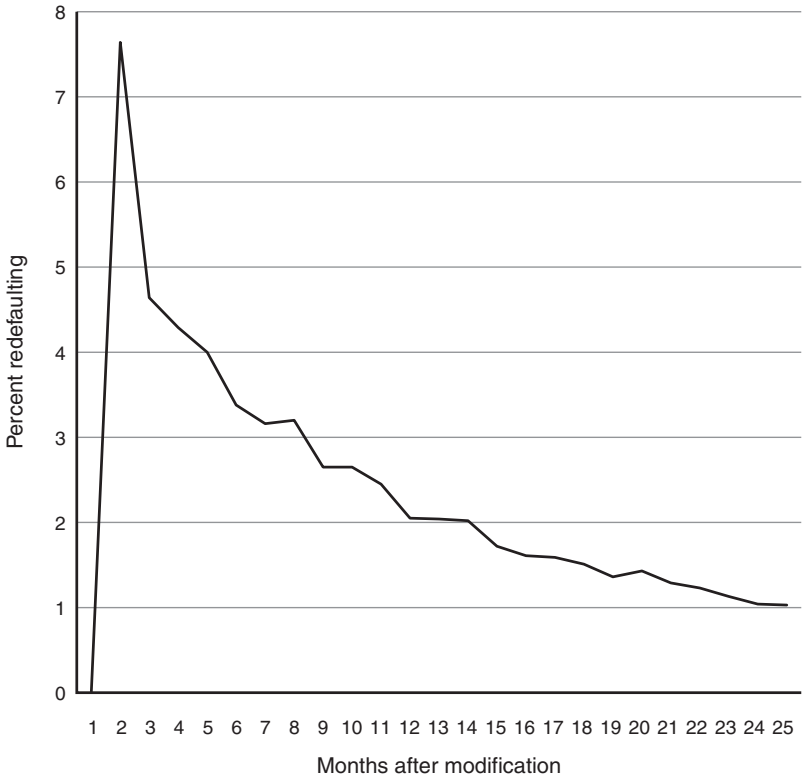
Figure 8
Turbulence Experiment: Percent Modifying Each Month



successful modifiers change the speed of recovery by delaying relocation to better job markets. Moreover, note that while the redefault rates settle down over time, there is still a persistent flow of new modifications even after the initial shock. This means that agents who do not lose their job in the initial period of job destruction do make use of the modification afterward.

These results also have implications for the recent change in the Beveridge curve. Specifically, many economists, including Hall (2010a and 2011), note that the Beveridge curve has shifted recently such that the efficiency of labor market matching is significantly lower today than in the past. While our model does not have a vacancy dimension, it is consistent with less efficient matching as

Figure 9
Turbulence Experiment: Modifications that Default



modified households choose to stay in relatively poor labor markets and thus may be consistent with a shifted Beveridge curve.

6. CONCLUSION

This paper has documented the slow recovery of the labor market from the Great Recession and has analyzed the impact of mortgage modification programs on why recovery has been delayed. These modification programs are means-tested, as the extent that mortgage payments are reduced by a modification depends on a borrower's current economic circumstances, including circumstances in which income is limited to unemployment benefits. Means-testing thus

changes the incentives for workers to relocate from relatively poor labor markets to better labor markets. Our findings indicate that these policies, as modeled in this version of the paper, can add about 0.5 percent to the unemployment rate and reduce per capita income by about 1 percent, which reflects both lower employment and lower worker productivity through skill erosion.

In terms of understanding why unemployment remains so high, it appears that other factors in addition to modifications are impacting current labor markets. Hall (2011) analyzes a model in which high real interest rates, combined with labor market rigidities, are important. It would be of interest to blend Hall's world with the modifications presented here, as well as to consider sectoral issues, given that some sectors of the economy, such as housing, have been more severely impacted.

REFERENCES

Adelino, M., K. Gerardi, and P. Willen. 2009. ''Why Don't Lenders Renegotiate More Home Mortgages? Redefaults, Self-Cures and Securitization.'' National Bureau of Economic Research Working Paper 15159.

American Securitization Forum. 2007. ''Streamlined Foreclosure and Loss Avoidance Framework for Securitized Subprime Adjustable Rate Mortgage Loans.'' December 6.

Bania, N., and L. Leete. 2009. ''Monthly Household Income Volatility in the U.S., 1991/92 vs. 2002/03.'' *Economic Bulletin* 29 (3): 2100–12.

Chatterjee, S., and B. Eyigungor. 2009. ''Foreclosures and House Price Dynamics: A Quantitative Analysis of the Mortgage Crisis and the Foreclosure Prevention Policy.'' Federal Reserve Bank of Philadelphia Working Paper 09-22.

Cole, H., and L. Ohanian. 2001. ''The Great Depression in the United States from a Neoclassical Perspective.'' In *Handbook of Monetary and Fiscal Policy,* ed. J. Rabin and G. L. Stevens, pp. 1035–64. Florence, KY: CRC Press.

———. 2004. ''New Deal Policies and the Persistence of the Great Depression: A General Equilibrium Analysis.'' *Journal of Political Economy* 112 (4): 779–816.

Congressional Budget Office. 2007. ''Historical Effective Federal Tax Rates: 1979 to 2005.'' Unpublished manuscript. December.

Corbae, D. and E. Quintin. 2010. ''Mortgage Innovation and the Foreclosure Boom.'' Working paper. University of Texas at Austin and Federal Reserve Bank of Dallas.

Davis, M., J. Fisher, and M. Veracierto. 2010. ''The Role of Housing in Labor Reallocation.'' Federal Reserve Bank of Chicago Working Paper 2010-18.

DiPasquale, D., and E. Glaeser. 1998. ''Incentives and Social Capital: Are Homeowners Better Citizens?'' NBER Working Paper 6363.

Fannie Mae. 2009. ''Fannie Mae HAMP Servicing Guide.'' November 2.

Ferreira, F., J. Gyourko, and J. Tracy. 2010. ''Housing Busts and Household Mobility.'' *Journal of Urban Economics* 68 (1): 34–45.

Gerardi, K., and W. Li. 2010. ''2010 Mortgage Foreclosure Prevention Efforts.'' *Economic Review* 95 (2): 1–13.
Green, R., and P. Hendershott. 2001. ''Home-ownership and the Duration of Unemployment: A Test of the Oswald Hypothesis.'' NBER Working Paper 10021.
Hall, R. 2011. ''The Long Slump.'' *American Economic Review* 101 (2): 431–69.
———. 2010a. ''The Labor Market in the Current Slump.'' Stanford University.
———. 2010b. ''Why Does the Economy Fall to Pieces after a Financial Crisis?'' *Journal of Economic Perspectives* 24 (4): 3–20.
Haughwout, A., E. Okah, and J. Tracy. 2009. ''Second Chances: Subprime Mortgage Modification and Re-Default.'' Federal Reserve Bank of New York Staff Reports 417. December 2009; revised August.
Herkenhoff, K. F., and L. E. Ohanian. 2011a. ''Modifications and the Employment Incentive Effect.'' Unpublished manuscript.
———. 2011b. ''Mortgage Modification Survey: Literature and Policies.'' Unpublished manuscript.
HOPE NOW. 2010. ''HOPE NOW Industry Extrapolations and Metrics (October 2010).'' December 5.
Hopenhayn, H. 1992. ''Entry, Exit, and Firm Dynamics in Long Run Equilibrium.'' *Econometrica* 60 (5): 1127–50.
Kennan, J., and J. Walker. 2011. ''The Effect of Expected Income on Individual Migration Decisions.'' *Econometrica* 79: 211–51.
Livshits, I., J. MacGee, and M. Tertilt.'' Consumer Bankruptcy: A Fresh Start.'' *American Economic Review* 97 (1): 402–18.
Ljungqvist, L., and T. Sargent. 1998. ''The European Unemployment Dilemma.'' *Journal of Political Economy* 106 (3): 514–50.
———. 2004. ''European Unemployment and Turbulence Revisited in a Matching Model.'' *Journal of the European Economic Association* 2 (2–3): 456–68.
Making Home Affordable. 2010. ''Handbook for Servicers of Non-GSE Mortgages—Version 3.0.'' December. http://www.HMPadmin.com
———. 2011. ''Making Home Affordable Program Servicer Performance Report through January 2011.'' January 21.
Mayer, C., K. Pence, and S. Sherlund. 2009. ''The Rise in Mortgage Defaults.'' *Journal of Economic Perspectives* 23 (1): 27–50.
Mincer, J. 1991. ''Human Capital, Technology, and the Wage Structure: What Do Time Series Show?'' NBER Working Paper 3581.
Mulligan, C. 2009. ''A Depressing Scenario: Mortgage Debt Becomes Unemployment Insurance.'' NBER Working Paper 14514.
———. 2010a. ''Aggregate Implications of Labor Market Distortions: The Recession of 2008-9 and Beyond.'' NBER Working Paper 15681.
———. 2010b. ''Foreclosures, Enforcement, and Collections under the Federal Mortgage Modification Guidelines.'' NBER Working Paper 15777.
Norris, F. 2009. ''Why Many Home Loan Modifications Fail.'' *New York Times*. December 4.
Ohanian, L. 2009. ''What—or Who—Started the Great Depression?'' *Journal of Economic Theory* 144 (6): 2310–35.
———. 2011. ''The Economic Crisis from a Neoclassical Perspective.'' *Journal of Economic Perspectives* 24 (4): 45–66.
Ohanian, L. E., and A. Raffo. 2011. ''Hours Worked over the Business Cycle in OECD Countries, 1960–2010.'' Unpublished manuscript.

Office of Thrift Supervision. Quarterly. ''OCC and OTS Mortgage Metrics Reports.'' http://www.ots.treas.gov/.

Oswald, A. 1996. ''A Conjecture on the Explanation for High Unemployment in the Industrialized Nations: Part 1.'' University of Warwick Economic Research Papers.

Sherlund, S. 2008. ''The Past, Present, and Future of Subprime Mortgages.'' Unpublished manuscript.

Schulhofer-Wohl, S. 2010. ''Negative Equity Does Not Reduce Homeowners' Mobility.'' Federal Reserve Bank of Minneapolis Working Paper 682. December.

Shimer, R. 2008. ''The Probability of Finding a Job.'' *American Economic Review* 98 (2): 268–73.

U.S. Census Bureau. 2011. ''American Housing Survey, 2003.'' April.

Vuuren, A. V. 2009. ''The Impact of Homeownership on Unemployment in the Netherlands.'' In *Homeownership and the Labour Market in Europe*, ed. C. van Ewijk and M. van Leuvensteijn. Oxford, UK: Oxford University Press.

Winkler, H. 2011. ''The Effect of Homeownership on Geographic Mobility and Labor Market Outcomes.'' UCLA working paper.

Comment

Robert E. Hall

The Herkenhoff-Ohanian paper has two purposes. The first is to demonstrate just how bad conditions have been in the labor market since the cyclical peak of employment in 2007. They compare the persistent shrinkage of employment to the similar, but much larger, shortfall in the Great Depression. The second part of the paper puts one program under a powerful microscope to see if its adverse effects on unemployment are an important part of the story of high recent unemployment.

Figures 1 through 3 and Tables 1, 2, and 4 of their paper make it clear how low employment growth has been in the recovery that began in mid-2010 compared to earlier recoveries, with the sole exception of the Great Depression. The authors diagnose ''labor market dysfunction.''

Their implicit hypothesis is that labor market factors account for the poor performance of the economy. The paper contains no mention of events in financial markets that figure so prominently in other discussions of the deep and persistent slump in the labor market that began in 2007. My own view—see Hall (2011)—assigns most of the blame for high unemployment on forces outside the labor market, notably the bulge in household capital and corresponding household debt inherited from the middle of the past decade and the paralysis of monetary policy resulting from its inability to depress the interest rate below zero. The paper's brief discussion of the Great Depression similarly omits the financial driving forces that others have emphasized.

That said, labor market dysfunction, or at least a decline in the efficiency of the hiring process given the widespread availability of

Robert E. Hall is the Robert and Carole McNeil Joint Hoover Senior Fellow and professor of economics at Stanford University.

willing workers, may well be part of the explanation of the explosion of unemployment. On this, see Davis, Faberman, and Haltiwanger (2010), which pursues very different ideas about the subject.

The Herkenhoff-Ohanian paper concentrates on one hypothesis about recent changes in the labor market. The hypothesis is that policies intended to help families deal with their inability to meet their mortgage payments may have the unintended consequence of limiting their incentives to look for jobs available in areas sufficiently remote to require moving houses. The policies unquestionably tie families to their existing homes and thus limit geographic mobility. The question is how much of the bulge in unemployment results from families' response to the incentives. The paper's answer is very little. Table 8 shows that their model, tuned to deliver an unemployment rate of 7.7 percent in the presence of the HAMP policy for mortgage assistance, predicts a rate of 7.4 percent without that program. In this respect, the paper supports the view that most of the rise in unemployment is the result of other forces, including diminished demand for labor arising from the financial crisis and diminished efficiency of reallocating unemployed workers to jobs for reasons other than HAMP.

The authors approach the task of quantifying the effects of HAMP on unemployment in a thoroughly modern way. First is a detailed description of HAMP and its statement in mathematical form. Second is embedding HAMP in a family dynamic program, laid out in wonderful detail in Section 4.1. Families assign a value to their current status, which depends on their employment opportunity and residential status (owner with original mortgage, owner with modified mortgage, or renter). A family chooses an action, such as applying for a mortgage modification through HAMP, when that action delivers a higher expected value than other actions available at the time.

The key interaction between mortgage modification and unemployment is that workers generally face better job opportunities in distant labor markets, thanks to mismatch in the locations of the unemployed and the location of jobs. Signing up for HAMP makes the choice to look for work in a distant market less likely because a move requires a homeowner to default on a mortgage, become a renter, and lose the benefit of HAMP.

The third step is to solve the model for its steady-state equilibrium (a nontrivial piece of computation). The reader is spared the details

of this aspect of the work. The final step is to adjust parameter values so that the equilibrium of the model matches known features of the labor and housing markets, as shown in Tables 7 and 8. Here the authors apply the econometric method of indirect inference. They are unable to provide information about the sampling accuracy of their parameter estimates—normally a standard feature of indirect inference—because they draw their reference moments from a variety of sources and thus lack the covariance matrix that would be needed for the calculation of sampling errors. That said, it would be desirable in future work to try to give some indication of the potential magnitude of sampling variation.

Armed with a complete computational model, the authors compare, in Table 8, an economy intended to resemble the actual economy, including HAMP, with a similar economy differing only in the absence of HAMP. In addition to the lower unemployment in the non-HAMP economy, the table shows more migration, more foreclosures, and much more propensity to rent in the non-HAMP economy.

The results in Table 8 should be compared to other estimates of the effect of HAMP and to other data on the current U.S. economy. The authors note the increase from 38 percent to 48 percent in the fraction of families renting their homes but do not go on to compare those figures to actual data. Prior to the crisis, 31 percent of U.S. families were renters, a figure that rose to almost 33 percent recently. Thus the model modestly overstates the incidence of renting in normal times (it says 38 percent) but seriously overstates any possible effect of HAMP because renting rose by about 3 percentage points from all the influences operating recently—all of which point upward—while the model predicts an increase of 10 percentage points from HAMP alone.

Other research has considered some issues that Herkenhoff and Ohanian take up. The evidence on any general decline in geographic mobility postcrisis is mixed, but it is fair to say it has not been large (nor was its level very large precrisis). Kothari, Eksten, and Yu (2011) show that mobility rates for homeowners fell by less than trend after the crisis, while rates for renters rose. The fact that the change in mobility for owners was less than for renters gives a bit of support to the hypothesis that geographic mobility among homeowners was impaired by recent events. Saks and Wozniak (2009) show as well that interstate mobility fell a small amount in the years since 2007.

Kothari et al. (2011) show that mobility among unemployed homeowners declined from 2006 to 2010. Saks and Wozniak (2009) show that, as a general matter, mobility has been lower in recent decades in years with high unemployment, suggesting that the recent declines in mobility may not be the result of programs such as HAMP that were not present during past periods of high unemployment.

Kothari et al. (2011) show that geographic moves for job reasons are generally low for both homeowners and renters, but lower for owners. Job-related mobility fell by more between 2006 and 2010 for renters than for owners.

Ferreira, Gyourko, and Tracy (2010) find that negative equity has a small but statistically unambiguous negative effect on mobility among homeowners. This finding supports the hypothesis that economic factors relating to housing, as studied by Herkenhoff and Ohanian, are a factor in mobility decisions. Schulhofer-Wohl (2011) presents a similar finding.

By necessity, the paper concentrates its detailed modeling on HAMP and the decisions of interest—signing up for HAMP, defaulting and becoming a renter with no further impediment to searching in a more favorable labor market, or staying put without HAMP and keeping an existing job. The interaction between housing and labor market decisions is nicely captured. On the other hand, many aspects of the labor market are streamlined relative to models that concentrate on that market and neglect housing decisions. In particular, the model lacks the key idea of the Diamond-Mortensen-Pissarides theory of unemployment: endogenous tightness. In that model, wage bargaining is central to the behavior of unemployment. When unemployment is high, the bargaining position of a worker is reduced because alternative jobs are hard to find. If the result is a lower wage, employers' incentives to create jobs are correspondingly higher, and unemployment returns to its normal level. If wages do not reflect the lower bargaining power of workers—if they are sticky instead—the self-correcting mechanism is less effective, and unemployment can be high and persistent. Because recent experience has shown that unemployment can, in fact, become high and then persist at high levels, the neglect of the feedback mechanism in the paper is probably not a major reason to question its findings.

Of course, the paper does not find that the labor market became dysfunctional as a result of a program, HAMP, that helped keep

people in their existing houses and dissuaded them from moving to places with more favorable job opportunities. It finds only a small effect of HAMP. The notion that the real harm to workers came from factors outside the labor market—the same factors that led to serious declines in consumption and investment spending—remains largely intact after the authors' careful examination of HAMP.

REFERENCES

Davis, S. J., R. J. Faberman, and J. C. Haltiwanger. 2010. ''The Establishment-Level Behavior of Vacancies and Hiring.'' NBER Working Paper 16265. August.

Ferreira, F., J. Gyourko, and J. Tracy. 2010. ''Housing Busts and Household Mobility.'' *Journal of Urban Economics* 68 (1): 34–45.

Hall, R. E. 2011. ''The Long Slump,'' *American Economic Review* 101 (2): 431–69.

Kothari, S., I. S. Eksten, and E. Yu. 2011. ''The (Un)importance of Mobility in the Great Recession.'' Department of Economics, Stanford University. May.

Saks, R. E., and A. Wozniak. 2009. ''Labor Reallocation over the Business Cycle: New Evidence from Internal Migration.'' Federal Reserve Board. January.

Schulhofer-Wohl, S. 2011. ''Negative Equity Does Not Reduce Homeowners' Mobility.'' NBER Working Paper 16701. January.

Comment

John V. Leahy

If one looks at the economy today, there are three big imbalances: unemployment, the housing sector, and government finances. The authors are to be commended for taking a step toward tackling the first two. My main criticism will be that, while the authors make a start, they did not go far enough. They chip off a small piece of the unemployment problem and a small piece of the housing problem. In reality, these problems are much larger than what we see in their model.

This critique is certainly unfair. The model is already very complicated and challenging to solve. Extending it would entail major effort. Still, I see this paper as an interesting first step, and it is useful to consider where further steps might take us.

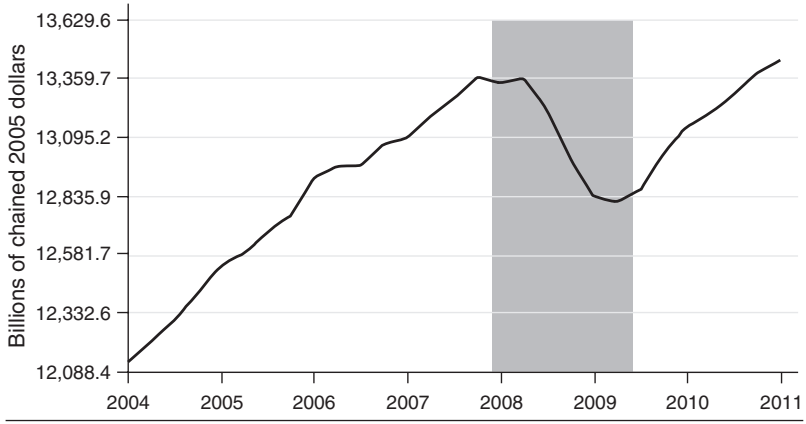
I will begin by briefly describing their argument and the setup of their model. This will give me a background on which to place my comments. I will close with some general comments about the state of housing and questions that still need to be answered. I believe that commenter Robert Hall's discussion will focus more on unemployment and migration.

THE MODEL

The paper begins with the observation that the current recession differs from the typical postwar recession: there has been little rebound toward trend after the initial drop-off in output. In this, the authors argue the current slump is more like a miniature Great

John V. Leahy is professor of economics at New York University.

Figure 1
Real Gross Domestic Product



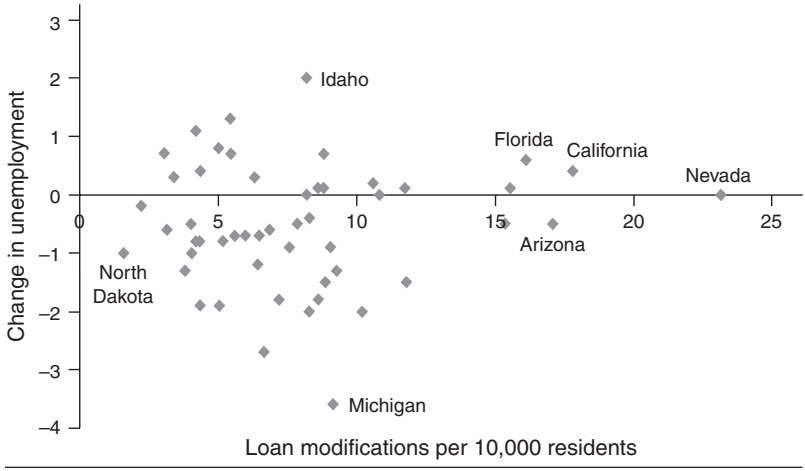
Source: U.S. Department of Commerce, Bureau of Ecconomic Analysis.
Note: Shaded area indicates recession.

Depression than a recession. Figure 1 shows the behavior of GDP since 2004. It is easy to see that there has been little mean reversion. The postrecession trend looks very similar to the prerecession trend. It is as if the economy experienced a permanent downward shift in its growth trajectory.

The remainder of the paper attempts to explain this shift and why the unemployment rate has remained so high for so long. In earlier work, the authors had some success in attributing the length of the Great Depression to the effects of government regulation. They search for a similar storyline in the current recession. This search leads them to consider housing, particularly mortgage modification. Their argument is that mortgage modification, by making it cheaper for borrowers to remain in their homes, may lead workers to remain in poor labor markets rather than move to where employment opportunities are better.

To get a quick check on the potential magnitude of this channel, I plotted data on unemployment and mortgage modifications at the state level. If one simply plots the level of mortgage modification on the level of unemployment, one gets a strong positive correlation that likely reflects the fact that poor economic conditions cause both

Figure 2
Loan Modifications and Changes in the Unemployment Rate



Source: Office of Thrift Supervision, ''OCC and OTS Mortgage Metrics Report, 3rd Quarter 2010,'' Table 50.

Note: Change in unemployment is from June 2009 to April 2011. Loan modifications are as of the third quarter of 2010.

to rise. Figure 2 instead plots the change in unemployment against mortgage modification at the state level. The x-axis is the number of modifications per 10,000 residents in the third quarter of 2010. The y-axis is the change in the unemployment rate between June 2009 and April 2011. The modification programs began in 2009 and ramped up during the year, so June seemed like a reasonable starting point. April 2011 was chosen as the end point because it was the latest data available at the time of the conference when these papers were presented. There is not much of a pattern in the figure. A regression reveals a slight upward slope that is statistically insignificant. If we take the estimates at face value, eliminating modifications would reduce unemployment by less than a quarter of a percent. This estimate is consistent with the small number that the authors' theoretical model generates. Still, this may be an overestimate, as we have not fully controlled for the effects of unemployment on modifications.

The author's theoretical analysis builds on a model of labor flows. The model has exogenous job-offer and -destruction rates, a skill ladder in which employed workers accumulate skills and unemployed workers decumulate skills, and three decisions: a decision to accept or reject a job; a decision to pay a mortgage, accept a modification, or default; and a decision to stay in one location or move to another. Modification involves a reduction in interest payments, and default involves a cost of entering a default state. The decision to move increases an agent's job-finding rate, but an agent must default in order to move, so in effect it is the decision to default that improves an agent's job prospects. Incomes of working agents fluctuate over time.

It is useful to think of the model as a large map. The locations on this map correspond to whether an agent is employed or unemployed, on the one hand, and has a mortgage that is current, modified, or in default, on the other. The model describes how agents move between these various states. In any given period, they may choose to move to a worse state (i.e., from employment to unemployment), or from a current mortgage to a modified mortgage, or to default. Movements from between employment and unemployment or from default to modification also happen by chance.

The main result of the analysis is that allowing modifications raises the unemployment rate. This follows directly from the assumption that those in default have a higher job-finding rate. There is also an amplification effect that comes from skill accumulation. The higher unemployment rate causes agents to lose skills, which further reduces their productivity and hence job prospects. All in all, the mechanism can explain about half of a percent rise in unemployment.

My first observation regarding the model concerns the plausibility of the mechanism. In the model, one has to default in order to move to a new location and improve one's job prospects. I do not know the data, but my intuition tells me that only a small fraction of agents who move have defaulted on their mortgages. The vast majority sell their homes, pay off their loans, and then move. The model shuts down this channel to focus on the lock-in effect of loan modification.

My second observation is that each agent moves independently across employment and mortgage states. Agents do not interact. There is no equilibrium. There is no house price that equalizes the supply and demand for housing. There is no wage that responds to

unemployment. There is no consumption-savings decision, with its effect on interest rates and capital accumulation.

Of these, the lack of a role for house prices is the most troubling, since house prices would appear to be at the center of our recent troubles. In the model, agents default to escape mortgage payments and to increase their chance of finding a job. There is no role for negative equity. Mortgage payments never exceed the value of a home. When I think of someone being locked into a poor labor market, I think of someone who cannot afford to sell his home, not someone who enjoys low interest on a modified mortgage.

Among the missing interactions are the spillover effects, in which some agents' attempts to sell their homes to get out from under their mortgages reduce the prices of all houses, and hence the position of other borrowers. Missing also are the effects of mass default, which may lead to fire sales as lenders attempt to reduce inventories of repossessed homes. Missing are the effects on the banks themselves, as they see their equity position eroded by a worsening mortgage portfolio. Plenty of good research focuses on specific issues at the expense of other concerns. The model shows that the direct effect of mortgage modification on mobility is small. The question remains whether the interaction with some of these other missing channels is more significant.

My third observation regards the welfare implications of the model. Default and modification are both good outcomes in this model. Every mortgage holder in the model wants to modify his loan, since modification reduces payments without any cost. The only reasons that mortgage holders do not modify immediately is that some are prevented by the debt-to-income threshold, while others are waiting to modify at even better terms in the future. It would seem that the model is missing some cost to modification. Maybe it is the implication of modification for the agent's credit score. Maybe it is the effect that modification will have on the agent's ability to borrow in the future. Maybe it is some notion of commitment or obligation to pay one's debts. Whatever it is, it does not appear to be an insignificant omission.

While modification is unambiguously good in the model, default comes with costs, but also a big benefit. One needs to default in order to improve one's chances of finding a job. The more agents default, the more quickly they find jobs, the lower is employment

and the higher is the skill level of the population. It would seem that the model is missing some costs of default as well. I have already mentioned the effect on the banking sector and the effect on house prices. One might also imagine an effect on the federal budget through government mortgage guarantees. Omitting these costs reduces the model's usefulness as a tool to evaluate modification as a policy.

My fourth observation regards what we learn about the mortgage modification as a policy. The main lessons are that modification has only modest effects on labor mobility and that there may be a surprisingly large option value to delaying modification. Everyone wants to modify, but only a few actually do. This may help to explain why so few take up these programs. Beyond this, we learn very little about modification as a policy. One might like to know how modification today will affect the decision to borrow in the future. Will borrowers assume that they will be bailed out again? One might want to know how it will affect banks. Will they demand higher interest rates as a cushion? One might want to know who is paying for this insurance and how it is being priced, if at all.

BUTTRESSING THE HOUSING MARKET

I want to close with a few comments about the housing market. The past few years have seen a massive effort by nearly all branches of government to support house prices. Fannie Mae, Freddie Mac, and the Federal Housing Administration now back over 90 percent of new mortgages (Inside Mortgage Finance Publications 2010). There have been large increases in conforming loan limits. The Federal Reserve has purchased large quantities of mortgage-backed securities. There has been a homebuyer tax credit and mortgage modification.

This has largely been an effort to contain the problems caused by declining house prices rather than a solution to these problems. The plan, if there is one, appears to be ''hope that the economy improves and the problem goes away.''

There is a lot that we do not understand about these efforts. The housing market is a very large market, and efforts to move it cannot be costless. What are the costs and benefits of these efforts? The loan guarantees and security purchases expose the government to significant risks. What is the fiscal exposure? Are there alternatives?

How do we extricate the government from this market and move to a more balanced and sustainable system?

The authors have written an interesting paper on an important topic. In my comments, I tried to point out some of the things that were missing. I am sure that none of this is news to the authors. All modeling efforts involve choices of where to focus attention and what to simplify. I look forward to seeing where they take this research in the next few years.

REFERENCES

Inside Mortgage Finance Publications, Inc. 2010. ''Mortgage Originations Surge in Third Quarter.'' *Inside Mortgage Finance* 27 (41).

Office of Thrift Supervision. 2010. ''OCC and OTS Mortgage Metrics Report: Disclosure of National Bank and Federal Thrift Mortgage Loan Data, Third Quarter.'' http://www.ots.treas.gov/.

Cato Institute

Founded in 1977, the Cato Institute is a public policy research foundation dedicated to broadening the parameters of policy debate to allow consideration of more options that are consistent with the traditional American principles of limited government, individual liberty, and peace. To that end, the Institute strives to achieve greater involvement of the intelligent, concerned lay public in questions of policy and the proper role of government.

The Institute is named for *Cato's Letters*, libertarian pamphlets that were widely read in the American Colonies in the early 18th century and played a major role in laying the philosophical foundation for the American Revolution.

Despite the achievement of the nation's Founders, today virtually no aspect of life is free from government encroachment. A pervasive intolerance for individual rights is shown by government's arbitrary intrusions into private economic transactions and its disregard for civil liberties.

To counter that trend, the Cato Institute undertakes an extensive publications program that addresses the complete spectrum of policy issues. Books, monographs, and shorter studies are commissioned to examine the federal budget, Social Security, regulation, military spending, international trade, and myriad other issues. Major policy conferences are held throughout the year, from which papers are published thrice yearly in the *Cato Journal.* The Institute also publishes the quarterly magazine *Regulation*.

In order to maintain its independence, the Cato Institute accepts no government funding. Contributions are received from foundations, corporations, and individuals, and other revenue is generated from the sale of publications. The Institute is a nonprofit, tax-exempt, educational foundation under Section 501(c)3 of the Internal Revenue Code.

CATO INSTITUTE
1000 Massachusetts Ave., N.W.
Washington, D.C. 20001
www.cato.org

$15.00

The Cato Institute is proud to add the Cato Papers on Public Policy to its collection of high-quality publications. Other Research & Analysis Includes

REGULATION
THE CATO REVIEW OF BUSINESS AND GOVERNMENT

Four times a year since 1977, *Regulation* has offered immediately usable insights about regulatory policies from leading economists, policy analysts, and legal experts. *Regulation* guarantees the objective in-depth analysis needed to stay on top of regulatory and economic policymaking in Washington, D.C.

THE CATO JOURNAL
AN INTERDISCIPLINARY JOURNAL OF PUBLIC POLICY ANALYSIS

America's leading free-market public policy journal since 1981, the *Cato Journal* provides insightful and engaging analysis of key issues by leading scholars and policy analysts three times each year. Its topics run the gamut of policy issues from foreign policy and economic freedom to domestic issues like health care and education.

ISBN 978-1-935308-4-8-5
51500
9 781935 308485

Subscribe today by calling 1-800-767-1241 or visiting Cato.org/subscribe.

CATO INSTITUTE
www.cato.org