Economic Quarterly—Volume 97, Number 2—Second Quarter 2011—Pages 111–131

Should Central Banks Raise Their Inflation Targets? Some Relevant Issues

Bennett T. McCallum

The purpose of this article is to consider the merits and demerits of the recently renewed suggestion that central banks should, because of the difficulty of providing additional monetary stimulus when the policy interest rate is at its zero lower bound, raise their inflation rate targets—for example, from 2 percent per annum to 4 percent. As is well known, this suggestion has been put forth by several economists over the years,[1] but has recently attracted special prominence as the result of a working paper coauthored by Olivier Blanchard, who is not only a leading macroeconomist but is also currently serving as director of research of the International Monetary Fund. The article by Blanchard and coauthors (Blanchard, Dell'Ariccia, and Mauro 2010) does not explicitly promote this suggestion but discusses it in a distinctly sympathetic manner.

In considering the issue, one apparently needs to compare the magnitude of the benefits from occasionally being able to provide additional monetary stimulus against the costs of maintaining inflation at a higher value (on average) than would be chosen in the absence of the zero lower bound (ZLB). An extensive and sophisticated analysis relating precisely to this topic has recently been provided by Schmitt-Grohé and Uribe (2010), an article that will

The author is affiliated with Carnegie Mellon University and the National Bureau of Economic Research and is a visiting scholar at the Federal Reserve Bank of Richmond. An earlier version of this paper was presented at the Federal Reserve Bank of Boston Conference "Revisiting Monetary Policy in a Low Inflation Environment," held October 14–15, 2010, in Boston. For helpful discussion and comments, the author is indebted to Marvin Goodfriend, Joseph Gagnon, Edward Nelson, and Julio Rotemberg. In addition, the author has benefitted from reports with useful suggestions provided by Pierre Sarte, John Walter, and Anne Davlin. Opinions expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: bmccallum@cmu.edu.

[1] Frequently cited examples are Summers (1991) and Fischer (1996). Also see Williams (2009) and Yellen (2009).

be drawn upon heavily in what follows. Our discussion begins in Section 1 with the usual starting point for such matters, the analysis underlying Milton Friedman's "optimal quantity of money" result, often termed "the Friedman Rule."[2] Next, in Section 2 attention is turned to the type of distortion emphasized more prominently in the mainstream New Keynesian literature of recent years, namely, resource misallocations brought about by the existence of nominal price stickiness that, in each period, affects some sellers but not others. Section 3 reports on the Schmitt-Grohé and Uribe (2010) analysis of one key question, namely, whether a steady inflation rate greater than 2 percent would be optimal, as indicated by recent formal analysis, when account is taken of the ZLB. Section 4 is concerned with suggestions to the effect that when the ZLB is a constraint on the usual one-period policy interest rate, other variables such as exchange rates or longer-term interest rates could be effectively used as the instrument variable. Then in Section 5 our focus shifts to a line of argument that contends that the ZLB is not in fact a necessary bound, i.e., that with modified institutions it would not be impossible for central banks to provide monetary stimulus even when the basic one-period interbank rate is zero. Finally, in Section 6 we take up matters not considered to that point, ones having to do with the essential role of central banks and other related political-economy issues. Section 7 concludes briefly.

1. TRADITIONAL THEORY: TAXATION DISTORTIONS

Most monetary economists are familiar with the basic idea of the Friedman-rule analysis: Valuable transaction-facilitating services are provided in larger amounts by larger holdings of real money balances, which will be chosen by agents when the interest-opportunity cost of holding money is lower. This cost can be varied by varying the ongoing inflation rate, which can be adjusted by varying the rate of nominal money creation. Accordingly, since different rates of (paper) money creation do not require different rates of usage of tangible resources, the rate should be chosen that leads agents to satiate themselves with the transaction-facilitating services provided by holdings of money balances.[3] This rate prevails when the opportunity cost of holding money is zero, i.e., when the real rate of return on money holdings is equal to the real rate of return

[2] It has been my preference to refer to this result as the "Chicago rule" because (i) there is a different "Friedman rule" that stipulates that the "total stock of money. . . rises month by month, and indeed, so far as possible, day by day, at an annual rate of X percent, where X is some number between 3 and 5" (Friedman 1962, 54) and (ii) the criterion of satiating the holders of money with the transaction-facilitating services of money balances had been put forth by Tolley (1957), who proposed interest on reserves—not deflation to drive the interest rate to zero—as the operative mechanism. In any event, Friedman's first clear statement of the optimal inflation rate rule appears in Friedman (1960, 70), not Friedman (1969).

[3] This argument does not, of course, require a model specification that implies monetary superneutrality.

on other assets. Since the real rate of return on money equals zero minus the inflation rate, this condition will prevail when the inflation rate plus the real rate of return equals zero, i.e., when the nominal interest rate equals zero.

This result is developed more formally, and under several assumptions, by Schmitt-Grohé and Uribe (2010).[4] Their analysis develops some points that involve variants of the basic reasoning and are perhaps unfamiliar to some readers. One of these concerns the role of nondistortionary (i.e., lump-sum) taxes. Schmitt-Grohé and Uribe emphasize that the basic Friedman result requires that the fiscal authority must maintain a negative value for the basic fiscal deficit minus bond sales to the public if it is to continually reduce the money supply by enough to bring about a deflation rate (one that yields a zero nominal interest rate). Thus, the basic reasoning presumes that the fiscal authority has available to it some form of lump-sum taxation.

Alternatively, suppose that some government consumption is essential to optimality and that only non-lump-sum taxes on income (of various types) are available. Then it is often argued that an inflation tax (i.e., an inflation rate above the Friedman-rule magnitude) is necessary since optimality requires that the distortionary cost per unit of revenue raised must be the same at the margin for all utilized sources of taxation. Schmitt-Grohé and Uribe dispute this conclusion, associated with Phelps (1973), on the basis of a finding that, under certain specified conditions, the optimal inflation rate continues to equal the Friedman-rule magnitude even when some distorting taxes must be used to finance government consumption: They argue that "in contrast to Phelps's conjecture, negative inflation emerges as optimal even in an environment in which the only source of revenue available to the government, other than seigniorage revenue, is distortionary income taxation. Remarkably, the optimality of the Friedman rule obtains independently of the financial needs of the government. . . " (Schmitt-Grohé and Uribe 2010, 15). This interesting result apparently requires, however, the following assumptions: (i) exactly constant returns to scale in production, (ii) factors paid their marginal products, (iii) all factor incomes taxed at the same rate, and (iv) zero transaction costs for government consumption. The first three are interesting baseline assumptions, but (iv) seems unattractive: Are the resources used by government in shopping not valuable?

Before moving on, it is germane to point out some unorthodox opinions concerning the Friedman line of analysis that are expressed in the frequently cited piece by Summers (1991), mentioned earlier. Summers states: "I think

[4] Their initial and simplest statement is as follows: "In monetary models in which the only nominal friction takes the form of a demand for fiat money for transaction purposes, optimal monetary policy calls for minimizing the opportunity cost of holding money by setting the nominal interest rate to zero. This policy, also known as the Friedman Rule, implies an optimal rate of inflation that is negative and equal in absolute value to the real rate of interest" (Schmitt-Grohé and Uribe 2010, 1).

the view that inflation is costly is correct, but it has nothing to do with optimal tax theory. . . . A valid case for low inflation must have to do with the inefficiencies caused by allowing the monetary standard to vary and by the instability that results when the inflation trend is changed. Standard optimal tax issues along Ramsey lines are nth-order considerations. Inflation as a Ramsey tax may be the most overstudied issue in macroeconomics" (1991, 626–7). That I have some sympathy with one aspect of Summers's position may become apparent below.

2. MAINSTREAM NEW KEYNESIAN THEORY: CALVO-MODEL DISTORTIONS

In recent years, due in large part to the huge influence of Woodford's (2003) opus, less attention has been devoted to "nominal frictions" of the type discussed in the previous section, i.e., those having to do with the medium-of-exchange role of money.[5] Instead, the frictions focused upon pertain to posited stickiness of nominal prices of goods and, in some cases, labor. While other models of gradual price adjustment have been put forth,[6] the clear leader in this regard is the basic discrete-time version of Calvo (1983). As is well known, the stylized friction is that in each period only a fraction $1 - \omega$ of the economy's sellers, randomly selected, have the opportunity to change their prices, while others continue with the same prices as in the previous period. In any period, accordingly, there are sellers (of goods that have the same production cost functions) charging different prices in a setting of monopolistic competition. These features imply a misallocation of productive resources among the various sellers, which is the social cost of the nominal friction implied by the Calvo price specification. Only if monetary policy generates an average inflation rate that makes the average price of currently reoptimizing sellers equal to those of the other sellers (who are stuck with their previous prices) will this distortion be eliminated. Consequently, the optimal inflation rate in this environment (with no other nominal friction) equals zero.[7]

To consider the compromise between the Friedman-rule and Calvo-model optimal inflation rates, Schmitt-Grohé and Uribe adopt a specification that includes both types of friction and obtain results under various calibration assumptions. By and large, their results suggest that the tradeoff is such that

[5] Actually, there is an important sense in which these are "real," not nominal, frictions. That is, typical specifications posit that real transaction costs (either shopping time or real resource usage) are reduced when purchasers keep on hand additional quantities of money in real—not nominal—terms.

[6] The "sticky information" formulation of Mankiw and Reis (2002) and Reis (2009) has attained a fairly substantial following; my own favorite is discussed in McCallum (2008).

[7] For fleshed-out discussions see Woodford (2003, 392–419) and Schmitt-Grohé and Uribe (2010, 38–42).

the optimal rate is close to zero, i.e., that the importance of the price-setting friction is quantitatively greater, to a considerable extent, than the medium-of-exchange aspect featured in the Friedman-rule analysis. Their summary statement is: "We conclude that for plausible calibrations the price-stickiness friction dominates the optimal choice of long-run inflation" (Schmitt-Grohé and Uribe 2010, 51).

In this regard, I would like to suggest that there is one feature of the standard Calvo model that is crucial for this finding and that should be considered questionable at best. In particular, I would argue that the basic version of the Calvo model is flawed, as a model of optimal price setting with the assumed type of friction, via its assumption that those sellers, who do not have an opportunity to reoptimize in a given period, leave their prices at the value charged in the previous period. This might make sense in a world in which the steady-state inflation rate is zero, but if that rate was (say) X percent per period, it would seem that a rational pricing policy would call for each seller who cannot reoptimize to have his selling price automatically rise from its previous level by an amount that implies an X percent increase.[8] For an example of one formulation of this type, but extended to non-steady-state conditions, see the pricing behavior assumed in Woodford (2008; 1,566–8).[9] Under such a formulation, the average dispersion of prices will be unaffected by the steady-state inflation rate, so the social optimum depends only on Friedman-rule considerations. To demonstrate this, suppose that the price-adjustment relationship is written, as in Woodford (2008), as

$$\Delta p_t - \bar{\pi}_t = \beta \left(E_t \Delta p_{t+1} - \bar{\pi}_{t+1}\right) + \kappa \left(mc_t\right), \tag{1}$$

where Δp_t is inflation in period t, $\bar{\pi}_t$ is the period-t expected value of the ongoing inflation rate in the economy under consideration with the monetary policy rule under consideration,[10] and mc_t is the fractional deviation of marginal production cost in t from its steady-state value. Also $0 < \beta < 1$, and $\kappa > 0$. That the cost of inflation in the usual version of the Calvo model, in which the $\bar{\pi}_t$ and $\bar{\pi}_{t+1}$ terms do not appear, is proportional to the unconditional expectation $E\pi_t = E\pi_{t+1}$ can be seen as follows. From a steady-state

[8] These automatic price adjustments would have been arranged earlier, on the basis of existing information concerning the prevailing steady-state inflation rate.

[9] It is well known that Yun (1996), in an important early paper, utilized a price adjustment model that is somewhat similar to the basic Calvo model but in terms of deviations of inflation from steady-state values. More recently, specifications embodying the same basic idea as (1), i.e., that automatic price adjustments for sellers unable to reoptimize should be part of an optimizing price strategy in the face of price stickiness, have been extensively developed by Calvo, Celasun, and Kumhoff (2003) and Freedman et al. (2010), among others.

[10] Woodford (2008; 1,568) notes that $\bar{\pi}_t$ can be operationally viewed as the Beveridge-Nelson (1981) value of the "stochastic trend rate of inflation," i.e., a stochastic growth rate for the price level, which is well defined so long as the inflation rate is difference-stationary with an unconditional mean of zero for its first difference, which is here assumed. This value is ultimately given by the central bank's inflation target.

perspective we have

$$\Delta p = \beta \Delta p + \kappa \, (mc) \, ; \tag{2}$$

so $mc = [(1-\beta)/\kappa]\Delta p$, which departs from zero in proportion to the ongoing inflation rate. Thus, in that setup, the costs are minimized if $\Delta p = 0$. In the modified model (1), however, we have—since $E\Delta p_t = E\Delta p_{t+1} = E\Delta p$ and $E\pi_t = E\pi_{t+1}$—the steady-state relation is

$$(1-\beta)\,(0) = (1-\beta)\,(0) + \kappa E\,(mc_t)\,, \tag{3}$$

which implies that the average dispersion cost equals zero for whatever steady-state rate prevails. With this specification, then, the steady-state cost of inflation depends (in the absence of the ZLB) only upon the Friedman-rule "shoe-leather" cost occasioned by non-satiation with the services of the medium of exchange.

3. OPTIMALITY IN THE PRESENCE OF THE ZERO LOWER BOUND

At this point we turn to the ZLB issue more directly. One significant accomplishment of the Schmitt-Grohé and Uribe (2010) article is to consider, by means of simulations of a rich calibrated model under various assumptions, quantitative aspects of the optimal rate of inflation in economies with more than one nominal friction. Indeed, one section of Schmitt-Grohé and Uribe (2010) is entitled, "Does the Zero Bound Provide a Rationale for Positive Inflation Targets?" A quotation from the article may be useful in judging the nature of their study:

> We believe... this argument is best evaluated in the context of an empirically realistic quantitative model of the business cycle. In Schmitt-Grohé and Uribe (2007b) we study Ramsey optimal monetary policy in an estimated medium-scale model of the macroeconomy. The theoretical framework employed there emphasizes the importance of combining nominal as well as real rigidities in explaining the propagation of macroeconomic shocks. Specifically, the model features four nominal frictions, sticky prices, sticky wages, a transactional demand for money by households, and a cash-in-advance constraint on the wage bill of firms, and four sources of real rigidities, investment adjustment costs, variable capacity utilization, habit formation, and imperfect competition in product and factor markets. Aggregate fluctuations are driven by three shocks: a permanent neutral labor-augmenting technology shock, a permanent investment-specific technology shock, and temporary variations in government spending (2010, 52).

Schmitt-Grohé and Uribe explain how parameter values used in these exercises are obtained and offer plausible justification. The basic finding is that

"the Ramsey optimal policy implies a mean inflation rate of −0.4 percent per year.... Under the Ramsey optimal monetary policy, the standard deviation of the nominal interest rate is only 0.9 percentage points at an annual rate... [while] the Ramsey optimal level of the nominal interest rate is 4.4 percent... [implying that] for the nominal interest rate to violate the zero bound, it must fall more than 4 standard deviations below its target level" (2010, 53). In this regard, the quoted results are for an assumed time-preference rate of 0.03 per year, smaller than that implied by the 0.99 quarterly discount factor that is typically used in monetary policy studies. Moreover, "lowering the subjective discount factor... to 1 percent per year results in a Ramsey-optimal nominal interest rate process that... [implies that]... the nominal interest rate must still fall by almost three standard deviations below its mean for the zero bound to be violated" (2010, 53).

The point is, then, that the Schmitt-Grohé and Uribe analysis suggests that the ZLB constraint will be binding so rarely that these authors are led "to conjecture that in an augmented version of the model that explicitly imposes the zero bound constraint, the optimal inflation target would be similar to the −0.4 percent per year that is optimal" in their model (2010, 53). In support of that view, Schmitt-Grohé and Uribe comment on the results of Adam and Billi (2006), as follows: "These authors compute the optimal monetary policy in a simpler version of the new Keynesian model.... An advantage of their approach is that they take explicitly into account the zero bound restriction in computing the optimal policy regime. They find that the optimal monetary policy does not imply positive inflation on average and that the zero bound binds infrequently.... We conjecture... that should a money demand be added to their framework, the average optimal rate of inflation would indeed be negative" (2010, 54).[11]

Results of the type cited in this section are optimistic in that they do not offer much—if any—support to the idea that raising the inflation-target objective (and with it the average inflation rate) would be desirable. Unfortunately, however, they are inherently open to challenge and/or reinterpretation.

A significant problem, for example, is the absence from the Schmitt-Grohé and Uribe analysis of the distinction between one-period interbank rates of interest and one-period rates of the "risk-free" or "purely intertemporal" variety that is relevant for intertemporal decisions. That is, in analysis that recognizes a banking sector that uses resources to make loans that finance its money issues—the central bank supplying this sector with base money—the discrepancy between these interbank and risk-free rates can be quite large. In

[11] Schmitt-Grohé and Uribe (2010, 54) also cite results of Reifschneider and Williams (2000) that are compatible. Recently, Williams (2009) has, by contrast, discussed considerations that are more favorable with respect to the proposal of a raised inflation target. Even more recently, Billi (2010) has demonstrated that an inability of the central bank to commit can raise the optimal inflation rate considerably.

the calibration of Goodfriend and McCallum (2007), for example, the difference between the (real) rates is $6.0 - 0.84 = 5.16$ percent per annum (2007; 1,492).[12] Since it is the lower interbank rate that is relevant for the ZLB problem, whereas the Schmitt-Grohé and Uribe analysis implicitly refers to the risk-free rate, recognition of this distinction could completely overturn the optimistic presumption that the analysis described above suggests that the ZLB would be binding only rarely.[13]

4. ALTERNATIVE MONETARY STRATEGIES

Before moving on to more drastic proposals, I should mention some proposed strategies for monetary policy management in the face of the ZLB constraint, taking it for granted that such a constraint exists. Here the prevailing view seems to be that of Eggertsson and Woodford (2003), who show that the output loss from a temporary ZLB constraint can be lessened by the use of a "history dependent" rule for the one-period policy interest rate, designed in a manner that has the effect of implying that policy will be kept more stimulative in the future than would otherwise (i.e., without the temporary ZLB constraint) be the case.

Alternatively, it has been argued by Svensson (2001) and McCallum (2000) that monetary demand management can be conducted effectively under ZLB conditions by appropriate exchange rate policies. The idea is that one-period risk-free bonds and foreign exchange are not perfect substitutes, presumably for reasons stressed in the "portfolio balance" literature of the 1970s.[14] Central bank purchases of foreign exchange will, accordingly, tend to depreciate the country's exchange rate. The central bank could then exploit that relationship to manage the (nominal) exchange rate in accordance with a policy rule expressed in terms of an exchange rate instrument—with the rate of exchange rate appreciation appearing in place of the policy interest rate in a Taylor-style rule.[15] Of course, real exchange rate depreciation appears in the "expectational IS" portion of a typical New Keynesian open-economy model, so with sticky prices this mode of policy behavior can have

[12] Here the 6.0 figure comes from assumed values of a 4 percent per annum (p.a.) time preference rate and a 2 percent p.a. growth rate of population. The 0.84 percent p.a. figure is close to the 1 percent that Campbell (1999; 1,241) reports for the real three-month Treasury bill rate for the United States over 1947.2–1996.4. (Somewhat confusingly, Campbell refers to this as the "risk-free" rate since he is also assuming the absence of costly banking.)

[13] Schmitt-Grohé and Uribe (2010) provide interesting analyses of several topics not mentioned in the present article, including the effects of foreign demand for domestic currency and of incorrectly estimated inflation rates.

[14] See, for example, Dornbusch (1980).

[15] To apply this rule the central bank would not need to know the specification of the portfolio balance relation between foreign exchange purchases and the rate of depreciation, just as Taylor-rule central banks do not need to know money demand functions to be able to implement interest-rate policy rules.

systematic effects on real aggregate demand in the economy under discussion, even with the one-period interest rate immobilized at zero. Simulations reported in McCallum (2000, 2003), for example, indicate that substantial stabilization can be effected in this manner.[16] Also, if the economy in question is small in relation to the world, the policy will not have "beggar-thy-neighbor" effects.

An argument against this position might seem to be implied by Woodford's (2005) comment on the suggestion by McGough, Rudebusch, and Williams (2005) in "Using a Long-Term Interest Rate as the Monetary Policy Instrument." Specifically, Woodford criticizes the long-rate strategy and mentions that "similar comments apply to the proposal by Svensson (2003) that the exchange rate be used as the instrument of policy when an economy is in a 'liquidity trap.'"[17] The problem is that rules based on multiperiod interest rates (or on exchange rates) cannot expand the set of possibilities without driving the one-period rate into the negative (and therefore infeasible) range. That argument is, however, based on an assumed term-structure model in which the longer-term interest rates are related to one-period rates by a relationship that depends only upon expected yields, with no included "portfolio" terms involving quantities, such as those mentioned above. This same statement applies, moreover, to the uncovered interest parity relationship involving exchange rates. Thus, Woodford's argument apparently does not refute the one made above, which does presume the presence of portfolio-balance departures from the counterpart of the expectations theory as applied to exchange rates, i.e., uncovered interest parity.[18] In addition, the argument made here would apply also to the use of long-term domestic interest rates if the term-structure relationship involves relative quantities of different-maturity bonds. Emphasis on the exchange rate case would seem to imply a belief that foreign one-period bonds are more highly imperfect substitutes for domestic one-period bonds than are domestic long-term bonds.

The discussion to this point has been based on orthodox analysis with rather standard models, even if those in the present section involve portfolio-balance considerations that are not made explicit or quantified. It would appear to be the case, unfortunately, that all such arguments are unlikely ever to be conclusive—primarily because of the smallness of inflation costs in the range under discussion, the infrequency of ZLB situations, and the huge

---

[16] For alternative results in much the same spirit, see Coenen and Wieland (2003).

[17] Presumably, the same objection would apply to the closely related proposal in McCallum (2000).

[18] My argument is, nevertheless, open to the objection that quantitative magnitudes relating to portfolio balance effects have not been established; e.g., my (McCallum 2003) simulations simply assume that the exchange rate depreciations called for by the policy rule can be implemented.

number of alternative model features that could be considered.[19] Accordingly, policy conclusions by many economists will be—and arguably should be—based largely upon their informal views concerning the basic role of a central bank and the social value of having a currency whose purchasing power does not change substantially over time. Such considerations are briefly treated below, in Section 6. Before turning to them, however, it will be appropriate to recognize, in Section 5, a different line of argument concerning the ZLB issue, one that questions the implicit assumption maintained above that the ZLB is an immutable aspect of reality.

5. IS THE ZLB ACTUALLY A GENUINE BOUND?

We ask, then, is it actually the case that zero represents a lower bound on nominal interest rates? Of course the precise lower bound may be slightly negative because of the cost of storing money, as mentioned by McCallum (2000, 875) and others, but this magnitude is small enough to be neglected. That is not the matter here under discussion. Instead, our concern now is the validity of the argument, developed by Goodfriend (2000, 2001) and Buiter (2003, 2010), that, with modern technology, institutions can be designed so as to permit payment of negative nominal interest on all forms of money, thereby making it possible to have negative (as well as positive) rates for the central bank's policy rate, and thereby *eliminating*—rather than surmounting—the putative problem of the ZLB. In this regard, Citi Research (2010, 5), presumably influenced strongly by Buiter (2009), states that "there are at least three administratively and technically feasible ways to eliminate the zero lower bound on nominal interest rates completely. . . The first is to abolish currency. The second is to. . . start paying interest, positive or negative, on currency. The third is to. . . end the fixed exchange rate. . . between currency and bank reserves or deposits with the central bank."

The abolishment of currency seems like an extremely radical step—almost unimaginable—until one contemplates it somewhat calmly. My own attitude has been influenced by a rather trivial aspect of my own routine—lunch each day at my university. Only a few years ago, my regular lunch companions and I used cash to pay for our lunches at the Carnegie Mellon Faculty Club, and I was annoyed when someone in line ahead of us chose to pay by credit card and thereby slowed the process noticeably. Then a new system for accepting credit card payments was adopted by the cashier, and the time needed for a credit card transaction decreased sharply. Next, a couple of years ago, I realized that one of my companions had adopted a routine of paying by credit card—and that this apparently involved no extra time at all. Finally, a

[19] It is, perhaps, this type of consideration that Summers (1991) had in mind in his provocative statement quoted in Section 1.

few months ago, I realized that all of my regular companions had switched to credit card payment as their usual mode of transaction—and that each of them was taking less of the cashier's time (and that of other customers) than I was imposing each day with my cash transaction! A second recognition was that taxi cabs now typically have facilities for accepting credit card payments, thereby eliminating an example that I used to mention in undergraduate classes as transactions for which one needed to carry cash.

More generally, I have been impressed by the point that approximately 75 percent (by value) of U.S. currency outstanding consists of $100 bills. These are notes of the largest denomination available, of course—which are of greatest use to "...the underground economy, the criminal community, that is, those engaged in tax evasion, money laundering and the financing of terrorism, and those wishing to store the proceeds from crime and the means to commit further crimes out of sight and reach of the authorities" (Buiter 2010, 224). In the case of the euro, 59 percent of the value of euro notes outstanding in April 2009 was in the denominations of 100, 200, or 500 euros, while less than 10 percent of the stock value was in the form of 5, 10, and 20 euro notes (Buiter 2010, 223). Partly on the basis of these facts, Buiter develops a strong argument for the elimination of (government) currency. An important part of the argument is the suggestion, made in Goodfriend (2000, 224), that the central bank make available free transaction accounts to all legal residents, accounts that could be administered through "commercial banks, post offices, and other retail facilities." In that case, it would not be true that the institutional change would be devastating for the poorer members of the (legal) population.

A second approach would involve taxation of currency. Buiter (2009) stresses, however, that there are inherent problems with the administration of positive tax rates (i.e., negative interest rates) on negotiable bearer instruments that sharply reduce the attractiveness of this approach. Goodfriend (2000; 1,016) has suggested that "...a carry tax could be imposed on currency by imbedding a magnetic strip in each bill. The magnetic strip could visibly record when a bill was last withdrawn from the banking system... [with a tax] deducted from each bill upon deposit according to how long the bill was in circulation since last withdrawn...." Perhaps such a system could become viable in the future, but with today's technology it would appear excessively expensive.

A third approach of Buiter's is to unbundle—divorce—the medium of exchange (MOE) and the medium of account (MOA). The MOE consists in part of currency and claims to currency; the MOA is the entity in terms of which prices are quoted. Governments do not invariably have full control over either of these, but can retain control over the MOE if government currency is not issued to excess. Furthermore, by requiring that transactions with the government must be denominated in terms of an appointed MOA, it can most

likely gain acceptance for its choice of the latter. Then in each period it can specify interest rates for both, with the MOE interest rate kept non-negative but with no such stipulation for the MOA rate, by issuing bonds in terms of both media. Then the central bank can conduct policy in terms of its instrument, the MOA interest rate. If prices in terms of this MOA are the prices that are relevant for market supplies and demands, then the central bank will continue to be able to influence aggregate demand by variations in the policy interest rate even when the MOE rate is immobilized at zero.

Buiter (2009) devotes many words to analysis of this third approach, but it seems that his preference is probably for the abolition of currency.[20] Actually, it should be said, it is the abolition of a government-issued currency that Buiter and Goodfriend have in mind. Both evidently would favor regulations that would not rule out the possibility of private issuers attempting to put their own currency-like vehicles into circulation.[21]

In any event, it would seem entirely appropriate that serious thought be given to the Buiter and Goodfriend proposals, if it transpires that the ZLB constraint is more of a problem than the Schmitt-Grohé and Uribe analysis suggests—or simply to prepare for possible future developments.

6. THE DUTIES OF A CENTRAL BANK

Before the financial crisis of 2008–2009, monetary economists had become rather proud of the development of their subject over the preceding 10–15 years. There had been great progress in formal analysis and also in the actual conduct of monetary policy. Analytically, the profession developed an approach to policy analysis that centers around a somewhat standardized dynamic model framework that is designed to be structural—that is, respectful of both theory and evidence—and therefore usable in principle for policy analysis. This framework includes a policy instrument that agrees with the one typically used in practice and recognizes that, for imperfectly understood reasons, nominal price adjustments do not take place immediately, in which case monetary policy actions will have significant consequences for the behavior of real aggregate variables such as output and employment. Indeed, models

[20] In this regard, Goodfriend has remarked in conversation that "currency is the most unsanitary object that most of us handle on a regular basis."

[21] Before continuing, it should be noted that there are similarities but also (crucial) differences between Buiter's third approach and what I would term the Yeager-Greenfield system. The latter has been developed in a number of articles by Leland Yeager (1983, 1992), plus others that are co-authored with Robert Greenfield (Greenfield and Yeager 1983). A brief discussion is provided in McCallum (2010). One major difference is that the Yeager-Greenfield system was originally designed as one intended to eliminate, or reduce as far as possible, governmental influence on monetary affairs. A second is that a major objective of the Yeager-Greenfield system is to achieve "stability," in the sense of constancy through time, of the price level, whereas Buiter's approach is more concerned with avoidance of recessions.

of this type were being used (in similar ways) by economists in both academia and in central banks, where several economic researchers had gained leading policymaking positions. Meanwhile, in terms of practice, most central banks had been much more successful than in previous decades in keeping inflation low while also avoiding major recessions (with a few exceptions) prior to 2008. Furthermore, these improvements in science and application had been interrelated: The "inflation targeting" style of policy practice that had been adopted by numerous important central banks—and that arguably had been practiced unofficially by the Federal Reserve—is strongly related in principle to the prevailing framework for analysis.[22] Accordingly, one keystone of the "consensus" view was that central bank control of the inflation rate is the central ingredient in successful monetary policy practice, and that this control called both for inflation "stability," in the sense of little variation from year to year, and for a low average level—often in the range of 1 percent to 2 percent per annum. The crisis has, however, damaged—if not destroyed—that consensus; the Blanchard, Dell'Ariccia, and Mauro (2010) article is evidence of that.

It would seem, however, that the recent crisis is highly inappropriate as a centerpiece for reconsideration of an economy's monetary policy. To a considerable extent, the crisis was precipitated by events in the United States. There the primary root of the crisis was, arguably, a genuine macroeconomic imbalance that required correction, namely, the housing price boom. What were its origins? The situation in housing was largely brought about by deliberate government action designed to stimulate homeownership even among—actually, especially among—families that could not afford homeownership.[23] This sectoral imbalance was then turned into a macroeconomic collapse by unwise regulations and practices in financial markets that led to the freezing-up of the latter. In that regard, numerous practices of private enterprises in the financial industry may have been appalling, but again much of the problem can be traced back to an unwise governmental framework—one prominent example being regulations that gave undue importance to the ratings of a few private firms in the credit-rating industry. The point is that none of these failures had much, if anything, to do with monetary policy.[24] To drastically alter the objectives of monetary policy in response to the crisis would seem, accordingly, to be lacking in logic.

---

[22] For an exposition that discusses this development, by an author who participated both as researcher and policymaker, see Goodfriend (2007).

[23] For short discussions, see Pinto (2010) and Wallison (2010).

[24] John Taylor (2009) has argued that monetary policy was unduly expansionary during the period 2003–2005 and that this mistake was an important cause of the crisis. I would agree that policy was inappropriate in that manner, but consider this policy mistake less egregious than the other items mentioned in the present paragraph. In any event, Taylor's argument would certainly appear to provide no support for an increase in the target inflation rate!

From a more general perspective, some lack of clarity about the monetary policy duties of a central bank has resulted from the drastic change in monetary arrangements—from metallic standards to fiat money arrangements—that took place during the 20th century. Under a metallic standard, a nation's central bank has basically no price-level duties so long as the standard does not break down. Behavior of the price level is governed primarily by the mint, whereas the central bank is just that—an intermediary intended to facilitate the financial activities of the nation's government. Under a fiat money arrangement, by contrast, price level trends are determined by the abundance of money in circulation relative to the quantity needed (i.e., useful) for conducting transactions, and modern central banks have been universally assigned the duty of price level management. For example, in the *Journal of Economic Literature's* recent "panel discussion" of Federal Reserve duties by Blinder (2010) and Feldstein (2010), both contributors take it for granted that central banks will be the makers of monetary policy and argue that they should have extensive independence in that role.[25]

In my opinion, one important justification for central bank independence is that generally—except in ZLB situations—the desirable effects of monetary policy loosening occur rapidly whereas the undesirable effects materialize only after a greater lag. When the central bank eases policy—i.e., makes monetary conditions more stimulative and aggregate demand stronger—the socially desirable effects arrive more promptly than do the undesirable effects. That is, there will normally be effects that can be thought of as expansions of output and employment (relative to what would have prevailed in the absence of the policy change) that will begin to occur within two or three months. Then after one or two years there will also occur upward pressures on the inflation rate. If instead the policy action is one that tightens policy, rather than loosening it, there will be relatively prompt reductions of output and employment, followed in a year or so by reductions in the inflation rate. Not surprisingly, it is the case that most economists, congressmen, commentators, and ordinary citizens consider expansions in the level of employment and output to be desirable and increases in the inflation rate to be undesirable. Accordingly, if monetary policy is required to be politically attractive, there is a tendency for policy to be more expansionary and inflationary the more *impatient* is the policymaker, i.e., the shorter is his effective time horizon. One way to avoid policies that give primary emphasis to short-run considerations is to place responsibility for monetary policy in an institution that is somewhat

[25] Consequently, the panel discussion referred to is mostly concerned with the regulatory responsibilities of a central bank. For an ambitious recent proposal for monetary policy strategy, see Goodfriend (2011).

sheltered from the stresses of day-to-day politics, and consequently able to take a longer-term perspective.[26]

An important ingredient in such a perspective is the understanding that there exists no usable long-run tradeoff between inflation and unemployment (or output)—i.e., that some version of the "natural rate hypothesis" is valid. Moreover, a major contribution of the "consensus" position of mainstream monetary economics that evolved in the 10–15 years prior to 2008 was the development of models that incorporated this natural-rate feature[27] while also reflecting the property that monetary policy has substantial short-term effects on the behavior of output and employment.[28]

But to adopt the position that the average ongoing inflation rate should be raised (as it certainly would be if the target were raised) in order to prevent or shorten recessions involving the ZLB, is to accept the notion that there does exist a type of long-run tradeoff that is usable and well-understood. It is based on a different mechanism than the Phillips Curve tradeoff, but in public debate and actual policy consideration this distinction would be lost. Thus, it might serve to overturn a basic message that the profession has been at great pains to present to policymakers, namely, that the overriding objective of monetary policy should be the prevention of inflation (positive or negative). This is the one important macroeconomic goal that the central bank—and only the central bank—has the power to deliver. The best thing that the central bank can do for employment and output, on a sustained basis, is to keep inflation close to a low and clearly specified target value. To some extent, this is an argument based on considerations of "communication," not science, but is nevertheless of great practical importance.

To some readers a move to a 4 percent inflation rate may seem entirely innocuous. To emphasize the contrary possibility, let us ask the following question: What would be the United States price level now, in November 2010, if a steady 4 percent inflation rate had prevailed since 1792, the year in which a United States monetary standard was first established?[29] Since $2{,}010 - 1{,}792 = 218$, the price level today would be $1.04^{218} = 5{,}167.3$ times the price level of 1792 if inflation had been 4 percent each year. In fact, the actual

---

[26] One obvious difference between central bankers and legislators is that typically the former are, by design, not elected officials and are appointed for rather lengthy terms.

[27] Actually, with the basic Calvo model of price adjustment, these models do not quite have the strict natural-rate property. The modification promoted in Section 5 does, however, satisfy a non-strict version as discussed by Andrés, López-Salido, and Nelson (2005).

[28] These effects are, however, poorly understood and are dependent on current values of the "output gap," which is not directly observable.

[29] Under the Articles of Confederation, the states did not share a national monetary standard. Implementation of the Constitutional provisions regarding money began with the Coinage Act of 1792.

consumer price index (CPI) today is only 23.54 times as high as in 1792.[30] Consequently, if a 4 percent inflation rate had prevailed since 1792, prices today would be 219.5 times as high as they actually are, on average. Of course, I would have to admit that in terms of pure economic analysis, this last fact alone is devoid of significance. At the same time, I believe (perhaps somewhat schizophrenically) that many citizens, even well-educated ones, are frequently confused in thinking about issues relating to inflation.[31] That being the case, it would seem highly desirable for a monetary system to have the property of being easy for an average citizen to understand and cope with. Under current conditions, a significant fraction of measured gross domestic product consists of the activities of persons seeking to profit from other individuals' lack of understanding of the causes and effects of inflation. A general, if somewhat elusive, discussion that emphasizes the medium-of-account role of money is provided by Niehans (1978, 118–31).

Finally, I would argue that in the United States, and also in many other countries, central banks have shown themselves in recent years to be the primary—indeed, only visible—source of intertemporal discipline in fiscal affairs. The point is that the overall government budget constraint implies that if the central bank maintains a low growth rate of the monetary base, it limits the extent to which the fiscal authority is able to engage in deficit finance.[32] If the Treasury seeks to exceed this limit by means of (excessive) borrowing (selling bonds), it will run into a constraint reflecting the implied violation of a transversality condition relevant for optimal behavior for private lenders. In this context, a switch to a higher target inflation rate would evidently represent one more move away from intertemporal discipline, a position that many economists would want to avoid.

7. CONCLUSION

A summary of the article's arguments can be presented briefly, as follows. First, in the absence of the ZLB, the optimal steady-state inflation rate, according to standard New Keynesian reasoning, lies somewhere between the

[30] The CPI, on the basis of a 100 value for 1982–1984, is reported by Measuring Worth (2010) to have equaled 9.72 in 1792, whereas the June 2010 value reported by the St. Louis Fed's FRED is 218.2.

[31] My own mother, who was the author of a well-respected work in U.S. history that was kept in print for two or three decades by a reputable university press, would occasionally express doubts that the inflation rate had recently fallen by stating that, "I know that [specific item] costs more now than it did at the same store a year ago."

[32] The fiscal deficit is identically equal to the amount of government revenue provided by bond sales plus the "inflation tax" revenue resulting from money issuance. The contention in the text presumes that the central bank is in fact given control of the monetary base, even when its desires conflict with those of the ministry of finance. It is my impression that this is the appropriate assumption for the United States and most other developed economies.

Friedman-rule value of deflation at the steady-state real rate of interest and the Calvo-model value of zero, with careful calibration indicating that the weight on the latter may be considerably larger. Second, however, an attractive modification of the Calvo model would imply that the weight on the second of these values should be zero, so that the Friedman-rule prescription would be optimal (in the absence of the ZLB). Third, even when the effects of the ZLB are added to the analysis, the optimal inflation rate is (according to this line of reasoning) probably negative. Fourth, there is probably some scope for activist monetary policy to be effective (via, e.g., an exchange rate channel) even when the one-period nominal interest rate is at the ZLB; but there exists professional disagreement on this matter. Fifth, while the ZLB is a genuine constraint under present institutional arrangements, these are not immutable. Elimination of traditional currency could be effected, in which case there would be no zero lower bound on one-period nominal interest rates and therefore no reason involving such losses for having an increased target rate of inflation. Sixth, increasing the target inflation rate for the purpose of avoiding occasional ZLB difficulties would constitute reversal of a central message, of recent monetary policy analysis, to the effect that there is no long-run benefit in terms of output or employment from the adoption of increased inflation rates. Seventh, such an increase in the target inflation rate would constitute an additional movement away from intertemporal discipline.

REFERENCES

Adam, Klaus, and Roberto M. Billi. 2006. "Optimal Monetary Policy under Commitment with a Zero Bound on Nominal Interest Rates." *Journal of Money, Credit and Banking* 38 (October): 1,877–905.

Andrés, Javier, J. David López-Salido, and Edward Nelson. 2005. "Sticky-Price Models and the Natural Rate Hypothesis." *Journal of Monetary Economics* 52 (July): 1,025–53.

Beveridge, Stephen, and Charles R. Nelson. 1981. "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle.'" *Journal of Monetary Economics* 7 (2): 151–74.

Billi, Roberto M. 2010. "Optimal Inflation for the U.S. Economy." Federal Reserve Bank of Kansas City Working Paper 07-03.

Blanchard, Olivier J., Giovanni Dell'Ariccia, and Paolo Mauro. 2010. "Rethinking Macroeconomic Policy." IMF Staff Position Note, International Monetary Fund. Also *Journal of Money, Credit and Banking* 42 (S1): 199–215.

Blinder, Alan S. 2010. "How Central Should the Central Bank Be?" *Journal of Economic Literature* 48 (March): 123–33.

Buiter, Willem H. 2003. "Helicopter Money: Irredeemable Fiat Money and the Liquidity Trap." Working Paper 10163. Cambridge, Mass.: National Bureau of Economic Research (December).

Buiter, Willem H. 2009. "Negative Nominal Interest Rates: Three Ways to Overcome the Zero Lower Bound." *The North American Journal of Economics and Finance* 20 (December): 213–38.

Calvo, Guillermo. 1983. "Staggered Prices in a Utility-Maximizing Framework." *Journal of Monetary Economics* 12 (September): 383–98.

Calvo, Guillermo A., Oya Celasun, and Michael Kumhof. 2003. "A Theory of Rational Inflationary Inertia." In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, edited by P. Aghion, R. Frydman, J. Stiglitz, and M. Woodford. Princeton, N.J.: Princeton University Press, 87–117.

Campbell, John Y. 1999. "Asset Prices, Consumption, and the Business Cycle." In *Handbook of Macroeconomics*, Vol. 1, edited by John B. Taylor and Michael Woodford. Amsterdam: Elsevier, 1,231–303.

Citi Research. 2010. "The Case for Raising the Inflation Target." Citigroup Global Markets, Global Macro View. Available at http://bx.businessweek.com/global-business/global-macro-view-05-march-2010-by-citi-research/6234371355931297299-fd3b1bf6b0e527571bc3eda957053c32/.

Coenen, Gunter, and Volker Wieland. 2003. "The Zero-Interest-Rate Bound and the Role of the Exchange Rate for Monetary Policy in Japan." *Journal of Monetary Economics* 50 (July): 1,071–101.

Dornbusch, Rudiger. 1980. "Exchange Rate Economics: Where Do We Stand?" *Brookings Papers on Economic Activity* 11 (1): 143–206.

Eggertsson, Gauti B., and Michael Woodford. 2003. "The Zero Bound on Interest Rates and Optimal Monetary Policy." *Brookings Papers on Economic Activity* 34 (1): 139–235.

Feldstein, Martin. 2010. "What Powers for the Federal Reserve?" *Journal of Economic Literature* 48 (March): 134–45.

Fischer, Stanley. 1996. "Why Are Central Banks Pursuing Long-Run Price Stability?" Federal Reserve Bank of Kansas City *Proceedings*, 7–34.

Freedman, Charles, Michael Kumhof, Douglas Laxton, Dirk Muir, and Susanna Mursula. 2010. "Global Effects of Fiscal Stimulus during the Crisis." *Journal of Monetary Economics* 57 (July): 506–26.

Friedman, Milton. 1960. *A Program for Monetary Stability*. New York: Fordham University Press.

Friedman, Milton. 1962. *Capitalism and Freedom*. Chicago: University of Chicago Press.

Friedman, Milton. 1969. "The Optimum Quantity of Money." In *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine Publishing Co., 1–50.

Goodfriend, Marvin. 2000. "Overcoming the Zero Bound on Interest Rate Policy." *Journal of Money, Credit and Banking* 32 (November): 1,007–35.

Goodfriend, Marvin. 2001. "Financial Stability, Deflation, and Monetary Policy." Bank of Japan *Monetary and Economic Studies* 19 (February): 143–67.

Goodfriend, Marvin. 2007. "How the World Achieved Consensus on Monetary Policy." *Journal of Economic Perspectives* 21 (Fall): 47–68.

Goodfriend, Marvin. 2011. "Central Banking in the Credit Turmoil: An Assessment of Federal Reserve Practice." *Journal of Monetary Economics* 58 (January): 1–12.

Goodfriend, Marvin, and Bennett T. McCallum. 2007. "Banking and Interest Rates in Monetary Policy Analysis: A Quantitative Exploration." *Journal of Monetary Economics* 54 (July): 1,480–507.

Greenfield, Robert L., and Leland B. Yeager. 1983. "A Laissez-Faire Approach to Monetary Stability." *Journal of Money, Credit and Banking* 15 (August): 302–15.

Mankiw, N. Gregory, and Ricardo Reis. 2002. "Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *The Quarterly Journal of Economics* 117 (November): 1,295–328.

McCallum, Bennett T. 2000. "Theoretical Analysis Regarding a Zero Lower Bound on Nominal Interest Rates." *Journal of Money, Credit and Banking* 32 (November): 870–904.

McCallum, Bennett T. 2003. "Japanese Monetary Policy, 1991–2001." Federal Reserve Bank of Richmond *Economic Quarterly* 89 (Winter): 1–31.

McCallum, Bennett T. 2008. "Reconsideration of the P-Bar Model of Gradual Price Adjustment." *European Economic Review* 52 (November): 1,480–93.

McCallum, Bennett T. 2010. "Alternatives to the Fed?" *Cato Journal* 30 (Fall): 439–49.

McGough, Bruce, Glenn D. Rudebusch, and John C. Williams. 2005. "Using a Long-Term Interest Rate as the Monetary Policy Instrument." *Journal of Monetary Economics* 52 (July): 855–79.

Measuring Worth. 2010. "The Annual Consumer Price Index for the United States, 1774–2008." Available at www.measuringworth.org/datasets/uscpi/result.php.

Niehans, Jurg. 1978. *The Theory of Money*. Baltimore, Md.: Johns Hopkins University Press.

Phelps, Edmund S. 1973. "Inflation in the Theory of Public Finance." *The Swedish Journal of Economics* 75 (March): 67–82.

Pinto, Edward. 2010. "The Future of Housing Finance." *The Wall Street Journal*, August 17.

Reifschneider, David, and John C. Williams. 2000. "Three Lessons for Monetary Policy in a Low-Inflation Era." *Journal of Money, Credit and Banking* 32 (November): 936–66.

Reis, Ricardo. 2009. "A Sticky-Information General Equilibrium Model for Policy Analysis." In *Monetary Policy Under Uncertainty and Learning*, edited by Klaus Schmidt-Hebbel and Carl E. Walsh. Santiago: Central Bank of Chile, 227–83.

Schmitt-Grohé, Stephanie, and Martín Uribe. 2010. "The Optimal Rate of Inflation." Working Paper 16054. Cambridge, Mass.: National Bureau of Economic Research (June). Also *Handbook of Monetary Economics*, Vol. 2, edited by Benjamin Friedman and Michael Woodford. Amsterdam, North-Holland Pub. Co.

Summers, Lawrence H. 1991. "How Should Long-Term Monetary Policy Be Determined?" *Journal of Money, Credit and Banking* 23 (August): 625–31.

Svensson, Lars E. O. 2001. "The Zero Bound in an Open Economy: A Foolproof Way of Escaping from a Liquidity Trap." Bank of Japan *Monetary and Economic Studies* 19 (February): 277–312.

Svensson, Lars E. O. 2003. "Escaping from a Liquidity Trap and Deflation: The Foolproof Way and Others." *Journal of Economic Perspectives* 17 (Fall): 145–66.

Taylor, John B. 2009. *Getting off Track: How Government Actions and Interventions Caused, Prolonged, and Worsened the Financial Crisis*. Stanford, Calif.: Hoover Institution Press.

Tolley, George S. 1957. "Providing for Growth of the Money Supply." *The Journal of Political Economy* 65: 465–85.

Wallison, Peter J. 2010. "Government Housing Policy and the Financial Crisis." *Cato Journal* 30 (Spring/Summer): 397–406.

Williams, John C. 2009. "Heeding Daedalus: Optimal Inflation and the Zero Lower Bound." *Brookings Papers on Economic Activity* 40 (Fall): 1–49.

Woodford, Michael. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton, N.J.: Princeton University Press.

Woodford, Michael. 2005. "Comment on: 'Using a Long-Term Interest Rate as the Monetary Policy Instrument.'" *Journal of Monetary Economics* 52 (July): 881–7.

Woodford, Michael. 2008. "How Important is Money in the Conduct of Monetary Policy?" *Journal of Money, Credit and Banking* 40 (December): 1,561–98.

Yeager, Leland B. 1983. "Stable Money and Free-Market Currencies." *Cato Journal* 3 (Spring): 305–33.

Yeager, Leland B. 1992. "Toward Forecast-Free Monetary Institutions." *Cato Journal* 12 (Spring/Summer): 53–80.

Yellen, Janet L. 2009. "Financial Markets and Monetary Policy." Panel discussion for the Federal Reserve Board/JMCB Conference, Washington, D.C. Available at http://frbsf.org/news/speeches/2009/0605.html (June 5).

Yun, Tack. 1996. "Nominal Price Rigidity, Money Supply Endogeneity, and Business Cycles." *Journal of Monetary Economics* 37 (April): 345–70.

Economic Quarterly—Volume 97, Number 2—Second Quarter 2011—Pages 133–152

Financial Firm Resolution Policy as a Time-Consistency Problem

Borys Grochulski

The financial crisis and the recession of 2007–2009 have shown the importance of government regulation and intervention in the financial services sector. During the crisis, governments in Europe and North America, among others, implemented a large variety of actions intended to mitigate the adverse impact of the financial sector disruptions on the macroeconomy as a whole. Soon after, broad-ranging reforms of the government oversight and regulation policies of the financial sector were introduced. One of the central objectives of these reforms is to improve government policy toward large financial institutions that face, or are in, the state of insolvency.

The problem of optimal design of government policy toward financial institutions in distress is very complex. This article is devoted to discussing one important aspect of this problem: the issue of time consistency. Our main objective in this article is to provide an elementary exposition of a fundamental economic insight of Kydland and Prescott (1977), which is that an ex ante optimal policy may require the government to tolerate inefficient outcomes ex post. In the context of policy toward insolvent institutions, this means that optimal policy may require the government to refrain from bailing out a firm despite the large adverse consequences that the firm's failure may have for the macroeconomy as a whole. To make this point, in the first part of this article, we build a simple model with a time-consistency problem associated with the government bailout policy toward large (systemically important) financial institutions.

The author would like to thank Kartik Athreya, Arantxa Jarque, Sabrina Pellerin, and Ned Prescott for their helpful comments. The views expressed in this article are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: borys.grochulski@rich.frb.org.

As a matter of practice, perhaps because of political-economy constraints, it may be impossible for a future, benevolent government to tolerate large adverse consequences of its inaction in a state of economic crisis, even if there are good reasons ex ante (prior to a possible crisis) to promise to not act ex post. As private-market participants are aware of this intolerance, the government will be expected to bail out large insolvent firms in order to protect the macroeconomy in crisis. This expectation may give private-sector investors an incentive to take on excessive risks, because large losses that excessive risk-taking can generate may be socialized, i.e., borne in part by the taxpayer, while large gains that can be realized will generally be retained by the investors. Recognizing this incentive, the government should take action ex ante to eliminate excessive risk-taking before the crisis state is induced.

In the second part of this article, we use the simple framework of our model to discuss some policies that have, in various formats, been proposed as means for either diminishing the probability of the next crisis or decreasing the severity of it. In particular, we discuss the following five types of mitigating measures: changes to the financial sector infrastructure that diminish the size of the spillover effects; direct government monitoring of risk-taking in the financial sector; regulations banning management and employee compensation practices consistent with firms' seeking excessive risks; levying a tax on extraordinary profits that may be attained in the financial sector under excessive risk-taking; and, finally, imposing binding capital requirements on financial firms. Given the simple structure of our model, we can discuss these measures only at a general, abstract level. Such a discussion, however, can be useful in organizing thoughts relevant to answering the concrete questions that policymaking and rule-writing authorities must confront in the face of the time-consistency problem of optimal government regulation of the financial sector.

Our analysis shows that government resolution policy toward large financial firms can be helpful in eliminating excessive risk-taking only to the extent that it decreases the negative spillover effect caused by failure of a large financial firm. Moreover, in order to have any effect, plans for ex post actions, like the resolution regime, must reduce the spillover effect to a level tolerable by a future government facing a state of crisis. Changes in policy that only marginally decrease the spillover effect do not have any impact on the equilibrium outcome. It is also clear in our model that the resolution regime should not be viewed as a commitment device for the government. Although institutional structure can have an effect on final outcomes, if future governments lack commitment, it will be impossible to achieve the commitment outcome by having the current government legislate that no bailouts are to be had in the future. Thus, the only way to eliminate inefficient bailouts is to remove the temptation for future (benevolent) governments to assume private investors'

losses. An efficient resolution mechanism can do it only if it can eliminate the negative spillover.

Assuming that spillover effects cannot be completely eliminated, a combination of binding capital requirements and monitoring of risk-taking and managerial incentives seem necessary to eliminate the inefficient bailout equilibrium. These policies can achieve the efficient outcome not by giving the government the power to commit, but rather by preventing the private sector from taking advantage of the government's lack of that power. In this way, current restrictions on the private sector's actions prevent damaging spillovers and, hence, remove future governments' temptation to use public funds to cover private losses. To be sure, capital requirements and monitoring of risk-taking are costly. Yet, these costs should be weighed against the cost of allowing the private sector to take on risks large enough to induce the next financial crisis.

The results of Kydland and Prescott (1977) spurred a large literature on the economics of optimal government policy with time-consistency constraints. This literature primarily focuses on fiscal and monetary policy of the government. Important contributions to the analysis of these problems include Barro and Gordon (1983); Lucas and Stokey (1983); Chari and Kehoe (1990); Chari, Christiano, and Eichenbaum (1998); and King and Wolman (2004). Beyond the applications to fiscal and monetary policy, Cochrane (1995) studies the provision of long-term health insurance and King (2006) discusses the time-consistency problem in the context of government policy toward floodplain development and the provision of insurance against catastrophic events.

An article most directly related to the topic of the present article is Chari and Kehoe (2009). Like we do here, they study the problem of time consistency of government bailout policy when bankruptcy is costly ex post. In addition to this problem, Chari and Kehoe (2009) simultaneously address the question of why optimal contracts within the firm lead to costly bankruptcies. Our model provides an exposition of the time-consistency issue at a more elementary level. In particular, we make strong assumptions on investors' preferences that give us a simple form of the firm's capital structure (debt versus equity financing). Also, we model the spillover effects of the firm's failure in reduced form.

This article is organized as follows. Section 1 presents the model. Section 2 studies equilibrium with government policy choices restricted to the singe option: bailout or firm failure. The time inconsistency of optimal policy is presented there. Section 3 considers additional policy tools and their potential for eliminating time inconsistency. Section 4 concludes.

1. THE MODEL

Consider a financial institution (firm) with total liabilities normalized to one. The firm's financial structure will be modelled as consisting of debt with face

value $1 - k$ and equity in the amount $k < 1$. (To keep a concrete number in mind, we can think of k as being equal to 0.05.)

There are three homogenous classes of agents in the model: financial institution equityholders, financial institutions debtholders, and the government. Uncertainty is represented in the model simply by two equally likely states of nature: a good state g, and a bad state b.

There are two possible projects that the institution can invest in: a prudent project P and a risky project R. Each project takes an up-front investment of size normalized to one. Given the funds available to the firm, only one project can be funded. The funded project represents the asset side of the firm's balance sheet.

The payoff structure of the projects is as follows. The prudent project P pays $1+k$ in state g, and $1-k$ in state b. Note that the expected return on P is $\mathbb{E}[P] = \frac{1+k}{2} + \frac{1-k}{2} = 1$. Thus, discounting with zero net interest rate, the net present value (NPV) of P is zero. Also note that the return on P is sufficient to cover the firm's debt face value in every state of nature. The risky project R pays $2 - \delta$ in state g, and 0 in state b, where $\delta > 0$. The expected return on R is $\mathbb{E}[R] = \frac{0}{2} + \frac{2-\delta}{2} = 1 - \frac{\delta}{2} < 1$, i.e., the NPV of R is negative. We will assume $2 - \delta > 1 + k$, i.e.,

$$1 - \delta > k, \tag{1}$$

which means that R has a higher best-case-scenario payoff than P. (For example, we could think of δ as being equal to 0.8.) Because both projects pay more in state g then they do in state b, we will call states g and b, respectively, good and bad.

We will assume that bond investors are risk-averse with preferences given by $-I + \min\{D_b, D_g\}$, where I is the amount invested and D_b and D_g are the returns on the investment I in state b and g, respectively. Thus, bond investors value only the riskless part of the state-contingent return vector (D_g, D_b). Given these preferences, the firm must provide a riskless gross return equal to one in order to float debt. Under the prudent project P, the firm's assets are worth at least $1 - k$ in every state of nature, but in state b they are not worth more than that. Thus, $1 - k$ is the largest face value of riskless debt that the firm should be able to float under project P without any anticipation of government bailout.

Equity investors are risk-neutral, discount at the same net interest rate of zero, and seek to maximize the return on equity.

The government has a loss function with penalty proportional to the losses suffered by debt investors, in case they suffer any. This loss function captures the adverse effects that an event of default of the financial institution would have on the broader economy. For every dollar lost by the bond investors, the adverse effect on the broader economy is M dollars. Thus, given that the loss sustained by debt is $\max\{I - D, 0\}$, the loss function of the government is

given by $M \max\{I - D, 0\}$. This loss function is motivated by the fact that debt investments may be leveraged and, thus, the firm's default may trigger deleveraging and a cycle of additional defaults in the economy. Also, debt defaults may cause disruptions in the secondary wholesale funding market, where fixed income instruments are used as collateral. We do not assume here that losses in equity value lead to such disruptions. Particularly interesting will be the case of $M > 1$. In this case, the loss to the larger economy exceeds the private loss debtholders sustain in case of firm default.

In the baseline model, agents take actions sequentially in three rounds. These rounds of moves describe the strategic interaction that takes place between debt investors, equity investors, and the government. First, the government announces its policy. Second, investors invest funds and select which project the firm will take on: P or R. Then, nature chooses the state of the world: g or b. Third, the government can take action. In the baseline case, we concentrate on the government policy choice consisting of bailing the debtholders out or not bailing them out in the event of firm default. Later on, we will extend this framework to consider other actions that the government could take, like making additional transfers, levying taxes, imposing capital requirements, applying a resolution policy toward the firm (if needed), etc. We will consider several specifications.

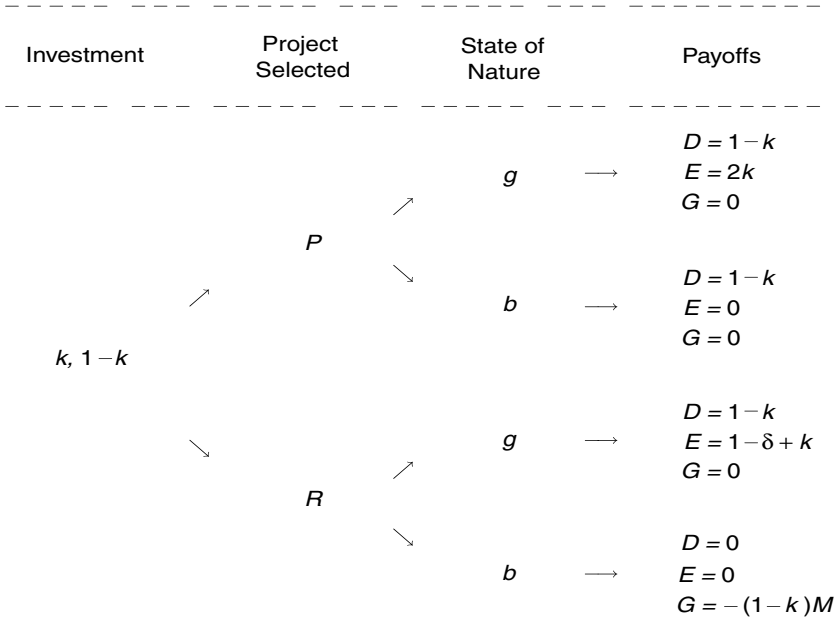
Denoting by D, E, and G the payoffs to, respectively, debt, equity, and the government, Figure 1 summarizes the three stages of interaction between the agents and presents terminal payoffs absent any government action. There are four terminal nodes in the game tree depicted in Figure 1: two representing payoffs under project P and two representing payoffs under project R. The firm is insolvent only if it selects the risky project R and the state of nature is bad. The payoffs in that case are given in the bottom node of the game tree. Since the firm's assets are worth zero in this state, so are its liabilities. Through channels that are not modelled here, bond default causes a disruption to the larger economy, which in turn gives the government a negative payoff of $-(1-k)M$. To emphasize, Figure 1 shows payoffs that are realized in the absence of any government transfers or other actions.

We want to analyze the effect of government policy on the final outcome of the game. This outcome will be determined by the actions that debt and equity investors choose under a given specification of government policy. An outcome in which the investors' actions maximize their profit will be called an equilibrium.

2. OPTIMAL GOVERNMENT ACTION WITH POLICY OPTIONS RESTRICTED TO BAILOUT OR NO BAILOUT

In this section, we consider the outcomes that strategic interaction given in the game form specified above will lead to under two possible government

Figure 1 The Game Tree with Payoffs



actions in the crisis/insolvency node of the event tree. Action one is to bail out the debtholders and avoid the disruption to the larger economy. Action two is the opposite: no bailout. This set of actions available to the government is very restrictive, and we relax it in the next section. In this section, we analyze this case to highlight the problem of time consistency.

We proceed by asking which action the government would prefer at the ex post stage if the firm is insolvent. If the disruption that the firm's default causes is large, bailing the firm out is optimal ex post. We then study the optimal policy from the ex ante perspective. We show that not bailing out is optimal ex ante.

Optimal Ex Post Bailout Policy

Given the payoff structure of Figure 1, let us consider the question of optimal government action in the "crisis" node in which the firm is insolvent. If the bottom payoff node of the game tree in Figure 1 is reached, the firm defaults, debtholders suffer a loss $1-k$, and the government suffers a loss $(1-k)M$. One action that the government can take in the crisis node is to bail the debtholders out, i.e., make them whole by, for example, buying the firm's debt at face

value. Because the assets backing up these claims are worthless in this node, the cost to bail debtholders out is $1-k$. But bailing debtholders out eliminates any disruption to the larger economy, the cost of which would be $(1-k)M$. Clearly, the optimal government bailout decision in the crisis state depends on the value of the larger economy disruption parameter M. If $M \leq 1$, the government's loss is smaller when the firm is allowed to fail, i.e., there is no incentive for the government to bail debtholders out in the crisis node. If $M > 1$, the best policy for the government in this node is to provide a full bailout to the debtholders.

Suppose that $M > 1$, which means the best action for the government ex post is to make debtholders whole if the firm is insolvent. Given this policy, what are the best investment decisions for the private-sector players? Because debtholders are fully repaid in every terminal node (either by the firm or by the government), debt investors provide funds $I = 1-k$ and collect the same payoff $D = 1-k$ in every state of nature, independent of which investment project is implemented. Under this payoff structure, therefore, debtholders have no incentive to influence the choice of the investment project, so the firm will choose the project that maximizes equityholders' value (i.e., debtholders do not monitor/discipline equityholders at all).

Which project will then the equityholders choose for the firm? Despite the fact that no bailout money is ever given to the equityholders, and despite the fact that the NPV of P exceeds the NPV of R, the equityholders prefer project R over project P. Indeed, if P is selected, the expected terminal value of equity is $\frac{0}{2} + \frac{2k}{2} = k$. If R is selected, the expected terminal value of equity is $\frac{0}{2} + \frac{1-\delta+k}{2}$, which is strictly larger than k due to (1). Thus, equity investors will select R.[1]

If $M > 1$ and the government uses its best ex post policy, the firm will be able to raise the required amount of debt and run the inefficient project R. This way, the expected value of the equity stake in the firm is maximized, and the expected loss to the government is $\frac{0}{2} + \frac{1-k}{2}$, which is strictly greater than zero.

Optimal Ex Ante Bailout Policy

Suppose now that the government can choose the action it will take ex post (i.e., after the project payoff is realized) already at the ex ante stage (i.e., before private-sector investors act). In particular, the government can announce ex ante whether or not it will make debt whole ex post in case the crisis node of the event tree is reached. We assume here that this announcement is credible, i.e., the announced policy will be adhered to with no reconsideration. What

[1] This is a version of the classic risk-shifting problem of Jensen and Meckling (1976). However, in our model, government plays an important role.

policy would the government choose to commit to? A bailout or letting the firm fail?

The answer depends on what the private-sector investors choose ex ante knowing what the government will do ex post. Consider first the policy in which the government announces it will bail the creditors out if the firm is insolvent. As we saw before, under this policy debtholders agree to provide funds and are indifferent in regard to which project is selected, as they receive the reservation return on their investment in any state of nature. Equityholders prefer the risky project R, so R will be implemented by the firm. The expected loss to the government is $\frac{1-k}{2}$.

Consider now the optimal actions of the private-sector players when the government announces ex ante that it will not bail the debtholders out in the crisis state. What is the response of the private agents to this pre-announced action? Equityholders still want to select R, as this selection continues to maximize the expected value of equity.[2] Debtholders, however, refuse to invest in the firm if R is to be selected, because without a government bailout the value of their investment turns out to be zero with a 50 percent chance.[3] Thus, if debt investors are to be attracted to the firm at all, the firm must convince debt investors that P will be selected. This can be done through private contracts, e.g., via an appropriate covenant attached to the debt contract. (This practice of debtholders monitoring the firm is often referred to as market discipline.) With P selected, the firm ends up solvent in both states of nature. The adverse effect for the larger economy, therefore, never occurs. The loss to the government is zero.

Which of the two options will the government choose? Its own loss is minimized, attaining the value of zero, with the pre-announced policy of no bailout. The social value of the outcome obtained under this policy, as well, is higher than the value obtained under the bailout policy. To see this, recall that project R has a negative expected NPV of $\frac{-\delta}{2}$. If the government bails out ex post, equity investors' equilibrium response is to implement R, and bond investors' equilibrium response is to not care if R or P is implemented. In this equilibrium, the expected NPV of debt investment is zero, the expected NPV of equity investment is

$$\frac{1-\delta+k}{2}-k=\frac{1-\delta-k}{2}, \tag{2}$$

[2] Note that even with the bailout, the ex post value of equity is zero in the crisis state. Thus, equity payoff is not decreased if there is no bailout in this state.

[3] Because we assumed that bond investors are infinitely risk-averse, a cashflow with any chance of zero payout is worth zero to them, and hence they will not invest in project R, absent government support. The infinite risk-aversion assumption is not critical here. With no bailout, project R will not be implemented if the risk premium that bond investors require is sufficiently high (not leaving any residual value to equity investors). High risk compensation will be required even if bond investors' risk aversion is large but finite.

and the expected NPV of the government loss is $\frac{1-k}{2}$. Adding the loss generated by R and the value of equity we have

$$\frac{\delta}{2} + \frac{1-\delta-k}{2} = \frac{1-k}{2}, \tag{3}$$

which equals the value of the government loss resulting from the bailout. It is important to note that under the bailout policy bailout money is paid to debtholders and the value of equity is "wiped out" in the crisis state. Yet, effectively, in addition to covering the expected loss of the wasteful project R selected by the firm, government bailout money makes for an indirect handout to equityholders in the form of the high expected return that equity investors earn in this equilibrium.

The social value obtained in equilibrium under the no-bailout policy, in turn, is non-negative because the NPV of project P is zero.[4] Under no bailout, the equilibrium expected NPV of debt investment and the government loss are both zero. The value of equity is equal to the NPV of P, which was normalized to zero. Thus, the pre-announced no-bailout policy leads to an efficient equilibrium outcome.

Time Consistency and Commitment

A key insight here, which goes back to Kydland and Prescott (1977), is that the optimal ex ante policy may be not time consistent. What does it mean? If $M > 1$, we notice that the optimal ex ante policy calls for a different action in the insolvency node than what the optimal ex post policy calls for once this node is reached. Thus, the optimal policy plan is not internally consistent as we move (in time) from the ex ante stage to the ex post stage.

With $M \leq 1$, there is no inconsistency. At the ex ante stage, the government can announce no bailout and, if the private sector selects project R and the state of nature is b, the government's best option ex post is in fact to not bail out. With $M \leq 1$, therefore, the government could announce nothing ex ante about what it will do and obtain the efficient outcome. This is because the private sector would be expecting no bailout knowing that the government faces no ex post incentive to bail out anybody.

With $M > 1$, the government does have an incentive to bail debtholders out in the insolvency state, and the private sector knows it. Can the government convince the private sector to stay away from the risky, inefficient project R when this project is exactly what maximizes the value of private claims? In the previous subsection, we assumed that the government can pre-announce

[4] It is straightforward to modify the example we have used here to have a strictly positive expected value of the prudent project P.

its policy and that the investors take it as given that the government always adheres to the pre-announced policy. This seems like a very strong assumption.

A weaker alternative assumption about the government's commitment power is that the threat of no bailouts is not a credible one. If the private sector does not believe that the government can follow through on a promise to let the firm fail, then the investors will select project R despite this threat. Without commitment, thus, the government's policy choices are reduced to those plans that are time consistent. In our model, with $M > 1$, this means that the government cannot promise to not provide a bailout in the crisis node, and, in effect, the inefficient project R is implemented in equilibrium.

It is worth pointing out here that the inefficient equilibrium would not exist if there were no government in our model. Debt investors agree to invest in the wasteful project R because they know that in the crisis state the government will bail them out. The government is benevolent, yet its lack of commitment power combined with the assumed spillover effect make investment in the inefficient project R a rational choice for private-sector investors. Because governments fulfill useful functions unrelated to bailouts of the financial sector, we take the existence of a benevolent, deep-pocketed government as given in our model.

In conclusion, the analysis in this section shows that a government whose policy options for resolution of a large financial institution are restricted to bailing out or allowing for failure is in a very difficult position when $M > 1$, i.e., when the spillover effects of the firm's failure are large. The ex ante optimal policy is not time consistent. Unless the government is committed to not bailing out failing firms regardless of the consequences, and the market participants actually believe this, excessive risks are taken and inefficient bailouts happen in equilibrium. In this equilibrium, although debt- not equityholders receive bailout funds, taxpayers make an indirect handout of value to equityholders.[5] In addition to this transfer, economic value is lost through firms' selection of inefficient investment projects whose large upside value makes the handout to equityholders possible. In the next section, we discuss several possibilities for how an expanded policy choice set can alleviate the problem of time inconsistency.

3. ALLEVIATING THE TIME-CONSISTENCY PROBLEM

The direct way of combating the time-inconsistency problem is to build an infrastructure acting as a so-called commitment device for the policymaker. A commitment device is something that makes it very costly, or—better yet—impossible, for the decisionmaker to deviate ex post from the pre-announced

[5] Note that even when the government provides a bailout, the equity value is "wiped out" in the bad state b. Despite this wiping out, equityholders do benefit from the bailout.

course of action. In simple decisionmaking settings with time-consistency problems, commitment devices may be available and effective.[6] In the case of the optimal government policy toward a large financial institution, because of political-process constraints, for example, such devices may be difficult or impossible to implement. In this section, we will therefore discuss alternative, indirect ways in which the time-consistency problem may be alleviated or altogether avoided in this model.

In particular, we will discuss the following five possibilities:

1. Decreasing the impact of firm failure on the larger economy.
2. Direct monitoring of the firm's risk-taking by the government.
3. Banning employee compensation practices indicative of excessive risk-taking.
4. Taxing extraordinary profits.
5. Imposing binding capital requirements on the firm.

In all these cases, our discussion changes, or goes beyond, the basic structure of the model we described in Section 1.

Decreasing M

It is clear from the previous discussion that there is no time-inconsistency problem when $M \leq 1$. Therefore, if there are actions that can be taken in practice to decrease the negative spillover effect of a single firm's default on the economy as a whole (i.e., decreasing M), these can clearly be useful to alleviate the time-consistency problem.

Efficient resolution policy under either the bankruptcy code or the Orderly Liquidation Authority of the Dodd-Frank Act can be consistent with reducing M. The question of what is the best way to reduce the negative ex post spillover effect is beyond the scope of this article. However, our analysis of the time-consistency problem provides an important observation. The inefficient bailout equilibrium can be eliminated not by legislating commitment, but rather by building a legal framework and financial infrastructure in which the government is no longer tempted to bail firms out as the spillover effects become relatively small. Therefore, government resolution policy toward large financial firms improves economic efficiency to the extent that it helps decrease spillovers. If the negative spillover effect of a financial firm

[6] A driver who tends to impulsively speed in certain road conditions may choose to drive an underpowered car in order to eliminate the possibility of acting on this impulse. In this case, the choice of a slow car is a commitment device.

default remains large, the government will be tempted to bail the firm out in a crisis state regardless of what the pre-announced resolution policy may say. Therefore, even if the pre-announced policy promises no bailouts, private sector investors may still expect a bailout to occur as long as spillovers are sufficiently large. As we have seen, such an expectation may lead to excessive risk-taking by the private sector.

At a deeper level, however, one could ask why M should be larger (greater than one) in the first place. Why is there a spillover effect?

We will not be able to address this question using the simple model at hand, but it is clear that there are two possibilities here. One is that the nature of the financial intermediation technology is such that having $M > 1$ is more productive than having $M \leq 1$. This may be because of increasing returns to scale in the provision of financial services, benefits of leverage, synergies from having different financial product lines under one corporate structure, etc. If this is the reason for $M > 1$, then changing resolution policies or forcing institutions into a shape in which their M is less than one has real costs. These costs should be weighted against the costs of moral hazard—excessive risk-taking—that arise in equilibrium with bailouts.

The other possibility is that there are no increasing returns to scale, but rather institutions become large/leveraged/interconnected precisely because, with $M > 1$, they receive an implicit and unpriced government guarantee for their debtholders. If firms become large (often in this context called "too big to fail") not for efficient production reasons but just so they can take on risk backed up by the unpriced government guarantee, then there is no economic cost to changing resolution policy or forcing the institutions into a shape in which their $M \leq 1$.

Given the scale of the time-consistency problem discussed here, further research directed at discerning which of these two possibilities is in fact a correct representation of reality is needed. This question is very difficult partly because, going back and gathering data, it is not easy to establish the exact policy regime that was in effect when the data were generated. Even more so, it is hard to know what the private sector's perception of the policy regime was at the time. As we have discussed, this perception has a strong impact on behavior and, thus, will have an impact on the data gathered.

Direct Monitoring of Project Choice

Clearly, if the government can directly monitor the firm and control the choice of the project, then it can ensure that the risky and wasteful project R is not implemented. This would eliminate the need for bailout because the firm is never insolvent if R is never implemented.

Monitoring and controlling firms' risk-taking is a traditional role of bank supervision, which of course is a costly activity. These costs, however, may be smaller then the expected present value of the ex post bailout costs.

We should note, however, that monitoring the firms may not be 100 percent effective, in the following sense. As long as risky projects R are out there and the government cannot commit to no bailouts, equity investors have an incentive to escape the control of the government by funding the project outside of the scope of any control mechanisms that the government may have in place at a particular point in time. (Recall the pre-crisis "shadow banking system.") Thus, in addition to monitoring the existing financial firms, the government should monitor risk-taking in general to minimize the possibility that project R is implemented under some other institutional arrangement that could have similar adverse consequences for the economy as a whole. This adds difficulty to the task of direct government monitoring of risk-taking.

Compensation Restrictions

One way in which the government could prevent the firm from implementing project R could be to limit the compensation practices that give the firm's employees incentives to take large risks.[7] In our simple model, we think of equityholders as those choosing the investment project, i.e., the project being a part of the definition of the firm. In practice, shareholders hire staff/management to operate the project. For large firms, in particular, hired managers control the operations and make investment decisions. In order to align the incentives of the managers with those of the equityholders, executive compensation packages make pay dependent on realized equity values. Executives, however, may be less inclined to take risk than the shareholders, as managers' exposures usually are large. If the shareholders desire to structure the firm in such a way that it takes on large risks similar to our risky project R, the compensation package for its manager may be different than that under which the manager would be implementing a prudent project similar to P.

Executive compensation schemes may be easier to examine than the whole portfolio of the firm's assets. Therefore, one way for the regulators to limit risk-taking may be to regulate executive compensation schemes at large financial institutions. Because of standard (not related to the possibility of government bailout) agency problems faced by financial firms, identifying compensation schemes that induce excessive risk-taking (as opposed to those that merely induce adequate managerial effort) may be very difficult in

[7] In the Unites States, as well as in Europe, regulatory agencies are currently working on rules restricting loan officers' and bank executives' compensation with that goal in mind. (For a rule proposal on incentive-based compensation arrangements in the United States see Federal Register, Vol. 76, No. 72, April 14, 2011. Loan officer compensation rules have been amended in Federal Register, Vol. 75, No. 185, September 24, 2010.)

practice. Phelan and Clement (2009) and Jarque and Prescott (2010) study optimal compensation of bankers using the tools of mechanism-design theory.

Taxing Extraordinary Profits

Suppose the government considers taxing the profits that the firm makes in the good state g in order to remove the incentive to invest in the risky project R. Can such a tax be effective? The answer to this question depends on whether or not the government knows if profits, when realized, are because of excessive risk-taking or not.

In the simple model studied here, even assuming that the government does not directly observe the project choice made by the firm, the government can tell with certainty which project was selected just by observing the realized payoff. This is because $2 - \delta \neq 1 + k$ and $0 \neq 1 - k$ and thus, the so-called full-support condition is not satisfied in our simple model. If the government sees payoff $2 - \delta$, it knows that R must have been selected by the firm. For this reason, it is possible in this model to tax the firm's returns only if project R was selected. Such a tax can in fact correct the incentives of the private sector.

Indeed, let the firm's profit in node (R, g), i.e., when project R is selected and the state of nature is g, and only in this node, be taxed in some amount $\tau(R, g)$. This amount can be thought of as a "windfall profit" tax—it occurs only if the realized payoff is $2-\delta > 1+k$, where $1+k$ represents the "normal" profit level that is realized in node (P, g) of the game tree. With this tax, the expected value of equity under R is $\frac{1}{2}0 + \frac{1}{2}(2 - \delta - \tau(R, G) - (1 - k))$, which is the same as without the tax less $\frac{\tau(R,g)}{2}$. The value of equity under P is unchanged, i.e., it remains equal to zero. Thus, when

$$\tau(R, g) \geq 1 - \delta - k,$$

the value of equity under R becomes less than under P, so this tax corrects the firm's incentive to invest.

In reality, however, it may be hard to tell even ex post if the firm's risk-taking behavior, represented here by the choice of the project, was prudent or risky. Suppose then that the government can observe the realized state of nature but not the firm's payoff. Any tax on profits in the good state g must be the same independent of which project was selected. It is immediate that such a tax, $\tau(g)$, decreases the value of equity under both P and R, by the amount $\frac{\tau(g)}{2}$. Since R was selected without the tax, i.e., when $\tau(g)$ was zero, it will continue to be selected with $\tau(g)$ different from zero. Thus, this tax cannot correct the risk-taking behavior and the inefficient project selection in equilibrium.

Further extensions of the model could include other, more complicated tax instruments. Even in a model in which the full-support condition is satisfied

so the government cannot detect the project selection ex post, likelihood ratios can be used to statically discriminate between risky and prudent firm behavior. This information can then be used to implement taxation and other ex post policies that discourage risky investment. Existing work on this problem includes Marshall and Prescott (2001, 2006).

Imposing Binding Capital Requirements

Finally, let us consider the effects of a binding capital requirement that government regulations could impose ex ante, i.e., before investors select the project and fund it. We will show that a large enough capital requirement eliminates the firm's incentive to take on excessive risk.

Up until now, we have assumed that equity investors provide k dollars in initial investment and bond investors provide $1 - k$. That amount is the minimum level of equity investment necessary to ensure that debt issued by the firm is riskless under project P with no government support. (Even in the bad state b the firm can fully repay the debtholders if the face value of debt is no larger than $1 - k$ because project P pays off $1 - k$ in state b.) In this section, we will consider other levels of initial equity capital investment than just k. Let us denote the initial equity investment amount by κ, which may not be equal to k. We will also suppose that the government mandates that κ not be smaller than some minimal amount $\underline{\kappa} \leq 1$. That is, government regulations impose on the firm a minimum capital requirement constraint

$$\kappa \geq \underline{\kappa}.$$

With equity investment κ, the amount of debt investment the firm must raise (and hence the face value of debt it issues) is $1 - \kappa$.

If the government chooses a required capital level $\underline{\kappa}$ such that $\underline{\kappa} \leq k$, then the minimum capital requirement constraint $\kappa \geq \underline{\kappa}$ is not binding, so it has no effect on equilibrium outcomes. We will therefore focus on binding capital requirements, $\underline{\kappa} > k$. We will also assume that the firm will not choose to hold more capital than the minimum level required. This is motivated by the notion, which is not explicitly modelled here, that debt financing is less expensive than equity financing.

Let us now examine the terminal payoffs to debt investors, equity investors, and the government when the capital requirement is binding. Equity investment is $\kappa = \underline{\kappa} > k$, and debt investment is $1 - \kappa = 1 - \underline{\kappa} < 1 - k$. Under project P, even in state b the firm's assets are worth more than the face value of debt, $1 - k > 1 - \underline{\kappa}$, so debt investors are repaid in full in both states. The expected payoff to equity is

$$\frac{1-k-(1-\underline{\kappa})}{2}+\frac{1+k-(1-\underline{\kappa})}{2} = \frac{\underline{\kappa}-k}{2}+\frac{\underline{\kappa}+k}{2} = \underline{\kappa},$$

which equals the initial investment, so equity investors break even in expectation. There is no government bailout when P is implemented, so the payoff/cost to the government is zero.

Under project R, the firm's assets are worthless in state b. The government bails out debtholders and incurs a loss in the amount $1-\underline{\kappa}$. The expected cost of the bailout is $\frac{1-\underline{\kappa}}{2}$. Comparing with (3), it is immediately clear that the cost of the bailout is smaller when capital investment is larger. Thus, imposing a capital requirement with $\underline{\kappa} > k$, even if the bailout equilibrium is not eliminated, decreases the expected cost to the taxpayer. Thus, in the inefficient bailout equilibrium, the cost to the taxpayer is lower when the firm's leverage is lower (equity to assets ratio is higher).

To show that a capital requirement $\underline{\kappa}$ can be chosen to eliminate the need for bailouts altogether, let us now calculate the NPV of equity in the bailout equilibrium as a function of $\underline{\kappa}$. Under R , in state g, the firm's assets are worth $2-\delta > 1-\underline{\kappa}$, so the firm is solvent and the payoff to equity in this state is $2-\delta-(1-\underline{\kappa}) = 1-\delta+\underline{\kappa}$. In state b, the firm is insolvent and equity is worthless. The NPV of equity, thus, is

$$\frac{1-\delta+\underline{\kappa}}{2}-\underline{\kappa} = \frac{1-\delta-\underline{\kappa}}{2}. \tag{4}$$

As before, cf. (3), the value of equity equals the expected cost to the government, $\frac{1-\underline{\kappa}}{2}$ here, less the amount of waste of value generated by project R, that is $\frac{\delta}{2}$. It is clear from (4) that if the capital requirement $\underline{\kappa}$ satisfies

$$\underline{\kappa} \geq 1-\delta, \tag{5}$$

then the NPV of equity is not greater than zero. This means that by selecting R, equity investors cannot do better than just break even. Thus, equity is not worth more under R than under P. For a sufficiently high capital requirement, we can see that equity investors' incentive to take on the wasteful project R is removed. The firm will not select the inefficient project R if doing so implies gambling with its own money (to a sufficiently high degree).

We should note here that under a capital requirement satisfying (5), the firm is still levered. In fact, if $\underline{\kappa} = 1-\delta$, the amount of debt the firm can raise is $\delta > 0$. In order to eliminate the bailout equilibrium, it is not necessary to require all-equity financing. However, the amount of leverage cannot be too high. (For example, if $\delta = 0.8$, the minimum capital requirement would be 20 percent of assets.)

It is interesting to note that sufficient capital requirements do not solve the time-consistency problem by making the government committed to no bailouts. With $\underline{\kappa} = 1-\delta$, if the firm were to select the risky project R, the

government would still provide a bailout to creditors in state b, in the amount δ. The expected loss to the government would still be positive in this case. That loss, in fact, equals the amount of value destroyed in the project R, which is $\frac{\delta}{2}$. The key here is the fact that, with a sufficient amount of their own capital on the line, by running project R, the firm cannot extract from the government more than that amount. If $\underline{\kappa}$ is strictly larger than $1-\delta$, even by a small bit, the bailout equilibrium ceases to exist because the amount of value lost in R is more than the amount of resources that can be extracted from the uncommitted government. Thus, we see that, rather than serving as a commitment device for the government, sufficient capital requirements provide the right risk-taking incentives for equity investors.

Given that, as we assume here, capital financing may be more costly than debt financing, regulators should be careful not to set capital requirements at an unnecessarily high level. However, the fact that capital is costly should not be used as a rationale for tolerating the inefficient bailout equilibrium. As our discussion shows, the costly bailout equilibrium can be eliminated by a correct capital requirement policy. The cost of equity financing is exogenous to this argument, i.e., government policy cannot change that. Given the restrictions faced by the government, it is important to draw a distinction between the problems government policy can and cannot solve. In our model, the inefficient bailout equilibrium is an example of the former and the high cost of equity capital is an example of the latter.

4. CONCLUSION

The main purpose of this article is to provide an elementary exposition of the time-consistency problem in the government's choice of its policy toward large financial firms that face insolvency. We use a simple model to show how this problem arises when the failure of a large financial institution has sufficiently large spillover effects for the economy as a whole, which the government is trying to prevent. The equilibrium outcome is efficient if the government is credibly committed to not bailing the firm out in case its excessive risk-taking backfires. Under the assumption of perfect government credibility, the private sector behaves prudently and the government's commitment is never tested in equilibrium.

A government's full commitment to letting a large firm fail and cause adverse spillover effects on the economy is a very strong assumption. Given political-economy constraints, it is probably impossible to achieve such a level of commitment in practice. The lack of time consistency of optimal policy therefore should be considered a very real and serious problem. In the context of our model, we discuss several ways in which excessive risk-taking and inefficient bailouts can be addressed despite the lack of commitment.

One such way involves making changes to the process of resolution of insolvent financial firms. These changes should be aimed at decreasing the negative spillover effects to sufficiently small levels. This approach may be a fruitful way of thinking about the solution, but it requires further fundamental research, given the incomplete understanding that we have of how exactly the spillover effects arise in the economy. It is also hard to know if efficient resolution policy can restrict spillovers to a degree sufficient to eliminate the government's ex post incentive to bail out systemically important firms.

The second way in which excessive risk-taking can be controlled in our model involves direct government supervision of the financial sector firms' actions. If the government, at a reasonable cost, can observe investment activities of the firms, it can enforce prudent risk-taking. This is the most direct way in which it is possible to prevent the equityholders from taking advantage of the implicit safety net that the government provides by being unable to commit to letting a large firm fail in a bad economy.

Although direct monitoring of investment is very straightforward in our simple model, it may be difficult and costly to effectively implement in practice. The costs of monitoring financial firms' compensation schemes may be lower than the costs of monitoring entire portfolios of assets. For that reason, restrictions on the structure of executive and financial firm employee compensation packages can be a cost-effective means of helping eliminate excessive risk-taking.

The fourth way we discuss involves imposing taxes on the extraordinary profits firms achieve if they take on an excessive amount of risk and are lucky to see that strategy pay off in high profits. That method, however, may be particularly hard to implement in practice given the government's imperfect information about what exactly constitutes an excessive amount of profit in a given macroeconomic state.

Finally, we discuss how binding capital requirements can eliminate bailouts. With enough own equity capital, firms will not seek excessive risk even if the government stands ready to bail them out in bad times. With firms behaving prudently, bad times are not bad enough to necessitate government support. In this way, bailouts are eliminated. Binding capital requirements are quite simple to implement in practice. Given our imperfect understanding of how a fully optimal resolution or supervision policy should be designed, capital requirements seem to be the most practical solution to the excessive risk-taking problem as of now.

The model discussed in this article can be extended in several directions. One is to study the interaction between moral hazard and commitment in a dynamic setting in which the government can manage private-sector expectations and its own reputation. Standard results in the repeated games literature state that, if the government is patient enough, equilibria can be constructed in which the government would not bail out in the (R, b) node of the event tree

and, thus, this node is never reached in equilibrium. In such an equilibrium, bailouts, although off-equilibrium events, would still constrain the level of efficiency attained in the equilibrium outcome. If the government's patience is sufficiently low (perhaps because of the electoral cycle), however, the efficient equilibrium cannot be sustained in the repeated-game setting. Given the political-economy constraints, effective reputation-based deterrents to moral hazard may be hard to establish.

Another extension of this model can involve relaxing the assumptions we made on the preferences of debt and equity investors. However, one should expect that the basic message of the model is not going to change if debt investors are modelled as having some risk-tolerance and equity investors as being somewhat risk-averse.

In this article, we look at a single large firm whose default, by assumption, can cause negative spillovers. In practice, firms of different sizes will have different potential for causing spillovers. Our model could be extended to discuss how the time-consistency problem and optimal government policy depend on the firm's size. Also, in our discussion we have not been specific about what kind of financial firm we have in mind. These firms can be commercial banks, broker-dealer units, or insurance companies, among others. Another way to extend our discussion is to consider the case of banks in particular, whose liabilities include federally insured deposits. In the case of insured deposits, the guarantee of an ex post bailout is explicit. In that case, discussion of optimal policy would include the issue of optimal ex ante pricing of federal deposit insurance.

Our analysis makes clear that the scale of the time-consistency problem of government policy toward large financial firms is determined by the spillover effects in which the failure of one large firm can have a significant impact on the macroeconomy as a whole. Optimal government regulation policies, including the optimal level of capital requirements, and, consequently, the optimal structure of the financial services sector will depend on what we can understand about the nature and size of this spillover effect. Further research on this important topic is needed.

REFERENCES

Barro, Robert J., and David B. Gordon. 1983. "Rules, Discretion, and Reputation in a Model of Monetary Policy." *Journal of Monetary Economics* 12: 101–21.

Chari, V. V., and Patrick Kehoe. 1990. "Sustainable Plans." *Journal of Political Economy* 98 (August): 783–802.

Chari, V. V., and Patrick Kehoe. 2009. "Bailouts, Time Inconsistency and Optimal Regulation." Federal Reserve Bank of Minneapolis Research Department Staff Report (November).

Chari, V. V., Lawrence J. Christiano, and Martin Eichenbaum. 1998. "Expectation Traps and Discretion." *Journal of Economic Theory* 81 (August): 462–92.

Cochrane, John H. 1995. "Time-Consistent Health Insurance." *Journal of Political Economy* 103 (June): 445–73.

Jarque, Arantxa, and Edward S. Prescott. 2010. "Optimal Bonuses and Deferred Pay for Bank Employees: Implications of Hidden Actions with Persistent Effects in Time." Federal Reserve Bank of Richmond Working Paper 10-16 (October).

Jensen, Michael C., and William H. Meckling. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *Journal of Financial Economics* 3 (October): 305–60.

King, Robert G. 2006. "Discretionary Policy and Multiple Equilibria." Federal Reserve Bank of Richmond *Economic Quarterly* 92 (Winter): 1–15.

King, Robert G., and Alexander L. Wolman. 2004. "Monetary Discretion, Pricing Complementarity, and Dynamic Multiple Equilibria." *The Quarterly Journal of Economics* 119 (November): 1,513–53.

Kydland, Finn E., and Edward C. Prescott. 1977. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy* 85 (June): 473–91.

Lucas, Robert E., Jr., and Nancy L. Stokey. 1983. "Optimal Fiscal and Monetary Policy in an Economy without Capital." *Journal of Monetary Economics* 12 (1): 55–93.

Marshall, David A., and Edward S. Prescott. 2001. "Bank Capital Regulation with and without State-Contingent Penalties." *Carnegie-Rochester Conference Series on Public Policy* 54 (June): 139–84.

Marshall, David A., and Edward S. Prescott. 2006. "State-Contingent Bank Regulation with Unobserved Actions and Unobserved Characteristics." *Journal of Economic Dynamics and Control* 30 (November): 2,015–49.

Phelan, Christopher, and Douglas Clement. 2009. "Incentive Compensation in the Banking Industry: Insights from Economic Theory." Federal Reserve Bank of Minneapolis Economic Policy Paper 09-1 (December).

Economic Quarterly—Volume 97, Number 2—Second Quarter 2011—Pages 153–173

Sectoral Disturbances and Aggregate Economic Activity

Nadezhda Malysheva and Pierre-Daniel G. Sarte

A key topic in the literature on business cycles concerns the origins of shocks underlying fluctuations in economic activity. One dimension of this topic focuses on whether we should think of aggregate economic fluctuations as being driven by disturbances that affect all areas of the economy simultaneously, or whether these movements are instead better thought of as arising from shocks to different sectors that affect economic activity by way of production complementarities such as input-output linkages. To the extent that sources of fluctuations include sectoral shocks, another key consideration then is the manner in which sectoral shocks potentially become amplified and propagate throughout the economy for a given degree of disaggregation.

A conventional wisdom argues that shocks to different sectors of the economy are unlikely to matter for aggregate fluctuations because they tend to average out in aggregation. Thus, positive shocks in some sectors will generally be offset by negative shocks in other sectors. This notion has in part led the bulk of the literature on business cycles to concentrate on the effects of different types of aggregate shocks. However, whether or not idiosyncratic sectoral shocks do average out in aggregation depends on various aspects of the economic environment. In particular, Gabaix (2011) describes how, when the economy comprises a handful of very large sectors, sectoral disturbances will not average out and contribute nontrivially to aggregate fluctuations. Horvath (1998) also makes the point that because of input-output linkages, shocks to particular sectors feed back into other sectors in a way that leads to significant

We wish to thank Kartik Athreya, Andreas Hornstein, Nika Lazaryan, and Zhu Wang for helpful comments. The views expressed in this article are those of the authors and do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System. All errors are our own. E-mail: pierre.sarte@rich.frb.org.

amplification and propagation of those shocks. This idea is further developed and analyzed from a network perspective in Carvalho (2007), and Acemoglu, Ozdaglar, and Tahbaz-Salehi (2010).

This article provides an overview of some key dimensions related to the effects of sectoral shocks on aggregate economic activity. It describes how the entire distribution of sectoral shares, or the weight of different sectors in aggregate activity, generally matters for the measured contribution of a sector to aggregate variability. It also illustrates how intersectoral linkages affect the propagation and amplification of sectoral idiosyncratic shocks. In particular, it summarizes sufficient conditions, carefully articulated in Dupor (1999), under which aggregate outcomes are invariant to sectoral disturbances, even in the presence of input-output linkages across sectors. A key condition requires that the matrix describing input-output linkages satisfies a particular structure according to which all sectors serve as equally important material providers to all other sectors.

To the degree that input-output linkages descriptive of U.S. production depart from Dupor's (1999) sufficient conditions for the irrelevance of sectoral shocks, it is generally not straightforward to characterize how this departure translates into sectoral contributions to aggregate variability.[1] As shown in Foerster, Sarte, and Watson (2011), the contribution of sectoral shocks to aggregate fluctuations is generally model-dependent and cannot be analytically characterized. Thus, using an actual input use matrix obtained from the Bureau of Economic Analysis (BEA) for 1997, and a two-digit level disaggregation of gross domestic product (GDP), this article describes key aspects of this calculation and provides estimates of the relative contribution of different sectors to aggregate variations given each sector's share in aggregate output. By and large, the manufacturing sector and the sector related to real estate, rental, and leasing tend to contribute the most to aggregate variations.

Given this article's emphasis on sectoral shocks, it also examines how these shocks propagate to other sectors and become amplified as a result of feedback effects resulting from intersectoral linkages. Using two canonical multisector growth models in the literature, specifically the foundational work of Long and Plosser (1983) and, its descendant, Horvath (1998), it illustrates how the propagation and amplification of sectoral shocks depend importantly on the details of the economic environment in which intersectoral linkages operate. Thus, it explains why using the share of the sector in which a disturbance occurs as a gauge of its effect on aggregate output constitutes, in general, a poor approximation. The article also shows that the effects of a given sectoral shock both on other sectors and on aggregate output will typically extend well beyond the life of the shock itself. In some sectors, because

[1] Carvalho (2007), as well as Acemoglu, Ozdaglar, and Tahbaz-Salehi (2010) make considerable progress along this dimension.

of feedback effects, things can get worse before improving even though the shock has already dissipated.

This article is not strictly concerned with accounting for the actual volatility of output because of sectoral co-movement (that is the main thrust of Foerster, Sarte, and Watson [2011]). Rather, this article deals with the way in which sectoral weights and input-output considerations affect the amplification and propagation of shocks. Therefore, unless otherwise noted, we use a stylized version of sectoral shocks, namely i.i.d. and uncorrelated across industries.

The rest of the article proceeds as follows. In Section 1, we describe how sectoral size relates to aggregate variability absent any complementarities in production. Section 2 provides an overview of how sectoral linkages influence the effects of sectoral shocks using the canonical multisector growth models of Long and Plosser (1983) and Horvath (1998). In Section 3, we use an unanticipated one-time decline in manufacturing total factor productivity (TFP) to illustrate how sectoral shocks propagate to other sectors, as well as their ultimate impact on aggregate output. Section 4 concludes.

1. SECTORAL SIZE AND AGGREGATE VARIABILITY

As a first approximation, it is natural to conjecture that sectoral shocks should not matter for aggregate economic activity because they will "average out." However, Gabaix (2011) carefully articulates the idea that this intuition does not hold if some sectors play a large role in economic activity, which he refers to as the "granular" hypothesis. In this view, idiosyncratic shocks to sectors with large shares have the potential to generate nontrivial disturbances in aggregate output. In particular, Gabaix (2011) shows that idiosyncratic i.i.d. shocks fail to average out in aggregation when the size distribution of sectors is sufficiently leptokurtic, or has "fat tails," as characterized for instance by the power law distribution. The nature of Gabaix's (2011) arguments relies on asymptotic calculations where the number of sectors, N, is large. In practice, however, N may not necessarily be very large if we think, for example, that real estate or manufacturing as a whole are being disrupted. The question then becomes: How do sectoral shares affect aggregate variability in practice?

Table 1 gives the two-digit sectoral decomposition of GDP with the industry code in the first column. The second column of Table 1 gives the value-added shares of each sector, as a percent of GDP, associated with this decomposition. To get an idea of how sectoral shares, or weights ω_i, affect aggregate variability, observe that aggregate output growth, denoted Δy_t at date t, can be (approximately) written as the following weighted average of

Table 1 Sectoral Shares and Contributions to the Variability of GDP, 1988–2010

Industry Name	NAICS Code	GDP Share, ω_i (Percent)	λ_i (Percent)	$\bar{\lambda}_i$ (Percent)
Agriculture, Forestry, Fishing, and Hunting	11	1.22	0.18	1.37
Mining	21	1.34	0.22	1.88
Utilities	22	2.05	0.51	1.55
Construction	23	4.29	2.23	6.08
Manufacturing	31–33	14.27	24.72	43.81
Wholesale Trade	42	5.98	4.33	5.83
Retail Trade	44–45	6.76	5.54	7.57
Transportation and Warehousing	48–49	2.99	1.09	2.42
Information	51	4.37	2.31	4.03
Finance and Insurance	52	7.26	6.40	9.35
Real Estate, Rental, and Leasing	53	12.44	18.79	5.21
Professional, Scientific, and Technical Services	54	6.32	4.85	4.04
Management of Companies and Enterprises	55	1.58	0.30	0.40
Administrative and Support Management	56	2.55	0.79	1.73
Educational Services	61	0.87	0.09	0.03
Health Care and Social Assistance	62	6.35	4.89	0.76
Arts, Entertainment, and Recreation	71	0.90	0.10	0.23
Accommodation and Food Services	72	2.73	0.91	1.17
Other Services (except Public Administration)	81	2.60	0.82	1.26
Government	92	13.13	20.92	1.28

sectoral output growth,

$$\Delta y_t = \sum_{i=1}^{N} \omega_i \Delta y_{it}, \tag{1}$$

where Δy_{it} represents output growth in sector i at t, N is the number of sectors, and $\sum_{i=1}^{N} \omega_i = 1$.[2] Suppose for now that output growth in each sector results directly from cross-sectionally unrelated i.i.d. shocks, ε_{it}, with identical variance, σ_ε^2, so that

$$\Delta y_{it} = \varepsilon_{it}, \text{ where } \Sigma_{\varepsilon\varepsilon} = \sigma_\varepsilon^2 I, \tag{2}$$

and $\Sigma_{\varepsilon\varepsilon}$ denotes the variance-covariance matrix of sectoral shocks. What can we say about the contribution of a given sector to the variance of aggregate output growth in this case?

Under the maintained assumptions, the variance of aggregate output is $\sigma_\varepsilon^2 \sum_{i=1}^{N} \omega_i^2$. Let λ_i denote the contribution to aggregate variance from sector i. Then, it follows that

$$\lambda_i = \frac{\omega_i^2}{\sum_{i=1}^{N} \omega_i^2}. \tag{3}$$

Observe that the size of the denominator in the above equation depends on the distribution of the ω_i's. Therefore, while we have assumed away the role of idiosyncratic volatility by assuming that all sectors are characterized by the same shocks, the entire sectoral size distribution nevertheless matters for the contribution of a given sector to aggregate volatility. The denominator in (3) is minimized when $\omega_i = 1/N \ \forall i$, so that the closer the ω_i's are to being evenly distributed, the lower the denominator will be. When $\omega_i = 1/N \ \forall i$, all sectors play an equally important role in aggregate output, $\lambda_i = 1/N \ \forall i$, and each sector's contribution to aggregate variance is equal to its share. In that case, sectoral disruptions will not be important for aggregate considerations as N becomes large.

Given the data in Table 1, where $N = 20$, we have that $\sum_{i=1}^{N} \omega_i^2 = 0.082$. The third column of Table 1 gives the contribution to aggregate variance of each sector under the assumption that idiosyncratic shocks are identically and independently distributed across sectors. For example, in the case of the construction sector, denoted by λ_c, we have that

$$\lambda_c = \frac{0.043^2}{0.082} = 0.022. \tag{4}$$

Therefore, under the maintained assumptions, construction contributes about 2 percent to the variability of aggregate GDP. For comparison, if all sectors

[2] To be specifc, ω_i in this case represents the mean share of sector i output as a percent of GDP over a given sample period.

were the same size, the contribution to aggregate variability from any one sector would be

$$\lambda_i = \frac{(1/20)^2}{20(1/20)^2} = \frac{1}{20} = 0.05. \quad (5)$$

Although construction is actually close to $\frac{1}{20}$ of GDP, its contribution to aggregate variability in this example is less than half of its share in GDP.[3] Put another way, the actual size distribution of sectors is such that it reduces the importance of construction relative to a distribution where all sectors have the same size. The reverse will be true for sectors that have large shares in GDP. For example, in the manufacturing sector, the contribution to aggregate variability, λ_m, implied by the share in Table 1 is

$$\lambda_m = \frac{0.143^2}{0.082} = 0.25. \quad (6)$$

Hence, although manufacturing represents 14 percent of GDP, when all sectors are subject to the same shocks, its contribution to aggregate variability is almost double its share. This gives one measure of the sense in which manufacturing might represent a key component of an economic recovery.

The basic calculations we have just outlined have ignored two important considerations. First, the size of sectoral shocks may be sector-dependent. Second, idiosyncratic shocks may be correlated across sectors. When the size of idiosyncratic shocks differs across sectors, the contribution of a given sector to aggregate variability also takes into account the volatility of that sector's output, $\sigma^2_{\varepsilon_i}$, relative to that of all other sectors,

$$\overline{\lambda}_i = \frac{\omega_i^2 \sigma^2_{\varepsilon_i}}{\sum_{i=1}^{N} \omega_i^2 \sigma^2_{\varepsilon_i}}. \quad (7)$$

Given equation (2), we have that $\sigma^2_{\varepsilon_i} = \text{var}(\Delta y_{it})$. The fourth column of Table 1 then gives the contribution to aggregate variability from each sector implied by equation (7). Importantly, this calculation continues to assume that idiosyncractic shocks are uncorrelated across sectors. Note that the contribution of the construction sector to aggregate variability now almost triples, from 2.2 percent to 6.1 percent. This contribution now exceeds construction's share of GDP. Similarly, manufacturing sees its contribution to aggregate variability jump from 25 percent to 44 percent. At the other extreme, the government

[3] Table 1 distinguishes between construction and real estate, rental, and leasing. The construction sector is comprised of establishments that are primarily engaged in the construction of buildings or engineering projects. Construction work may include new work, additions, alterations, or maintenance and repairs. The real estate, rental, and leasing sector is comprised of establishments that are primarily engaged in leasing and renting, and establishments providing related services. Also included are establishments primarily engaged in appraising real estate and the management of real estate for others (e.g., renting, selling, or buying real estate), as well as owner-occupied real estate.

sector's contribution to aggregate volatility falls dramatically from 21 percent, in the third column of Table 1, to just 1.3 percent in the fourth column. This result stems from the fact that while government is a relatively large share of GDP, its output is very smooth relative to that of other sectors.

While we have thus far ignored the fact that sectoral shocks may be cross-sectionally correlated, it is important to recognize that the presence of input-output linkages between sectors is likely to create some degree of cross-sectional dependence. In Table 1 for example, mining is likely to use the output of manufacturing, utilities, and construction as inputs. In general, the effect of a shock to a given sector on aggregate output will reflect not only that sector's share, ω_i, but also its degree of connection to all other sectors. In particular, all else equal, a shock to a sector that produces inputs for many other sectors will have a larger effect on aggregate output. Put differently, the presence of input-output linkages creates additional propagation from sectoral disturbances that amplify their effect on aggregate output. The next section addresses key aspects of the mechanisms by which this additional amplification and propagation takes place.

2. SECTORAL SHOCKS AND SECTORAL LINKAGES: IMPLICATIONS FOR AGGREGATE ACTIVITY

This section explores the role of sectoral linkages in amplifying and propagating sector-specific shocks. In other words, these linkages may, effectively, transform shocks that are specific to a particular sector into shocks that affect all sectors and, therefore, amplify variations in aggregate output. Because this analysis requires a model that incorporates linkages between sectors, this section uses two canonical models in the literature. The first model reflects the foundational work of Long and Plosser (1983), which explicitly considers each sector as potentially using materials produced in other sectors. The second model is that of Horvath (1998), also discussed in Dupor (1999), which allows the effects of sectoral shocks to be propagated over time through capital accumulation. A key lesson in this section is that, conditional on a given set of sectoral linkages, conclusions about the effects of sectoral shocks may differ depending on other aspects of the model in which these linkages operate.

Long and Plosser (1983)

Consider an economy composed of N distinct sectors of production indexed by $j = 1, ..., N$. Each sector j produces the quantity $Y_{j,t}$ of good j at date t using labor, $L_{j,t-1}$, and materials produced in sector $i = 1, ..., N$, $M_{ij,t-1}$, according to the Cobb-Douglas technology

$$Y_{j,t} = A_{j,t} L_{j,t-1}^{\alpha_j} \Pi_{i=1}^{N} M_{ij,t-1}^{\gamma_{ij}}, \tag{8}$$

where $A_{j,t}$ is a productivity index for sector j. Note that the technology features a version of time-to-build in the sense that production is subject to a one-period lag.

The fact that each sector potentially uses materials produced in other sectors represents a source of interconnectedness in the model. An input-output matrix for this economy is an $N \times N$ matrix Γ with typical element γ_{ij}. The column sums of Γ give the degree of returns to scale in materials in each sector. The row sums of Γ measure the importance of each sector's output as materials to all other sectors. Put simply, one can think of the rows and columns of Γ as "sell to" and "buy from," respectively, for each sector.

Let $A_t = (A_{1,t}, A_{2,t}, ..., A_{N,t})^T$ denote a vector of productivity indices that follow a random walk,

$$\ln \mathbf{A}_t = \ln \mathbf{A}_{t-1} + \varepsilon_t, \tag{9}$$

where $\varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, ..., \varepsilon_{N,t})^T$ has covariance matrix $\Sigma_{\varepsilon\varepsilon}$.

A representative household derives utility from the consumption of these N goods and leisure, Z_t, according to

$$E_t \sum_{t=0}^{\infty} \beta^t \left\{ \theta_0 \ln Z_t + \sum_{i=1}^{N} \theta_j \ln(C_{jt}) \right\}. \tag{10}$$

In addition, each sector is subject to the following resource constraints,

$$Z_t + \sum_{j=1}^{N} L_{j,t} = 1 \tag{11}$$

$$C_{jt} + \sum_{i=1}^{N} M_{ji,t} = Y_{j,t}, \;\; j = 1, ..., N. \tag{12}$$

Let Δy_t denote the vector of sectoral output growth, $(\Delta y_{1,t}, \Delta y_{2,t}, ..., \Delta y_{N,t})^T$. Then, Long and Plosser (1983) show that the solution to the planner's problem is given by

$$\Delta \mathbf{y}_t = \Gamma^T \Delta \mathbf{y}_{t-1} + \varepsilon_t. \tag{13}$$

Letting $\boldsymbol{\omega} = (\omega_1, \omega_2, ..., \omega_N)^T$ represent the vector of sectoral shares in Table 1, an expression for aggregate output growth is

$$\Delta y_t = \boldsymbol{\omega}^T \Delta \mathbf{y}_t. \tag{14}$$

Let σ_y^2 denote the variance of aggregate output growth. Then, given equation (14), we have that

$$\sigma_y^2 = \boldsymbol{\omega}^T \Sigma_{yy} \boldsymbol{\omega}, \tag{15}$$

where Σ_{yy} is the variance-covariance matrix of sectoral output growth.

For given N, and given equation (13), an analytical expression for the variance of output growth in the Long and Plosser (1983) model is given by (15) where[4]

$$\text{vec}(\Sigma_{yy}) = (I_{N^2} - \Gamma^T \otimes \Gamma^T)^{-1}\text{vec}(\Sigma_{\varepsilon\varepsilon}). \tag{16}$$

For the purpose of calibration, the matrix Γ in this article is based on estimates of the 1997 Input-Output use table constructed by the Bureau of Economic Analysis (BEA). The BEA constructs the use table based on data from the Economic Census conducted by the Bureau of the Census every five years. The table shows the value of commodities (given by commodity codes) used as inputs by intermediate and final users (represented by industry codes). By matching commodity and industry codes for the 20 industries, we create an input use table showing the value of commodities from each industry used by all other industries. A row sum of the use table represents the total value of materials provided by a given industry to all 20 industries. A column sum of the use table shows the total expenses of a given industry on the inputs from all sectors. Input shares, γ_{ij}, are the payments from sector j to sector i as a fraction of the total value of production in sector j.

We saw earlier that when sectoral shocks have unit variance, the variance of aggregate output growth absent sectoral linkages is $\sigma_y^2 = 0.082$, slightly larger that $N^{-1} = (\frac{1}{20})$ predicted under uniform sectoral shares. When sectoral linkages are taken into account in the model of Long and Plosser (1983), and using the input-output matrix corresponding to the sectoral decomposition in Table 1, the variance of aggregate output growth is approximately 0.12 or about one and a half times larger.

One can also obtain some measure of the contribution of individual sectors to aggregate variability. To calculate the relative effect of sector i on σ_y^2, let $\widetilde{\Sigma}_{\varepsilon\varepsilon}$ denote a diagonal matrix whose diagonal is (0, 0, ..., 1, ..., 0) where the "1" is located in the i^{th} position. Then, we can calculate what the variance of output growth would be with sectoral linkages if the model were driven exclusively by shocks to sector i:

$$\widetilde{\sigma}_y^2 = \boldsymbol{\omega}^T \widetilde{\Sigma}_{yy} \boldsymbol{\omega},$$

where

$$\text{vec}(\widetilde{\Sigma}_{yy}) = (I_{N^2} - \Gamma^T \otimes \Gamma^T)^{-1}\text{vec}(\widetilde{\Sigma}_{\varepsilon\varepsilon}). \tag{17}$$

In that case, the contribution of sector i to aggregate variability, λ_i, is

$$\widetilde{\lambda}_i = \frac{\widetilde{\sigma}_y^2}{\sigma_y^2}. \tag{18}$$

[4] This result follows from the fact that for any matrices A, B, and C, such that the product ABC exists, $\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B)$.

Table 2 Input-Output and Contributions to the Variability of GDP, 1988–2010

Industry Name	NAICS Code	$\widetilde{\lambda}_i^{LP}$ (Percent)	$\widetilde{\lambda}_i^{HD}$ (Percent)
Agriculture, Forestry, Fishing, and Hunting	11	0.32	0.60
Mining	21	0.33	0.51
Utilities	22	0.50	0.56
Construction	23	1.71	0.69
Manufacturing	31–33	35.31	42.77
Wholesale Trade	42	3.69	2.66
Retail Trade	44–45	4.18	1.49
Transportation and Warehousing	48–49	1.34	1.48
Information	51	2.11	1.87
Finance and Insurance	52	6.16	6.06
Real Estate, Rental, and Leasing	53	15.77	25.51
Professional, Scientific, and Technical Services	54	5.80	6.80
Management of Companies and Enterprises	55	0.63	0.72
Administrative and Support Management	56	1.15	1.37
Educational Services	61	0.07	0.02
Health Care and Social Assistance	62	3.66	1.02
Arts, Entertainment, and Recreation	71	0.08	0.05
Accommodation and Food Services	72	0.72	0.38
Other Services (except Public Administration)	81	0.71	0.53
Government	92	15.68	4.90

Notes: $\widetilde{\lambda}_i^{LP}$ is computed using Long and Plosser (1983) and uncorrelated sectoral shocks with unit variance. Similarly, $\widetilde{\lambda}_i^{HD}$ is computed using Horvath (1998) or Dupor (1999).

Table 2 shows $\widetilde{\lambda}_i$ for the sectors considered in this paper using both the Long and Plosser (1983) and the Horvath (1998) frameworks. Under the maintained assumptions, the only difference between the third column in Table 1 and the second column in Table 2 relates to input linkages across sectors. In both cases, shares are taken into account in the calculations and sectors have homogenous variances.[5] Although input-output linkages generally increase overall variance by slowing down the averaging that takes place in aggregation, Table 2 indicates that the relative importance of any one sector may increase or decrease depending in part on how important it is as an input provider to other sectors. For example, manufacturing contributes 25 percent of aggregate variability absent input-output linkages. However, when input-output linkages are taken into account, this contribution increases to 35 percent using the Long and Plosser (1983) framework. Manufacturing, therefore, plays an important

[5] This calculation highlights the importance of input-output linkages only. As shown in Foerster, Sarte, and Watson (2011), in practice, the relative magniture of shocks across sectors also matters.

role as an input provider to other sectors. In contrast, retail trade explains roughly 6 percent of aggregate variations based solely on its share in total output. Once input-output linkages are considered, the contribution of retail trade to aggregate variability falls to 4 percent. Thus, linkages of retail trade to other sectors play somewhat minor roles relative to those of other sectors.

Horvath (1998)

The model in Horvath (1998) is very similar to that of Long and Plosser (1983) but adds sectoral capital. Specifically, production in sector j now takes the form

$$Y_{j,t} = A_{j,t} K_{j,t}^{\alpha_j} (\Pi_{i=1}^{N} M_{ij,t}^{\gamma_{ij}}) L_{j,t}^{1-\alpha_j-\Sigma_{i=1}^{N}\gamma_{ij}}, \tag{19}$$

while each sector's resource constraint now reads as

$$C_{jt} + \sum_{i=1}^{N} M_{ji,t} + K_{jt+1} = Y_{j,t}, \;\; j = 1, ..., N. \tag{20}$$

Horvath's (1998) model makes two key concessions to realism for the sake of analytical tractability. First, capital is assumed to depreciate entirely within the period. In that sense, the distinction between materials and capital is more one of timing than any other consideration. Second, each sector produces its own capital. Under these assumptions, the solution for sectoral output growth is now given by

$$\Delta \mathbf{y}_t = Z^T \boldsymbol{\alpha}_d \Delta \mathbf{y}_{t-1} + Z^T \boldsymbol{\varepsilon}_t, \tag{21}$$

where $\boldsymbol{\alpha}_d$ is a diagonal matrix with the vector of sectoral capital shares, $(\alpha_1, \alpha_2, ..., \alpha_N)$ along its diagonal and $Z = (I - \Gamma)^{-1}$. This vector is based on the estimates of other value added (rents on capital) from the BEA's use table.

Similar to Long and Plosser (1983), an analytical expression for the variance of aggregate output growth is given by equation (15):

$$\sigma_y^2 = \boldsymbol{\omega}^T \Sigma_{yy} \boldsymbol{\omega},$$

where Σ_{yy} now satisfies

$$\text{vec}(\Sigma_{yy}) = (I_{N^2} - Z^T \boldsymbol{\alpha}_d \otimes Z^T \boldsymbol{\alpha}_d)^{-1} \text{vec}(Z^T \Sigma_{\varepsilon\varepsilon} Z). \tag{22}$$

There are two key differences that distinguish Horvath's (1998) framework from that of Long and Plosser (1983). First, a shock to sector i immediately propagates to other sectors by way of input-output linkages, as captured by the term $Z^T \varepsilon_t$ in (21) rather than just ε_t in (13). This follows from the fact that Horvath's (1998) model loses the one-period time-to-build feature of Long and Plosser (1983). Second, sectoral shocks propagate through time by way of capital accumulation and thus are scaled by the matrix of capital shares, as

captured by the autoregressive coefficient $Z^T\alpha_d$. Both of these features will change the variance decompositions carried out earlier as well as the nature of the propagation of sector-specific shocks.

Recall that under Long and Plosser (1983) and unit variance sectoral shocks, aggregate variability was amplified one and a half times relative to the case without sectoral linkages. Under Horvath (1998), aggregate output variance increases to 0.42 or a five-time increase relative to the case without sectoral linkages. The third column of Table 2 shows $\widetilde{\lambda}$ the contribution from different sectors to aggregate variability using the Horvath (1998) model. By and large, the sectors that contribute most to aggregate variability are the same as those in the first column of the table using the Long and Plosser (1983) framework. However, the importance of the sectors with extensive sectoral linkages is amplified in Horvath (1998). Thus, manufacturing's share of aggregate variability increases from 35 percent to 43 percent. Similarly, real estate, rental, and leasing sees its contribution to aggregate variance increase from 16 percent to roughly 26 percent. As we shall see below, intersectoral linkages take on a greater role in Horvath (1998) because sectoral shocks get propagated by way of not only the input-output matrix but also internal capital accumulation, $Z^T\alpha_d$, where α_d is the matrix of capital shares.

Some Key Assumptions and the Irrelevance of Sectoral Shocks

We suggested earlier that sectoral shocks can fail to average out as N becomes large when the distribution of sectoral shares is sufficiently leptokurtic. Aside from this consideration, one might also ask whether sectoral linkages necessarily prevent sectoral shocks from being irrelevant at the aggregate level. To that end, Dupor (1999) uses Horvath's (1998) framework to analyze the conditions under which sectoral shocks average out even in the presence of sectoral linkages. In particular, Dupor's work relies on three key conditions:

(A1) $\omega = N^{-1}\mathbf{h}$, where $\mathbf{h}$ is a vector of ones, $(1, 1, ..., 1)^T$.

(A2) $\Gamma\mathbf{h} = \kappa\mathbf{h}$, where κ is a positive scalar. Put another way, $\mathbf{h}$ is an eigenvector of the $N \times N$ matrix Γ with corresponding eigenvalue, κ. This assumption implies that all rows of Γ sum up to κ, so that all sectors serve as equally intense material input providers to all other sectors.

(A3) $\Sigma_{\varepsilon\varepsilon} = I$.

It turns out that under these assumptions, the role of sectoral shocks vanishes in aggregation not only in the environment studied by Dupor (1999), but

also in other canonical versions of the multisector growth model including Long and Plosser (1983) and, more recently, Carvalho (2007).

In the case of Long and Plosser (1983), assumption (A1) and equation (13) imply that the variance of aggregate output growth (15) can be expressed as

$$\begin{aligned}\sigma_y^2 &= N^{-2}\mathbf{h}^T\Sigma_{yy}\mathbf{h} \\ &= N^{-2}\mathbf{h}^T\Gamma^T\Sigma_{yy}\Gamma\mathbf{h} + N^{-2}\mathbf{h}^T\Sigma_{\varepsilon\varepsilon}\mathbf{h}.\end{aligned}$$

When (A2) and (A3) hold, the first term on the right-hand side of this last equation simplifies as follows,

$$N^{-2}\mathbf{h}^T\Gamma^T\Sigma_{yy}\Gamma\mathbf{h} = N^{-2}\kappa^2\mathbf{h}^T\Sigma_{yy}\mathbf{h},$$

while the second term becomes

$$\begin{aligned}N^{-2}\mathbf{h}^T\Sigma_{\varepsilon\varepsilon}\mathbf{h} &= N^{-2}\mathbf{h}^T\mathbf{h} \\ &= N^{-1}.\end{aligned}$$

It immediately follows that

$$\sigma_y^2 = N^{-1}(1-\kappa^2)^{-1}, \tag{23}$$

which indeed converges to zero at rate N.

Using the same assumptions, and by following similar steps, the variance of aggregate output growth in Horvath (1998) becomes

$$\sigma_y^2 = N^{-1}[(1-\kappa-\alpha)(1-\kappa+\alpha)]^{-1}, \tag{24}$$

which also converges to zero at rate N.

Several observations are worth noting with respect to equations (23) and (24). Under the maintained assumptions, the irrelevance of sectoral shocks holds in the limit. Hence, a question arises as to what the relevant level of disaggregation is in practice. To us, this question is mainly one that relates to technology and the nature of shocks under consideration. In particular, it is likely befitting that manufacturing and retail trade should be thought of as characterized by fundamentally different technologies, and hence affected by fundamentally different shocks, but it may also be the case that within manufacturing, "iron and steel products" should be treated differently than "metalworking machinery."

In general, the Census uses two criteria for making industry classifications. The first is the economic significance of the industry, which refers to the size of an industry at the highest level of disaggregation relative to the average size of industries in its particular division. For example, breaking up "iron and steel products" within manufacturing into two separate industries, "iron products" and "steel products" would involve comparing the size of each industry individually to the average manufacturing industry size. The notion

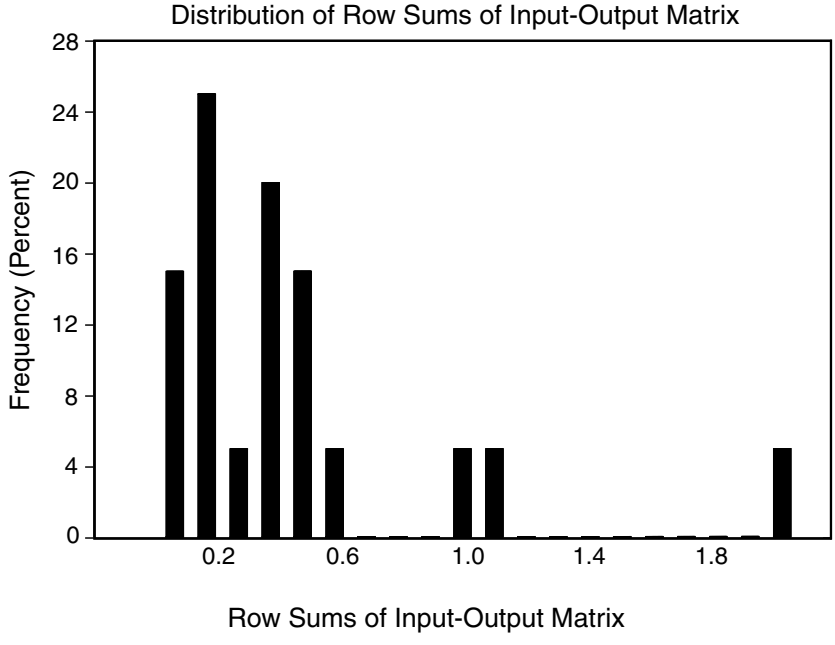
of size, or economic significance, considers five main characteristics: the number of establishments in the industry, the industry's number of employees, its payroll, its value added, and its value of shipments. A weighted average is then constructed from these five measurements, which are expressed relative to a similarly computed measure of the average size of existing industries in the pertinent division. Once a given economic significance score is reached, the industry potentially qualifies as a new classification at the highest level of disaggregation. The second criterion is based on specialization and coverage ratios. These ratios combine to measure the share of production and shipments of an industry's primary products in the economy. Conditional on meeting the first criterion, an industry is then recognized only if each of these ratios reaches a threshold level.

In an exercise that focuses on U.S. industrial production, Foerster, Sarte, and Watson (2011) consider up to 117 sectors, the highest level of disaggregation for which input-output tables from the BEA can be matched to sectoral output data. Interestingly, the authors find that the relevance of sectoral shocks for aggregate variability appear robust to the level of disaggregation. This finding arises in part because, as an empirical matter, sectoral variability increases with the level of disaggregation, thus offsetting the "averaging out" effect of N^{-1} in equations (23) and (24).

The conclusions in this section rely crucially on assumption (A2), $\Gamma\mathbf{h} = \kappa\mathbf{h}$, so that all rows of Γ must sum up to the same scalar. Put differently, this condition requires the input-output matrix to be such that all sectors serve as equally important material input providers to all other sectors. Figure 1 shows the row sums of the input-output matrix, Γ, associated with our two-digit decomposition. The figure indicates that the row sums, $\Gamma\mathbf{h}$, can differ considerably from one another in practice. Using a four-digit decomposition of industrial production, Foerster, Sarte, and Watson (2011) show that when output is disaggregated further, $\Gamma\mathbf{h}$ further displays pronounced skewness. This skewness is consistent with the notion emphasized in Carvalho (2007) that a few sectors play crucial roles as input providers. Thus, the key step that allows for aggregation despite sectoral linkages, assumption (A2), does not appear to be consistent with our sectoral data.

That said, one should be clear that the assumptions outlined in this section represent sufficient conditions for the asymptotic irrelevance of sectoral shocks. To the degree that $\Gamma\mathbf{h}$ differs from $\kappa\mathbf{h}$, so that not all sectors serve as equally important material providers to other sectors, the implications of this difference for the contribution of sectoral shocks to aggregate variability is not immediately clear. In particular, the way in which a given sectoral shock becomes amplified and propagates to other sectors, and thus affects aggregate output, generally needs to be computed numerically for a given input-output matrix, Γ, and sectoral shares, ω. It is to this consideration that we next turn our attention.

Figure 1 Importance of Sectors as Material Input Providers to Other Sectors



3. SECTORAL SHOCK PROPAGATION WITH SECTORAL LINKAGES

Given the Long and Plosser (1983) model solution in (13), the effects of sectoral shocks arising at t, ε_t, on sectoral output growth at date $t + j$ are given by

$$\frac{\partial \Delta \mathbf{y}_{t+j}}{\partial \boldsymbol{\varepsilon}_t} = (\Gamma^T)^j, \tag{25}$$

and the resulting change in aggregate output growth is

$$\boldsymbol{\omega}^T \frac{\partial \Delta \mathbf{y}_{t+j}}{\partial \boldsymbol{\varepsilon}_t} = \boldsymbol{\omega}^T (\Gamma^T)^j.$$

Consider the effects of a negative shock to a given sector, say manufacturing, denoted by ε_{mt}, so that $\boldsymbol{\varepsilon}_t = (0, 0, ..., \varepsilon_{mt}, ..., 0)^T$. Two noteworthy observations arise.[6]

[6] Observe that under conditions (A1) and (A2) in the previous section, $\omega^T \frac{\partial \Delta \mathbf{y}_{t+j}}{\partial \boldsymbol{\varepsilon}_t} = N^{-1}\kappa$ $\forall j$, which goes to zero as N becomes large.

First, because $(\Gamma^T)^0 = I$, the shock to manufacturing will only affect itself in the period in which the shock occurs. In other words $\partial \Delta y_t / \partial \varepsilon_{it} = 0$ for all sectors that are not manufacturing. There is no propagation of the shock to other sectors in the period in which the shock occurs. Hence, the contemporaneous effect of the manufacturing shock on aggregate output growth is simply $\omega_m \partial \Delta y_t / \partial \varepsilon_{mt}$, where ω_m is the share of the manufacturing sector in GDP. This result may be interpreted as the formal justification for the notion that the aggregate effects of sectoral shocks can be judged from the output share of the sector in which the shock occurs. As we shall see shortly, however, this is generally not the case.

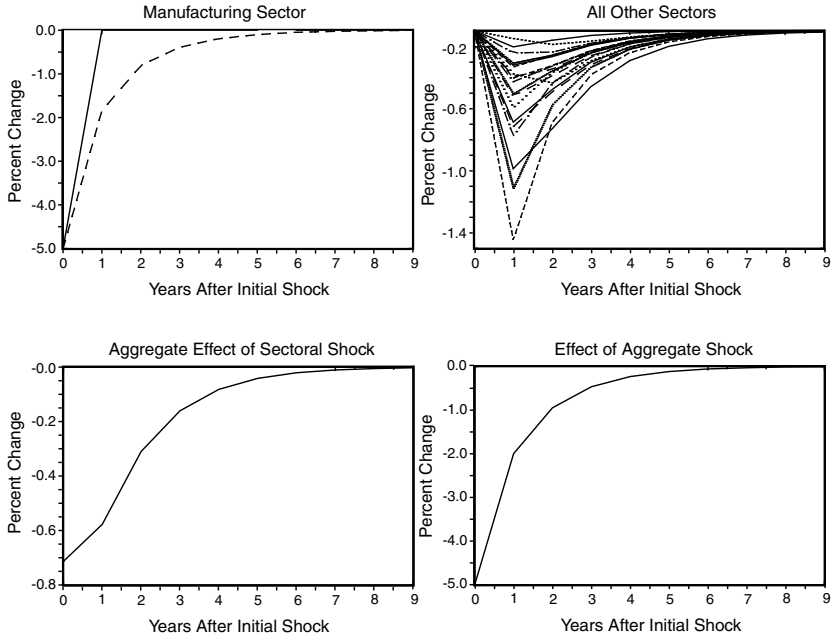
Second, the effect of the shock to manufacturing on any other sector j in the following period is given by

$$\frac{\partial \Delta y_{j,t+1}}{\partial \varepsilon_{m,t}} = \gamma_{mj}. \tag{26}$$

In other words, in Long and Plosser (1983), a shock to the manufacturing sector begins to propagate to another sector j, in the period after the shock, by exactly γ_{mj}, the amount that sector j spends on materials produced in the manufacturing sector as a fraction of sector j's total spending on inputs. Therefore, the less sector j spends on materials from sector m, the lower will be the effect of a shock to sector m on sector j.

Figure 2 depicts impulse responses to an unanticipated exogenous one-time 5 percent fall in manufacturing total factor productivity (TFP) in the Long and Plosser (1983) economy. The solid line in the top left-hand panel of Figure 2 shows the time path of the shock. The dashed line in that panel shows the response in output growth in the manufacturing sector. By design, the shock to manufacturing TFP has dissipated after one period. As discussed above, because the shock does not propagate to other sectors contemporaneously, thus also preventing a feedback effect from those sectors back into manufacturing, the initial decline in manufacturing output growth is exactly equal to the size of the decline in TFP, or 5 percent. Moreover, we can see that the effect of the shock on manufacturing output growth is considerably longer lived than the shock itself. The top right-hand panel of Figure 2 explains why. That panel depicts the effects of the fall in TFP in manufacturing on all other sectors. As suggested above, the initial effect of the manufacturing TFP decline on all other sectors is zero. However, in the period after the manufacturing disturbance occurs, the shock has spread to all other sectors by way of input-output linkages so that these all experience a decline in output growth. The size of this decline in the different sectors reflects the degree to which they rely on manufacturing as an input provider, $\gamma_{mj} > 0$, which then feeds back into manufacturing in so far as it uses the output of those sectors as inputs, $\gamma_{im} > 0$. In the top right-hand panel of Figure 2, the largest decline in output in the period following the manufacturing TFP shock occurs in the construction sector at -1.4 percent.

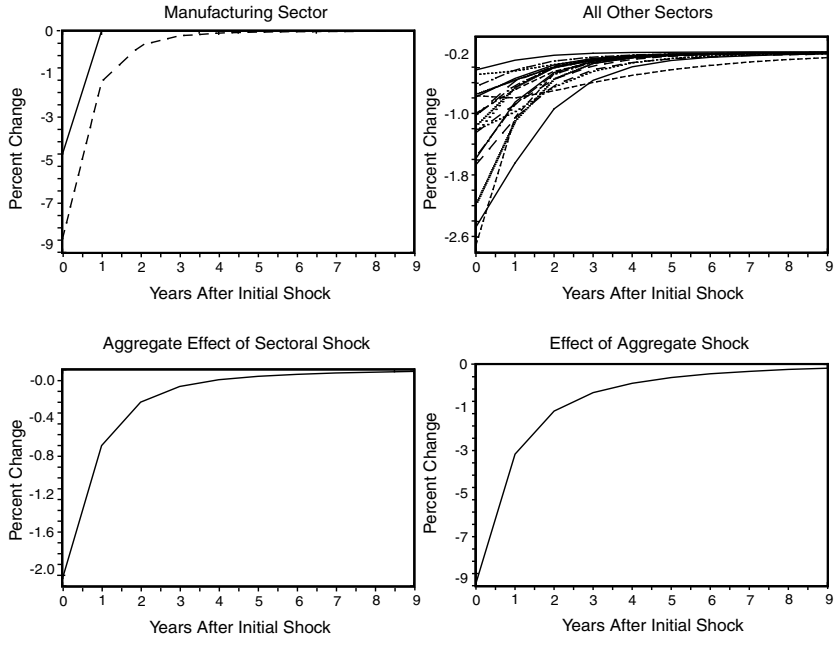
Figure 2 Effects of One-Time Unanticipated Shocks in Long and Plosser (1983)



The bottom left-hand panel of Figure 2 illustrates the aggregate effect of the manufacturing shock on output. As indicated above, because the share of manufacturing in GDP is approximately 0.14, the initial effect of the shock on aggregate output in the Long and Plosser (1983) model is roughly 0.14×5 or a 0.7 percent decline in GDP. In addition, note that this aggregate effect is considerably more persistent than the initial one-time decline in manufacturing TFP. As before, this feature arises mainly from the propagation of the shock to other sectors which, by way of a feedback effect, induces persistence in the output decline of all sectors and, therefore, at the aggregate level as well.[7] For comparison, the bottom right-hand panel of Figure 2 shows the effect on aggregate output of an unanticipated one-time decline in TFP in all sectors, which may be interpreted as an unanticipated aggregate TFP shock. As in the case of sectoral shocks, the effect on aggregate output is considerably

[7] Recall that the model in Long and Plosser (1983) contains no internal propagation mechanism, such as might occur through capital accumulation, other than the one-period delivery lag in materials. Strictly speaking, the induced persistence in the impulse responses to a one-time sectoral shock in Figure 2 stems from the combination of that lag with sectoral linkages.

Figure 3 Effects of One-Time Unanticipated Shocks in Horvath (1998)



longer lived than the shock itself. However, because the 5 percent fall in TFP now applies to all sectors, the size of the output decline is considerably more pronounced.

In Horvath (1998), the effects of sectoral shocks arising at t, ε_t, on sectoral output growth at date $t + j$ are given by

$$\frac{\partial \Delta \mathbf{y}_{t+j}}{\partial \boldsymbol{\varepsilon}_t} = (Z^T \boldsymbol{\alpha}_d)^j Z^T. \tag{27}$$

In this case, a negative shock to the manufacturing sector immediately propagates to other sectors by way of input-output linkages, as embodied in $Z^T = (I - \Gamma^T)^{-1}$, because materials are used within the period. This is the source of the notable amplification of sectoral shocks in Horvath (1998) relative to one without input-output linkages. In particular, the variance-covariance matrix of sectoral output growth (absent any propagation) is $Z^T \Sigma_{\varepsilon\varepsilon} Z$ rather than just $\Sigma_{\varepsilon\varepsilon}$. In addition, sectoral shocks further propagate over time through their effects of capital accumulation by way of input-output linkages, $Z^T \boldsymbol{\alpha}_d$. In other words, the model contains an internal propagation mechanism that potentially extends the life of the original shock on aggregate economic activity.

Analogous to Figure 2, Figure 3 shows the effects of a 5 percent unanticipated one-time decline in manufacturing TFP, but this time in the Horvath

(1998) economy. The impulse responses in Figure 3 highlight several key differences with those that obtain in the Long and Plosser (1983) model. First, the effect of the fall in manufacturing TFP is immediately amplified through input-output linkages. In particular, while TFP falls by 5 percent in the top left-hand panel of Figure 3, manufacturing output growth falls by 9 percent or nearly double the size of the shock. This stems from the fact that materials are used within the period in Horvath (1998). As pointed out earlier, in the solution for sectoral output growth (21), output growth at time t reflects the effects of contemporaneous sectoral links, $(I - \Gamma^T)^{-1}\varepsilon_t$, instead of the effects of sectoral disturbances alone, ε_t, in the solution to the Long and Plosser (1983) model, (13). The top right-hand panel of Figure 3 illustrates this feature and, unlike Figure 2, shows that output growth falls in all sectors at the time that manufacturing TFP declines. In addition, observe that the output decline in all sectors is considerably larger than that in the period after the shock in the Long and Plosser (1983) economy in Figure 2. Second, the top right-hand panel of Figure 3 suggests that impulse responses in some sectors are slightly non-monotonic so that, in those sectors, the outlook gets worse before it gets better even though the manufacturing TFP shock itself has already dissipated.[8] Third, and related to this last observation, the effects of the one-time decline in TFP is somewhat more persistent than in the Long and Plosser (1983) framework. Finally, because of contemporaneous intersectoral linkages, the effect of the decline in manufacturing TFP on aggregate output is now noticeably more pronounced than in Figure 2. Specifically, aggregate output growth declines by 2 percent on impact and continues to be below its steady state well after the shock has dissipated. We saw earlier that the contribution of manufacturing's share to the fall in aggregate output is roughly 0.7 percent in the bottom left-hand panel of Figure 2. Therefore, in this case, contemporaneous input-output links add about 1.3 percent to the decline in aggregate output on impact.

The basic lesson of this section is that input-output linkages, and potentially other forms of complementarities in production, propagate and amplify the effects of sectoral disturbances. Therefore, using the share of the sector in which a disturbance occurs as a gauge of its effect on aggregate output constitutes, in general, a relatively poor approximation. However, the extent of the amplification and propagation mechanism that results from intersectoral linkages depends on the particular economic environment in which these linkages operate. In their recent article, Foerster, Sarte, and Watson (2011) extend the analysis in this section to include intersectoral linkages in investment goods (so that some sectors produce new capital goods for other sectors), less than full capital depreciation within the period, and allow for aggregate

[8] Given our calibration of Γ and α_d based on the input use tables from the BEA, $Z^T\alpha_d$ has complex eigenvalues.

shocks. They find that the importance of sectoral disturbances in explaining aggregate fluctuations has noticeably increased over time and that, over the period 1984–2007, these disturbances explain half the variation in U.S. industrial production. However, although the nature of intersectoral production has changed over time, the authors also find that changes in the input-output matrix reflecting new sectoral links has not led to greater propagation of shocks.

4. CONCLUDING REMARKS

This article has provided an overview of some key aspects of the effects of sectoral shocks on aggregate economic activity. It discussed the role of sectoral shares in determining each sector's contribution to aggregate variations. It also illustrated how input-output linkages in production influenced the amplification and propagation of sectoral shocks. The mechanisms by which this amplification and propagation take place depend importantly on the details of the economic environment in which intersectoral linkages operate. In general, because of input-output linkages across sectors, using the share of the sector in which a disturbance occurs as a gauge of its effect on aggregate output constitutes a poor approximation. In addition, the key condition required of the input-output matrix that lead sectoral shocks to average out in aggregation, carefully articulated in Dupor (1999), does not appear to apply in practice. Using an input use matrix obtained from the BEA for 1997, as well as a two-digit level disaggregation of GDP, suggests that manufacturing and real estate, rental, and leasing contribute the most to aggregate variations.

REFERENCES

Acemoglu, Daron, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. 2010. "Cascades in Networks and Aggregate Volatility." MIT Working Paper (October).

Carvalho, Vasco M. 2007. "Aggregate Fluctuations and the Network Structure of Intersectoral Trade." Manuscript, University of Chicago.

Dupor, Bill. 1999. "Aggregation and Irrelevance in Multi-sector Models." *Journal of Monetary Economics* 43 (April): 391–409.

Foerster, Andrew T., Pierre-Daniel G. Sarte, and Mark W. Watson. 2011. "Sectoral versus Aggregate Shocks: A Structural Analysis of Industrial Production." *Journal of Political Economy* 119 (February): 1–38.

Gabaix, Xavier. 2011. "The Granular Origins of Aggregate Fluctuations." *Econometrica* 79 (May): 733–72.

Horvath, Michael. 1998. "Cyclicality and Sectoral Linkages: Aggregate Fluctuations from Independent Sectoral Shocks." *Review of Economic Dynamics* 1 (October): 781–808.

Long, John B., Jr., and Charles I. Plosser. 1983. "Real Business Cycles." *Journal of Political Economy* 91 (February): 39–69.

Economic Quarterly—Volume 97, Number 2—Second Quarter 2011—Pages 175–187

Legal Protection to Foreign Investors

Juan Carlos Hatchondo and Leonardo Martinez

Governments have the ability to affect the return of foreign investments. The typical expropriation is one in which a government unilaterally transfers the property right of a firm without compensating the previous owners. However, governments can also expropriate through discriminatory taxation or regulation. For instance, governments can impose a high differential tax rate on a firm's benefits, limit the prices or locations at which a firm may sell its products, limit royalty payments, etc. Chifor (2002) notes that indirect expropriation through taxation and regulation has supplanted direct takings as the most common type of expropriation. Governments can also expropriate by defaulting on their debt. Borensztein and Panizza (2008) report 114 default episodes during the last 30 years. Sovereign defaults have been common in developing countries, though we have observed in 2010 and 2011 a significant increase in the perceived probability of a sovereign default in some European countries.

Governments could resort to expropriations to increase current fiscal resources or as a way to avoid implementing unpopular policies that could, for example, avert a sovereign default. However, expropriations can be costly in the long run. For instance, expropriations may be followed by lower capital inflows, expropriated firms may be run less efficiently, or expropriations may distort the behavior of firms that were not directly affected but fear being so in the future. When the long-run benefits derived from foreign investment or from better borrowing terms in international capital markets offset the potential short-run gains from expropriation, the welfare of domestic households could be increased by limiting the government's ability to expropriate foreign

Hatchondo is an economist with the Federal Reserve Bank of Richmond. Martinez is with the International Monetary Fund. For helpful comments, we thank Borys Grochulski, Andreas Hornstein, Nika Lazaryan, and Felipe Schwartzman. The views expressed herein are those of the authors and should not be attributed to the IMF, its Executive Board, or its management; the Federal Reserve Bank of Richmond; or the Federal Reserve System. E-mail: juancarlos.hatchondo@rich.frb.org.

investors.[1] One mechanism that could be used to limit expropriation risk is to pass national laws that explicitly grant protection to investors. Yet, the fact that the authorities in charge of enforcing the law are the same ones that may violate investors' property rights casts doubt about the degree of protection that can be offered by local legal systems. Instead, one mechanism that is used to discipline current and future governments is to reduce the degree of sovereignty by increasing the government's exposure to foreign courts. This second mechanism is the focus of the present article.

One way sovereigns involve foreign courts is by signing international investment agreements that grant foreign investors the right to settle a dispute in international arbitration tribunals. Governments have also ratified international conventions that bind them to recognize arbitration tribunals' decisions concerning investment—and commercial—disputes. As far as the success of litigating investors is concerned, several authors (see, for example, Dolzer and Stevens [1995]; Reed, Paulsson, and Blackaby [2004]; and Baldwin, Kantor, and Nolan [2006]) argue that governments have tended to comply with unfavorable rulings in international tribunals. This has been so despite investors' limited legal means available to enforce reparation payments. The fact that governments have complied with unfavorable rulings suggests the presence of other types of costs. For example, ignoring unfavorable rulings may send a negative signal about the government's commitment to respect investors' property rights, which may have adverse aggregate consequences on capital inflows. But the apparent success could also be contaminated by the presence of sample bias in the set of cases that has been submitted to international tribunals. Investors who expected difficulties in collecting compensation payments may have decided not to bear the costs of litigation.[2] Note also that investors' past success in litigations may not be a good predictor of future success.

International investment agreements were originally designed with the intent to promote foreign direct investment, but they have gradually adjusted to extend protection to other types of investment. In part, this may explain the fact that not all investment agreements explicitly protect holders of sovereign debt.

In order to protect foreign lenders, governments have increasingly chosen to issue debt in international financial centers such as New York. This practice exposes defaulting governments to litigations in foreign national courts.

---

[1] Some authors have argued that the risk of losing political support could serve as an enforcement mechanism that protects domestic residents. Hatchondo and Martinez (2010) present a survey on the politics of sovereign defaults.

[2] A country that anticipates an unfavorable tribunal decision could also withdraw from an international investment agreement. Bolivia did so in 2007, shortly after a Dutch-based subsidiary of Telecom Italia filed a claim seeking arbitration in an alleged case of expropriation of a telecommunications investment.

Holders of bonds in default that were issued in foreign countries can enforce repayment in courts by diverting some type of sovereign assets located outside the defaulting country. However, defaulting governments have, in general, succeeded in locating those assets outside the reach of creditors. It should be mentioned that even when holders of debt in default do not succeed in collecting payments, they may be imposing a cost to the defaulting sovereign. This occurs because, in order to keep their assets outside the reach of creditors, governments in default may not be able to issue debt in international financial markets.

The rest of the article is organized as follows. Section 1 discusses international investment agreements. Section 2 discusses the protection granted to lenders by the issuance of sovereign debt in international financial centers. Section 3 concludes.

1. INTERNATIONAL INVESTMENT AGREEMENTS

This section discusses the legal protection that international investment agreements grant to a broad class of foreign investments. The typical investment agreement takes the form of a reciprocal bilateral investment treaty in which two countries agree on a set of conditions under which the nationals of one country may seek compensation if their investments in the other country are affected. There are also a few multilateral investment agreements, like Chapter 11 of the North American Free Trade Agreement between Canada, Mexico, and the United States.

It should be mentioned that international investment agreements do allow for states to expropriate foreign investors under certain circumstances, namely that the expropriation is done for a public purpose, in accordance with the law, in a nondiscriminatory manner, and after paying a prompt and adequate compensation to the property owner (see Dolzer and Stevens [1995] and Organization for Economic Cooperation and Development [2004]). The exact description of the conditions under which investors are granted the right to request compensation varies across treaties.

International investment agreements specify the arbitration rules that investors and governments can follow to settle disputes. The most common arbitration rules are specified by the International Center for the Settlement of Investment Disputes (ICSID) and the United Nations Commission on International Trade Law (UNCITRAL).[3] In what follows, we describe how these arbitration rules work and the enforcement mechanisms available to investors.

[3] According to the United Nations Conference on Trade and Development (UNCTAD 2009), of the 317 investor-state disputes outstanding in international tribunals, 201 had been filed under the ICSID arbitration rules and 83 under UNCITRAL arbitration rules.

Arbitration under ICSID Rules

The ICSID was established in 1965 under the ICSID Convention.[4] The ICSID Convention was sponsored by the World Bank with the objective of promoting the flow of foreign direct investments; by the end of 2010, it had been ratified by 157 countries. Countries that ratify the Convention agree to abide by the ICSID arbitration rules, including the enforcement of the decisions of its tribunals. Reed, Paulsson, and Blackaby (2004) and the United Nations Conference on Trade and Development (UNCTAD 2009) have noted that the increase in the number of bilateral investment agreements observed in the last two decades has been accompanied by a parallel increase in the number of arbitrations conducted under ICSID rules.

The ICSID, which is one of the five organizations that make up the World Bank group, provides facilities for the resolution of investment disputes through conciliation or arbitration. For instance, it assists in the constitution of tribunals, it administers the funds necessary to cover the costs of the proceedings, it produces publications to contribute to the understanding of international investment laws, etc. The ICSID is not in charge of conducting arbitration proceedings.

With respect to the enforcement of arbitration tribunals' decisions, Article 54 of the ICSID Convention states that final decisions of ICSID tribunals must be considered equivalent to "final judgments" of local courts in countries that have signed the ICSID Convention.[5] Baldwin, Kantor, and Nolan (2006) point out that this clause may not necessarily imply that final decisions of ICSID tribunals cannot be challenged in local courts because in some countries the legal system allows, under some circumstances, for challenges to local equivalents of final judgments. In fact, Baldwin, Kantor, and Nolan (2006) review four cases in which ICSID rulings were challenged in local courts. Even though some received favorable judgments in lower courts, eventually all challenges were unsuccessful. Another important implication of Article 54 of the ICSID Convention is that final decisions of ICSID tribunals not only bind in the country that hosted the expropriated investment, but also in all countries that have signed the Convention.[6] This implies that investors could seek reparation in any country that has signed the Convention and not only

---

[4] Convention refers to an agreement among countries that establishes obligations to the countries that ratify it.

[5] Article 54(1) of the Convention states that: "[e]ach Contracting State shall recognize an award rendered pursuant to this Convention as binding and enforce the pecuniary obligations imposed by that award within its territories as if it were a final judgment of a court in that State. A Contracting State with a federal constitution may enforce such an award in or through its federal courts and may provide that such courts shall treat the award as if it were a final judgment of the courts of a constituent state."

[6] As a clarification, Reed, Paulsson, and Blackaby (2004) point out that "...[i]n the context of ICSID arbitration, enforcement is generally indistinguishable from recognition. The two terms are used in a single phrase—recognition and enforcement—that broadly refers to all steps leading

in the country where the expropriation took place. However, Reed, Paulsson, and Blackaby (2004) and Baldwin, Kantor, and Nolan (2006) note that it is assumed that under the ICSID Convention, signatory states shall enforce the judgments according to their own national law, which has two implications. First, sovereign immunity laws protect (some) government assets from foreign investors when investors attempt to seize government assets in jurisdictions different from the one in which the expropriation took place.[7] Second, the Convention does not obligate a signatory country to enforce the compensation of investors after a favorable arbitration decision if the local law does not allow enforcement of compensation of equivalent local court judgments.

Baldwin, Kantor, and Nolan (2006) discuss that if a government refuses to honor a tribunal's decision, the affected investor could resort to the International Court of Justice (the primary judicial body of the United Nations). But this alternative also presents its own difficulties. First, the International Court of Justice only accepts disputes between two states, which means that the affected investor should request its government to sponsor such a claim. There are political and economic reasons why government authorities may decide not to sponsor claims of individual investors against another state. Second, it is unclear that the International Court of Justice will accept jurisdiction over the dispute without a consent of the government that refused to honor the arbitration tribunal's decision. Third, even in the case of a favorable decision in the International Court of Justice, the means to collect the payments are limited. Potentially, a state could take the issue before the Security Council, but it is highly unlikely that the Security Council would decide to enforce the claims.

The previous discussion suggests that the actual legal protection enjoyed by investors is somewhat limited. Despite that, Dolzer and Stevens (1995); Reed, Paulsson, and Blackaby (2004); and Baldwin, Kantor, and Nolan (2006) state that, with a few exceptions, governments have complied with ICSID tribunals' decisions. Besides, on some occasions, the parties reached a settlement before a final decision was made and, on other occasions, before a case was submitted for arbitration. The authors above argue that there may be various reasons why governments comply. First, it could be that countries that have ratified the Convention are the ones that try to attract foreign investments and backing out of honoring the decisions of arbitration tribunals may discourage future investors. That said, the tradeoffs or preferences of government

---

up to, but stopping short of, actual execution of an award." This meaning is different from the meaning that the term enforcement is typically assigned in economics.

[7] For instance, the French company Liberian Easter Timber Corporation (LETCO) obtained in 1986 an arbitration against Liberia for breach of a forestry concession. LETCO first tried in a New York court to obtain the right to seize registration fees and taxes owed to the government of Liberia, but the court ruled against LETCO based on the U.S. Foreign Sovereign Immunities Act. Later LETCO tried in a court in Washington, D.C., to obtain the right to seize bank accounts of the Liberian Embassy in the United States and the court also ruled against LETCO.

authorities that were in office when the country ratified the Convention may differ from the ones of government authorities that are supposed to enforce an unfavorable arbitration decision, and from the ones of future governments. Second, given that the ICSID is part of the World Bank, it may be expected that the World Bank could withhold benefits—like extending new loans—to countries that refuse to comply.

Arbitration under UNCITRAL Rules

The UNCITRAL was established in 1966 by the United Nations with the objective to help harmonize and unify the law of international trade. Since then, the UNCITRAL has prepared several conventions, model laws, and other instruments related to laws of trade transactions. Among the contributions that UNCITRAL has developed are rules for arbitration of commercial disputes (see UNCTAD [2003]), which were designed to offer a well-specified international arbitration procedure that could be used in a variety of disputes, including disputes concerning the expropriation of foreign investments.

The enforcement of an arbitration tribunal decision that acted according to the UNCITRAL rules depends on the conventions ratified by the countries of the parties in dispute. The most common instrument governing the enforcement of international arbitrations is the United Nations Convention on Recognition and Enforcement of Foreign Arbitral Awards of 1958, also known as the New York Convention. The New York Convention, which had been ratified by 145 countries by the end of 2010, requires that the states that have ratified it recognize and enforce international arbitration agreements and foreign arbitral decisions issued in other contracting states, subject to certain exceptions. This means that two parties can decide to locate their disputes in a third, neutral country, knowing that the tribunal's decision can be enforced in any country that has adhered to the Convention. There are also regional conventions, like the Inter-American Convention on International Commercial Arbitration, that can be invoked to pursue the enforcement of international arbitration decisions. Reed, Paulsson, and Blackaby (2004) argue that tribunals' decisions enforced under the ICSID Convention are more favorable to recognition than the ones enforced under the New York Convention or regional conventions, as the latter allow for challenges in local courts under more circumstances than do the former. The enforcement limitations described for the case of the ICSID Convention also apply to the New York Convention and other regional conventions.

2. SOVEREIGN DEBT

This section reviews the legal protection enjoyed by holders of debt in default.[8] Holders of sovereign bonds issued in New York, London, or other financial centers can resort to courts in those jurisdictions in order to enforce repayment (subject to certain conditions such as the majority enforcement provision in collective action clauses). That said, the bondholders' ability to enforce courts' rulings is uncertain and the absence of a well-specified international bankruptcy procedure and successive law changes have generated a significant degree of heterogeneity in the success of litigations of holders of defaulted sovereign bonds (see Panizza, Sturzenegger, and Zettelmeyer [2009] and the references therein). The discussion below describes that, de facto, bondholders' ability to enforce debt repayment through the judicial system has been quite limited.

Buchheit (1995) explains that, until the first half of the twentieth century, most countries (including the United States) recognized an "absolute" theory of sovereign immunity, which implied that sovereigns could not be sued in foreign courts without their consent. The United States began to recognize a "restrictive" theory of sovereign immunity in 1952, which limited sovereigns' immunity for commercial activities carried on outside sovereigns' territories. This principle turned into law in 1976 with the approval of the Foreign Sovereign Immunities Act. That law specifies that sovereigns can be judged in U.S. courts for their commercial contracts signed with foreign counterparties, and several court decisions have confirmed that bond issuances in U.S. markets are to be considered commercial activities. A similar law was approved in the United Kingdom in 1978 (the State Immunity Act), and most countries now have similar laws (see Buchheit [1995]).[9]

In spite of the more limited sovereign immunity, creditors who tried to collect sovereign debt through judicial systems experienced mixed results (see Sturzenegger and Zettelmeyer [2006] and Panizza, Sturzenegger, and Zettelmeyer [2009]). This casts doubt on the degree of protection granted by issuing debt in developed countries. The challenge that litigators face is not so much to obtain judgments against a sovereign debtor but to enforce that judgment. For instance, the Foreign Sovereign Immunities Act grants creditors the right to seize sovereigns' property in the United States though litigators

[8] The anonymity of bondholders limits governments' ability to default only on foreigners. In contrast, it may be easier for governments to target foreign firms for expropriation, especially in developing countries with underdeveloped stock markets.

[9] The legal protection granted to debtors was raised by Bulow and Rogoff (1990) as a potential source of the excessive borrowing that led to the debt crisis of the 1980s. As a result, Bulow and Rogoff (1990) propose to augment sovereign immunity for debt liabilities. This would also induce governments in developing countries to improve domestic institutions that determine the enforceability of contracts or the accountability of government authorities. They argue that those reforms would enable developing countries to attract foreign capital flows while also helping incoming capital flows to be allocated to better projects.

can only seize property that "is or was used for the commercial activity upon which the claim is based" (Foreign Sovereign Immunities Act 1976). Given that sovereigns usually do not need to use any of their property located in the United States to issue debt, and that the financial assets obtained at the time of the bond issuances are no longer located in the United States, the repayment that creditors may expect to obtain through that route is minimal. Creditors have also tried to attach international reserves of the country in default but with limited success (see Panizza, Sturzenegger, and Zettelmeyer [2009]). Of course, a sovereign would only choose to default when there are no significant assets investors could attach.

One of the most prominent cases in which creditors were able to induce repayment was that of *Elliott Associates, L.P., v. Banco de la Nacion and the Republic of Peru*.[10] In 1996 the "vulture fund" Elliott purchased, in the secondary market, loans that had been extended to Banco de la Nacion and Banco Popular del Peru and that had been guaranteed by the Peruvian government.[11] The loans were bought for $11.4 million and had a face value of $20.7 million. Those bonds were part of government debt that was scheduled to be included in the Brady Plan restructuring. The Brady restructuring agreement was finalized in March 1997, and was accompanied by a promise not to provide any preferential treatment to creditors who had not participated in the agreement. For that reason, Peruvian authorities refused Elliott's demands for full repayment. Elliott started litigations in a New York court. In 2000, it obtained authorization to recover $55.7 million from the government of Peru for the principal and past due interests up to such date and post-judgment interest. Even though Elliott did not manage to confiscate property belonging to Peru's government, it obtained a court authorization to intercept and attach the first payment that Peru's government was about to make through the Chase Manhattan Bank in New York to creditors who had participated in the Brady restructuring agreement. Elliott was also able to obtain enforcement orders from courts in Luxembourg, the United Kingdom, Germany, and Canada. In response to that, Peru's government decided to channel the Brady bonds payment through Euroclear: a financial company that operates in Brussels and that provides domestic and international securities services. Elliott succeeded in convincing the Brussels Court of Appeals to suspend those payments. After that, Peru's government decided to settle by paying Elliott $58.4 million and not risk defaulting on its new debt by not being able to pay on time creditors who had participated in the restructuring agreement. Defaulting on Brady

[10] See Gulati and Klee (2001), Nolan (2001), and Singh (2003).

[11] A vulture fund typically refers to an investment company that purchases debt claims in secondary markets at a relatively large discount because the debtor has defaulted or there is a high chance of default. In the event of a default, these investors have the legal expertise to litigate and are willing to hold those debt claims for many years until they reach a settlement with the debtor.

bonds would have triggered the right of all Brady bondholders to demand full repayment of their securities at that time. Ex-post, Elliott made a return of around 400 percent in four years for that investment (without including legal fees).

Panizza, Sturzenegger, and Zettelmeyer (2009) note that, for several reasons, the success of Elliott's strategy proved to be more of an exception than a rule. First, the legal argument used by Elliott was weak and relied on a controversial interpretation of the *pari passu* clause (see Gulati and Klee [2001]).[12] The argument presented by Elliott at the Brussels Court of Appeals was that Peru's government was trying to use Euroclear to violate the right of equal treatment of creditors, and that right was entitled to Elliott since the loans it owned contained the *pari passu* clause. That interpretation of the *pari passu* clause was rejected in courts in several subsequent litigations. Second, the law changed to avoid other cases like *Elliott v. Peru*. For instance, Belgium passed a law that tries to prevent creditors from obtaining court orders that could intercept payments from a sovereign to its bondholders. Third, sovereigns could move preemptively by settling payments within their legal jurisdiction or by using the Bank of International Settlements, which would prevent litigators from intercepting those payments.

One notorious case in which holders of debt in default have not been able to induce repayment through the judicial system is the 2001 Argentine default. For bonds issued in Argentina, the government decided in 2002 to change the currency of denomination (from U.S. dollars to Argentine pesos). The pesification of government debt was done using an exchange rate below its market value and Sturzenegger and Zettelmeyer (2006) estimate a mean recovery rate of 64 percent across bonds. For debt issued in foreign countries, the Argentine government proposed in 2004 to exchange those bonds with three new securities from which bondholders could choose. The exchange took place in 2005 with a participation rate of 76 percent and with a recovery rate ranging between 25 percent and 29 percent, according to Sturzenegger and Zettelmeyer (2005). In addition, the Argentine government passed a law that forbids the executive branch from negotiating with creditors who do not participate in the exchange and from incurring in transactions with bondholders arising from any court order. In spite of that, some creditors who did not participate in the exchange (holdouts) litigated in the United States and other developed countries' courts. They managed to obtain judgment orders against Argentina's assets but they have not succeeded in confiscating assets. It must be said that the limited success of bondholders does not necessarily mean that the litigation process has been costless for Argentina. Holdouts may have

[12] Many sovereign bonds include a *pari passu* clause that states that bondholders rank equally in priority of payments. The clause limits the ability of sovereigns to dilute past claims by issuing new debt that ranks senior to previous bond issuances.

barred Argentina from international capital markets because the government may be unable to receive the proceeds of bond issuances before holdouts are paid off. That may have motivated Argentine authorities to open up negotiations with holdouts in 2010, after Congress passed a law interrupting, for one year, the ban to negotiate with holdouts.

In terms of the implications for the future, Buchheit and Gulati (2010) and others note that there has been an increased use of collective action clauses in sovereign bond contracts in recent years.[13] This may curb the ability of bondholders to hold out and not accept the terms of restructuring agreements with the hope that they may obtain a better deal after litigating. In addition to that, Buchheit and Gulati (2010) mention that legislative initiatives have been considered in the United States, United Kingdom, and other developed countries to reduce "vulture creditor activity." These developments may facilitate debt restructuring processes, but if that makes defaults and subsequent renegotiations less costly, it may deteriorate the terms at which sovereigns can borrow.

International Arbitration and Sovereign Debt

Are holders of sovereign debt in default entitled to seek reparation in arbitration tribunals? Griffin and Farren (2005) and Cross (2006) argue that a higher recovery may be expected after arbitration in an ICSID tribunal than after litigation in a national court located in the country where the bonds were originally issued. This statement is partially based on the fact that countries have complied with ICSID rulings. In addition, resorting to the ICSID may be more efficient given that its decisions are equivalent to final judgments in all ICSID member states, whereas national court judgments must be validated in other countries.

In line with this reasoning, in 2006 a group of 170,000 Italian holders of Argentine defaulted debt requested arbitration under the ICSID Convention, invoking the bilateral investment agreement between Italy and Argentina. This request was followed by similar requests of two other groups of Italian bondholders. These cases are still pending and some experts believe that it is unlikely that the arbitration tribunal will accept jurisdiction (see Waibel [2007]). Litigations in ICSID tribunals might become a more widespread strategy in coming years if ICSID tribunals' rulings enable bondholders to recover a higher fraction of their claims.

[13] Collective action clauses specify that if a certain percentage of bondholders agree on a debt restructuring plan, that plan binds for all bondholders, including those who opposed it.

3. CONCLUSIONS

This article illustrates that foreign investors enjoy legal protection, but this protection is imperfect. Several analysts argue that governments have tended to comply with unfavorable rulings of international arbitration courts. This may also be consistent with the fact that sovereign default episodes observed in recent years were followed by relatively friendly debt restructuring agreements (see Sturzenegger and Zettelmeyer [2005]). The case of Argentina has been more exceptional and illustrates the limited legal protection available when a sovereign debtor decides not to repay bondholders who did not participate in the debt restructuring agreement.[14]

The fact that expropriated investors may have difficulties in being repaired does not mean that there are no costs associated with ignoring foreign court or tribunal decisions. For instance, the absence of Argentine sovereign debt issuances in financial centers—because of the risk that bondholders of Argentine debt in default may divert the receipts from those issuances—may have imposed a cost to the Argentine government. However, it is unclear how significant that cost may be. In the case of investment disputes, a potential cost of not complying with unfavorable rulings is that it may send a negative signal about the government's commitment to respect investors' property rights, which may have aggregate negative effects on capital inflows.

REFERENCES

Baldwin, Edward, Mark Kantor, and Michael Nolan. 2006. "Limits to Enforcement of ICSID Awards." *Journal of International Arbitration* 23 (1): 1–24.

Borensztein, Eduardo, and Ugo Panizza. 2008. "The Costs of Sovereign Default." IMF Working Paper 08/238 (October).

Buchheit, Lee C. 1995. "The Sovereign Client." *Journal of International Affairs* 48 (January): 527–40.

Buchheit, Lee C., and G. Mitu Gulati. 2010. "Responsible Sovereign Lending and Borrowing." United Nations Conference on Trade and Development Discussion Paper 198 (April).

[14] UNCTAD (2011) reports that Argentina is the country with the largest number of current pending investment disputes in international arbitration tribunals. As discussed in this article, eventual rulings favorable to creditors would not guarantee full repayment.

Bulow, Jeremy, and Kenneth Rogoff. 1990. "Cleaning up Third World Debt Without Getting Taken to the Cleaners." *Journal of Economic Perspectives* 4 (Winter): 31–42.

Chifor, George. 2002. "Caveat Emptor: Developing International Disciplines for Deterring Third Party Investment in Unlawfully Expropriated Property." *Law and Policy in International Business* 33 (January): 179–85.

Cross, Karen. 2006. "Arbitration as a Means of Resolving Sovereign Debt Disputes." *American Review of International Arbitration* 17 (3): 335.

Dolzer, Rudolf, and Margrete Stevens. 1995. *Bilateral Investment Treaties*. The Hague, The Netherlands: Kluwer Law International and The International Centre for Settlement of Investment Disputes.

Foreign Sovereign Immunities Act of 1976. 1976. *U.S. Code*. Title 28, sec. 1330, 1332, 1391f, 1441d, and 1602–11.

Griffin, Peter, and Ania Farren. 2005. "How ICSID Can Protect Sovereign Bondholders." *International Financial Law Review* 24 (September): 21–4.

Gulati, G. Mitu, and Kenneth N. Klee. 2001. "Sovereign Piracy." *The Business Lawyer* 56 (February): 635–51.

Hatchondo, Juan Carlos, and Leonardo Martinez. 2010. "The Politics of Sovereign Defaults." Federal Reserve Bank of Richmond *Economic Quarterly* 96 (3): 291–317.

Nolan, John. 2001. "Emerging Market Debt & Vulture Hedge Funds: Free-Ridership, Legal & Market Remedies." Special Policy Report 3. Derivatives Study Center, Financial Policy Forum, Washington, D.C. (September 29).

Organization for Economic Cooperation and Development. 2004. "'Indirect Expropriation' and the 'Right to Regulate' in International Investment Law." OECD Working Papers on International Investment No. 2004/4 (September).

Panizza, Ugo, Federico Sturzenegger, and Jeromin Zettelmeyer. 2009. "The Economics and Law of Sovereign Debt and Default." *Journal of Economic Literature* 47 (September): 651–98.

Reed, Lucy, Jan Paulsson, and Nigel Blackaby. 2004. *Guide to ICSID Arbitration*. The Hague, The Netherlands: Kluwer Law International.

Singh, Manmohan. 2003. "Recovery Rates from Distressed Debt: Empirical Evidence from Chapter 11 Filings, International Litigation, and Recent Sovereign Debt Restructurings." IMF Working Paper 03/161 (August).

Sturzenegger, Federico, and Jeromin Zettelmeyer. 2005. "Haircuts: Estimating Investor Losses in Sovereign Debt Restructurings, 1998–2005." IMF Working Paper 05/137 (July).

Sturzenegger, Federico, and Jeromin Zettelmeyer. 2006. "Defaults in the 90s." Mimeo, Universidad Torcuato Di Tella.

UNCTAD. 2003. "Course on Dispute Settlement in International Trade, Investment and Intellectual Property." Geneva, Switzerland: United Nations Conference On Trade and Development.

UNCTAD. 2009. "Latest Developments in Investor-State Dispute Settlement." IIA Monitor No. 1. Geneva, Switzerland: United Nations Conference On Trade and Development.

UNCTAD. 2011. "Latest Developments in Investor State Dispute Settlement." IIA Monitor No. 1. Geneva, Switzerland: United Nations Conference On Trade and Development.

Waibel, Michael. 2007. "Opening Pandora's Box: Sovereign Bonds In International Arbitration." *American Journal of International Law* 101 (4): 711–59.