

Some Theoretical Considerations Regarding Net Asset Values for Money Market Funds

Huberto M. Ennis

On Tuesday, September 16, 2008, the day after Lehman Brothers filed for bankruptcy, the Reserve Primary Fund, a large prime money market fund, announced that it would not be able to redeem investors' funds one for one. The fund had “broken the buck” mainly due to losses on its holdings of Lehman's debt instruments. In the days that followed, outflows from prime money funds spiked, with investors withdrawing, in the space of a week, approximately \$300 billion—roughly 15 percent of total assets invested in these funds at the time (Financial Stability Oversight Council 2012). By Friday of that week, the U.S. Treasury and the Federal Reserve would decide to implement several major interventions aimed at stabilizing the money market funds industry. While outflows did, in fact, slow down in the following weeks, money funds continued divesting large amounts of commercial paper and other assets for some time.

The interventions announced by the U.S. Treasury and the Federal Reserve on September 19, 2008, were broad and unprecedented. The Temporary Guarantee Program adopted by the Treasury Department guaranteed that shareholders of those funds opting to participate would receive the fund's stable net asset value (NAV) per share were the fund to suspend redemptions and fully liquidate. At the same time, the

■ I would like to thank Todd Keister, Jeff Lacker, Jon Lecznar, Ned Prescott, Zhu Wang, and Alex Wolman for comments on an earlier draft. All errors and imprecisions are of course my exclusive responsibility. The views expressed in this article are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: huberto.ennis@rich.frb.org.

Federal Reserve created the Asset-Backed Commercial Paper Money Market Mutual Funds Liquidity Facility that was used to extend central bank credit to banks buying high-quality asset-backed commercial paper from money market funds (see Duygan-Bump et al. [2013]).

Money market funds (or, money funds, for short) are open-end mutual funds that invest in short-term high-credit-quality debt instruments such as commercial paper, large certificates of deposit, Treasury bonds, and repurchase agreements. Most money funds maintain a stable redemption value of shares, usually set at a value equal to one, and pay dividends that reflect the prevailing short-term interest rates. As of September 2012, there were 632 money market funds in the United States with total assets under management of approximately \$2.9 trillion. In comparison, deposits at banking institutions amount to about \$11 trillion. So, the size of the U.S. money market fund industry is significant.

SEC rule 2a-7 pursuant to the Investment Company Act of 1940 provides the regulatory framework for these funds. The rule permits funds to use the amortized cost method of valuation to compute net asset values and allows the funds to round such value to the nearest 1 percent.¹ The possibility of stable net asset values is a consequence of these provisions. At the same time, the rule puts limitations on the type of assets that the funds can hold: Funds must hold low-risk investment instruments with remaining maturity no longer than a given maximum date.

Within the broader category of money market funds, there are different sub-categories based on the main investments taken by the funds. Prime money funds hold predominantly private debt instruments. Government funds, instead, are restricted to invest only in government-issued securities. Prime money funds tend to be more exposed to credit risk (Rosengren 2012) and they are the ones that experienced serious financial distress during the second half of 2008.

In February 2010, partly as a response to the problems with prime money funds during the crisis, the Securities and Exchange Commission (SEC) adopted amendments to rule 2a-7 intended to make money funds more resilient and less likely to break the buck. The changes tightened restrictions on the amount of risk that money funds can assume and, for the first time, required that money funds maintain liquidity buffers to help them withstand sudden demands for redemptions. The new

¹ The amortized cost method allows the funds to value assets at their acquisition cost rather than market value, and interest earned on the asset is accrued uniformly over the maturity of the asset (adjusting for amortization of any premium or accretion of any discount involved upon purchase).

rules also enhanced information disclosure by funds and provided a framework for the liquidation of funds that break the buck and suspend redemptions.

Even after the wide-ranging revisions of rule 2a-7 in 2010, many policymakers and interested parties believe that a more comprehensive reform of the money funds industry is still necessary. In November 2012, the Financial Stability Oversight Council (FSOC) made public a set of proposed recommendations to the SEC for further reform (Financial Stability Oversight Council 2012). The Council proposed three different avenues for reform. The first alternative is to remove the valuation and pricing provisions in rule 2a-7 and to require money market funds to have a floating NAV that reflects the market value of their assets.

The second alternative is to require funds to maintain a buffer of assets in excess of the value implied by a fixed (and stable) NAV on outstanding shares. This buffer would be combined with a minimum balance at risk—in certain circumstances a small percentage of each investor's shares would be made available for redemption only on a delayed basis (see McCabe et al. [2012] for a detailed analysis of the minimum balance at risk idea). Finally, the third proposal is to require funds to hold a risk-based buffer and combine it with requirements on portfolio diversification, liquidity, and disclosure.²

To assess the Council's proposals, or any other reform proposal, it seems crucial first to be able to discern what is the ultimate function that money funds perform in the economy and how appropriate regulations depend on that. There are (at least) two possible ways to think about this issue. On one hand, some observers have argued that money funds provide useful maturity transformation by issuing claims (shares) that can be redeemed on demand while, at the same time, investing in longer-term financial instruments. Even though the funds' portfolios are concentrated in relatively short-term instruments, funds stand ready to redeem shares on demand and, hence, are exposed to a maturity mismatch and the threat of illiquidity.

On the other hand, it may be that the main role of money funds is to manage the portion of investors' portfolios intended to be allocated to relatively short-term money market instruments. In other words, according to this view, money funds are expert "cash" managers and, for this reason, it is efficient for investors to delegate to them the administration of part of their short-term and liquid investment strategy.

² See the FSOC document for a thorough description and evaluation of the reform proposals (Financial Stability Oversight Council 2012). The document also provides a good summary of the institutional details of the U.S. money market funds industry.

Assessing which of the two alternative views best describes the economic value associated with money funds is important for choosing the appropriate design of a regulatory framework. In particular, how redemption values should be computed often depends on this assessment. The aim of this article is to illustrate this point by presenting and comparing the implications of using different methods for computing NAVs in two very simple models that capture, in a stark way, the two aforementioned views about the function of money funds.

The first model is a version of the canonical maturity transformation framework introduced by Diamond and Dybvig in 1983. We find that, to the extent that NAV regulations are designed in a way that still allow funds to fulfill their basic function, then illiquidity and potential instability are likely to remain an integral feature of the money fund business. Furthermore, from this standpoint, computing appropriate market-sensitive NAVs requires an estimation of the amount of withdrawals that the fund can be expected to face. This process of anticipation is especially difficult because it involves predicting economic behavior that depends on agents' expectations about the decisions of others.³

The second model maintains many of the structural features of the first model, but is modified so that the motives investors have to deposit money with the fund are different. In particular, investors no longer derive value from maturity transformation but, instead, they rely on the funds exclusively to manage their investments.⁴ In this case, we find different implications relative to the first model. Computing NAVs that accurately reflect market valuations is perfectly compatible with the role played by the funds and can actually make the funds more stable. The model also illustrates how a wave of withdrawals from a poorly performing fund may just be the way that the system has to implement the best possible allocation of resources. Trying to stop that process would, in fact, be detrimental to economic efficiency.

Obviously, it is hard to determine which is the main function that money funds are performing in the economy, or even if they are essential organizations to pursue the highest attainable welfare of society. This article considers two candidate functions, one at a time. However, it is certainly possible that money funds perform, at least to a certain extent, these and potentially other functions simultaneously. Sorting these issues out is essentially an empirical undertaking,

³ Chen, Goldstein, and Jiang (2010, Appendix A) study a different, yet related model of a mutual fund where the redemption strategies of agents are also interdependent in equilibrium and can generate the conditions for fund instability.

⁴ The recent article by Parlatore Siritto (2013) also studies a model where the main function of money funds is to manage the assets of investors.

beyond the scope of our study. The objective in this article is, instead, rather theoretical. The point we want to illustrate is that once one has taken a stand on the answer to the empirical question, some theoretical implications follow that can help guide the design of an appropriate regulatory policy for money funds.

In principle, the models we present could be extended and modified to evaluate the other reform proposals currently being considered. For example, to understand the implications of requiring a buffer of assets one would need to take a stand on the way the buffer is being funded and model the objectives of the agents providing such funding. While this is potentially a productive activity, it would complicate the models in a way that would reduce the clarity of the results related to NAV policies. For this reason, we choose to limit our discussions to the NAV proposals.⁵

Before turning to the models, we should mention here that there is, in fact, a third commonly held perspective on the role of money funds in the economy, which we will not discuss in this article. The money funds industry developed and grew briskly in the 1970s, a period when banks were subject to strict interest rate ceilings imposed by regulation. These restrictions on the ability of banks to pay competitive rates did not apply to money funds and allowed money funds to become a natural alternative to banks (see Rosen and Katz [1983] for example). Even though the restrictions have been mostly removed now, funds may still be a vehicle for regulatory arbitrage to the extent that they are not subject to strict capital requirements and other regulations faced by banks.

The rest of the article is organized as follows. In the next two sections, we study two alternative frameworks that can be used to think about the problem of setting the appropriate redemption value of shares in a mutual fund. The first model, presented in Section 1, considers the case in which the role of the fund is to perform a maturity transformation function. The second model, in which the fund is just an investment vehicle that performs no essential maturity transformation, is the subject of Section 2. We close the article in Section 3 with a brief conclusion.

⁵ Another aspect left unexplored in this article is the possibility of contingent support from an institutional sponsor when the fund experiences financial distress. Sponsor support has played a significant role in the recent history of U.S. money market funds (Rosengren 2012). For a theoretical analysis of the issue, see Parlatore Siritto (2013).

1. MATURITY TRANSFORMATION

The canonical framework for studying maturity transformation in financial economics is the Diamond and Dybvig (1983) model of banking. A way to obtain desirable allocations in such an environment is to allow for an institutional arrangement that resembles a mutual fund. In this section, we analyze the implications of this theory for the determination of the fund's net asset value.⁶

The Model

There is a continuum of agents of mass 1. Agents are risk averse and each owns one unit of resources at the beginning of time. Time is denoted by $t = 0, 1$. Agents are homogeneous ex ante, but in period 0 a proportion q of the agents gets a preference shock and needs to consume at that time to be able to get any utility. We call these agents impatient and the $1 - q$ remaining agents, patient. Patient agents are indifferent about consuming at time 0 or 1. There is a productive technology that returns $R > 1$ units of resources in period 1 per unit of resources (not consumed and) invested in period 0. Resources can be taken out of the production technology during period 0 at a one-for-one basis (one unit per unit invested); in other words, there are no liquidation "costs" from interrupting the production process at an early stage.

A Benchmark Optimal Allocation

Since $R > 1$, there is a clear benefit from delaying consumption in this economy. For this reason, it is generally optimal to have patient agents consume only in period 1. Impatient agents, however, must consume in period 0.

Consider the solution (c_0^*, c_1^*) to the following planning problem:

$$\max_{c_0, c_1} qu(c_0) + (1 - q)u(c_1) \quad (\text{PP1})$$

subject to

$$(1 - q)c_1 = R(1 - qc_0).$$

We take such a solution as a benchmark optimal allocation in this environment. It is the allocation that maximizes the sum of the total

⁶ There is an extensive literature dedicated to the study of possible extensions of the Diamond-Dybvig model (see, for example, Freixas and Rochet [2008]). We use the simplest version of the model that allows us to illustrate the general points we are trying to make. Studying the implications for money funds of extensions of the model in various directions is a potentially fruitful activity. We consider this section a first step in that direction.

utility of both groups of agents, patient and impatient, subject to the resource constraint. In this constraint, $1 - qc_0$ is the amount of resources left after making a payment of value c_0 to each of the q impatient agents. This amount remains invested in the productive technology and is multiplied by the return R after waiting until period 1. In period 1, the resulting resources are divided between the remaining $1 - q$ patient agents and each of them gets an amount equal to c_1 .

When investors' coefficient of relative risk aversion is greater than one it can be shown that

$$1 < c_0^* < c_1^* < R.$$

The thing to notice here is that patient and impatient agents share the return from the productive investment in the optimal allocation. This is a form of insurance. Impatient agents get more than their initial resources even though the productive investment has not yielded any returns at the time that these agents wish to consume. This insurance is possible because only a proportion of the agents is expected to be impatient.

Institutions: An Open-End Mutual Fund

There are two main categories of mutual funds: those that are open-end and those that are closed-end. Open-end mutual funds stand ready to redeem shares held by investors at an announced net asset value. Closed-end mutual funds, instead, issue a fixed number of shares that in principle trade in a securities market but do not redeem shares on demand. Money market funds in the United States are predominantly open-end funds. Given the focus of our study, we restrict attention to this arrangement in the main body of the article. The reasons for the prevalence of open-end funds is the subject of active academic research (see, for example, Stein [2005]). We do not address the issue here but we present a brief analysis in the Appendix of how a closed-end fund would work in this environment.⁷

Suppose that at the beginning of time agents form an open-end mutual fund and deposit their endowment with the fund. The fund then invests the resources and sets dividend payments and a NAV that determines how much an agent is entitled to withdraw from the fund at each time.

⁷ There are many complex issues associated with the economics of closed-end mutual funds. For a survey of the subject see Lee, Shleifer, and Thaler (1990). Cherkas, Sagi, and Stanton (2008) is an interesting recent contribution.

One way for the fund to implement the optimal allocation (c_0^*, c_1^*) is to set a NAV equal to 1 and assign $c_0^* - 1$ new shares to each investor in period 0 in the form of a dividend payment. At that point, then, each agent has in their account c_0^* shares of the fund. If only the proportion q of agents that need to consume early decide to withdraw from the fund, then total withdrawals from the fund equal qc_0^* and there will be enough resources to pay the rest (a proportion $1 - q$) of the agents an amount equal to c_1^* in period 1. Since $c_1^* > c_0^*$, an investor that expects these payments and does not need to consume early will be willing to wait to withdraw. For this reason, the optimal allocation is a possible outcome associated with this mutual fund scheme.

As is well-known from the bank-run literature, when withdrawals from the fund happen sequentially, there is another possible outcome associated with this scheme (see Diamond [2007] for a simple exposition). Given that c_0^* is greater than unity, if all agents attempt to withdraw at time 0 then the fund would not have enough resources to cover all the required payments. As a result, if agents expect that all other agents will attempt to withdraw from the fund, then they also have incentives to try to withdraw, creating a situation that would resemble a run on the fund.⁸

It is also well-known from the bank-run literature that a scheme that allows the suspension of redemptions after q withdrawals will be able to costlessly rule out the run equilibrium. In reality, money funds can *and have* asked the SEC to authorize them to suspend redemptions after experiencing a wave of withdrawals. However, the authorization is usually granted under the assumption that the fund will fully liquidate and terminate operations after that. To the extent that the requirement of full liquidation still imposes costs on the fund, the suspension becomes less effective in limiting the incidence of runs.

In the model, the possibility of runs arises because, after the fund has distributed the new shares as dividends, if all agents are expected to want to withdraw from the fund at time 0, then the current value of fund assets is not sufficient to justify a NAV equal to 1. In particular, at time 0 total assets in the fund have a current (liquidation) value of 1. Agents, however, own $c_0^* > 1$ shares which, with a NAV of 1, entitle them to total time-0 payments that are greater than the current (liquidation) value of assets (one unit). An obvious solution to this problem

⁸ The fact that withdrawals take place sequentially during time 0 implies that the fund initially makes payments without knowing the total number of time-0 withdrawals that will ultimately happen. If the fund would be able to observe the total number of withdrawal requests before making any actual payments, then it is easy to show that the fund would adjust the value of those payments in such a way that runs could not happen in equilibrium.

is not allowing the fund to allocate new shares in the form of dividends before the actual returns are realized. However, the “early” dividends are essential for implementing the benchmark optimal allocation when the NAV is set to equal 1.⁹

In general, however, the fund may not want to value assets at their liquidation value (i.e., using a NAV equal to 1). Suppose, instead, that the fund sets a NAV equal to the *future discounted value* of the cash flow from the assets (FDV for short). If the manager of the fund (or some regulator) looks at the assets currently in the fund and disregards the withdrawal issue, following FDV would require setting a NAV equal to $\frac{R}{1+r}$, where r is an appropriate discount rate.

Since we are considering a situation without discounting, one possibility would be to take $r = 0$. In this case, the fund’s NAV will be set to equal R . We know, however, that if agents withdrawing at time 0 get a payment equal to R , then the optimal allocation will not be implemented (since $c_0^* < R$). Furthermore, if q agents get R in period 0, then there will not be enough resources to pay R or c_1^* to those agents withdrawing (and consuming) at time 1. If withdrawals from the fund happen sequentially, the only optimal withdrawal strategy for *all* investors under these payments is to try to withdraw early in a situation resembling a run.

Given that the rate of return on investment between $t = 0$ and $t = 1$ is equal to R , another possibility would be to use $1 + r = R$ as the appropriate discounting to compute the FDV. In this case, then, the fund’s NAV will be set to equal unity and again, without an early distribution of shares in the form of dividends, the optimal allocation would not be obtained. An attractive aspect of setting this value for the NAV is that the unique equilibrium in this case is for only impatient agents to withdraw at $t = 0$. While this conveys a sense of stability to the fund, it is also the case that impatient agents consume only one unit (not c_0^*) in this situation and, hence, the fund no longer performs the maturity transformation function that was the purpose of its creation.

It is unclear the extent to which money funds in reality are able to make higher payments to investors in anticipation of future expected returns. In the model, implementing a value of c_0 greater than 1 requires such anticipation. Money funds may not be performing the type of maturity transformation suggested by this model. We will consider an alternative model in the next section.

⁹ Initially each agent owns one share with a NAV equal to 1. As impatient agents need to consume $c_0^* > 1$ to conform with the benchmark optimal allocation, an entitlement of extra shares needs to be assigned to agents in period 0 so that impatient agents can actually consume an amount greater than 1 (c_0^*) at the appropriate time.

Even if the model in this section is the relevant one, it could be that due to legal (or “best practice”) restrictions, money funds do not perform the function described here. For example, suppose that the law requires that the fund pays dividends only after returns have been realized and always sets the NAV at the current liquidation value of the assets. In that case, the fund would set a NAV equal to 1 in period 0 and the payments would be given by $c_0 = 1$ and $c_1 = R$. This payment scheme, again, makes the fund immune to runs even when withdrawals are restricted to happen in a sequential manner.

The main insight thus far is that the maturity transformation function may involve a tradeoff between efficiency and stability. Some schemes result in a system that is immune to runs but does not provide beneficial insurance to impatient agents. Other schemes transfer resources appropriately among agents but make funds open to instability. The setting of the NAV plays a crucial role in the design of these schemes.

Variable Liquidation Terms

Suppose the fund is not able to liquidate and recover the invested resources one for one at time 0. Instead, the fund can only get ξ per each unit initially invested and later liquidated during period 0. In principle, the value of ξ may depend on the amount x being liquidated early. That is, ξ is a function of x .

An optimal arrangement is one that delivers the consumption allocation (c_0^*, c_1^*) obtained by solving the following problem:

$$\max_{c_0, c_1, x} qu(c_0) + (1 - q)u(c_1) \quad (\text{PP2})$$

subject to

$$\begin{aligned} qc_0 &= x\xi(x), \\ (1 - q)c_1 &= R(1 - x). \end{aligned}$$

Here, the first constraint indicates that to make a payment of c_0 to each of the q impatient agents, the fund needs to liquidate x units of investment, which allows it to obtain $x\xi(x)$ units of resources at time 0 when the payments to impatient agents need to occur. After liquidating x units of resources, $1 - x$ units are left in the productive technology and, hence, result in $R(1 - x)$ available resources at time 1. The second constraint, then, says that these resources will be used to pay an amount c_1 to each of the $1 - q$ patient agents.

It is easy to see that if $\xi(x) = R$ for all x , then $c_0^* = c_1^* = R$. In this case, a NAV equal to R per share implements the optimal allocation.¹⁰ However, if $\xi(x) < R$ for some x then it becomes less obvious how to compute an appropriate NAV. For example, if $\xi(x) = \tilde{\xi} < R$ for all x then $c_0^* < c_1^* < R$ and a fund trying to implement the best arrangement for its investors could need to set a NAV that would expose it to instability. The benchmark situation we studied before is the particular case when $\tilde{\xi} = 1$.

When funds liquidate, they usually sell assets in the market. It is often argued that the price of the assets may depend on how much is being liquidated. In our simple framework, liquidation at time 0 does not involve market prices but rather the direct technological costs of liquidating productive investment. Still, using the flexibility of the function ξ we can consider some cases that produce valuable insights about the more complex situation in which market prices play a role during liquidation. In particular, consider the case in which $\xi(x) = R$ as long as $x \leq q$ and $\xi(x) = \tilde{\xi} < R$ if x is greater than q . Here, again, the appropriate NAV would depend on the expected number of withdrawals. Suppose that the fund expects to have q withdrawals. Then, using a NAV equal to R allows the fund to implement the allocation $c_0^* = c_1^* = R$ with only impatient agents withdrawing from the fund at time 0.

However, if unexpected extra withdrawals were to happen (that is, if more than q agents decide to withdraw at time 0), the NAV would have to be drastically adjusted. Evidently, a crucial issue is how soon in the withdrawal process would the fund realize that withdrawals will be higher than q . If this realization comes after the first q withdrawals have already happened, then the fund will have to adjust the NAV at that point. The appropriate value of the NAV would depend on how many more withdrawals are expected after the first q . Suppose that after seeing that withdrawals continue beyond the first q the fund expects $q' > q$ withdrawals. Then, setting a NAV equal to $\tilde{\xi}$ would make the fund solvent but would destroy any insurance possibilities that the fund could still try to exploit given that q' is expected to be lower than 1.

This extension of the model captures in a stylized manner the technological (or market-based) costs that are often associated with the

¹⁰ Notice here that when $\xi(x) = R$ for all x , the fund has the ability to come up with resources immediately at no cost. For each unit of resources that the fund invests in the productive technology, it can get R units immediately, without waiting or bearing any risk. For this reason, the case of $\xi(x) = R$ seems of limited applicability for understanding actual real life investment situations.

early liquidation of an investment position. The analysis clearly illustrates that liquidation costs, in interaction with expectations about the number of early withdrawals, significantly complicate the setting of an appropriate NAV.

Portfolio Choice: Adding a Liquid Asset

Suppose now that in the setup just studied the liquidation value of the productive technology is $\xi(x) = \tilde{\xi} < 1$ for all x . This situation may seem peculiar since some *costly* liquidation is taking place even though it is completely predictable. In other words, given that the fund is expecting at least q redemptions, it would be better to invest some resources in an asset that, while less productive, avoids any significant liquidation costs (i.e., a more liquid asset).

To address this issue, we extend the previous setup to include an alternative technology that returns, per unit invested at the beginning of time 0, one unit of resources at any time. Then, an optimal arrangement would produce the allocation that solves the following problem:

$$\max_{c_0, c_1, \gamma, x} qu(c_0) + (1 - q)u(c_1) \quad (\text{PP3})$$

subject to

$$\begin{aligned} qc_0 &= \gamma + x\tilde{\xi}, \\ (1 - q)c_1 &= R(1 - \gamma - x), \end{aligned}$$

where γ is the portion invested in the liquid asset and x , again, is the amount liquidated at time 0 of the fund's investment in the productive technology, $1 - \gamma$. As before, the two constraints are resource constraints on payments at time 0 and 1, respectively. The first constraint shows that the investment γ in the liquid asset is fully used to make payments to impatient agents. In the second constraint, total unliquidated productive investment is now equal to $1 - \gamma - x$. Multiplying this amount by $R > 1$, we obtain the total available resources at time 1 that can be used to make payments of value c_1 to each of the $1 - q$ patient agents.

When $\tilde{\xi} < 1$ and the fund expects that exactly q agents will withdraw at $t = 0$, it is optimal to choose $x^* = 0$ and $\gamma^* = qc_0^*$. Furthermore, the optimal values of c_0 and c_1 are given by the same c_0^* and c_1^* obtained in the benchmark optimal allocation (problem PP1). The perfect predictability of the number of withdrawals, combined with the fund's access to a liquid asset, implies that costly liquidation never happens.

How should the fund compute its NAV at time 0? Here, again, combining the payment of early dividends with a NAV equal to 1 would

be consistent with obtaining the optimal allocation as an equilibrium outcome. The alternative approach based on calculating a FDV with a discount rate $r = 0$ would result in a value of the NAV equal to $\gamma^*1 + (1 - \gamma^*)R$. While the FDV method is often considered natural, it is easy to show that in this case the implied NAV is greater than c_0^* and, hence, it would provide too much consumption to those agents withdrawing in period 0 (relative to the optimal allocation).¹¹

The fact that the fund can perfectly predict the amount of withdrawals is important and may be considered unrealistic. Uncertainty over q significantly complicates the calculations. To gain some perspective on this issue, consider a situation where the fund was expecting q withdrawals but instead $\tilde{q} > q$ withdrawals happen. After making the first q payments the fund would have to reassess the rest of its planned payments. Suppose that after making the first q payments the fund immediately discovers that the number of withdrawals will be $\tilde{q} > q$. Then, the optimal continuation payments would solve the following problem:

$$\max (\tilde{q} - q) u (c'_0) + (1 - \tilde{q}) u (c'_1) \tag{PP4}$$

subject to

$$(\tilde{q} - q) c'_0 = x\tilde{\xi},$$

$$(1 - \tilde{q}) c'_1 = R(1 - \gamma^* - x).$$

The first constraint indicates that for the fund to be able to make a payment of value c'_0 to $\tilde{q} - q$ agents in period 0 it will have to liquidate an amount x of productive investment that, given liquidation costs, results in $x\tilde{\xi}$ available resources. It is important to realize here that the fund has already made q payments of size c_0^* , and since $\gamma^* = qc_0^*$, there are no more liquid assets available to make extra payments in period 0. The second constraint (over payments in period 1) is similar to that in the previous problem. Let us denote by c'^*_0 and c'^*_1 the solution to problem PP4.¹²

Setting the appropriate continuation NAV in this case is again a difficult issue. Note that there are only $(1 - \gamma^*)$ units of the asset left at the fund after the initial q withdrawals. These assets can be liquidated at a rate of $\tilde{\xi} < 1$ and the fund has to still make $1 - q$ payments. In

¹¹ We know that $c_0^* < c_1^*$, $c_0^* = \gamma^*/q$, and $c_1^* = R(1 - \gamma^*)/(1 - q)$. Then, we have that $\gamma^*/q < R(1 - \gamma^*)/(1 - q)$, which can be rearranged to $\gamma^* + (1 - \gamma^*)R > \gamma^*/q = c_0^*$.

¹² We do not discuss here whether the fund managers would have the incentives at this point to redesign payments so as to maximize the remaining investors' utility. Perhaps reputational issues could be brought to bear in explaining a behavior of the fund in line with that suggested by the optimal continuation payments studied here.

principle, using current values of the assets, the fund would set a NAV equal to $(1 - \gamma^*)\tilde{\xi}/(1 - q)$ and it can be shown that c_0^* is actually greater than this number. The reason for the discrepancy between the optimal continuation payment c_0^* and the NAV computed using current valuations is essentially the same as we discussed before: The fund does not expect to have to liquidate all assets (as long as $\tilde{q} < 1$) and, as a consequence, it can still provide some insurance (maturity transformation) to the agents requesting early redemptions. In the optimal continuation, the fund's payments to these agents are such that they receive a portion of the returns coming from the productive investment that will be held to maturity.

This last extension of the model shows that when the fund holds a portfolio of investments, some more liquid than others (as it would want to do, given that it expects some withdrawals to happen early and some to happen late), the standard methods for computing NAVs again may fail to deliver the most desirable allocations. In summary, then, setting appropriate values for NAVs within the maturity transformation paradigm often involves a tradeoff between efficiency and stability. This is the case in the simplest version of the model and it remains true even when we consider liquidation costs and a non-trivial portfolio choice available to the fund.

2. INVESTMENT MANAGEMENT

In this section, we study a model in which the mutual fund performs the function of investment management. The underlying justification is an assumption that the fund can administer the allocation of funds to productive activities more efficiently than individual investors. For this reason, then, investors delegate management functions to the fund by investing directly in it. The model is again very simple. We attempt to stay as close as possible to the formal analysis of the previous section but introduce some modifications that produce a different perspective on the recent experiences with money funds.

The Model

There is a mass 1 of risk averse agents and each of them own one unit of resources at the beginning of time. Time is again given by $t = 0, 1$. Different from the model in the previous section, here all agents are patient (that is, they are indifferent between consuming at either time 0 or 1). There is a risky productive technology that returns a random amount R of resources in period 1 per unit of resources invested in period 0. The value of R gets realized after investment in this risky technology

has taken place. However, resources can be removed from the risky productive technology at any time during period 0 on a one-for-one basis. Agents can also invest in an alternative riskless technology at any time during period 0 that returns a fix gross return $R_z > 1$ in period 1 per unit of resources invested in period 0. Call z the amount invested in this alternative riskless technology.

A Benchmark Optimal Allocation

Since z can be decided after observing the realization of R , it is optimal to make z a function of R . The optimal allocation of resources solves the following planning problem:

$$\max_{c(R), z(R)} E[u(c(R))], \quad (\text{PP5})$$

subject to

$$c(R) = R[1 - z(R)] + R_z z(R)$$

and

$$0 \leq z(R) \leq 1 \quad \text{for all } R.$$

The expectation in the objective function is taken with respect to the random variable R . The first constraint is a resource constraint that must hold pointwise, for each possible value of R . It says that consumption is equal to the return on the portfolio of investment implied by $z(R)$. The second constraint reflects natural non-negativity requirements on the amount invested in each of the two technologies.

Let us denote by $z^*(R)$ the optimal investment strategy implied by the solution to this problem. We have that $z^*(R) = 1$ whenever $R < R_z$ and $z^*(R) = 0$ when $R > R_z$. If $R_z = R$, then the value of z^* is not pinned down by this problem and it is irrelevant for payoffs. Just for concreteness assume that $z^*(R_z) = 0$.

Institutions: An Investment Fund

Since all agents are equally exposed to the underlying uncertainty in the environment, risk-sharing is no longer a reason for them to pool resources in a fund. Assume, however, that only the fund has the necessary infrastructure (expertise) to be able to invest in the technology with random return R . Agents have to decide whether to invest in the fund before the value of R is realized. Let e be the amount of the initial resources that each agent decides to keep outside the fund. Hence, the amount $1 - e$ of resources is invested in the fund.

Once the value of R is realized and observed, agents may want to withdraw some of the resources initially invested in the fund. At that time, the fund calculates a NAV and allows withdrawals according to that value. Suppose R can take a finite number of possible values. We use the subindex $j \in J$ to indicate the different values of R , where J is a finite set. Let p_j be the probability that $R = R_j$ for each $j \in J$ and, of course, $\sum_{j \in J} p_j = 1$. Denote by h_j and z_j the NAV set by the fund and the amount that an agent withdraws from the fund, respectively, when $R = R_j$. Then, the optimization problem faced by an investor is the following:

$$\max_{e, \{c_j, z_j\}_{j \in J}} \sum_{j \in J} p_j u(c_j) \quad (\text{IP})$$

subject to

$$c_j = R_j (1 - e - z_j) + R_z (h_j z_j + e)$$

and $0 \leq z_j \leq 1 - e$ for all $j \in J$, and $0 \leq e \leq 1$. Agents initially invest $1 - e$ at the fund and then withdraw z_j after they discover that returns will be equal to R_j . The shares z_j withdrawn from the fund are valued at a NAV equal to h_j and, hence, the total amount withdrawn equals $h_j z_j$. Agents re-invest this amount in the alternative riskless technology, together with the previously invested amount e . Hence, total consumption equals the sum of resources obtained from the fund, $R_j (1 - e - z_j)$, and from the riskless technology, $R_z (h_j z_j + e)$.

The Case of a Fixed NAV Equal to One

Since the fund can physically liquidate investment one for one, setting $h_j = 1$ for all j is feasible. When $R_j > R_z$ for some $j \in J$ and the fund sets $h_j = 1$ for all j , agents will be willing to invest all their endowment in the fund at the beginning of time. To see this, define $z'_j = z_j + e$ for all $j \in J$ and note that now we can write $c_j = R_j (1 - z'_j) + R_z z'_j$ since $h_j = 1$ for all j . Given that we still have the constraint $z_j \leq 1 - e$ as a requirement, choosing $e = 0$ relaxes the domain constraints on z_j and, consequently, can only improve the solution to the agent's problem. In particular, note that when $h_j = 1$ and $e = 0$ the problem of the agent is the same as the planning problem for the benchmark optimal allocation (PP5), but where now $z(R_j) = z_j$ stands for withdrawals from the fund in state j . Parallel to the solution of problem (PP5), then, whenever R_j is less than R_z the optimal value of z_j equals 1 and agents withdraw all their investments from the fund. Even though this event could look like a run on the fund, it is actually part of the process involved in obtaining an optimal allocation of resources.

This result provides an interesting perspective on some proposals to reform the regulatory framework for money market funds. Specifically, some reform proposals are designed to provide investors with a disincentive to withdraw from a troubled fund. The objective is to reduce the incidence of runs. However, we see here that limiting the ability of investors to reallocate resources at certain points in time could stand in the way of economic efficiency.

Note that we have considered only the case when investment in the fund actually constitutes a risky alternative for the agents. It is often the case, however, that money funds are considered a relatively safe investment alternative. It would not be hard to modify the model so that R_z is random and R is a fixed (safe) return. While the results have a similar flavor, some of the interpretations may not be as natural. For example, investors would want to withdraw from the fund at those times when R_z is relatively high. In other words, run-like episodes in relatively safe funds would tend to be associated with “good times” (high returns) for investors.

Variable Liquidation Terms

So far, we have studied a situation where the fund can liquidate investment one for one. More generally, suppose that the fund can obtain resources equal to ξ_j per unit liquidated of the risky productive technology, with $j \in J$. To simplify the calculations in what follows, assume that $J = \{L, H\}$ with $R_H > R_L$ and $p_L = p$ (so that $1 - p$ is the probability that $R = R_H$).

An optimal arrangement in this case produces an allocation that solves the following problem:

$$\max_{e, \{c_j, z_j\}_{j=L,H}} pu(c_L) + (1 - p)u(c_H) \tag{PP6}$$

subject to

$$c_j = R_j(1 - e - z_j) + R_z(\xi_j z_j + e)$$

and $0 \leq z_j \leq 1 - e$ for $j = L, H$, and $0 \leq e \leq 1$.

In principle, the liquidation values could be independent of the observed value of R . When $\xi_L = \xi_H = 1$, problem (PP6) is equivalent to problem (IP) with $h_j = 1$ for all j . Then, when $R_H > R_z$, it is optimal to set e equal to zero (recall that e must be chosen before the realization of R can be observed). More generally, however, when $\xi_L = \xi_H = \xi$ for some value of $\xi \in (0, 1)$ and $R_L < R_z$, it is possible to have an optimal value of e that is different from zero. There are two cases to consider, depending on whether ξR_z is greater or less than R_L .

When $\xi R_z < R_L$, it is never optimal to liquidate investments in the funds, and the expressions for consumption are given by:

$$\begin{aligned} c_L &= R_L + (R_z - R_L) e, \\ c_H &= R_H - (R_H - R_z) e. \end{aligned} \tag{NL}$$

It is clear here that there is a tradeoff involved in choosing the optimal value of e . Investing more in the fund (lower e) increases consumption when returns are high (when $R = R_H$) but decreases consumption when returns are low (when $R = R_L < R_z$). For some parameter values the optimal value of e is positive.

When $\xi R_z > R_L$, it is optimal to liquidate investments when the realization of R is known to be equal to R_L . Given this, the expressions for consumption are now given by:

$$\begin{aligned} c_L &= \xi R_z + (R_z - \xi R_z) e, \\ c_H &= R_H - (R_H - R_z) e. \end{aligned} \tag{FL}$$

Notice the similarities with respect to the previous expressions, (NL). As a result, it is not hard to see that a similar logic applies and that for certain parameter values the way to balance the tradeoff of returns is to choose an interior (positive) value of e .

It is important to realize here that, given the information constraints implied by the environment, this situation reflects ex ante efficient choices. However, when $R = R_L$, costly liquidation takes place. This liquidation may be regarded as a regrettable outcome ex post but it should be understood that trying to avoid it through regulation could be detrimental to ex ante welfare.

Even though we do not model explicitly a market for assets we can use the model, as in the previous section, to help us think about a situation in which the fund is liquidating assets by selling them (potentially at a discount) in the market. To this end, let us consider the case in which ξ_j is positively correlated with R_j . One particular, simple version of this correlation is when $\xi_j = \xi R_j$ for $j = L, H$. This assumption implies that the liquidation value of assets reflects immediately the deterioration in prospective future returns, as one would expect would happen in a market. We turn to the study of this case next.

First, it is easy to see that if $\xi R_z > 1$ then it is always optimal to set $z_L = z_H = 1 - e$ and liquidate all investments from the fund immediately after making them. This seems an implausible situation, mainly due to the stark timing in the model. Hence, we will proceed here under the assumption that $\xi R_z \leq 1$.

When $\xi R_z < 1$ it is optimal to set $z_L = z_H = 0$ and the expressions for consumption are the same as those labeled (NL) above. As before,

then, the choice of e reflects a tradeoff between lower returns in good times and higher returns in bad times.¹³

Comparing the problem for the optimal arrangement, (PP6), with the problem of the private investor, (IP), we can see that by setting $h_j = \xi R_j$ for $j = L, H$ the fund would be able to provide the agents with the optimal contract. Under this arrangement, agents do not liquidate any of their investments in the fund, regardless of the state of asset returns. That is, agents choose $z_L = z_H = 0$ and the fund never experiences a wave of withdrawals.

The key to understanding this result is to note that when the return R is expected to be low, the NAV set by the fund immediately adjusts to reflect the lower valuation of the fund's assets. By the time the investors get a chance to withdraw, the losses are already reflected in the withdrawal values. There is no way in which withdrawing from the fund can be used by investors as a way to "escape" the expected losses associated with the low returns from the fund's assets.

Delays in Adjusting the NAV

Suppose, as before, that $\xi_j = \xi R_j$ for $j = L, H$. Now, however, assume that the fund is not able to immediately adjust the NAV when the news about the returns of the assets are first revealed. As an example, suppose that the fund initially sets an (unconditional) redemption value of shares h equal to one (before any information about returns have been revealed) and that the fund is only able to adjust h after q investors have had an opportunity to withdraw from the fund.¹⁴

The payments to the first q investors are now given by:

$$c_L = R_L(1 - e - z_L) + R_z z_L + R_z e,$$

$$c_H = R_H(1 - e - z_H) + R_z z_H + R_z e,$$

and it is optimal for these investors to set $z_L = 1 - e$ and $z_H = 0$. In other words, those investors that are able to withdraw from the fund at a NAV equal to 1 will withdraw all their investments when the return on the assets is expected to be low and will leave all their investments in the fund if the return on the assets is expected to be high.

When $R = R_L$, after the first q agents have redeemed their shares, the fund will be able to reset its NAV. At that point, the fund would

¹³ Under constant relative risk aversion, it is easy to show that the amount invested in the fund $1 - e$ is increasing in the average return R and decreasing on the (mean-preserving) variance of R .

¹⁴ This timing can perhaps be motivated by thinking of a gradual process of diffusion of information, whereby only some agents find out that returns will be low before the fund is able to (or willing to) adjust redemption values.

have already liquidated $s = q(1 - e) / \xi R_L$ units of the initial $(1 - e)$ investments and the payoff to the remaining investors would have to be recalculated. In particular, if the fund sets a NAV equal to ξR_L , the payoff to these agents from withdrawing from the fund equals

$$\xi R_L \frac{1 - e - s}{1 - q} R_z.$$

The payoff from not withdrawing equals

$$R_L \frac{1 - e - s}{1 - q}.$$

Given that $\xi R_z < 1$, these agents will prefer not to withdraw.

This example illustrates how delays in updating the NAV of an investment fund may create the conditions for an initial rush of withdrawals resembling a run, which only stops after the NAV has been appropriately adjusted. Within the context of this interpretation about the nature of money funds, *floating* NAVs that adjust every time an investor has an opportunity to withdraw could be helpful in reducing fund instability.

At this point, it is natural to ask why delays in the adjustment of NAVs would happen. Current regulation allows money funds not to reflect in their redemption value deviations from the market value of their assets as long as they are small (fewer than 50 basis points). Furthermore, it seems possible that announcing changes in redemption values that were otherwise expected to be relatively constant would raise awareness and doubts among investors. If fund managers perceive a threshold-like effect from making these announcements they would have incentives to delay them on the hope that new information arrives and reverts the negative news previously received.

3. CONCLUSION

Money market funds experienced considerable distress in 2008 during the U.S. financial crisis. Their resiliency was questioned again in 2011 during the European sovereign crisis (see Chernenko and Sunderam [2012] and Rosengren [2012]). Currently, a generalized concern exists that the instability of money funds may have systemic consequences (Financial Stability Oversight Council 2012). For these reasons, there is a heated ongoing debate about the appropriate reform of the regulatory framework that applies to these funds.

In this article, we have presented two models that represent, in a stylized manner, two possible alternative interpretations of the economic function fulfilled by money funds. In both models, money funds may experience waves of withdrawals that resemble runs. The

frameworks, however, are not flexible enough to address systemic concerns such as contagion and economy-wide disruptions triggered by the troubles in the money funds industry. Still, some important insights about fund stability and regulation arise from the analysis. One of the main lessons of the article is that the appropriate regulation of money market funds depends on the stand taken with respect to the fundamental economic function performed by the funds.

In particular, if money funds are mainly providers of maturity transformation services, then the setting of the redemption value of shares needs to take into account the optimal insurance component involved in this kind of arrangement. Extreme versions of floating net asset values may undermine this function, just as narrow banking tends to undermine the maturity transformation function of banks. Perhaps some instability is inextricably associated with maturity transformation, and trying to completely rule out instability translates into ruling out any degree of maturity transformation. Under this view, stable money funds can, in effect, be redundant institutions.

However, in the second model we presented in this article, we took on the interpretation that money funds are instead investment managers that are able to access, select, and implement beneficial asset-allocation strategies for their investors. Under this view, money funds do not perform any maturity transformation. We learned that in this case a timely adjustment of the fund's redemption value of shares (such as a floating NAV) may be conducive to stability and is compatible with the fund's intended function. To a certain extent, then, alternative reform-proposals involving NAVs indirectly reflect different perspectives about the main function that money funds perform in the economy.

APPENDIX

In this appendix we study an arrangement resembling a closed-end fund in the environment presented in Section 1. We can interpret this arrangement as a version of the financial intermediation system proposed by Jacklin (1987).

Suppose that at the beginning of time, investors form a fund that issues shares in exchange for investors' endowment. The fund, then, invests in a productive technology with return R . The value of each share is set to equal 1 and each share pays a dividend d_t at $t = 0, 1$. In other words, each share represents the right to a dividend stream.

At time $t = 0$ investors holding a share receive the dividend d_0 and a market for ex-dividend shares opens. *Redemptions of shares are not allowed* at time $t = 0$ (i.e., it is a closed-end fund).

Clearly, the q impatient agents will want to sell their shares. If the fund sets $d_0^* = qc_0^*$ and $d_1^* = R(1 - qc_0^*)$, we have that market clearing in the shares market is given by

$$(1 - q)d_0^* = vq,$$

where v is the price of a share and $(1 - q)d_0^*$ is the total amount of resources in the hands of patient agents that can be used to buy the q shares of the impatient agents. The equilibrium price is given by $v^* = (1 - q)c_0^*$. Note that, for each share, patient agents pay $(1 - q)c_0^*$ and receive in the following period $R(1 - qc_0^*)$. Since $R(1 - qc_0^*) = (1 - q)c_1^* > (1 - q)c_0^*$, patient agents want to buy the shares at the price v^* . Patient agents, as a group, then consume d_1^* since they own all the shares in period $t = 1$ and each of them consume

$$\frac{d_1^*}{1 - q} = \frac{R}{1 - q}(1 - qc_0^*) = c_1^*.$$

Impatient agents consume $d_0^* + v^*$ (the dividend plus the proceeds from selling the shares) and we have that

$$d_0^* + v^* = qc_0^* + \frac{(1 - q)qc_0^*}{q} = c_0^*.$$

We see here, then, that a closed-end fund could also implement the optimal allocation in this environment. In fact, this arrangement would make the fund immune to runs. The reasons for why funds choose to be open-end were left unmodeled in this article. See Stein (2005) for a general discussion of this issue and for a possible explanation.

REFERENCES

- Chen, Qi, Itay Goldstein, and Wei Jiang. 2010. "Payoff Complementarities and Financial Fragility: Evidence from Mutual Fund Outflows." *Journal of Financial Economics* 97 (August): 239–62.
- Cherkes, Martin, Jacob Sagi, and Richard Stanton. 2008. "A Liquidity-Based Theory of Closed-End Funds." *Review of Financial Studies* 22 (April): 257–97.

- Chernenko, Sergey, and Adi Sunderam. 2012. "Frictions in Shadow Banking: Evidence from the Lending Behavior of Money Market Funds." Fisher College of Business Working Paper 2012-4 (September).
- Diamond, Douglas W. 2007. "Banks and Liquidity Creation: A Simple Exposition of the Diamond-Dybvig Model." Federal Reserve Bank of Richmond *Economic Quarterly* 93 (Spring): 189–200.
- Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91 (June): 401–19.
- Duygan-Bump, Burcu, Patrick M. Parkinson, Eric S. Rosengren, Gustavo A. Suarez, and Paul S. Willen. 2013. "How Effective Were the Federal Reserve Emergency Liquidity Facilities? Evidence from the Asset-Backed Commercial Paper Money Market Mutual Fund Liquidity Facility." *Journal of Finance* 68 (April): 715–37.
- Financial Stability Oversight Council. 2012. "Proposed Recommendations Regarding Money Market Mutual Fund Reform." Washington, D.C.: U.S. Department of the Treasury (November).
- Freixas, Xavier, and Jean-Charles Rochet. 2008. *Microeconomics of Banking*. Cambridge, Mass.: The MIT Press.
- Jacklin, Charles. 1987. "Demand Deposits, Trading Restrictions, and Risk Sharing." In *Contractual Arrangements for Intertemporal Trade*, edited by E. Prescott and N. Wallace. Minneapolis: University of Minnesota Press, 26–47.
- Lee, Charles M. C., Andrei Shleifer, and Richard H. Thaler. 1990. "Anomalies. Closed-End Mutual Funds." *Journal of Economic Perspectives* 4 (Fall): 153–64.
- McCabe, Patrick E., Marco Cipriani, Michael Holscher, and Antoine Martin. 2012. "The Minimum Balance at Risk: A Proposal to Mitigate the Systemic Risks Posed by Money Market Funds." Federal Reserve Bank of New York Staff Report No. 564 (July).
- Parlatore Siritto, Cecilia. 2013. "The Regulation of Money Market Funds: Adding Discipline to the Policy Debate." Manuscript, New York University.
- Rosen, Kenneth T., and Larry Katz. 1983. "Money Market Mutual Funds: An Experiment in Ad Hoc Deregulation: A Note." *Journal of Finance* 38 (June): 1,011–7.

- Rosengren, Eric S. 2012. "Money Market Mutual Funds and Financial Stability." Speech given at Federal Reserve Bank of Atlanta 2012 Financial Markets Conference, Stone Mountain, Ga., April 11.
- U.S. Securities and Exchange Commission. 2010. "Money Market Fund Reform: Final Rule." Available at www.sec.gov/rules/final/2010/ic-29132.pdf.
- Stein, Jeremy C. 2005. "Why Are Most Funds Open-End? Competition and the Limits of Arbitrage." *Quarterly Journal of Economics* 120 (February): 247–72.

Debt Default and the Insurance of Labor Income Risk

Kartik B. Athreya, Xuan S. Tam, and Eric R. Young

Recent research (e.g., Chatterjee et al. 2007, Livshits, MacGee, and Tertilt 2007) has found that allowing for debt default, such as through the relatively lenient U.S. bankruptcy code, is likely to improve *ex ante* welfare relative to more strict forms of debt forgiveness. The welfare gains come from improved consumption insurance provided by the option to not repay debt in some circumstances. Thus far, however, *all* instances where quantitative work finds a beneficial role for default have been ones with large and transitory shocks directly to household consumption *expenditures*. It is clear therefore that these “expense shocks” that lead to involuntary reductions in net worth are sufficient, given the specification of non-expense-related income risk in current models, to justify debt relief in forms resembling U.S. personal bankruptcy provisions.

The availability of bankruptcy, and more generally, default, will be reflected in the pricing on consumer debt, and so will affect households’ ability to smooth consumption across dates and states of nature. It is therefore important to note that a significant amount of the risk to lifetime household resources may come from persistent shocks to *labor income* (Huggett, Ventura, and Yaron 2010). As a result, to the extent

■ Athreya is an economist at the Richmond Fed; Tam is affiliated with the University of Cambridge; Young is an economist at the University of Virginia. This article previously circulated under the title “Are Harsh Punishments for Default Really Better?” We would like to thank seminar and conference participants at UT-Austin, the Board of Governors, the Cleveland Fed, Georgetown University, the Philadelphia Fed, and Queen’s University for comments on the earlier versions. We thank the EQ committee, especially Huberto Ennis, for detailed comments. Tam thanks the John Olin Foundation for financial support. The opinions expressed here do not reflect those of the Federal Reserve System or the Federal Reserve Bank of Richmond. All errors are the responsibility of the authors. E-mail: kartik.athreya@rich.frb.org.

that one might be able to locate other, more targeted, ways of insuring expense shocks, it is useful to better understand how effective debt forgiveness is for managing income risk in isolation.

In this article, we evaluate in detail the role of debt forgiveness in altering the impact of income risk in the absence of expense shocks. The experiments we present can be thought of as asking: “If we insure the out-of-pocket expenses that constitute expenditure shocks, is there still a role of debt relief as a form of insurance against ‘pure labor income risk’?” We address this question by studying a range of specifications for households’ attitudes toward the intra- and intertemporal properties of income, when expense shocks are not present. Our main finding is that, absent expenditure shocks, the ability to default very generally hinders the ability of households to protect themselves against labor income risk.

Despite the nature of our results, we stress that our work is not to be taken as a strong statement about the overall desirability of U.S. personal bankruptcy law, for two reasons. First, to the extent the expense shocks are a feature of reality, our model is missing a feature known to be capable of justifying bankruptcy protection. Second, *informal* default or “delinquency” whereby a borrower simply ceases making payments (and leaves themselves open to legally protected collections efforts) may simply increase if formal bankruptcy is made stricter or disallowed altogether. Indeed, in ongoing work (Athreya et al. 2013), we find that this channel is quantitatively relevant. These related, and coexisting, options to avoid debt repayment are not modeled here. Instead, our results apply more narrowly: They suggest that labor income risk alone may not provide a strong rationale for allowing households to default. In other words, our findings suggest that the scope of shocks that debt forgiveness is providing insurance against is limited, perhaps limited principally to relatively catastrophic outcomes.

It is interesting to note that similar results are now being located in the literature on sovereign debt. Namely, it has proved very difficult to find plausible circumstances in which the benefits to being able to repudiate debts (or perhaps more accurately, the costs of being unable to commit to repayment of sovereign debt) are positive. The reasons for the similarity of the results are natural. Most importantly, the models themselves are largely isomorphic in the optimization problems they lead to, and do not differ substantially enough in their quantitative specification of either preferences or risk. Moreover, even though sovereign debt models differ somewhat in the interpretation of the debt itself (i.e., that is public debt, not private), the standard assumption in that literature is that government is benevolent and seeks to borrow on behalf of households who themselves wish to smooth consumption.

This blurs the distinction between the path of public debt the government chose and that which households would have chosen.

Our results come from comparing allocations arising from two underlying trading environments. First, we study allocations arising from what we will refer to as the textbook, or “standard model” (SM), of consumption and saving in which households face uninsurable earnings risks with persistent and transitory components. In this model, households can only borrow using nondefaultable debt and also face liquidity constraints. Canonical examples of SM include those laid out in Deaton (1992, chapter 7) and Carroll (1997). To be consistent with the view that borrowing limits should be endogenously determined by repayment incentives, under SM, we investigate primarily the so-called “natural borrowing limit” case.¹

The second trading arrangement we consider is one where, as before, households face life-cycle consumption/savings problems in which they encounter identical risks as in SM, but can issue defaultable debt. We will refer to this as the “default model” (DM). Benchmarks in this literature are Chatterjee et al. (2007) and Livshits, MacGee, and Tertilt (2007). Following these articles, default in the DM will be represented as a procedure whereby those with negative net worth can stop paying obligations, subject to any costs that may be present. The two trading arrangements we consider are thus clearly different. Nonetheless, they are related in a simple way: SM is the limiting case of DM as default becomes prohibitively costly.

To focus directly on the role of default in insuring labor income risk relative to the SM, we take two steps. First, as already noted, we deliberately set aside expenditure shocks. The presence of such shocks rules out the comparison of models with default against the standard model as budget sets would be empty for some dates and states were it not for the possibility of default. Second, we will examine a wider array of household preferences than has been done in the literature thus far. Specifically, we (i) separate risk aversion from the intertemporal willingness of households to substitute consumption, and (ii) evaluate the role of ambiguity aversion (or uncertainty aversion) when households are unsure of the stochastic environment they populate.

Both the separation of risk aversion from intertemporal elasticities and the possibility of ambiguity have been previously identified with a beneficial role for debt default. However, neither has been studied formally. The logic for suspecting that they may be important in delivering a welfare-enhancing role for default is as follows.

¹ See, e.g., Ljungqvist and Sargent (2004, p. 577).

First, the tradeoff between intertemporal and intratemporal smoothing was first suggested in Livshits, MacGee, and Tertilt (2007) in a life-cycle model of personal default. Assessing the relative importance of these motives therefore requires allowing for preferences in which the two attitudes can be distinct, irrespective of the uncertainty surrounding income. However, prior work has employed constant relative risk aversion (CRRA) preferences that conflate the two aspects of household preferences. In contrast, we employ Epstein-Zin recursive utility (Epstein and Zin 1989), which we select because of its tractability and demonstrated ability to improve the performance of asset pricing models, of which defaultable debt is a special case.

Second, with respect to the role of ambiguity in determining the value of an option to default, the legal and political history of bankruptcy law suggests that allowing for the release of debtors subject only to modest penalties is a policy that improves welfare if households are not perfectly sure of the probabilistic structure of income risk (see Jackson [2001] for one example).² This view is not confined to legal experts. As noted as early as Friedman (1957), agents will typically be unsure about the process that generates their labor income shocks, instead accepting that a family of potential distributions that may be difficult to distinguish are possible. Within this class of preferences, an agent who displays ambiguity aversion (Epstein and Schneider 2003) will solve a max-min problem—the agent will choose the member of the class that makes utility lowest and then choose consumption and savings in order to deliver the highest utility in this worst case.³ It is precisely this feature of the problem that will allow for a more nuanced understanding of how penalties can be “excessive” and thereby welfare-reducing: Eliminating default through harsh penalties may leave the agent unwilling to borrow at all. As a result, such a policy could perversely inhibit both intertemporal and intratemporal consumption smoothing, despite “mechanically” alleviating the limited commitment problem that the young and poor face. U.S. bankruptcy law, for instance, appears directly predicated on the idea that penalties can indeed be excessive, in the sense that they may leave would-be borrowers unwilling to do so (see Jackson [2001]).

The potential role for ambiguity in altering the welfare implications of having defaultable debt is also suggested by the observation that, in

² Miao and Wang (2009) study the decision to exercise an option under ambiguity. Due to the presence of fixed costs, bankruptcy has option value. We focus on a related setting but are interested in the quantitative aspects associated with household consumption smoothing.

³ These preferences are a special case of the more general ambiguity-averse preferences axiomatized by Klibanoff, Marinacci, and Mukerji (2009).

all extant work on consumer default, the relative gains seen in the SM relative to DM strongly depend on the “worst case” for household income. In particular, the large welfare losses in the DM relative to SM stem from the ability of young agents to borrow out to the natural debt limit. The natural debt limit is, however, extremely sensitive to small changes in the value of the worst-possible labor income realization, particularly for (i) young agents for whom the annuity value of future labor income is particularly high, and (ii) all agents when the risk-free borrowing rate is low.⁴ This lower bound is difficult to estimate accurately (see Deaton [1992] or Pemberton [1998]) and the worst-case outcomes are the primary focus of ambiguity-averse agents; thus, it seems important to understand whether the superiority of SM hinges entirely on the lowest value of income.

Our main finding along these dimensions is that even in the presence of very high levels of uninsurable labor income risk, high risk aversion, an unwillingness to substitute intertemporally, and the presence of ambiguity, the ability of households to default on debt leads to allocations that all households prefer less than the outcome that arises when they retain full commitment to repay. The intuition for our welfare results involves the relationship between the current economic situation of the borrower and the price of debt. When short-term debt is used in a setting with household labor income risk that is persistent, limited commitment to debt repayment will make credit expensive anytime the household experiences a negative shock; pricing “moves against” the unlucky borrower. (In Athreya, Tam, and Young [2009], we argue that unsecured credit markets are not insurance markets for precisely this reason.) As a result, agents who most “need” debt to smooth consumption are exactly those that find themselves unable to obtain it, because they also pose the highest risk of default. Tam (2009) extends this result to longer-term arrangements; specifically, he finds that competitively priced longer-period debt (in which the pricing function is held fixed over a number of periods) is welfare-dominated by one-period debt.

In contrast, the possibility of welfare gains from lowering penalties by enough to yield default in equilibrium was first suggested by Dubey,

⁴ Denoting by $y_{\min} > 0$ the lowest realization of potential labor income and r the risk-free interest rate on debt, the natural borrowing limit for an infinitely lived agent is given by $\underline{b}_{nat} \equiv -\frac{y_{\min}}{r}$, a function that asymptotes to $-\infty$ as interest rates go to zero. Assuming a credit card interest rate of 14 percent (the modal interest rate in Survey of Consumer Finances data in 1983 adjusted for a measure of realized inflation), the natural debt limit moves roughly seven times as much as the minimum income level. For good borrowers, for whom interest rate discounts have recently appeared (Furletti 2003; Livshits, MacGee, and Tertilt 2008), the natural debt limit will be even more sensitive.

Geanakoplos, and Shubik (2005). Theirs was a setting where borrowers of differential default risk were pooled together and thereby *did not* pay the individually actuarially fair price for their debt issuance. As a result of the stylized nature of their two-period model, it is not suitable for determining whether defaultable debt is welfare-improving in a more quantitatively oriented model economy. In some quantitative settings where pooling is imposed exogenously, Athreya (2002) and Mateos-Planas and Seccia (2006) find that welfare is higher in SM than DM. More recently, in a setting where private information allows for equilibrium pooling, the findings of Athreya, Tam, and Young (2009) suggest again that, as a quantitative matter, short-term defaultable debt is unlikely to be able to function as a form of insurance. Viewing these findings as a whole, they support the notion that the benefits of slacker borrowing constraints outweigh the costs of having no default option.

Lastly, with respect to political support for a policy allowing debt default, in addition to the welfare gains from having defaultable debt available in the presence of expense shocks, it seems possible that such provisions would enjoy support even in their absence. One obvious possibility is that the current regime may simply reflect objectives other than the maximization of the welfare of newborn agents. We therefore ask if *ex post* welfare can account for the evident political support enjoyed by proponents of relatively lax rules on default. Specifically, we ask whether model agents would choose to allow the option to default on debt in an economy where it was not already present (taking into account all changes resulting from the policy change). We find some support for such a change, but it falls well short of a majority. Support for the default option comes from relatively unlucky middle-aged college graduates: These are agents who borrowed a lot when young, in (rational) anticipation of higher income in middle age. When realized income did not materialize as expected, such households have significant debt as they approach retirement, and so will benefit from having debt obligations removed. Young agents, by contrast, are almost uniformly opposed to allowing defaultable debt, and even less-educated workers do not generally support it.

1. MODEL

Households in the model economy live for a maximum of $J < \infty$ periods. We assume that the economy is small and open, so that the risk-free interest rate is exogenous, while the wage rate is still

determined by a factor price condition.⁵ As a result, our welfare calculations will be biased toward finding a positive role for bankruptcy, since any lost resources arising from the implementation of default procedures like bankruptcy courts and legal costs will be ignored.

Households

Each household of age j has a probability $\psi_j < 1$ of surviving to age $j + 1$ and has a pure time discount factor $\beta < 1$. Households value consumption per household member $\frac{c_j}{n_j}$ and attach a negative value $\lambda_{j,y}$ (in terms of a percentage of consumption) to all nonpecuniary costs of defaulting, which depend on type y to be defined below. Their preferences are represented by a recursive utility function $U\left(\left\{\frac{c_j}{n_j}\right\}_{j=1}^J\right)$ that we detail below. Households retire exogenously at age $j^* < J$.

We follow Chatterjee et al. (2007) in allowing for household-level costs from default that are primarily nonpecuniary in nature. The existence of nonpecuniary costs of default are also suggested by the calculations and evidence in Fay, Hurst, and White (1998) and Gross and Souleles (2002), respectively. The former article shows that a large measure of households would have “financially benefited” from debt default via personal bankruptcy but did not file for protection, while both articles document significant unexplained variability in the probability of default across households even after controlling for a large number of observables. These results suggest the presence of implicit unobserved collateral that is heterogeneous across households, including (but not limited to) any “stigma” associated with default along with any other costs that are not explicitly pecuniary in nature (as in Athreya [2004]). We will therefore sometimes refer to $\lambda_{j,y}$ as stigma in what follows, although we intend it to be more encompassing.

The household budget constraint during working age is given by

$$c_j + q(b_j, I)b_j + \Delta \mathbf{1}(d_j = 1) \leq a_j + (1 - \tau)W\omega_{j,y}ye\nu, \quad (1)$$

where q is an individual-specific bond price that depends on bond issuance b_j and a vector of individual characteristics I . Net worth *after* the current-period default decision is denoted a_j , and therefore satisfies $a_j = b_{j-1}$ if the household does not default and $a_j = 0$ otherwise; Δ is the pecuniary cost of filing for default. The last term is after-tax

⁵ In our previous work we introduce a class of “special” agents who hold large amounts of capital for the purpose of endogenously obtaining a low, risk-free rate in the presence of low asset holdings for the median agent. Here we ignore the general equilibrium determination of returns and thus drop the special households from the model because their presence is irrelevant to the question at hand.

current labor income (τ is the tax rate). Log labor income is the sum of five terms: the aggregate wage index W , a permanent shock y realized prior to entry into the labor market, a deterministic age term $\omega_{j,y}$, a persistent shock e that evolves as an AR(1):

$$\log(e') = \varsigma \log(e) + \epsilon', \quad (2)$$

and a purely transitory shock $\log(\nu)$. Both e and $\log(\nu)$ are independent mean zero normal random variables with variances that are y -dependent.⁶ The budget constraint during retirement is

$$c_j + q(b_j, I) b_j \leq a_j + \Delta \mathbf{1}(d_j = 1) + vW\omega_{j^*-1,y}ye_{j^*-1}\nu_{j^*-1} + \Upsilon W, \quad (3)$$

where, for simplicity, we assume that pension benefits are composed of a fraction $v \in (0, 1)$ of income in the last period of working life plus a fraction Υ of average income (which is normalized to 1).

The survival probabilities $\psi_{j,y}$ and the deterministic age-income terms $\omega_{j,y}$ differ according to the realization of the permanent shock. We interpret y as differentiating between non-high school, high school, and college education levels, as in Hubbard, Skinner, and Zeldes (1994), and the differences in these life-cycle parameters will generate different incentives to borrow across types. In particular, college workers will have higher survival rates and a steeper hump in earnings; the second is critically important as it generates a strong desire to borrow early in the life cycle. Less importantly, they also face slightly smaller shocks than the other two education groups. The life-cycle aspect of our model is key—in the data, defaults are skewed toward young households (who borrow at least in part for purely intertemporal reasons), particularly those who do not report medical expenses as a main contributor to their default.⁷

Nonpecuniary costs, λ , follow a two-state Markov chain with realizations $\{\lambda_{L,y}, \lambda_{H,y}\}$ that are independent across households, but serially dependent with transition matrix

$$\Pi_\lambda = \begin{bmatrix} \pi & 1 - \pi \\ 1 - \pi & \pi \end{bmatrix}.$$

Due to data limitations, we assume that the transition probability matrix is symmetric and type-invariant, so the only difference across types in terms of stigma costs are their realizations. Our parametrization is more flexible than we used in previous work (Athreya, Tam, and Young 2009, 2012) so that we can match the default rates across education groups. As we show in a subsequent section, the process is still not

⁶ We approximate both e and ν with finite-state Markov chains. This approximation has the convenient property that income is bounded.

⁷ See Sullivan, Warren, and Westbrook (2000).

flexible enough to match all the targets of interest, although it does a reasonable job. Households cannot borrow or save during the period in which they declare default; however, they face no restriction in any subsequent period.⁸

Loan Pricing

We focus throughout on competitive domestic lending. There exists a competitive market of intermediaries who offer one-period debt contracts and utilize available information to offer individualized credit pricing. Let I denote the information set for a lender and $\hat{\pi} : b \times I \rightarrow [0, 1]$ denote the function that assigns a probability of default to a loan of size b given information I ; $\hat{\pi}(b, I)$ is identically zero for positive levels of net worth and is equal to 1 for some sufficiently large debt level. The break-even pricing function $q(\cdot)$ satisfies

$$q_j(b, I) = \begin{cases} \frac{1}{1+r} & \text{if } b \geq 0 \\ \frac{(1-\hat{\pi}(b, I))\psi_j}{1+r+\phi} & \text{if } b < 0 \end{cases} \quad (4)$$

given $\hat{\pi}(b, I)$.

In terms of loan pricing, some remarks are in order. In earlier work, Athreya (2002) specified an exogenous credit limit and then limited the sensitivity of loan pricing by forcing all loans to be priced identically. This approach has the benefit of plausibly capturing the “optionality” of the typical unsecured debt contract, whereby households can count on being able to borrow at a predetermined interest rate up to a predetermined credit limit, i.e., a credit “line.” A second benefit from this approach is that it might allow a shortcut to analyzing pooling outcomes that arise from private information on borrower characteristics. However, there are clear drawbacks to this approach as well. First, for the counterfactuals we are interested in, we desire a setting in which both the supply side of the credit market and prices jointly respond to changes in borrowing and repayment incentives. By contrast, in Athreya (2002), only prices responded. For large changes in default incentives, such as what we will examine, this is not a desirable limitation. More recently, Mateos-Planas and Seccia (2006) extended the approach of Athreya (2002) to allow for changes in credit limits, but both it and Athreya (2002) in the end employ a framework substantially different enough to make the comparison to the existing models described at the outset difficult. Second, from even a purely empirical perspective, there are reasons to avoid the use of pooling contracts. As

⁸ That is, exclusion from credit markets beyond the initial period is not sustainable as a punishment.

documented in Livshits et al. (2012), and Athreya, Tam, and Young (2012), among others, the variation in unsecured credit terms is now large and appears sensitive to household-level conditions. Lastly, while not directly observable, it is plausible that while individual credit contracts are best characterized by a single interest rate and credit limit, the proper interpretation of credit in the model is the sum of all credit available to the household. In this case, then, the question is the extent to which the household would have to pay more, sooner or later, to acquire additional credit. Our chosen approach features pricing that responds to default in a manner that yields supply-side effects and makes the marginal cost of credit an increasing function.

Returning to the model, r is the exogenous risk-free saving rate and ϕ is a transaction cost for lending, so that $r + \phi$ is the risk-free borrowing rate; the pricing function takes into account the automatic default by those households that die at the end of the period.⁹ We assume I contains the entire state vector for the household: $I = (a, y, e, \nu, \lambda, j)$. Zero profit for the intermediary requires that the probability of default used to price debt must be consistent with that observed in the stationary equilibrium, implying that

$$\hat{\pi}(b, I) = \sum_{e', \nu', \lambda'} \pi_e(e'|e) \pi_\nu(\nu') \pi_\lambda(\lambda'|\lambda) d(b(a, y, e, \nu, \lambda, j), e', \nu', \lambda'). \quad (5)$$

Since $d(b, e', \nu', \lambda')$ is the probability that the agent will default in state (e', ν', λ') tomorrow at debt level b , integrating over all such events *tomorrow* produces the relevant default risk. This expression also makes clear that knowledge of the persistent component e is critical for predicting default probabilities; the more persistent e is, the more useful it becomes in assessing default risk.

Government

The only purpose of government in this model is to fund pension payments to retirees. The government budget constraint is

$$\begin{aligned} \tau W \int y \omega_{j,y} e \nu \Gamma(a, y, e, \nu, \lambda, j < j^*) = \\ W \int (\nu \omega_{j^*-1, y} y e_{j^*-1} \nu_{j^*-1} + \Upsilon) \Gamma(a, y, e, \nu, \lambda, j \geq j^*). \end{aligned}$$

The left-hand side is the total revenue obtained by levy of a flat tax rate τ on all working agents, where the distribution of working

⁹ We assume any savings of households who die is taxed at 100 percent and used to fund wasteful government spending.

households (those for whom $j < j^*$) over productivity levels and age is given by $\Gamma(\cdot)$. The right-hand side is the total expenditure on retirees (those for whom $j \geq j^*$). Recall that to provide a tractable representation of social security and retirement benefits, we assume that retirement income is composed of a fraction $v \in (0, 1)$ of income in the last period of working life plus a fraction Υ of average income (which is normalized to 1).

Price Determination

We assume that the risk-free rate r is exogenous and determined by the world market for credit. Given r , profit maximization by domestic production firms implies that

$$W = (1 - \alpha) \left(\frac{r}{\alpha} \right)^{\frac{\alpha}{\alpha-1}},$$

where α is capital's share of income in a Cobb-Douglas aggregate production technology. Our assumption that the risk-free rate is exogenous deserves discussion. It is certainly reasonable to assume that the U.S. capital market is open, so empirically it is not implausible. Furthermore, if we close the economy we confront the high concentration of wealth puzzle directly—the median-wealth agent in the United States has little or no wealth and thus cares about default policy, since they may borrow in the future if unlucky, while the mean agent holds substantial wealth and is unlikely to be concerned with the default policy in place.¹⁰ There is a caveat, however. Li and Sarte (2006) is an early article that establishes a role for general equilibrium feedback effects that overturn partial equilibrium implications. Though we suspect our findings are robust to the determination of the risk-free rate via general equilibrium restrictions, it is not known for sure whether this is the case.

Preferences

Here we present the recursive representations of the preferences we study.

¹⁰ Chatterjee et al. (2007) calibrate their model to match the wealth distribution in the United States in a dynastic setting. As we have argued, life-cycle considerations are important for assessing the welfare effects of bankruptcy.

Constant Relative Risk Aversion

The agent's problem is standard under CRRA preferences, with the Bellman equation for a household of age j given by

$$v(a, y, e, \nu, \lambda, j) = \max_{b, d(e', \nu', \lambda') \in \{0, 1\}} \left\{ \frac{n_j}{\rho} \left(\frac{c_j}{n_j} \right)^\rho + \beta \psi_{j,y}(EU) \right\}$$

$$EU = \sum_{e', \nu', \lambda'} \pi_e(e'|e) \times \pi_\nu(\nu') \pi_\lambda(\lambda'|\lambda) V \left(\begin{array}{c} b, y, e', \nu', \lambda', j+ \\ 1 \end{array} \right)$$

$$V(b, y, e', \nu', \lambda', j+1) = (1 - d(e', \nu', \lambda')) v(b, y, e', \nu', \lambda', j+1) + d(e', \nu', \lambda') v^D(0, y, e', \nu', \lambda', j+1), \quad (6)$$

subject to the budget constraint given in (1) and (3), depending on their age.

The value function for a household that defaulted in the current period is given by

$$v^D(0, y, e, \nu, \lambda, j) = \max \left\{ \frac{n_j}{\rho} \left(\lambda \frac{c_j}{n_j} \right)^\rho + \beta \psi_{j,y}(EU) \right\}$$

$$EU = \sum_{e', \nu', \lambda'} \pi_e(e'|e) \times \pi_\nu(\nu') \pi_\lambda(\lambda'|\lambda) v \left(\begin{array}{c} 0, y, e', \nu', \lambda', j+ \\ 1 \end{array} \right). \quad (7)$$

$1 - \rho \geq 0$ is the coefficient of relative risk aversion and also the inverse of the elasticity of intertemporal substitution. Given our assumptions, the budget constraints remain the same as for all other agent types, aside from current net worth being zero as a result of the default.

Epstein-Zin

Under Epstein-Zin preferences, a household of age j solves the dynamic programming problem

$$\begin{aligned}
 v(a, y, e, \nu, \lambda, j) &= \max_{b, d(e', \nu', \lambda') \in \{0, 1\}} \left\{ n_j \left(\frac{c_j}{n_j} \right)^\rho + \beta \psi_{j, y} (EU)^{\frac{\rho}{1-\sigma}} \right\}^{\frac{1}{\rho}} \\
 EU &= \sum_{e', \nu', \lambda'} \pi_e(e'|e) \times \\
 &\quad \pi_\nu(\nu') \pi_\lambda(\lambda'|\lambda) V(b, y, e', \nu', \lambda', j+1) \\
 V(b, y, e', \nu', \lambda', j+1) &= (1 - d(e', \nu', \lambda')) \times \\
 v \left(\begin{matrix} b, y, e', \nu', \lambda', j+ \\ 1 \end{matrix} \right)^{1-\sigma} &+ d(e', \nu', \lambda') \times \\
 &\quad v^D(0, y, e', \nu', \lambda', j+1)^{1-\sigma}, \tag{8}
 \end{aligned}$$

subject to the usual budget constraints, and where

$$\begin{aligned}
 v^D(0, y, e, \nu, \lambda, j) &= \max \left\{ n_j \left(\lambda \frac{c_j}{n_j} \right)^\rho + \beta \psi_{j, y} (EU)^{\frac{\rho}{1-\sigma}} \right\}^{\frac{1}{\rho}} \\
 EU &= \sum_{e', \nu', \lambda'} \pi_e(e'|e) \pi_\nu(\nu') \times \\
 &\quad \pi_\lambda(\lambda'|\lambda) v \left(\begin{matrix} 0, y, e', \nu', \lambda', j+ \\ 1 \end{matrix} \right) \tag{9}
 \end{aligned}$$

is the value of default. $\sigma \geq 0$ governs the household’s aversion to fluctuations in utility across states of nature while $\rho \leq 1$ controls the substitutability between current and future utility; specifically, σ is the coefficient of relative risk aversion with respect to gambles over future consumption and $\frac{1}{1-\rho}$ is the elasticity of intertemporal substitution in consumption. When $\rho = 1 - \sigma$, these preferences generate the same ordering over stochastic streams of consumption as expected utility does.

2. RESULTS

The results are organized into two subsections. First, we study the roles played by aversion to fluctuations in consumption over time and across states-of-nature. We begin with expected utility preferences. We then relax this by employing Epstein-Zin preferences. Throughout this subsection, we consider parameter values that lie near the values implied by the benchmark calibration; these values ensure that model

outcomes remain in congruence with cross-sectional facts on consumption and income inequality. We show that welfare under the default option is lower, at least *ex ante*. Second, based on this result, we ask the “inverse” question: Are there economies in which welfare in the standard model is worse? In this subsection, we no longer restrict ourselves to parameters dictated by U.S. data; rather, our goal is to understand whether any parameterizations within the parametric classes we study are capable of generating lax default as a welfare-improving policy. Specifically, we consider shocks with counterfactually large persistent and transitory components and preferences that display ambiguity aversion.

As noted at the outset, our approach throughout will shut down expense shocks in an otherwise standard consumption smoothing problem. A caveat is in order. While we have argued that this is informative about a case in which insurance is introduced where it was previously missing, it should be recognized that this is not necessarily identical to that case. In particular, the most direct route to addressing the question of whether default would remain useful if society located a way to insure what are presently uninsurable expenses is to explicitly model such an option. We opt for a simpler approach here in part because the form of such insurance, were it to become available, is not obvious a priori. This is primarily because it is unlikely to be provided privately, given that it has not emerged to this point. As a result, the form it takes will likely be as part of a tax-transfer scheme. Our model lacks the detail needed to address the associated incentive-related costs. Our approach is therefore similar to the thought-experiment of Lucas (1987) in which the costless removal of risks was employed as a benchmark for the gains from business cycle stabilization. Still, the reader should keep in mind the indirect nature of our approach and the limits it places on the interpretation of our results. In particular, our approach leads us to calibrate more than once, sometimes with only partial success, depending on the case under study, as opposed to calibrating once at the outset. We acknowledge this limitation and leave the alternate approach for future work.

Does Default Help Insure Labor Income Risk?

In this subsection, we evaluate the implications of default relative to the standard model for a variety of empirically plausible values for agent attitudes toward intra- and intertemporal consumption smoothing. Before evaluating these alternatives, we present our argument for why default regimes must be a matter of policy rather than an endogenous

outcome of decentralized trading arrangements. The most prevalent form of explicitly unsecured credit is that arising from the open-ended revolving debt plan offered by credit card lenders. Credit card lending, in turn, has been (certainly since the mid-1990s) extremely competitive.¹¹ The relevance of the competitiveness of the U.S. unsecured lending industry is that the credit market cannot be punitive in its treatment of those who default. That is, no single firm would be willing to treat an individual borrower any worse than the current assessment of their state would justify. As a result, a household contemplating default in such a setting can safely rule out being “punished” for it. In the case where default conveys no additional information to a lender than what it was able to observe *ex ante*, there is literally no change in terms that are “caused” by the act of renegeing on a payment obligation. Conversely, when default does reveal information, the change in terms is again not “punitive” in nature, but instead reflects an updated assessment of default risk. As a result, “high” *ex post* interest rates following default are implausibly ascribed to deadweight loss-inducing penalties. In the symmetric-information and competitive setting we study, punishments that are *ex post* inefficient will not be sustainable. Even if any single lender could withhold credit after default, the presence of other lenders would undermine the possibility of anything purely punitive. As a result, default costs capable of sustaining unsecured credit markets are likely to require intervention by policymaking authorities.¹² Thus, in the market for unsecured consumer debt, it is likely that *any costs of default filing that are in any way punitive have to be policies*.¹³

At the outset, we noted that for plausible parameterizations of preferences that admit an expected utility representation, the

¹¹ The average interest rate on credit card balances is high—currently 14 percent—relative to more secured forms of debt. As Evans and Schmalensee (2005) have pointed out, however, it is straightforward to account for the interest rate after funding costs, transactions costs, and, most crucially, default costs are taken into account, without relying on market power distortions.

¹² Most dynamic contracting models of limited borrower commitment, for example, currently use implicit or explicit appeals to public institutions with commitment to punish, in order to motivate penalties for the value of autarky. In recent work, Krueger and Uhlig (2006) show that the inability of the supply side of the credit market to commit to punishments can have severe implications for the existence of the market itself. In the “normal” case, Krueger and Uhlig (2006) show that competition in fact collapses credit and insurance markets completely even without informational frictions.

¹³ We want to be clear that what we call “penalties” differs from the usage in Ausubel and Dawsey (2008), where rates imposed after late or missed payments are labeled punitive. They attribute the high values of such rates to a common agency problem. Modeling the bilateral contracting problem that would arise in the presence of noncompetitive intermediation is well beyond the goals for this article. We are pursuing the endogenous determination of interest rate hikes for delinquent borrowers in other work.

standard model typically maximizes welfare. Our first step is to understand whether this argument against default obtains only because of the restriction to expected utility or is a more fundamental property of models of life-cycle consumption smoothing. To collapse the model to the standard model, the specific quantitative experiment we consider is the imposition of a cost of default Δ that is large enough to eliminate all default on the equilibrium path.¹⁴ Before proceeding, we note the following property of our model.

Proposition 1 *For each (a, y, e, ν, j) there exists Δ large enough that $\hat{\pi}(b) = 0$.*

This result relies on the nonnegativity condition for consumption— if Δ exceeds the labor income of the household in the current period, default cannot occur since consumption would have to be negative. Given that total labor income is bounded (by assumption) and borrowing is proscribed in the period of default, we can always impose a cost of filing sufficient to generate zero default along the equilibrium path. We then compute the change in lifetime utility for each individual given a Δ that exceeds the maximum required; in the absence of general equilibrium effects, we can compute these changes for each individual, rather than simply for newborns, without the need to track transitional dynamics. We will focus in general on *ex ante* welfare of newborns.

Calibration

We consider a benchmark case of expected utility, where $\rho = 1 - \sigma = -1$. We choose $(\beta, \lambda_{L,y}, \lambda_{H,y}, \pi)$ to match the default rates of each type y , the measure of negative net worth as a fraction of gross domestic product for each type y , the fraction of borrowers, and the discharge ratio (mean debt removed via default divided by mean income at time of filing). Table 1 contains the constellation of parameters that fits best (when viewed as exactly identified generalized method of moments with an identity weighting matrix). Other parameters are identical to those in Athreya, Tam, and Young (2009)—these include the resource cost of default Δ , the income processes faced by each type, the measure of each type, and the parameters of the retirement system (θ, Θ) .¹⁵

¹⁴ Similar results would obtain if the government could impose “shame” on households by choosing values for λ , provided it could make λ large enough to guarantee zero default on the equilibrium path. In our model, the Inada condition on consumption implies that such a λ always exists.

¹⁵ Specifically, we set $\nu = 0.35$, $\Upsilon = 0.2$, $\phi = 0.03$, $\Delta = 0.03$, $\zeta = 0.95$, $\sigma_{n,\epsilon}^2 = 0.033$, $\sigma_{n,\nu}^2 = 0.04$, $\sigma_{h,\epsilon}^2 = 0.025$, $\sigma_{h,\nu}^2 = 0.021$, $\sigma_{c,\epsilon}^2 = 0.016$, and $\sigma_{c,\nu}^2 = 0.014$.

Table 1 Calibration

Case Parameter, Target	$\rho = -1, \sigma = 2$		$\rho = -0.5, \sigma = 2$		$\rho = -1, \sigma = 5$	
	Parameter	Outcome	Parameter	Outcome	Parameter	Outcome
$\lambda_{nhs}^h, \pi_{nhs} = 1.03\%$	0.8972	0.31%	0.8668	1.24%	0.9376	0.51%
$\lambda_{nhs}^l, E(\frac{b}{y} b < 0)_{nhs} = 0.1552$	0.7624	0.2071	0.6929	0.2104	0.7538	0.1561
$\lambda_{hs}^h, \pi_{hs} = 1.11\%$	0.8832	0.97%	0.8064	1.29%	0.8872	1.31%
$\lambda_{hs}^l, E(\frac{b}{y} b < 0)_{hs} = 0.5801$	0.7135	0.1835	0.6933	0.1825	0.6236	0.2553
$\lambda_{coll}^h, \pi_{coll} = 0.57\%$	0.7067	0.63%	0.7136	0.79%	0.7055	0.76%
$\lambda_{coll}^l, E(\frac{b}{y} b < 0)_{coll} = 0.7251$	0.5698	0.1504	0.6352	0.1506	0.4205	0.2194
$\beta, \Pr(b < 0) = 12.5\%$	0.9765	17.5%	0.9895	13.3%	0.9532	12.5%
$\rho_\lambda, \frac{E(b d=1)}{E(y d=1)} = 0.56$	0.8597	0.3986	0.6655	0.4073	0.7658	0.4630

Our model is not capable of exactly matching the entire set of moments—for example, we underpredict default rates and discharge, generally underpredict debt-to-income ratios, and overpredict the measure of borrowers. This inability arises because the model actually places very tight links between some variables, restricting the minimization routine’s ability to independently vary them.¹⁶ In the end, one either accepts that expense shocks do indeed play a very dominant role in default data, or one is left with a puzzle relative to standard consumption-savings models. Still, we note that the qualitative findings from our analysis do not appear to depend on our specification of the stochastic process for λ .¹⁷

Expected Utility and Ex Ante Welfare

We consider two environments—one with the calibrated value for Δ and one with a cost Δ sufficient to eliminate default on the equilibrium path. Table 2 contains the welfare gain from the standard model in which it is infeasible for any household to declare default. Consistent with our previous work, we find that welfare is higher in the standard model *ex ante* for every newborn (independent of type). College types benefit the most from the change, and their welfare gain is substantial (1.2 percent of lifetime consumption). To aid the discussion in subsequent sections where we alter preference parameters, we quickly summarize the reasons for the welfare gains here.

In the standard model, the loss of resources generated by the filing cost is not present. Since we do not impose an economy-wide resource constraint, these lost resources are not important. Instead, the welfare gain is driven by an improved allocation of consumption. By the law of total variance, the variance of consumption over the life cycle can be decomposed into two components:

$$\text{Var}(\log(c)) = \text{Var}(E[\log(c) | \text{age}]) + E[\text{Var}(\log(c) | \text{age})].$$

¹⁶ Consider an attempt to improve the model’s prediction for the measure of borrowers by increasing β . Holding all other parameters constant, this reduces default rates and debt-to-income ratios for all types (and these variables are generally already too small). To counteract this effect, one might then move λ for each type and each state. Consider first increasing both λ_i^H and λ_i^L for one type i . While this change would increase the default rate—default becomes less costly—it would via a supply-side effect tend to reduce debt levels (see Athreya [2004]). By contrast, suppose we increase λ_i^H and decrease λ_i^L ; this change has countervailing effects on both default rates and debt levels and default rates could rise because it becomes cheaper for H types, but fall as it becomes more expensive for L types. A similar tension exists for debt-to-income ratios—driving it up for one type tends to drive it down for the other.

¹⁷ In the real world, “stigma” may also be a function of aggregate default rates (an agent cares less about default if everyone else is defaulting), in which case this invariance may break. To analyze this case would be of interest, but it poses some challenges with respect to calibration. We therefore defer it to future work.

Table 2 Welfare Gains (without Recalibration)

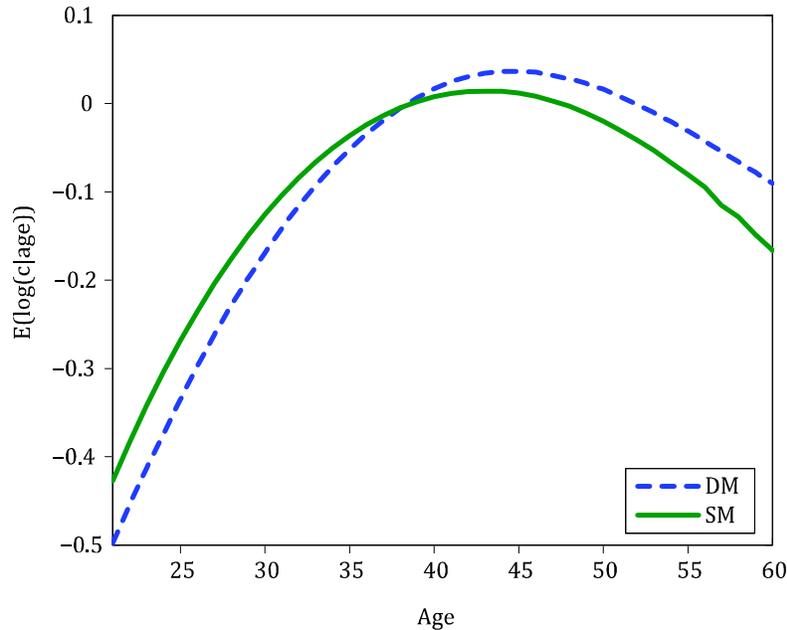
$\sigma = 2$ & $EIS = 0.5$	Coll	HS	NHS
$DM \rightarrow SM$	1.21%	0.54%	0.52%
$\sigma = 2$ & $EIS = 0.67$	Coll	HS	NHS
$DM \rightarrow SM$	0.58%	0.21%	0.13%
$\sigma = 5$ & $EIS = 0.5$	Coll	HS	NHS
$DM \rightarrow SM$	0.47%	0.16%	0.13%

We label the first term the “intertemporal” component of consumption smoothing; it represents how expected consumption differs across time periods. The second term is the “intra-temporal” component; it measures how much consumption varies across agents of a given age. Roughly speaking, how costly the first component is in terms of welfare depends on the elasticity of intertemporal substitution, because it measures the deterministic variance of consumption over time, whereas the welfare cost of the second part is governed by static risk aversion. In Figure 1 we see that the standard model, or “no-default” case (SM), improves intertemporal smoothing (the curve gets flatter) because all lending becomes risk-free. Thus, as we noted in the introduction, the only debt limit that is relevant is the natural debt limit, which is very large in our model for newborn agents. Turning to the intra-temporal component, in Figure 2 we see that the SM improves this as well, restating the analysis in Athreya, Tam, and Young (2009) that unsecured credit markets do not provide insurance. Here, bad shocks trigger tightening of credit constraints, making consumption smoothing across states of nature more difficult. As a result, young agents are unable to respond effectively to bad income realizations when they can default, causing their consumption to be highly volatile. Under the SM, the natural debt limit is sufficient to protect them against adverse shocks; by middle age, default has ceased to be relevant and thus the two cases largely coincide.¹⁸

The differences in outcomes across the DM and SM cases are given in Figures 3, 4, and 5, and are driven by changes in the pricing functions agents face. In Figure 3 we show the pricing functions in the low costs of default environment facing a young college agent across realizations of the persistent shock e . The initial flat segment is driven by Δ and is increasing in the current realization of the persistent shock e . As debt

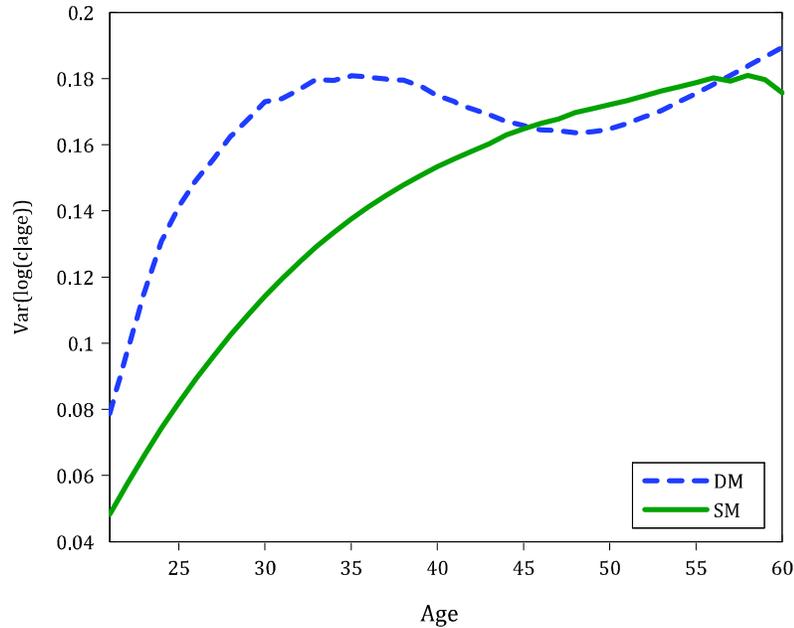
¹⁸ The figures are drawn for the aggregate, since the results are the same for each type qualitatively. Figures decomposed by type are available from the authors upon request.

Figure 1 Intertemporal Consumption Smoothing, Expected Utility



increases, more realizations of e' would trigger default, causing q to decline until it reaches zero; looking across e values we see that higher e realizations permit more borrowing. Of course, higher e realizations in our model are typically associated with less, not more, borrowing, so these increased debt limits are not particularly valuable; instead, the tightening of credit limits when e is low generates substantial costs for poor agents. In contrast, under SM pricing is flat out to the natural debt limit. Crucially, transitory shocks do not impact pricing; because ν' cannot be predicted using ν , the current transitory shock has no effect on the default decision tomorrow conditional on b (b is changed by the transitory shock, however).

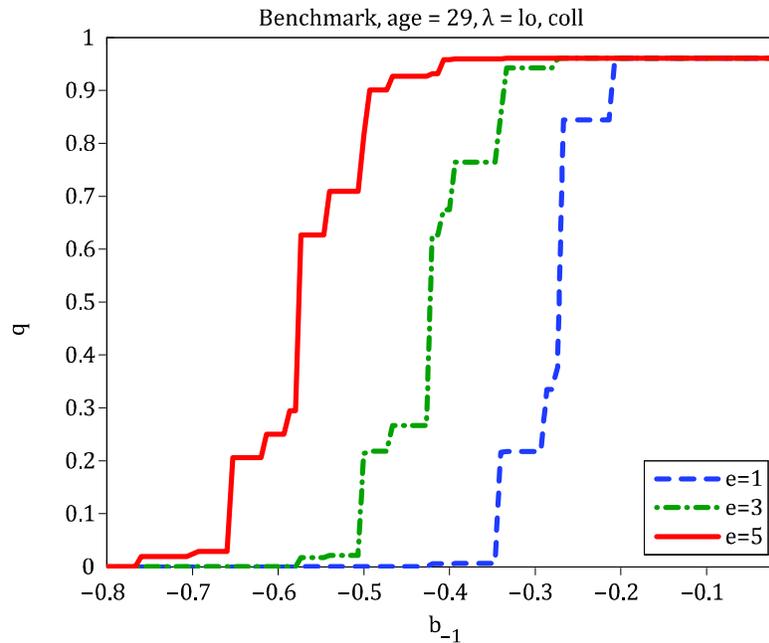
The potential tradeoff between the two components of smoothing motivated the life-cycle analysis of Livshits, MacGee, and Tertilt (2007) and Athreya (2008), so why doesn't default generate this tradeoff? As discussed in Athreya, Tam, and Young (2009), default can either help or hinder intratemporal smoothing, depending on which agent you ask. An agent facing an income process with low intertemporal variance

Figure 2 Intratemporal Consumption Smoothing, Expected Utility

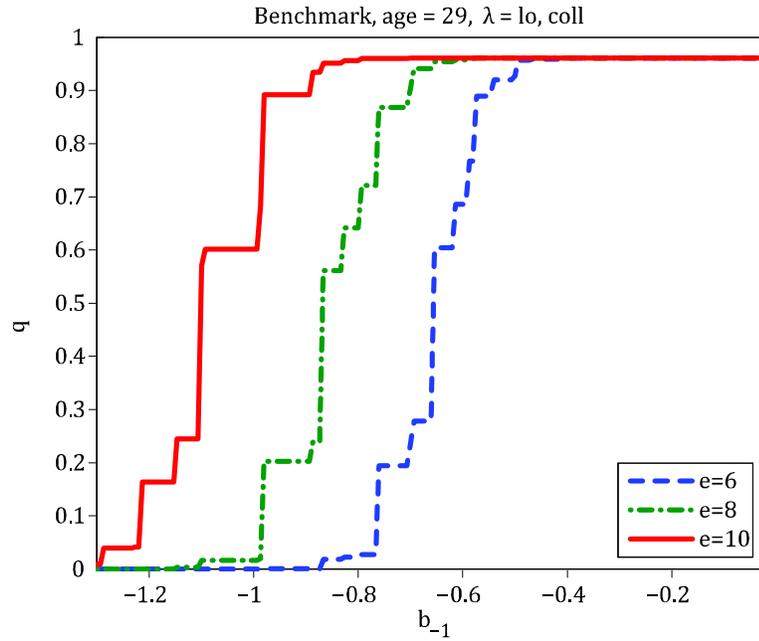
but high intratemporal variance—that is, tomorrow’s expected income is close to current income but tomorrow’s income has substantial risk—may benefit from default; the intertemporal distortion is minimal while the potential to truncate the consumption distribution at the low end conveys significant benefits (even once pricing is taken into account). In contrast, an agent facing the opposite process—income that grows over time and is relatively safe—generally does not benefit; default is not used because pricing prevents it and the intertemporal distortion is substantial, leading to significant welfare losses. In our model, a young agent is of the second type, especially a college-educated one, while older households are members of the first type.

Ex Post Welfare—Voting over Default Policy

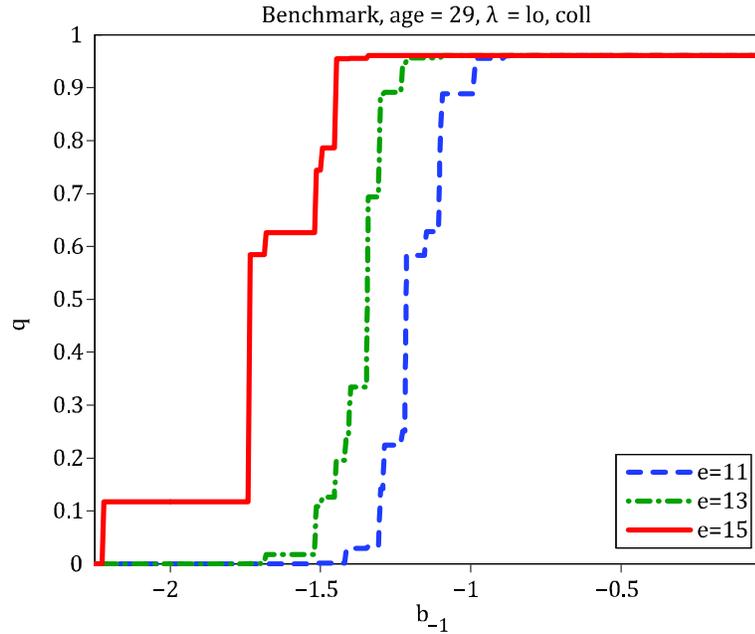
Because we study a small open-economy model in which the risk-free rate is fixed, but also allow all pricing to be individualized, there are

Figure 3 Pricing, Expected Utility

no “pecuniary” externalities. We can therefore compute the welfare consequences of policy changes for any agent at any point in the state space; since the distribution plays no role in pricing (and therefore no role in welfare), we do not need to calculate the transitional dynamics of the model to get the welfare changes. We ask agents of a given age and type whether, conditional on their current state, they would be in favor of eliminating the option to declare default. Figure 6 displays the measure of each type, conditional on age, that would support retaining default with the calibrated Δ . A substantial portion of college types oppose elimination, but they are all middle-aged and have experienced histories of bad shocks; the peak in opposition occurs earlier for high-school types and later for non-high-school types, with correspondingly fewer such households opposing overall. For the convenience of the reader, Table 3 presents the aggregate measures of each type that oppose eliminating default (the column labeled “DM Regime”); they are small for each education group. Furthermore, as is clear from the figures, almost no newborns oppose eliminating the option.

Figure 4 Pricing, Expected Utility

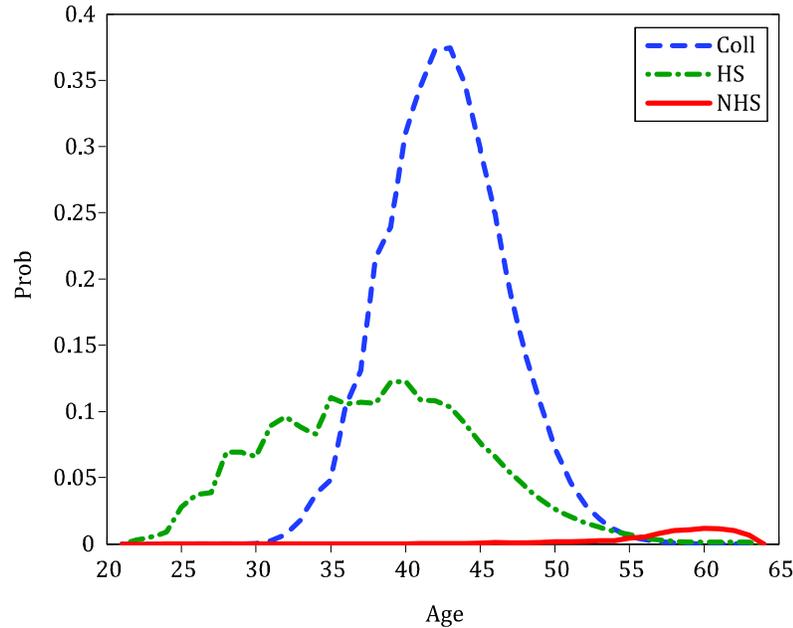
We now consider the inverse of the preceding experiment: Agents of different ages and types are asked if they would prefer to *introduce* default (again, with Δ set to its calibrated value) into a setting in which it is currently prohibited. As seen in Figure 7, a nontrivial fraction of agents would like to introduce default. The intuition here is that the no-default case allows significant borrowing at the risk-free rate. As a result, many households, especially the college-educated, borrow when young in anticipation of higher earnings. The relatively unlucky among them then find themselves indebted by middle age and thereby will benefit from the discharge of debts. Moreover, by virtue of being middle-aged, these households place relatively low value on being able to access the cheap unsecured debt later in life. This effect is especially strong for the college-educated, for whom purely intertemporal consumption smoothing motives dictate a strong effort to save for retirement beyond middle age. As a result, a substantial proportion of high-school- and college-educated household groups would support the introduction of default when they reach middle age. In contrast,

Figure 5 Pricing, Expected Utility

those who have not completed high school support the introduction of default only late in working life, when the subsequent increase in borrowing costs is not long-lasting. However, as Table 3 shows (the column “SM Regime”), the aggregate number of agents who vote in favor of introduction falls well short of majority status.

Separating Risk Aversion from Intertemporal Substitution

As discussed above, the two pieces of the variance decomposition have welfare costs that depend (mainly) on different aspects of preferences. Our benchmark case using CRRA expected utility restricted these two aspects of preferences to be reciprocals of each other. Here, we relax that requirement by using the Epstein-Zin preference structure, and consider two particular deviations. First, we make households more tolerant of intertemporal variance than in the expected utility benchmark by employing a high value for ρ . Second, the default option

Figure 6 Fraction Supporting Bankruptcy, BK Regime

may shrink the volatility of intratemporal consumption, at least for some ages. Given this, making intratemporal variance more painful to households may help us explain the presence of low default costs. We therefore select a relatively high value for σ . It is important to note that this particular combination of insensitivity to the timing of consumption and sensitivity to the income state in which it occurs is the arrangement that gives default its best chance of improving *ex ante* welfare and does not lie within the class of expected utility preferences.

The specific experiments we investigate involve changing ρ and σ without recalibrating the entire model. This type of change generates two effects—an effect conditional on borrowing (which we call the price effect) and an effect caused by changes in the number of borrowers (the extensive effect). We then compare the results with cases where the model is recalibrated (to the extent that is possible) in an attempt to isolate the two effects.

We first consider changes in ρ . To understand how this change affects welfare, it is helpful to first consider the extreme case of $\rho = 1$,

Table 3 Measure of Agents in Favor of Bankruptcy

Education	DM Regime	SM Regime
College	6.45%	4.09%
High School	4.05%	3.26%
Non-High School	0.16%	0.24%
Total	4.05%	2.98%

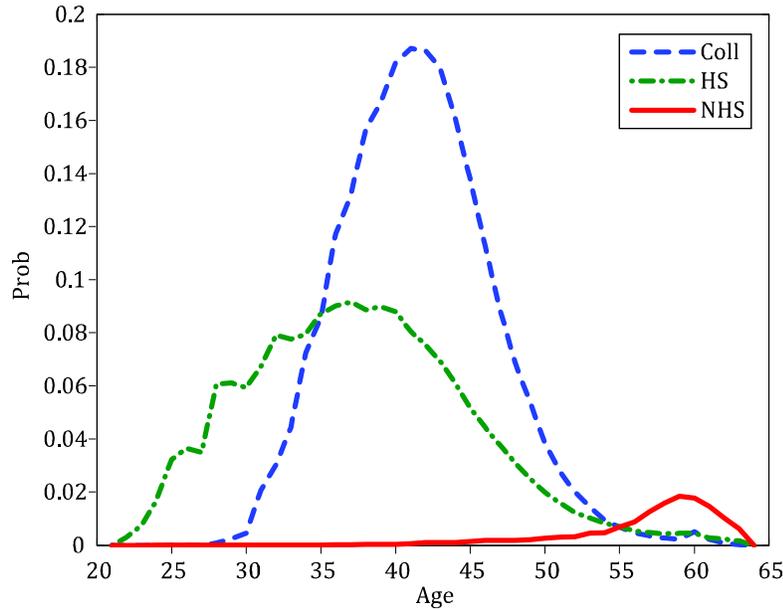
making the household infinitely willing to move consumption deterministically through time. As $\rho \rightarrow 1$, the Bellman equation converges to the form

$$v(a, y, e, \nu, \lambda, j) = \max \left(c_j + \beta \psi_{j,y} \left(\frac{\sum_{e', \nu', \lambda'} \pi_e(e'|e) \pi_\nu(\nu') \times}{\pi_\lambda(\lambda'|\lambda)} V(b, y, e', \nu', \lambda', j+1) \right)^{\frac{1}{1-\sigma}} \right).$$

Here, the household will either completely frontload or backload consumption, depending on the relationship between the discount factor and the interest rate. For the parametrization we use, the effective discount factor (β times the survival probability) lies between the risk-free saving and borrowing rates for almost every age, meaning that households don't wish to borrow and, critically, do not value the default option *no matter how risk averse* they are. For some older households, whose survival probabilities are relatively low, the effective discount factor is sufficiently low that they want to borrow and "frontload" their consumption; the option to default makes borrowing expensive enough to render complete frontloading impossible. This, in turn, reduces the welfare of these households—since they face no uncertainty, default is either probability zero or one and pricing therefore eliminates it. Obviously such extreme consumption behavior is inconsistent with U.S. cross-sectional facts; in particular, the model with $\rho = 1$ would miss very badly on the life-cycle pattern of consumption inequality, which in the data is substantially smaller than income inequality.

Returning to less extreme values for ρ , Figure 8 displays the pricing function across several different values of ρ and demonstrates the effect on loan prices. As ρ increases, the pricing function shifts downward because at any given level of debt an agent with a higher ρ is more willing to default. The intuition for this result is not straightforward. When ρ increases, the household is more willing to accept variability in consumption across time. If a household enters the current period with some debt and wishes not to lower debt, they have two options: (i) borrow more if possible or (ii) default and void those obligations. Borrowing more is only feasible if there is a reasonable commitment to repay. But since a bad shock would lead to low mean consumption,

Figure 7 Fraction Supporting Bankruptcy, NBK Regime



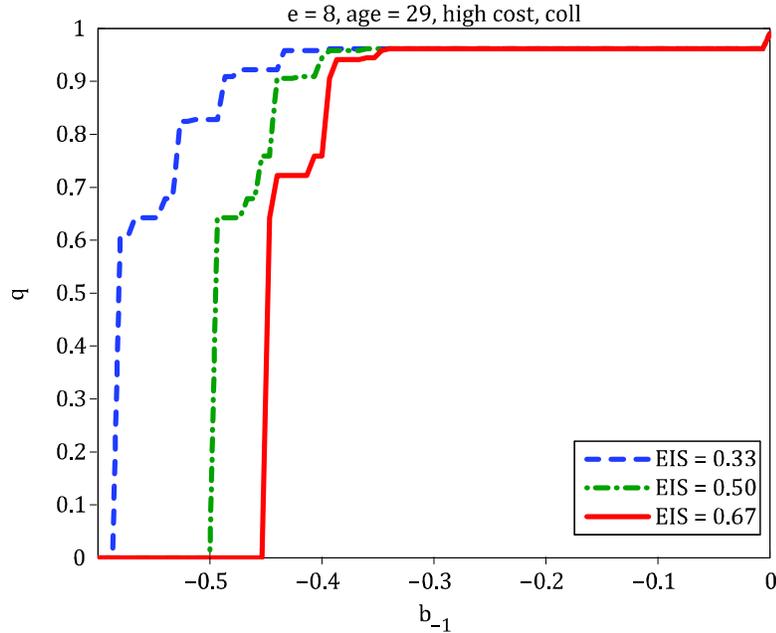
default becomes attractive, and households lack strong commitment to repay debt. As a result, they cannot borrow easily. For the cases with “intermediate” values for ρ , the creation of strong default incentives makes intertemporal smoothing more costly, but the latter is relatively unimportant.

Consider next an experiment where σ , the risk aversion with respect to gambles over future utility, is increased. Again, turning first to the polar case, let $\sigma \rightarrow \infty$, so that the household becomes infinitely risk averse. In this case, the limiting household Bellman equation takes the form

$$v(a, y, e, \nu, \lambda, j) = \max \left\{ \begin{array}{l} n_j \left(\frac{c_j}{n_j} \right)^\rho + \\ \beta \psi_{j,y} \min_{e', \nu', \lambda'} \{ V(b, y, e', \nu', \lambda', j+1) \}^\rho \end{array} \right\}^{\frac{1}{\rho}},$$

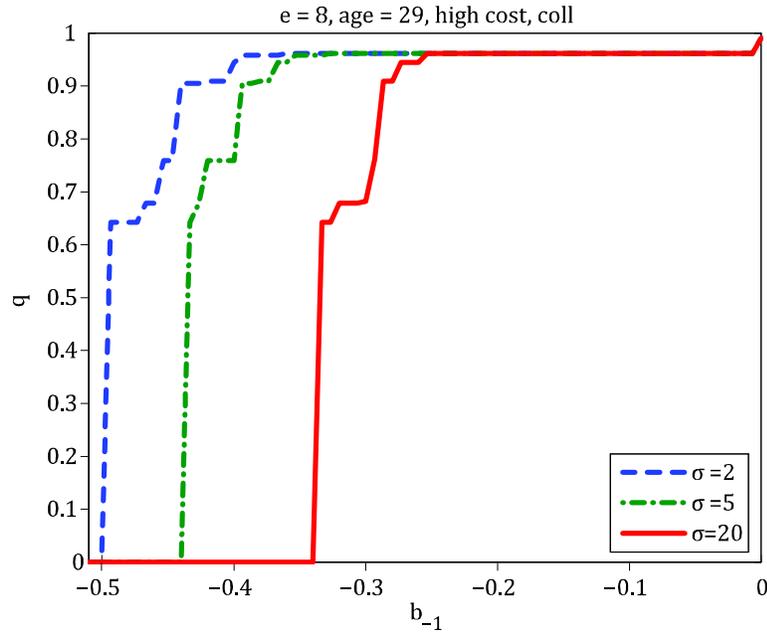
subject to the usual budget constraints seen earlier in equations (1) and (3).

When households are infinitely risk averse, they choose not to borrow for the reasons outlined in Athreya, Tam, and Young (2009)—

Figure 8 Pricing, Epstein-Zin with Different EIS

unsecured credit markets do not provide insurance and thus agents will be unwilling to pay the transaction cost to borrow. As a result, there is a welfare gain to living in the standard model, as no household has negative net worth. Again, extreme preferences render the model grossly inconsistent with cross-sectional facts; here, consumption inequality would be essentially zero over all ages.

Returning again to more intermediate cases, we see that changes in risk aversion generates two effects. The extensive margin effect is similar to increasing ρ , but for different reasons. When σ is large, households have a strong demand for precautionary savings; for $\sigma = 5$, for example, we see a clear decline in the measure of total borrowers, again making default overall less damaging. The pricing effect is also similar; by increasing risk aversion, we make the household less willing to have consumption differ across states of the world tomorrow. Conditional on borrowing, the pricing functions reveal a stronger desire to default—for any given b , the price of debt is decreasing in σ (see Figure 9). As above, there are only two options for a household with

Figure 9 Pricing, Epstein-Zin with Different Risk Aversion

debt; since even a moderately bad outcome will cause a highly risk-averse agent to default, commitment is not possible, leaving default as the only option for smoothing consumption across states.¹⁹ Combining these results into one statement, we see that no combination of (ρ, σ) leads to default being a welfare-improving policy, although for extreme cases it will be nearly innocuous.

Table 2 shows that welfare is higher (for newborns) in the standard model (SM), but that the gains from (imposing the high Δ) decline with risk aversion and elasticity of intertemporal substitution (EIS). $\rho > 1 - \sigma$, which is satisfied when either parameter increases, implies the household has a preference for early resolution of uncertainty; thus, default appears to be least damaging when households prefer to resolve their risk early rather than late.

¹⁹ Our model satisfies the conditions noted in Chatterjee et al. (2007) that imply default occurs only if current debt cannot be rolled over: If $d(\epsilon', \nu', \lambda') > 0$ for some $\epsilon', \nu', \lambda'$, then there does not exist b such that $a + y - q(b, Y)b > 0$ for total income Y .

Table 4 Welfare Gains (with Recalibration)

$\sigma = 2$ & $EIS = 0.5$	College	High School	Non-High School
$DM \rightarrow SM$	1.21%	0.54%	0.52%
$\sigma = 2$ & $EIS = 0.67$	College	High School	Non-High School
$DM \rightarrow SM$	0.28%	0.05%	0.04%
$\sigma = 5$ & $EIS = 0.5$	College	High School	Non-High School
$DM \rightarrow SM$	1.28%	0.57%	0.56%

All of these results are obtained without recalibrating the model. To ensure that our findings are not particularly sensitive to this strategy, we also recalibrate the model for different values of ρ and σ , to the extent that this recalibration is possible; Table 1 contains the new parameter values that best fit the targets under alternative settings. By doing so, we attempt to shut off the extensive margin, although we are not completely successful. When we recalibrate, we find that with high EIS all welfare gains from eliminating default are substantially reduced, with both noncollege types now barely benefiting at all (see Table 4), while for high risk aversion the welfare gains increase slightly. As noted above, this welfare change is entirely due to the shifts in the pricing function that higher EIS and/or higher risk aversion engender. Thus, for no parameter combination that we consider do we observe welfare gains from retaining the default option.

A summary of findings thus far is that default significantly worsens allocations for income risk and preference parameters that are empirically plausible for U.S. data, as well as for more extreme values of preference parameters within the class of Epstein-Zin non-expected utility preferences. We turn now to the question of whether such policies continue to remain desirable under two additional (and more substantial) departures from the settings studied so far.

Is the Standard Model Ever Worse?

We begin this section by allowing for the underlying volatility of income to be driven by relatively more and less persistent income shocks. For this experiment, we hold the unconditional variance of labor income fixed and vary the relative contributions of the persistent component e and the transitory component ν . We then ask whether a relaxation in the household's understanding of the probabilistic structure of earnings risk can open the door for welfare-improving default. For this

experiment, we allow for households to display ambiguity aversion in the sense of Klibanoff, Marinacci, and Mukerji (2009).²⁰

The Roles of Persistent and Transitory Income Risk

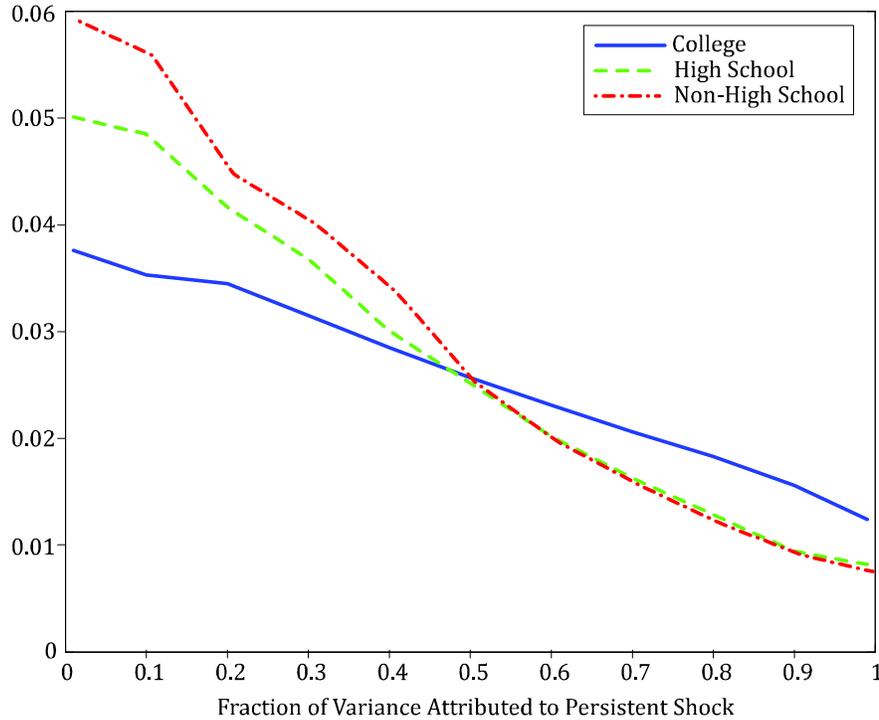
It has long been known that self-insurance, and therefore also the benefit of insurance markets, hinges critically on the persistence of the risks facing households. As a general rule, the more persistent are shocks, the more difficult they are to deal with via the accumulation of assets in good times and decumulation and borrowing in bad times. In contrast, purely transitory income shocks can typically be smoothed effectively. In a pure life-cycle model, however, there are additional impediments to self-insurance: Young households are born with no wealth and often face incentives to borrow arising from purely intertemporal considerations. In particular, those with relatively high levels of human capital, especially the college-educated, can expect age-earnings profiles with a significant upward slope into late middle age. As a result, such households would like to borrow even in the absence of any shocks to income, often substantially, against their growing expected future income. In contrast, those households with low human capital face a far less income-rich future, and as a result borrow primarily to deal with transitory income risk.

In order to understand the role that the persistence of income risk plays in the welfare gains or losses arising from U.S.-style bankruptcy and delinquency, we now evaluate the effects of changes in the persistent component of household income risk for all three classes of households. However, in order to avoid conflating persistence and overall income volatility, we adjust the variance of transitory income volatility such that the overall variance of log labor income remains constant.²¹ Figure 10 and Tables 5 and 6 present the welfare and consumption smoothing implications of the standard model under varying income shock persistence. The first column of each table documents the fraction of total variance contributed by the persistent component.

Normatively, three findings are noteworthy. First, and perhaps most importantly, the standard model displays higher welfare irrespective of the nature of shocks accounting for observed income

²⁰ There are connections between ambiguity aversion and the concept of Knightian uncertainty from Bewley (2002), although the latter concept does not permit preferences to be represented by a utility function and is therefore hard to analyze quantitatively. There are also connections between ambiguity aversion and robust decisionmaking as defined by Hansen and Sargent (2007).

²¹ Athreya, Tam, and Young (2009) are primarily concerned with the role of income variance in models of default.

Figure 10 Welfare Gains from Eliminating Bankruptcy

volatility. This result strengthens our findings thus far, and it further suggests that defaultable debt is simply unlikely to be useful to households. It is also a particularly important form of robustness, given both the general importance of persistence for the efficacy of self-insurance and borrowing and because estimates of income shock persistence vary dramatically—see Guvenen (2007), Hryshko (2008), or Guvenen and Smith (2009) for discussions of the debate between so-called “restricted income profiles” (RIP), in which all households draw earnings from a single stochastic process, and “heterogeneous income profiles” (HIP), in which households vary in the processes from which they derive earnings. This debate has implications for models like ours because these two models differ, sometimes strongly, in the persistence of earnings shocks their structure implies. Most recent work now suggests that income-process parameters vary over the life cycle as well (Karahan and Ozkan 2009).

Table 5 Consumption Smoothing (DM)

	Intra			Inter			Total		
	Coll	HS	NHS	Coll	HS	NHS	Coll	HS	NHS
1.0%	0.0306	0.0462	0.0575	0.0359	0.0364	0.0386	0.0665	0.0826	0.0961
10.0%	0.0377	0.0561	0.0872	0.0343	0.0367	0.0357	0.0720	0.0938	0.1229
20.0%	0.0459	0.0807	0.1092	0.0336	0.0347	0.0325	0.0795	0.1154	0.1417
30.0%	0.0538	0.0884	0.1367	0.0327	0.0327	0.0297	0.0865	0.1211	0.1664
40.0%	0.0619	0.1013	0.1472	0.0316	0.0301	0.0280	0.0925	0.1314	0.1752
50.0%	0.0700	0.1146	0.1613	0.0305	0.0284	0.0263	0.1005	0.1430	0.1876
60.0%	0.0779	0.1280	0.1797	0.0294	0.0264	0.0241	0.1065	0.1544	0.2038
70.0%	0.0859	0.1413	0.1992	0.0283	0.0247	0.0224	0.1141	0.1660	0.2216
80.0%	0.0946	0.1543	0.2182	0.0272	0.0231	0.0211	0.1218	0.1774	0.2393
90.0%	0.1053	0.1681	0.2368	0.0258	0.0212	0.0199	0.1311	0.1893	0.2567
99.0%	0.1248	0.1863	0.2566	0.0235	0.0187	0.0180	0.1483	0.2050	0.2680

Table 6 Consumption Smoothing (SM)

	Intra			Inter			Total		
	Coll	HS	NHS	Coll	HS	NHS	Coll	HS	NHS
1.0%	0.0196	0.0307	0.0474	0.0318	0.0314	0.0120	0.0514	0.0621	0.0594
10.0%	0.0271	0.0397	0.0577	0.0315	0.0298	0.0124	0.0586	0.0695	0.0801
20.0%	0.0360	0.0541	0.0771	0.0311	0.0290	0.0131	0.0671	0.0831	0.0902
30.0%	0.0444	0.0683	0.0971	0.0306	0.0284	0.0137	0.0750	0.0967	0.1108
40.0%	0.0524	0.0820	0.1173	0.0300	0.0277	0.0144	0.0824	0.1097	0.1317
50.0%	0.0600	0.0951	0.1364	0.0295	0.0271	0.0151	0.0895	0.1222	0.1515
60.0%	0.0673	0.1076	0.1550	0.0291	0.0267	0.0158	0.0964	0.1343	0.1708
70.0%	0.0743	0.1197	0.1729	0.0288	0.0262	0.0164	0.1031	0.1495	0.1893
80.0%	0.0811	0.1314	0.1903	0.0285	0.0258	0.0171	0.1096	0.1627	0.2075
90.0%	0.0878	0.1428	0.2072	0.0282	0.0255	0.0178	0.1160	0.1638	0.2250
99.0%	0.0935	0.1528	0.2218	0.0280	0.0253	0.0182	0.1215	0.1781	0.2400

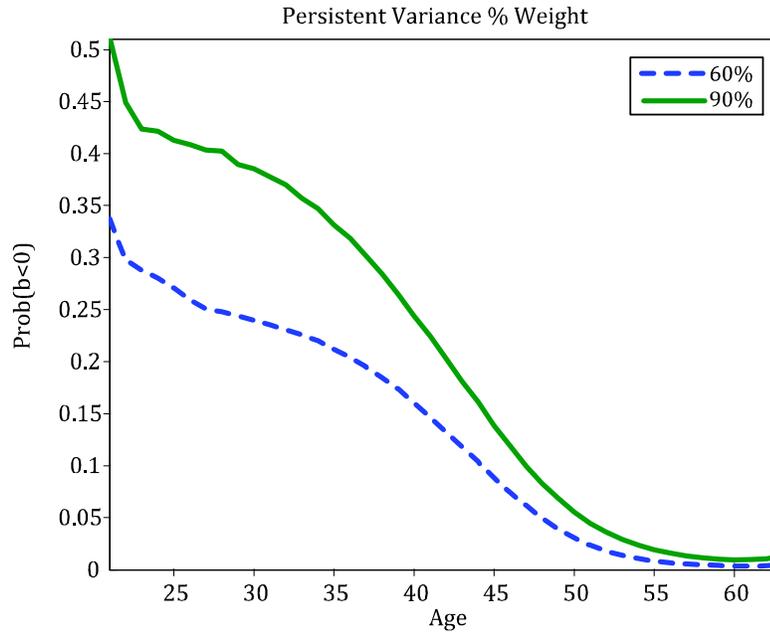
Second, the effect of the contribution of persistent shocks to income volatility depends on the education level of households. In particular, when volatility is driven primarily by persistent shocks, the relatively well-educated benefit from the elimination of default substantially more than their less-educated counterparts. Conversely, when most income variability is driven by large but transitory shocks, it is the relatively less-educated who benefit most from the elimination of the default option. The intuition for this result comes from the nature of borrowing: College types borrow primarily to use future expected income today while less-educated types borrow to smooth shocks.

Third, within each educational class, the welfare losses from default decline monotonically as the relative contribution of the persistence of the shock grows; default on debt is least (most) useful when income volatility is driven primarily by shocks that are transitory (persistent). What is surprising, but in keeping with the main theme of our results, is that in *no* case is it true that U.S.-style default is *ex ante* more desirable than allocations obtaining under the standard model. Moreover, even in the case where essentially all income risk is delivered in the form of persistent shocks where credit markets are least useful in dealing with income risk, outcomes that allow for default are worse for agents than those arising in the standard model. The welfare in the standard model is non-trivially higher, at up to 1.24 percent of consumption for college-educated households (as seen in Figure 10).

In Figures 11 and 12 we display the measure of borrowers at each age and the conditional mean of debt among those who borrow for two levels of the importance of persistent income risk.²² The fact that the losses from allowing default rise for all agent types with the importance of transitory shocks is a consequence of the increased usefulness of credit in dealing with transitory income risk. Conversely, when shocks are primarily persistent, a negative realization requires more frequent borrowing and leads, all else equal, to more debt in middle age; the combination is ultimately unable to stem the transfer of income risk to consumption volatility. In Tables 5 and 6, we see that, irrespective of default policy, persistence translates into higher consumption volatility, and that the presence of lax default policy seen in Table 6 does little to stem the flow of income risk into consumption risk (echoing our previous result in Athreya, Tam, and Young [2009]).

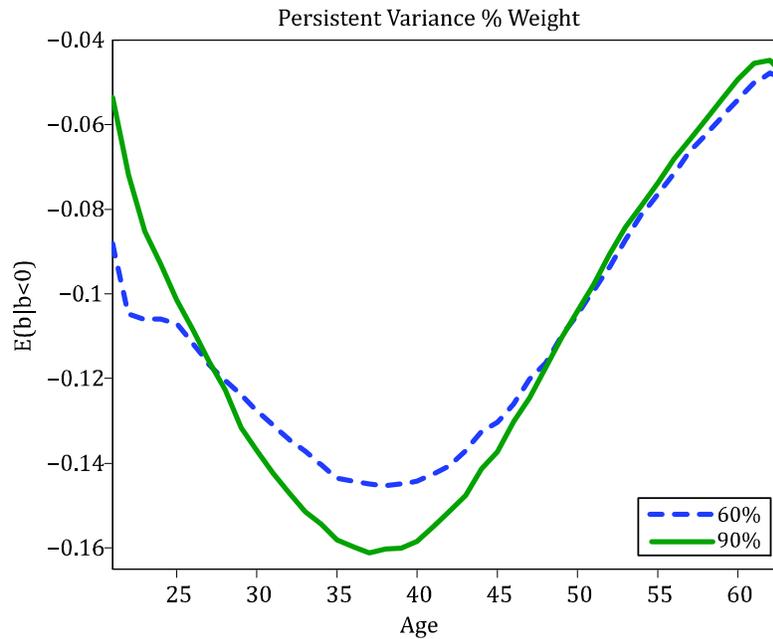
We turn next to the relationship between shock persistence and equilibrium default rates, displayed in Figure 13. Default is “U-shaped,” with high default rates at both ends. To understand this shape, con-

²² From the perspective of a newborn, the measure of borrowers of a given age equals the probability of the newborn borrowing at that age.

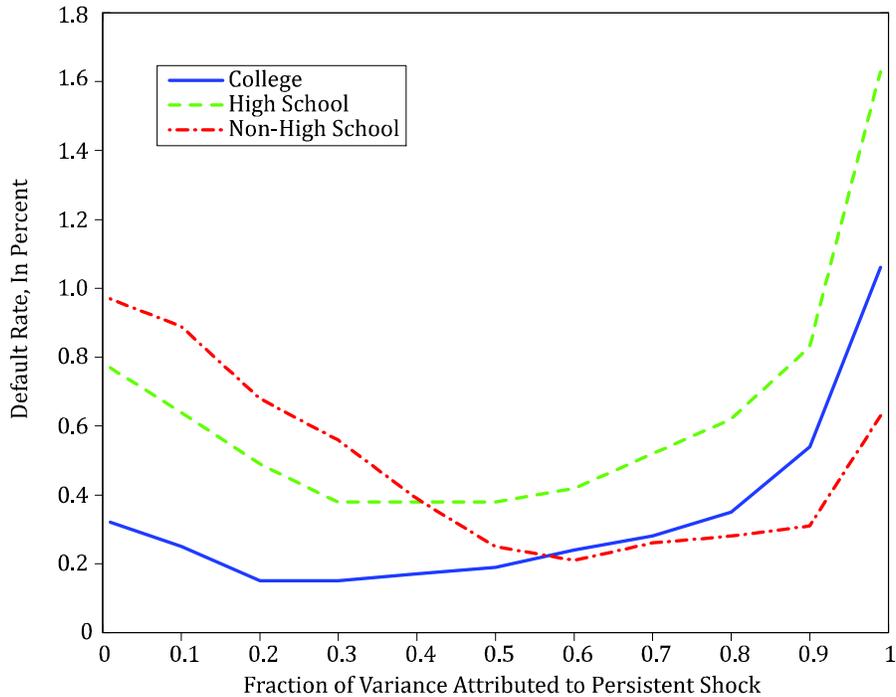
Figure 11 Fraction of Borrowers

sider first the case where the labor income shocks are nearly all transitory (the left side of the graph). Here, agents can generally manage their risk effectively via saving and dissaving, but they choose to augment the self-insurance mechanism with default at higher rates than they do in the benchmark setting. The reason they do so is that risk-based pricing is not effective here, because there is no useful information contained in the current labor income of the borrower that would identify future bad risks. In contrast, in the case where labor income is driven entirely by the persistent component (the right side of the graph), high default is the result of agents being generally unable to smooth consumption; persistent shocks are hard to smooth using assets alone (and if permanent are in fact impossible). As a result, despite the pricing effects, borrowers will use default relatively often (and pay the costs to do so). The middle parts of the graph, where default is lowest, balance these two effects.

Intuitively, in the standard model, borrowers realize that debt must be repaid, and under high persistence, heavy borrowing in response

Figure 12 Mean Debt of Borrowers

to a negative shock makes low future consumption relatively likely. Nonetheless, credit markets are willing to lend to such households at the risk-free rate (adjusted for any transactions costs of intermediation), making total debt rise. When default is available, borrowing today to deal with persistent income risk does not expose the borrower to severe consumption risk in the long term as default offers an “escape valve,” but it does expose lenders to severe credit risk in the near term. Creditors then price debt accordingly; as seen in Figure 14, when shocks are primarily persistent, as the current shock deteriorates so do the terms at which borrowers can access credit. Moreover, under a bad current realization of income, households facing persistent risk see a disproportionate decline in the price of any debt they may issue, while the reverse occurs in the event of a good current realization of income; the pricing functions essentially “switch places.” Yet, despite the increased sensitivity of loan pricing to the borrower’s current income state under relatively high persistence, the welfare gains under the SM, though still positive, fall. This result obtains because of the reduction

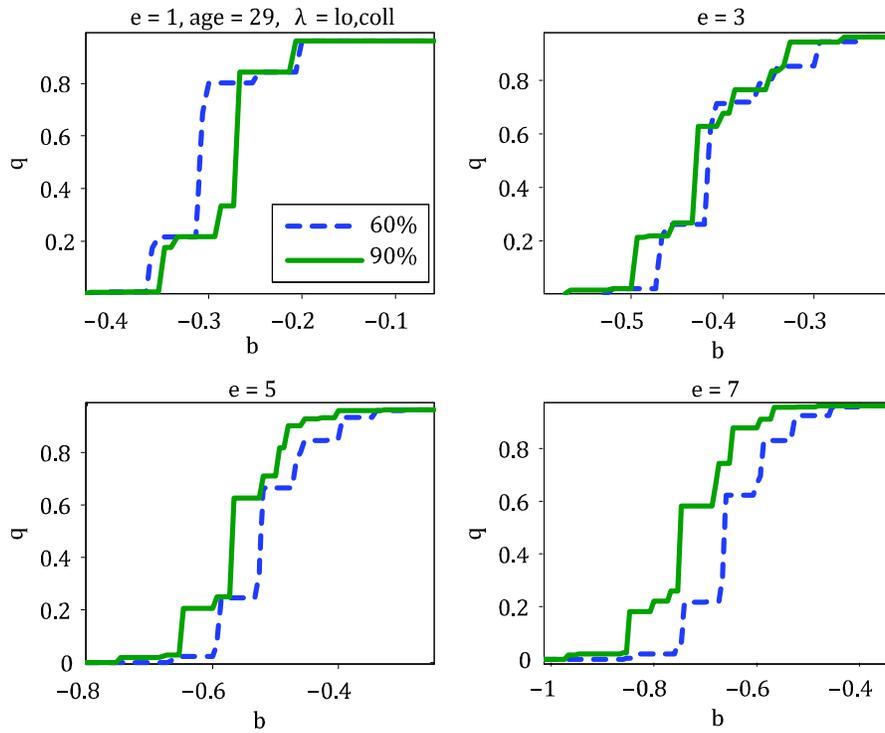
Figure 13 Default Rates

in the ability of self-insurance, inclusive of borrowing, to prevent income fluctuations from affecting consumption. To sum up, income risk is quantitatively relevant in governing the losses conferred by default, but irrelevant for altering the qualitative welfare property that, in the absence of expense shocks, the default option lowers welfare.

Ambiguity Aversion

We turn next to the question of whether default can improve outcomes when households are not perfectly certain about the probabilistic structure of income risk. Households that face ambiguity are uncertain about the probability process for their incomes; if ambiguity-averse, these households behave pessimistically and therefore adopt views about their income that would, for example, imply that it would mean-revert more slowly from low realizations. In such a situation, borrowing to smooth away temporary falls may not be optimal, since asset

Figure 14 Pricing Functions



decumulation is not effective against permanent shocks, and therefore in the absence of a default option households may be unwilling to do so. In contrast, if default is an option, the household may be willing to borrow since, even if their pessimism is validated, consumption can be protected via discharge. We formalize this idea, as in Klibanoff, Marinacci, and Mukerji (2009), by assuming agents are averse to ambiguity. In this formulation, a household of age j solves the dynamic

programming problem

$$v(a, y, e, \nu, \lambda, j) = \max_{b, d(e', \nu', \lambda') \in \{0, 1\}} \left\{ \frac{n_j}{1-\sigma} \left(\frac{c_j}{n_j}\right)^{1-\sigma} + \beta \psi_{j,y} \sum_{e', \nu'} p(e', \nu' | e, \nu) \times \Phi(EU) \right\}$$

$$EU = \sum_{e', \nu', \lambda'} \pi_e(e' | e) \pi_\nu(\nu') \pi_\lambda(\lambda' | \lambda) \times V(b, y, e', \nu', \lambda', j + 1), \tag{10}$$

subject to budget constraints, (1) and (3), where $\Phi(\cdot)$ is given as follows:

$$\Phi(x) = \begin{cases} \frac{1 - \exp(-\eta x)}{1 - \exp(-\eta)} & \text{if } \eta > 0 \\ x & \text{if } \eta = 0 \end{cases}$$

determines preferences over ambiguity. $\eta \geq 0$ controls the attitude toward ambiguity; as η increases, the household becomes more averse to ambiguous stochastic processes. The restrictions on the choices of $p(e', \nu' | e, \nu)$ are that they must sum to 1 for each (e, ν) and every element must lie in some set $\mathcal{P} \subset [0, 1]$; we nest the standard model by setting the \mathcal{P} to be an arbitrarily small interval around the objective probabilities.²³ We use π to denote objective probabilities and p to denote subjective ones; note that households are assumed to be uncertain only about the distribution of income shocks, not the process for λ .

Because we are interested in these preferences only to the extent that they may provide an environment in which relatively low-cost default and debt discharge are welfare-enhancing, we will deliberately take the most extreme case of $\eta = \infty$, yielding the max-min specification from Epstein and Schneider (2003):

$$v(a, y, e, \nu, \lambda, j) = \max_{b, d(e', \nu', \lambda') \in \{0, 1\}} \left\{ \frac{n_j}{1-\sigma} \left(\frac{c_j}{n_j}\right)^{1-\sigma} + \beta \psi_{j,y} \min_{p(e', \nu' | e, \nu)} EU \right\}$$

$$EU = \sum_{e', \nu', \lambda'} p(e', \nu' | e, \nu) \times \pi_\lambda(\lambda' | \lambda) V(b, y, e', \nu', \lambda', j + 1)$$

$$V(b, y, e', \nu', \lambda', j + 1) = (1 - d(e', \nu', \lambda')) v(b, y, e', \nu', \lambda', j + 1) + d(e', \nu', \lambda') v^D(0, y, e', \nu', \lambda', j + 1), \tag{11}$$

²³ We do not require that the household assume that the probabilities of the independent events are independent in every distribution that is considered. That is, the household may be concerned that the independence property is misspecified and therefore select a worst-case distribution in which the events are correlated.

where

$$v^D(0, y, e, \nu, \lambda, j+1) = \max \left\{ \frac{n_j}{1-\sigma} \left(\lambda \frac{c_j}{n_j} \right)^{1-\sigma} + \beta \psi_{j,y} \min_{p(e', \nu' | e, \nu)} EU \right\}$$

$$EU = \sum_{e', \nu', \lambda'} p(e', \nu' | e, \nu) \times \pi_\lambda(\lambda' | \lambda) v(0, y, e', \nu', \lambda', j+1) \quad (12)$$

is the value of default.

The min operator that appears in front of the summation reflects the agent's aversion to uncertainty; as shown by Epstein and Schneider (2003), a household who is infinitely uncertainty-averse chooses the subjective distribution of future events that is least favorable and then makes their decisions based on that subjective distribution. The size of the set of possible processes \mathcal{P} measures the amount of ambiguity agents face; a typical p_{ij} element lies in the interval $[\mathbf{p}_1^{ij}, \mathbf{p}_2^{ij}] \subset [0, 1]$.²⁴

Standard ambiguity aversion models imply that households will learn over time and reject stochastic processes that are inconsistent with observed data (for example, a household who initially entertains the possibility of permanently receiving the worst possible income level forever will dismiss this process as soon as one non-worst realization occurs). For simplicity, we will focus our attention on a special case of extreme ambiguity aversion in which this learning does not occur; if default is not useful in this environment, it is likely of less use to households than when they face less uncertainty over time. The intuition is that the income process we buffet agents with is a non-unit process. To the extent that households would realize by a certain age that the data they've received makes unit-root earnings unlikely, they would be able to rule out such a persistent process and thereby smooth more effectively, and as a result, may not value default as much as someone viewing shocks as permanent.

Given the qualifications and considerations discussed above, we now evaluate outcomes in the standard model in the case where $\mathcal{P} = [0, 1]$, the most extreme case possible (households behave as if the minimum income draw will be realized with probability 1 next period). The intuition is that such a case offers the possibility, discussed at the outset, that lax penalties for default might actually encourage the use of credit for consumption in a setting where the agent's aversion to ambiguity would otherwise preclude becoming indebted. And in fact, we *do* find that this case delivers default as welfare-improving for some

²⁴ Hansen and Sargent (2007) provide an interpretation of \mathcal{P} in terms of detection probabilities.

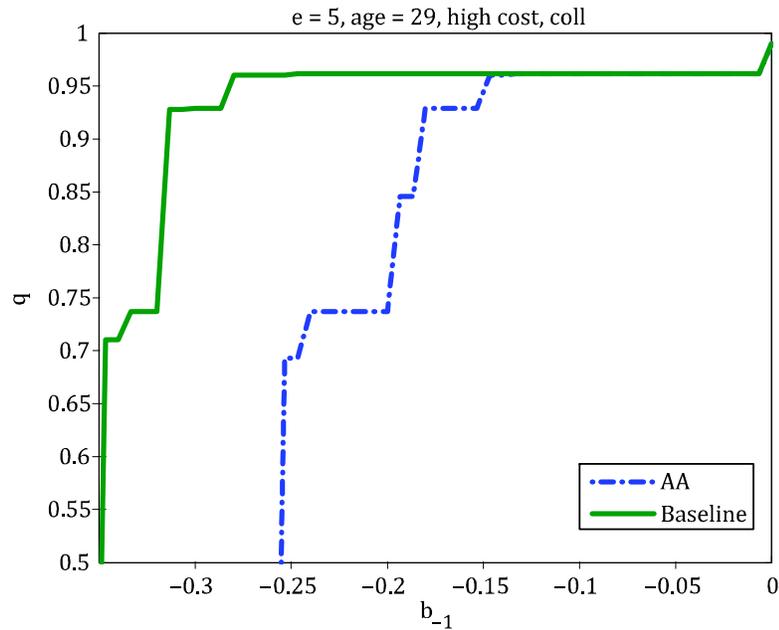
Table 7 Welfare Effects Under Ambiguity Aversion

$\mathcal{P} = [0, 1]$	Non-High School	High School	College
DM \rightarrow SM	0.215%	0.189%	-0.185%
$\mathcal{P} = \min(1, \pi + 0.5)$	Non-High School	High School	College
DM \rightarrow SM	0.296%	0.219%	0.044%

agents (see Table 7). However, this finding is very limited: Benchmark default costs improve welfare for only the college type and the welfare gain is tiny (under 0.2 percent of consumption). As a result, unconditional *ex ante* welfare is negative since college types are not a large enough group to overcome the losses to the remainder of the population. It is interesting to see, however, that the welfare changes from allowing default are now reversed—the largest gains are experienced by the most educated, while the least educated suffer more. Part of the intuition for this result is that it is the best educated who face the steepest mean age-earnings profiles. Therefore, these agents would have the strongest purely intertemporal motives to borrow, absent any ambiguity. Low default costs mitigate the effect of ambiguity and allow for states in which a temporarily unlucky college-educated agent would find borrowing desirable.

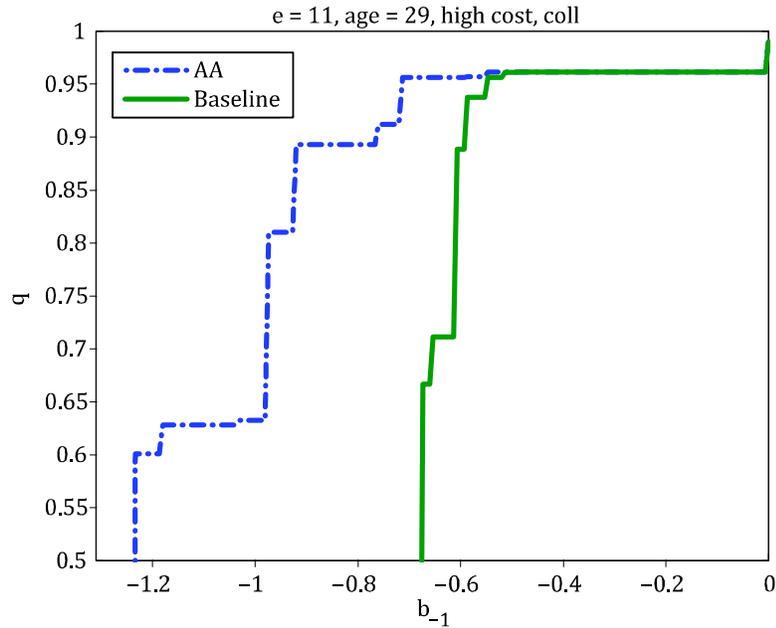
Pricing is presented in Figures 15 and 16. Notice that for the low realization of e , the pricing function under ambiguity aversion is everywhere below the baseline expected utility case, but for the higher realization they switch places; ambiguity-averse agents with high income actually pose *less* of a default risk. The difference in pricing stems only from a difference in the households' willingness to default next period for a given b . Since default has a fixed cost component (Δ), households want to time their usage of default; in particular, households must balance the gains from defaulting tomorrow from those arising from waiting until additional shocks have been realized. This fact places the expectations of income in periods after tomorrow at the heart of the timing of default decisions, and here households who face ambiguity about the income process act quite differently from those in the benchmark economy.²⁵

²⁵ The exposition is simpler if we refer to the expectations of the households facing ambiguity as coinciding with the choice of p , because the ambiguity-averse agents act *as if* those probabilities were the objective ones. Of course, if one were to ask ambiguity-averse agents about their forecasts of future income, they would use the true objective probabilities; they just do not use these probabilities for decisions. The proper phrasing of our statement “ambiguity-averse agents expect low future income” would be the more cumbersome “ambiguity-averse agents act as if they expect low future income.” We

Figure 15 Pricing, Ambiguity Aversion, Low e 

Take first the household with low e . For a “rational expectations” household, income in the distant future is expected to be better than whatever is realized tomorrow, as e is persistent but mean-reverting; for the household facing ambiguity, however, income is actually expected to be no better, or even worse, than tomorrow’s realization. Since ambiguity-averse households do not think the future will be better, they may as well default next period if the realization of income is bad; lenders must therefore offer them higher interest rates to break even. In contrast, the ambiguity-averse household with higher e views a realization near the mean for next period as unexpectedly *good*, but does not expect better times in the more-distant future. Default in the next period is therefore not as valuable as waiting for a future period when those bad states are expected to occur. In contrast, without ambiguity a bad realization will induce the household to substantially

abuse the notion of expectation slightly as a result, and beg for the reader’s indulgence on this matter.

Figure 16 Pricing, Ambiguity Aversion, High e 

revise their future expectations downward, making default today more attractive (the decline in future income makes the fixed cost of default worth paying).²⁶ The result is that ambiguity-averse households with high current income obtain better terms.

Is such extreme ambiguity aversion “reasonable?” It seems highly unlikely that households entertain a stochastic process in which they receive the worst possible outcome forever with probability one as reasonable, at least not for long—after all, they need only observe the fact that their income is occasionally higher than the lower bound to discard this process empirically. As we noted above, we could introduce this learning into the model—since the households are simply learning about an exogenous process, it can be done “offline”—but it is computationally quite burdensome to condition the set of

²⁶ The median e has the pricing functions crossing, so that agents who face ambiguity are more likely to default on small debts but less likely to default on large ones.

permissive stochastic processes on the history of observations.²⁷ It is also the case that this extreme ambiguity leads to a discrepancy between model and data in terms of borrowing patterns; there is far too little debt, which lessens our interest in making this economy “more realistic.” If we consider smaller limits for \mathcal{P} , such as 10 percent above or below the objective value, we find that default is welfare-reducing for all education levels. Thus, while ambiguity aversion provides a theoretical foundation for default options, it does not appear to provide an empirically tenable one.

3. CONCLUDING REMARKS

We have studied the efficacy of default in helping households better insure labor income risk in a large range of settings in which risk aversion, intertemporal smoothing motives, income risk, and uncertainty—and attitudes to uncertainty—over income risk itself were all varied. Our findings here suggest that within the broad class of models used thus far to develop quantitative theory for unsecured consumer credit and default, relatively generous U.S.-style default does not appear to be capable of providing protection against labor income risk.

Despite the fact that we find that labor income risk is not well hedged from the *ex ante* perspective, we also show that there are *ex post* beneficiaries from allowing default as it currently is; specifically, we show that the standard model generates a positive measure of agents *ex post* who would vote to introduce default. Our calibrated model predicts that these agents do not constitute a majority, though, since they are primarily college-educated middle-aged households who have been unlucky enough to still have significant debt. This result warrants further investigation since it may help explain why default penalties are becoming less stringent over time (with the exception of some aspects of the most recent reform).

Our results also suggest that “expense” shocks or catastrophic movements in net worth are likely to be essential to justify the view of default as a welfare-improving social institution. To the extent that uninsured, catastrophically large, and “involuntary” expenditures are indeed a feature of the data, a natural question is whether consumer default is the best way to deal with such events. Given the nature of resource transfers created by default and the constraints that it imposes

²⁷ Since this learning is not Bayesian, it can be quite difficult to write recursively, and, in any case, learning about discrete processes generally involves a large number of states. Campanale (2008) investigates non-Bayesian learning in a two-state model where the approach taken introduces only one additional state.

on the young, who disproportionately account for *both* the income-poor and uninsured, this statement seems unlikely.

With respect to future work, it is worth stressing that since expense shocks and their absence seem so important to the implications of the class of models considered here, the value of purely empirical work better documenting the nature of expense shocks, and their (a priori plausible) connection to income shocks (for example, job loss leading to insurance loss, which in turn exposes households to out of pocket expenditures), is high. Relatedly, the pivotal role played by borrowing costs “moving against” unlucky borrowers seems important to independently substantiate. In the absence of such work, it remains a possibility that the welfare findings of this article (and essentially all others) hinges too much on an institutional arrangement for borrowing that is inaccurate. Use of detailed household level credit card pricing and income information seems productive.

In addition to the preceding, in light of the findings of this article and the larger quantitative theory of consumer default, two directions seem particularly useful. First, a more “normative” approach that asks if observed default procedures can arise an optimal arrangement under plausible frictions, may yield different conclusions. One interesting example of the latter approach is the theoretical work of Grochulski (2010), where default is shown to be one method for decentralizing a constrained Pareto optimum in the presence of private information. Quantifying models with default and endogenously derived asset market structures may lead to better understanding of policy choices in this area (such as why Europe has chosen to make default available under very strict conditions, and social insurance generous, while the United States has chosen the opposite).

Second, with respect to the experiments we studied, we were led to allow for two specific preference extensions beyond CRRA expected utility in order to accurately assess the particular tradeoffs created by default. While we emphatically did not attempt to turn the article into a survey of any larger variety of non-expected utility preferences, some further extensions seem potentially important: disappointment aversion (Gul [1991] or Routledge and Zin [2010]), deviations from geometric discounting (Laibson 1997), habit formation (Constantinides 1990), and loss aversion (Barberis, Huang, and Santos 2001). Why these preferences specifically? In each case, the more general preference structure breaks the link between risk aversion and intertemporal substitution (and generally makes risk aversion state-dependent), and some (such as nongeometric discounting and loss aversion) provide arguments for government intervention; there is also extensive empirical work supporting many of them. A recent contribution to this literature

is Nakajima (2012), who investigates whether the temptation preferences of Gul and Pesendorfer (2001) alter the consequences of default reform.²⁸ We suspect other work will follow.

APPENDIX: COMPUTATIONAL CONSIDERATIONS

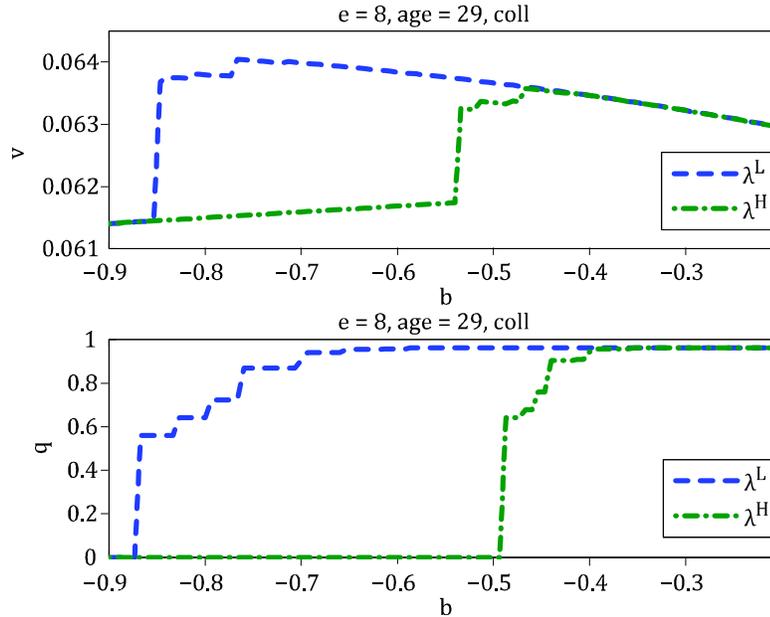
We make some brief points here regarding the computation of the model. The model is burdensome to calibrate, and all programs are implemented using Fortran95 with OpenMP messaging.

In all the models we study, the objective function (the right-hand side of the Bellman equation) is not globally concave, since the discrete nature of the bankruptcy decision introduces convex segments around the point where the default option is exercised (we find that, as in Chatterjee et al. [2007], the default decision encompasses an interval and in our case it extends to $b = -\infty$ as Δ is smaller than even the worst income realization). The nonconcavity poses a problem for local optimization routines, so we approach it using a global strategy. We use linear splines to extend the value function to the real line and a golden section search to find the optimum, with some adjustments to guarantee that we bracket the global solution rather than the local one. It is straightforward to detect whether we have converged to the local maximum at any point in the state space, as the resulting price function will typically have an upward jump.

For the ambiguity aversion case we have a saddlepoint problem to solve. By the saddlepoint theorem we can do the maximization and minimization in any order; the minimization (conditional on b and d) is a linear program that we solve using a standard simplex method conditional on some b (as in Routledge and Zin [2009]). We then nest this minimization within our golden section search, again with adjustments to deal with the presence of the local maximum. For our model, this linear program turns out to be extremely simple to solve—the household puts as much weight as allowed on the worst possible outcome, then as much weight as allowed on the next worst, and so on.

To impose boundedness on the realizations of income, we approximate both e and ν by Markov chains using the approach in Flodén (2008). Having income be bounded above is convenient since it implies

²⁸ Nakajima (2009) finds that increasing borrowing constraints in a model with quasi-geometric discounting is not always welfare-improving, similar to Obiols-Homs (2011).

Figure 17 Optimal Choice of b given q 

that there always exists a cost of default Δ such that bankruptcy is completely eliminated because it becomes infeasible. Quite naturally, bankruptcy is also likely not to occur when Δ is high enough even if filing is feasible for some types; in general, households with high income are not interested in the default option in our model.²⁹

Figure 17 shows a typical objective function for a household in our benchmark case (expected utility with $\sigma = \rho^{-1} = 2$). The objective function has three distinct segments. The first segment is at the far right, where the values for both the low- and high-cost types coincide. In this region, default is suboptimal because borrowing either does not or barely exceeds Δ . The second segment is at the other end, where $q(b) = 0$; although impossible to see in the picture, the low-cost de-

²⁹ Households with high income realizations do not want to pay the stigma cost (which is proportionally higher for them) even if they are currently carrying a large amount of debt (which is very rare due to persistence). Thus, our model does not predict any “strategic” default, which can arise in models that rely on exclusion as a punishment for bankruptcy.

fault experiences slightly more utility in this region since default is less painful. The action is all in the middle segment. For this particular individual, the high-cost type (λ_L) borrows significantly more than the low-cost type; this extra borrowing reflects primarily the pricing function (as seen in the lower panel) and not any particular desire to borrow. High-cost types have more implicit collateral and are less likely to default at any given debt level, so they face lower interest rates. As a result, high-type borrowers today who become low-type borrowers tomorrow are a main source of default in our model—they both have debts and are not particularly averse to disposing of those debts through the legal system. Since type is persistent, low-type borrowers today will not generally make the same choice—the supply side of their credit market will contract.

REFERENCES

- Athreya, Kartik. 2002. “Welfare Implications of the Bankruptcy Reform Act of 1999.” *Journal of Monetary Economics* 49 (November): 1,567–95.
- Athreya, Kartik. 2004. “Shame As It Ever Was: Stigma and Personal Bankruptcy.” Federal Reserve Bank of Richmond *Economic Quarterly* 90 (Spring): 1–19.
- Athreya, Kartik. 2008. “Default, Insurance, and Debt over the Life-Cycle.” *Journal of Monetary Economics* 55 (May): 752–74.
- Athreya, Kartik, Juan M. Sánchez, Xuan Tam, and Eric R. Young. 2013. “Bankruptcy and Delinquency in a Model of Unsecured Debt.” Federal Reserve Bank of St. Louis Working Paper 13-2.
- Athreya, Kartik, Xuan S. Tam, and Eric R. Young. 2009. “Unsecured Credit Markets Are Not Insurance Markets.” *Journal of Monetary Economics* 56 (January): 83–103.
- Athreya, Kartik, Xuan S. Tam, and Eric R. Young. 2012. “A Quantitative Theory of Information and Unsecured Credit.” *American Economic Journal: Macroeconomics* 4 (July): 153–83.
- Ausubel, Lawrence M., and Amanda E. Dawsey. 2008. “Penalty Interest Rates, Universal Default, and the Common Pool Problem of Credit Card Debt.” Unpublished manuscript, University of Maryland and University of Montana.

- Barberis, Nicholas, Ming Huang, and Tano Santos. 2001. "Prospect Theory and Asset Prices." *Quarterly Journal of Economics* 116 (February): 1–53.
- Bewley, Truman F. 2002. "Knightian Decision Theory, Part I." *Decisions in Economics and Finance* 25 (November): 79–110.
- Campanale, Claudio. 2008. "Learning, Ambiguity, and Life-Cycle Portfolio Allocation." Mimeo, Universidad de Alicante.
- Carroll, Christopher D. 1997. "Buffer-Stock Saving and the Life Cycle/Permanent Income Hypothesis." *The Quarterly Journal of Economics* 112 (February): 1–55.
- Chatterjee, Satyajit, P. Dean Corbae, Makoto Nakajima, and José-Víctor Ríos-Rull. 2007. "A Quantitative Theory of Unsecured Consumer Credit with Risk of Default." *Econometrica* 75 (November): 1,525–90.
- Constantinides, George M. 1990. "Habit Formation: A Resolution of the Equity Premium Puzzle." *Journal of Political Economy* 98 (June): 519–43.
- Deaton, Angus. 1992. *Understanding Consumption*. New York: Oxford University Press.
- Dubey, Pradeep, John Geanakoplos, and Martin Shubik. 2005. "Default and Punishment in General Equilibrium." *Econometrica* 73 (January): 1–37.
- Epstein, Larry G., and Martin Schneider. 2003. "Recursive Multiple-Priors." *Journal of Economic Theory* 113 (November): 1–31.
- Epstein, Larry G., and Stanley E. Zin. 1989. "Substitution, Risk Aversion, and the Temporal Behavior of Consumption Growth and Asset Returns I: A Theoretical Framework." *Econometrica* 57 (July): 937–69.
- Evans, David S., and Richard L. Schmalensee. 2005. *Paying with Plastic: The Digital Revolution in Buying and Borrowing*. Cambridge, Mass.: MIT Press.
- Fay, Scott A., Erik Hurst, and Michelle J. White. 1998. "The Bankruptcy Decision: Does Stigma Matter?" University of Michigan Working Paper 98-01 (January).
- Flodén, Martin. 2008. "A Note on the Accuracy of Markov-Chain Approximations to Highly Persistent AR(1) Processes." *Economics Letters* 99 (June): 516–20.

- Friedman, Milton. 1957. *A Theory of the Consumption Function*. Princeton, N. J.: Princeton University Press.
- Furletti, Mark. 2003. "Credit Card Pricing Developments and Their Disclosure." Federal Reserve Bank of Philadelphia Payment Cards Center Discussion Paper 03-02 (January).
- Grochulski, Borys. 2010. "Optimal Personal Bankruptcy Design under Moral Hazard." *Review of Economic Dynamics* 13 (April): 350–78.
- Gross, David B., and Nicholas S. Souleles. 2002. "An Empirical Analysis of Personal Bankruptcy and Delinquency." *Review of Financial Studies* 15 (1): 319–47.
- Gul, Faruk. 1991. "A Theory of Disappointment Aversion." *Econometrica* 59 (May): 667–86.
- Gul, Faruk, and Wolfgang Pesendorfer. 2001. "Temptation and Self-Control." *Econometrica* 69 (November): 1,403–35.
- Guvenen, Fatih. 2007. "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?" *American Economic Review* 97 (June): 687–712.
- Guvenen, Fatih, and Anthony A. Smith, Jr. 2009. "Inferring Labor Income Risk from Economic Choices: An Indirect Inference Approach." Mimeo, University of Minnesota and Yale University.
- Hansen, Lars Peter, and Thomas J. Sargent. 2007. *Robustness*. Princeton, N. J.: Princeton University Press.
- Hryshko, Dmytro. 2008. "RIP to HIP: The Data Reject Heterogeneous Labor Income Profiles." Mimeo, University of Alberta.
- Hubbard, R. Glenn, Jonathan Skinner, and Stephen P. Zeldes. 1994. "The Importance of Precautionary Motives in Explaining Individual and Aggregate Saving." *Carnegie-Rochester Conference Series on Public Policy* 40 (June): 59–126.
- Huggett, Mark, Gustavo Ventura, and Amir Yaron. 2011. "Sources of Lifetime Inequality." *American Economic Review* 101 (December): 2,923–54.
- Jackson, Thomas H. 2001. *The Logic and Limits of Bankruptcy Law*. Hopkins, Minn.: Beard Books.
- Karahan, Fatih, and Serdar Ozkan. 2009. "On the Persistence of Income Shocks over the Life Cycle: Evidence and Implications." Penn Institute for Economic Research PIER Working Paper 09-05 (December).

- Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2009. "Recursive Smooth Ambiguity Preferences." *Journal of Economic Theory* 144 (May): 930–76.
- Krueger, Dirk, and Harald Uhlig. 2006. "Competitive Risk Sharing Contracts with One-Sided Commitment." *Journal of Monetary Economics* 53 (October): 1,661–91.
- Laibson, David A. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112 (May): 443–77.
- Li, Wenli, and Pierre-Daniel Sarte. 2006. "U.S. Consumer Bankruptcy Choice: The Importance of General Equilibrium Effects." *Journal of Monetary Economics* 53 (April): 613–31.
- Livshits, Igor, James MacGee, and Michèle Tertilt. 2011. "Costly Contracts and Consumer Credit." Working Paper 17448. Cambridge, Mass.: National Bureau of Economic Research (September).
- Ljungqvist, Lars, and Thomas J. Sargent. 2004. *Recursive Macroeconomic Theory*. Cambridge, Mass.: MIT Press.
- Livshits, Igor, James MacGee, and Michèle Tertilt. 2007. "Consumer Bankruptcy: A Fresh Start." *American Economic Review* 97 (March): 402–18.
- Livshits, Igor, James MacGee, and Michèle Tertilt. 2008. "Costly Contracts and Consumer Credit." Mimeo, University of Western Ontario and Stanford University.
- Lucas, Robert E., Jr. 1987. *Models of Business Cycles*. Oxford: Basil-Blackwell Ltd.
- Mateos-Planas, Xavier, and Giulio Seccia. 2006. "Welfare Implications of Endogenous Credit Limits with Bankruptcy." *Journal of Economic Dynamics and Control* 30 (November): 2,081–115.
- Miao, Jianjun, and Neng Wang. 2009. "Risk, Uncertainty, and Option Exercise." Manuscript, Boston University and Columbia University.
- Nakajima, Makoto. 2012. "Rising Indebtedness and Temptation: A Welfare Analysis." *Quantitative Economics* 3 (July): 257–88.
- Obiols-Homs, Francesc. 2011. "On Borrowing Limits and Welfare." *Review of Economic Dynamics* 14 (April): 279–94.
- Pemberton, James. 1998. "Income Catastrophes and Precautionary Saving." Mimeo, University of Reading.

Routledge, Bryan R., and Stanley E. Zin. 2009. "Model Uncertainty and Liquidity." *Review of Economic Dynamics* 12 (October): 543–66.

Routledge, Bryan R., and Stanley E. Zin. 2010. "Generalized Disappointment Aversion and Asset Prices." *Journal of Finance* 65 (August): 1,303–32.

Sullivan, Teresa A., Elizabeth Warren, and Jay Lawrence Westbrook. 2000. *The Fragile Middle Class: Americans in Debt*. New Haven, Conn.: Yale University Press.

Tam, Xuan S. 2009. "Long-Term Contracts in Unsecured Credit Markets." Mimeo, University of Virginia.

Regulation and the Composition of CEO Pay

Arantxa Jarque and Brian Gaines

It is well known that the use of stock options for compensating executives in large U.S. companies was widespread during the last 15 years. But were all firms using them with equal intensity? We are interested in the answer to this question because option grants are different from other compensation instruments in the type of incentives they provide, how transparent they are to investors, and the level of insider trading that they allow. In this article, we provide an empirical examination of the trends in the last two decades of the use of different compensation instruments, mainly focusing on restricted stock grants and option grants. We find that there have been important changes, and that they coincide in time with two changes in regulation: the modifications to reporting requirements for option grants introduced by the passage of the Sarbanes-Oxley Act in 2002, and the 2006 adoption of revised accounting standards from the Financial Accounting Standards Board (FASB) included in statement no. 123R (FAS 123R), which mandated the expensing of option grants.

Today, companies pay their top executives through some or all of the following instruments: a salary, a bonus program, stock grants (usually with restrictions on the ability to sell them), grants of options on the stock of the firm, and perks and long-term incentive plans that specify retirement and severance payments, as well as pension plans and deferred benefits. The most accepted explanation for the inclusion of compensation instruments that are contingent on the performance of the firm is the existence of a moral hazard problem: The separation of ownership and control of the firm implies the need to provide incentives

■ The views expressed here do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: arantxa.jarque@rich.frb.org.

to the chief executive officer (CEO) that align his interests with those of the firm owners.

In the presence of moral hazard, the optimal contract prescribes that the pay of the executive should vary with the results of the firm. However, in spite of the need for incentives, limited funds on the part of CEOs or risk aversion considerations imply that exposing the CEO to the same risk as shareholders is typically either an unfeasible or an inefficient arrangement. The optimal contract should balance incentives and insurance. Some part of the compensation should not be subject to risk, like the annual salary, providing some insurance to the CEO against bad performance of the firm over which he does not have control. However, some part of the compensation should be variable and tied to some measure of performance of the firm. The main variable pay instruments can be classified in three categories. First, bonus plans, which make annual pay dependent on yearly accounting results. Second, grants of stock of the firm (often referred to as “restricted stock,” since the executive cannot sell them for some time after they are granted, typically about three or four years); these make pay in the longer term dependent on the results of the firm over a longer time horizon. Third, grants of stock options, which allow the executive to purchase stock of the firm at a pre-established price (the “exercise price”) and also typically are granted with restrictions as to how soon they can be exercised; these also provide incentives for longer-term performance, but they only pay off for the executive if the stock price of the firm is above the exercise price.¹

These different compensation instruments differ in how transparent they make compensation to shareholders or outside investors. For example, bonus schemes that are based on both objective and subjective performance targets may be more difficult for an outside investor to evaluate than a plain restricted stock grant. These instruments also differ on how robust they are to insider trading and other opportunistic behavior; the exercise of stock options or the sales of vested stock can potentially be timed by the CEO to the disclosure of particularly good or bad news on the prospects of the firm, for example, and bonuses may be sensitive to creative accounting practices where some annual results are made to look better by using the degree of freedom present in accounting standards, or by fraudulent misrepresentation of financial results.

¹ When options are granted with the exercise price equal to the stock market price at the date of grants, they are called “at the money;” this is the most popular practice, although some options are occasionally also granted “in the money” (with exercise price below market price) or “out of the money” (with exercise price above market price).

Another important factor is that various compensation instruments are treated differently for taxation purposes and are subject to different disclosure requirements and accounting standards. As an example of heterogeneity in tax treatment, non-qualified option grants, which have been the most popular type of option grant in the last two decades, trigger a tax deduction for the company when they are exercised by an employee; salaries or any other compensation that is not performance-based (like plain restricted stock awards) in excess of one million dollars, instead, do not qualify for a deduction.²

As an example of heterogeneity in disclosure requirements, compensation that is given to executives in the form of perks does not need to be detailed in the compensation disclosure tables of proxy statements if its value is less than \$10,000; when the value exceeds that sum, the disclosure is only in a footnote. Salary, bonus, stock, and option grants are disclosed in the mandatory compensation table instead.³

These differences have historical origins, and are likely subject to political pressures. One cannot ignore, however, the distortion that the tax, disclosure, or accounting treatment may potentially have on the choice of instruments, and through that—as we just argued—on the efficiency of incentives and the transparency of compensation practices to shareholders. Hence, in this article we ask the following questions: Are firms in certain industries or of larger size more likely to use option grants? Are firms that use options more likely to pay higher compensation to their CEOs? Have restricted stock grants replaced option grants after expensing and reporting rules were changed in 2002 with Sarbanes-Oxley, and then again in 2006 with the adoption of Statement of Financial Accounting Standards 123R (SFAS 123R) accounting standards? Has the relative importance of salary and bonuses decreased in recent years in favor of option or stock grants?

Regulatory Changes

In this article we consider two major changes in the U.S. regulation of compensation practices. The first change is the Sarbanes-Oxley Act of 2002, which aimed to improve corporate governance after several earnings management scandals surfaced in the early 2000s. The second change is a revision of accounting rules introduced by the SFAS 123R in 2006, which for the first time mandated a positive expense for options

² This differential tax treatment was introduced in 1993, IRC section 162(m). See Hall and Liebman (2000) for an analysis of the taxation of executive compensation. See Meyers (2012) for a recent explanation of requirements on stock awards that are considered performance-based and qualify for a tax deduction.

³ See Securities and Exchange Commission (2006).

awarded “at the money” (with exercise price equal to the stock market price at the date of the grant).

We use available data on executive compensation from 1993 to 2010 to evaluate the effect of these two regulation changes in the choice of compensation instruments of large public U.S. firms. Note that the Dodd-Frank Act, which was motivated by the financial crises of 2008, was passed in 2010 and it affected financial firms only. It would certainly be interesting to know how the increased scrutiny of incentive schemes that the Act mandates (both for executives and lower level employees) is affecting pay practices at large financial institutions. However, we do not have enough observations in our sample to deal with that regulatory change in this article.

The Sarbanes-Oxley Act, as part of its effort to improve transparency to shareholders, decreased the time window allowed for the disclosure of insider trades to two business days.⁴ Before the Act, firms had until the end of the fiscal year to report any type of insider trading, including the grants of options to employees of the firm. As it became apparent after some investigations, a number of firms were able to exploit the lax reporting requirements to engage in “backdating,” the (illegal) practice of artificially changing the grant date of options to the day with the lowest stock price in the time window allowed for reporting.^{5,6} The benefits of this practice were twofold, and hinged on both the accounting standards and the tax treatment of options. First, it allowed the firm to report higher earnings. At the time, accounting standards under SFAS 123 allowed firms to expense grant options according to their “intrinsic value,” which is zero for options granted at the money. Instead, the intrinsic value of an option in the money (which is what was being effectively granted without the backdating) would have been positive, and hence a compensation expense would have been deducted from the firm’s income, resulting in lower reported earnings. Second, it allowed a larger tax deduction for the firm at the time that the option was exercised. Under Internal Revenue Code (IRC) section 162(m), firms can deduct from their tax liability any compensation costs that originate in incentive pay. In contrast,

⁴ “Ownership Reports and Trading by Officers, Directors and Principal Security Holders,” Release No. 34-46421 (Aug. 27, 2002) [56 FR 56461] at Section II.B.

⁵ Investigations pointing to the existence of backdating became well-known only in 2005. Since Sarbanes-Oxley was passed in 2002, it may be the case that the change in reporting requirements was not directly aimed at preventing backdating. In other words, in trying to improve corporate governance in general, the Act inadvertently limited the possibility of backdating.

⁶ For a discussion of the issues and anecdotal evidence, see the *Wall Street Journal* article “The Perfect Payday” (March 18, 2006). For an academic evaluation of the backdating practice, see Heron and Lie (2007) and references therein.

there is a limit of one million dollars for deducting compensation that is not tied to incentives. Hence (provided the employee was already receiving one million dollars in non-incentive compensation), the tax deduction would have been lower for an option in the money, since the difference between the stock price at the time of grant and the exercise price would not have been considered incentive pay.⁷ Backdating options without proper disclosure, then, implied both misreporting to investors the amount of incentive pay given to employees, and engaging in fraudulent accounting to save on taxes.⁸ The Act, by decreasing the time window allowed to report the granting of options to two business days after the trade takes place, constrained the firms' ability to misreport the actual date of the grant, and hence made options a less attractive compensation instrument for firms that were backdating, or for those that were considering the possibility of doing it at some point.

The second piece of regulation that we consider is SFAS 123R, a revision to accounting standards SFAS 123, which was adopted by the Security and Exchange Commission (SEC) in 2006. The main change introduced by the revision was a homogenized method of valuation of options to "fair value" calculations, such as Black and Scholes. Previously, the "intrinsic valuation" method was allowed, which attributed a zero value to options granted at the money. Because option grants are accounted for as expenses in the income statement of the firm, this change in valuation method effectively eliminated the possibility of not charging any expense of compensation for options granted at the money.⁹ The general view on this piece of regulation is that, after its adoption, companies were no longer able to "hide" the dent of option grants on their accounting profit. This view is supported by the numerous complaints by large U.S. corporations when the measure was first proposed, arguing that lower earnings per share would hurt, for example, their ability to borrow and grow, hindering innovation and job creation. However, under the disclosure requirements in SFAS 123 before 2006, firms were already required to report (in a footnote in

⁷ This tax treatment applies to "non-qualified" option grants, which are the most common in executive compensation packages during the time period that we study. Firms are also allowed to grant "qualified" options, or "incentive stock options," to their employees, which are limited to a maximum value of \$100,000, and hence are not usually granted to executives. See Bickley (2012) for details on the taxation of employee stock grants.

⁸ Because backdating implies a violation of the SEC's disclosure rules, a violation of accounting rules, and a violation of tax laws, the SEC has sued a number of companies suspected to have engaged in this practice. See, for example, the testimony of Christopher Cox as Chairman of the SEC on September 6, 2006 (available at www.sec.gov), where he states that charges related to this matter were made as early as 2003.

⁹ Accounting standards and a detailed description of accepted "intrinsic value" calculations can be found in APB 25, from the FASB.

their proxy statement) enough information about their grants of employee options for any interested shareholder to compute the cost of these (using, for example, the Black and Scholes valuation). Hence, the economic impact of this change in regulation remains unclear, and it somehow hinges on the assumption that the information disclosed in the footnotes was somewhat less available to the public than after it was officially included as an expense in the income statement.¹⁰

It is important to note that the first proposal for the expensing of options was drafted as far back as 1993. Due to strong opposition from the corporate sector and other political forces, the final recommendations in FASB 123 issued in 1995 merely recommended the expense, but did not mandate it. The public debate about the pros and cons of expensing, which involved senators, congressmen, the SEC, and lobbyists from the corporate sector, was ongoing for more than a decade. Finally, in 2006, the SEC endorsed the revision SFAS 123R, which mandates expensing. It is worth noting that many large public firms started the expensing on a volunteer basis as early as 2002; some commentators have noted that this voluntary adherence, and the final political push that led to the mandatory requirement, were rooted in the Enron and other accounting scandals in 2002.¹¹ Hence, the effect of SFAS 123R is potentially present as early as the passage of the Sarbanes-Oxley Act, preventing the separate identification in the data of the effect of the two regulations. Nevertheless, in our analysis we find significant changes in the patterns of usage of stock and option grants coinciding with both changes in regulation.

Outline

In this article, we start by describing the data. We provide a motivating example that illustrates the primary difficulties in using the currently available data on CEO compensation to answer the main questions of interest to us. In Section 2 we briefly review some previous attempts in the academic literature to shed light on similar issues, and the differences with the approach we take here. We proceed with our main analysis in two parts: First, in Section 3, we document facts related to the extensive margin (i.e., when and by which firms are stock and option grants used), and second, in Section 4, we discuss facts related to the intensive margin (i.e., what is the relative importance of

¹⁰ See Guay, Kothari, and Sloan (2003) and Guay, Larcker, and Core (2005) for a clear exposition of these issues.

¹¹ See, for example, Brown and Lee (2011), or “Reporting Employee Stock Option Expenses: Is the Debate Over?” by Paulette A. Ratliff (www.nysscpa.org/cpajournal/2005/1105/essentials/p38.htm).

stock, options, and other forms of pay for the firms that use them). We document the change in compensation practices across the different regulatory regimes. We also explore the correlation of other firm characteristics, like size, industry classification, and executive characteristics, like age, tenure, and gender, with the choice and importance of the different available compensation instruments. We also examine the relationship of usage of stock and option grants with the level of pay. We conclude in Section 5.

1. SAMPLE DESCRIPTION AND DATA INTERPRETATION ISSUES

Thanks to disclosure requirements by the SEC, we have data available on pay to the top executives of public U.S. companies starting in 1992. This data is collected systematically by Compustat into a database called Execucomp. Many academic studies have used Execucomp and other available data to document the regularities in the level of pay and its sensitivity to firm performance, across time and also for firm characteristics like size and industry.¹²

The Execucomp data set is published by Compustat four times per year. Each release includes the new information for companies that filed their proxy statements with the SEC in that period (companies can decide when their fiscal years start, and hence there is variation in when annual proxies are filed). Execucomp tries to collect data on the firms that are listed in the S&P 1500 index, which roughly corresponds to the 1,500 largest U.S. firms by market capitalization. This article uses the information on CEO pay of the October 2011 edition of the Execucomp data, which covers 1992 to 2010, for a total of 19 complete fiscal years. We exclude observations in year 1992, since there are very few and they may not be representative. We exclude CEOs who own a large fraction of the firm's stock, since presumably pay is not set to provide incentives for these owner-CEOs. Next, we elaborate on the issues in choosing the threshold value for this selection.

¹² For the analysis of sensitivity of pay to performance, see the seminal contributions of Jensen and Murphy (1990), Rosen (1992), and Hall and Liebman (1998). For the relationship of pay level and sensitivity to firm size in the cross section, see Schaefer (1998) and Baker and Hall (2004). A more recent study of the variation of the level of pay over time and its potential relationship to firm size is Gabaix and Landier (2008). Frydman and Saks (2010) provides a comprehensive historical overview of both level and sensitivity of pay facts using a small sample of firms over an unusually long period, from 1936 to 2005.

Ownership, Incentives, and Steve Jobs

In this article we are interested in the decisions of firms to use or not use a given compensation instrument. One potential concern with this analysis is that the choice of a firm of not using stock or options may be explained by the fact that its CEO is a founder of the company, or that he or she is very vested in the firm already. An example of this would be Steve Jobs, who is, in our sample from 1997 to 2010, listed as the CEO of Apple, Inc.

Jobs's history of compensation over 12 years is easily summarized. In 1997, the year he took the CEO position, Jobs received, as a director of the company, 30,000 stock options with an exercise price of \$23, to be vested proportionally over a three-year period.¹³ The salary of Jobs was \$1 for all the years we observe him in the sample. He received sporadic bonus and "other compensation" payments, stock in 2003, and options in 1997, 2000, and 2002.^{14,15} In the company's own words:

"In 2010, Mr. Jobs's compensation consisted of a \$1 annual salary. Mr. Jobs owns approximately 5.5 million shares of the Company's common stock. Since rejoining the Company in 1997, Mr. Jobs has not sold any of his shares of the Company's stock. Mr. Jobs holds no unvested equity awards. The Company recognizes that Mr. Jobs's level of stock ownership significantly aligns his interests with shareholders' interests. From time to time, the Compensation Committee may consider additional compensation arrangements for Mr. Jobs given his continuing contributions and leadership."¹⁶

Jobs's ownership shares are only reported in Execucomp, combined with option holdings, for four of the years, and they never exceed 1.35 percent of the total shares outstanding, which is about the 67th percentile ownership in the original sample of CEOs. Even if one may be tempted to think that Jobs was not an "agent" for the shareholders of Apple due to the great value of the stock that he owned (especially after the 2003 grant, valued at more than \$80 billion at the grant date), a closer look at the evolution of his ownership shows that he went from owning one share in 1997 to owning 5.5 million shares mainly as a result of his compensation packages. Moreover, Jobs had a considerable amount of wealth from his investment in Pixar, and one could argue

¹³ See Apple's Definitive Proxy statement on March 16, 1998.

¹⁴ According to Execucomp, Jobs received a bonus payment in 2001 and 2002, and two big sums as "other compensation" in 2001 and 2002.

¹⁵ Given the compensation pattern of Jobs, it is interesting to note that Apple stated in April 2003 its decision to voluntarily expense option grants to its employees according to FASB recommendations

¹⁶ See Apple's Definitive Proxy statement on January 7, 2011.

that stakes had to be necessarily high in order to provide him with adequate incentives. Finally, when Jobs's illness was made public, markets reacted, providing proof that the value that Jobs was bringing to the company was real.

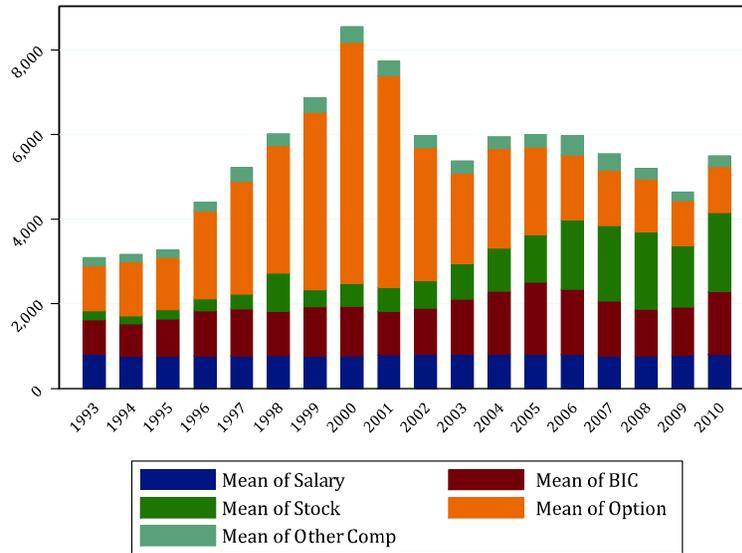
The case of Steve Jobs is easy to check and understand, but in general the data on ownership in Execucomp shows some inconsistencies, and there are many missing values, since ownership is recorded only if it is over 1 percent. Hence, a back-of-the-envelope calculation of the value of the stock held by the CEO is not always available, and even if it were it would be hard to determine when the CEO is subject to a moral hazard problem based on those numbers.¹⁷ From our analysis of the Jobs case, however, we conclude that we cannot rule out that ownership, in our sample, is a result of dynamic incentives provided by the firm. Hence, we are most comfortable adopting a conservative criteria of only dropping CEOs from our sample if their ownership reaches 50 percent in any of the years that they worked for a given firm, as opposed to more restrictive selection criteria in the literature.¹⁸ Our final sample includes information on 6,146 different executives, and 3,248 firms, which amounts to 6,416 unique executive-firm pairs. In the year 1993, we observe 1,147 firms, and every year after that the number is at least 1,500, with a maximum of 2,010 firms in the year 2007.

Compensation Measures

Our focus in this article is on the choice of compensation instruments by the firm, and we use the information readily available in Execucomp about each of the components of total compensation: salary, bonus and incentive compensation, stock and option grants, and "other compensation" such as pension plans, life insurance premiums, or perks. Note that to avoid discontinuity issues with the "bonus" and "incentive compensation" variables due to changes in reporting requirements in 2006, we sum these two to construct a single series, which we refer to as BIC throughout the article. Also, in spite of the different accounting standards during the sample period, Execucomp contains the Black and Scholes valuation of option grants for the whole period: Companies that used alternative valuations prior to SFAS 123R were required

¹⁷ In spite of the sparse availability, we did construct a value of shares owned for the CEOs for which we had data: The average value for those that we classified as non-owners was \$1,928,000, compared to a mean total compensation of \$2,507,000.

¹⁸ Clementi and Cooley (2010) used the more restrictive threshold of 1 percent ownership. We conducted a robustness check of our main analysis by dropping all CEOs who owned 3 percent or more shares on average over their tenure and results did not change qualitatively.

Figure 1 Average CEO Compensation

Notes: “BIC” stands for bonus and incentive compensation.

to provide the parameters necessary to calculate the Black and Scholes value. Whenever we need a measure of total compensation, we use the sum of these components (the variable $TDC1$ in Execucomp).¹⁹

Figure 1 presents the evolution of the mean total compensation in our sample over time, and its components. All the amounts here and in the rest of the article are normalized to thousands of 2010 dollars using the consumer price index. The year 2000 stands out as the peak in our measure of compensation, with an average of \$8,553,690 and a median of \$3,107,580. The year 2009 seems to be the last one of a decreasing compensation trend coinciding with the financial crisis: Mean compensation for this year was \$4,637,950, while the median was \$3,030,940.

The most salient fact about the composition of pay in Figure 1 is that the variation of pay with the business cycle is implemented through

¹⁹ For recent studies that use this same measure of total compensation, see Gabaix and Landier (2008); Frydman and Saks (2010); and Cheng, Hong, and Scheinkman (2012).

the grants of stock and options, rather than through salary, bonus, or other compensation. For example, the graph shows that the decline in average total compensation between 2000 and 2003 is driven by a decline in the value of stock options. However, after 2002, the category BIC becomes somewhat cyclical as well. It is important to keep in mind that, of these components of total compensation, only bonus and incentive payments are mechanically related to the results of the firm. For example, the amount used to construct Figure 1 is the expected value of the grant at the time when it was awarded. Hence, the fact that compensation was the highest in the year 2000 is not due to a high value of past grants driven by a stock market boom, but rather to a conscious decision by the firms to increase the value of compensation for their CEOs.²⁰

2. PREVIOUS LITERATURE

Before we start our analysis of the data, we review the relevant literature and explain our contribution.

In an influential chapter of the *Handbook of Labor Economics*, Murphy (1999) provides some suggestive evidence for a sample of firms between 1992 and 1996 that the importance of the different compensation instruments in pay packages (salary, bonus, stock, and option grants) varies across firms according to their size and the industry to which they belong.²¹

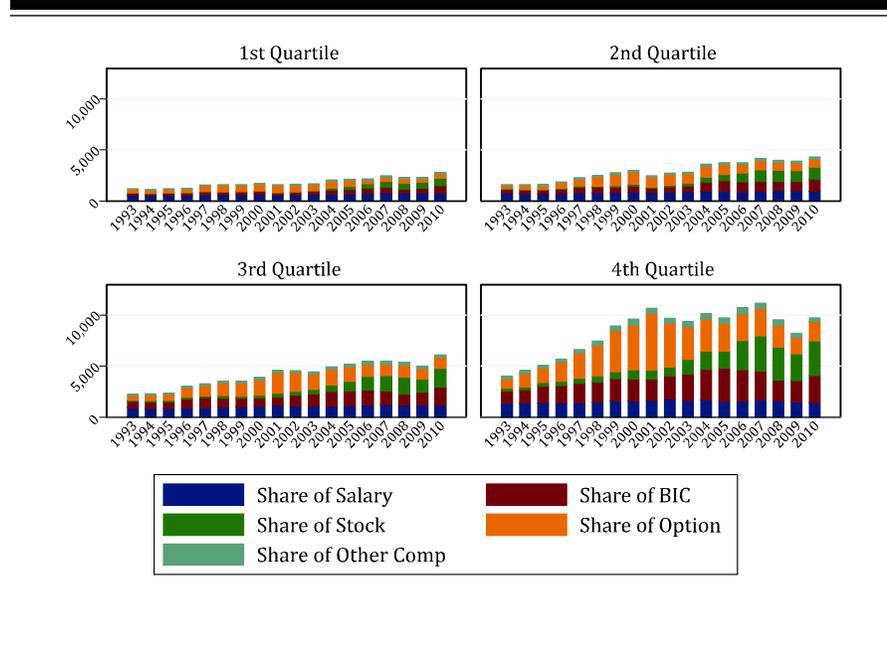
In his graphical analysis for the effect of size, Murphy compares S&P 500 industrials, mid-cap industrials, and small-cap industrials. We replicate and extend his analysis (including data up to 2010) in Figure 2, where we classify firms in our sample, year by year, into four quantiles according to their volume of sales.

The most striking fact that emerges from Figure 2 is that firms with larger sales figures have higher levels of pay. The variation in the relative importance of the different compensation instruments is difficult to evaluate in a systematic manner, although it is clear that larger firms have a larger portion of their pay given in stock and options. Also, the increase in the relative importance of options in the late 1990s that has been frequently commented on both in the academic and the popular press seems to have been disproportionately concentrated in the quantile of the largest firms.

²⁰ Note that firms amortize the expense from these grants over their vesting period, and hence compensation expenses are actually smoothed out over time by the firms.

²¹ See Figures 2 and 3 in Murphy (1999).

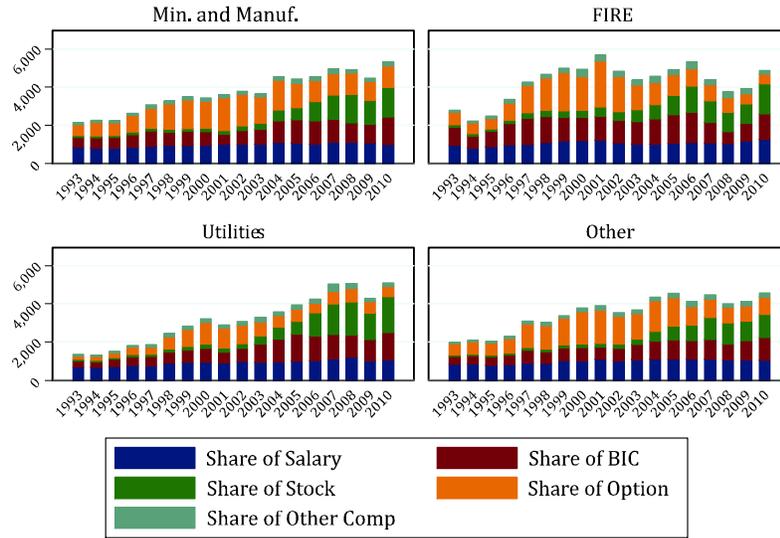
Figure 2 Average Compensation and its Components, by Quartiles of Sales Volume



Murphy's graphical analysis for the regularities across different industries is limited to S&P 500 firms, and it uses a classification of SIC industries in four groups: mining and manufacturing firms, financial services firms, utilities firms, and other industries. In Figure 3 we replicate this evidence, again extending the sample to include data from 1992 to 2010, as well as all firms in the S&P 1500. Figure 3 does not allow us to draw any clear conclusions. If anything, it seems to suggest that firms in utilities seem to rely more on restricted stock than option grants. In a related study, and for the period 1992–2001, Murphy (2003) classifies firms into “new economy” versus “old economy” according to the industry sector they belong to, and he finds that new economy firms (those competing in the computer, software, internet, telecommunications, or networking fields) use stock-based compensation (both restricted stock and options) more often and to a larger extent.

One important shortcoming of the simple facts reported in Murphy (1999, 2003) is that they do not inform us about the relationship between combinations of individual characteristics (industry and size together, for example) and usage of instruments. Also, the information about the variation in the cross-section is lost in the graphs. Our

Figure 3 Average Compensation and its Components, by Industry Group



contribution in this article consists of analyzing the data according to firm characteristics by running some simple regressions. Our analysis is still partial, since we are not exploiting the panel component in the data, but we are able to provide a more accurate description of the facts by controlling for several individual firm characteristics. We also split our analysis into the extensive margin (which compensation instruments are used) and the intensive margin (given a set of instruments that is being used, what is their individual share of total compensation).

In addition to answering the questions posed above about the trends in the usage of different compensation instruments, we explore whether factors other than firm size or industry classification may be associated with the usage of certain instruments. For example, given the limits on tax deductions imposed on salaries, firms that—for reasons other than their industry and size—choose to compensate their CEO with a larger sum of money may benefit more from issuing non-qualified option grants or restricted stock grants. As another example, executives who have longer tenures may need fewer restricted stock grants if they already hold a large number of shares of the firm from previous grants.

This last point, which is an interesting one, refers to the dynamic nature of incentives for CEOs. There have been important efforts in the literature of CEO compensation that track the evolution of the portfolio of grants of the executives, so that at each point in time we have a better understanding of how the executive's wealth would vary with a particular realization of the firm's results. Some important examples are Hall and Liebman (1998), Core and Guay (1999), and, more recently, Clementi and Cooley (2010). These measures of incentives are a way of controlling for outstanding past issues of stock and option grants. The focus of these studies, however, has not generally been the trends in the usage of compensation instruments. An important exception is Core and Guay (1999), who study this in detail for a shorter time period than the one we are analyzing here. They construct a model of the optimal level of stock holdings of the CEO, for incentives purposes. They find evidence that new grants (combining stock and options) are aimed at maintaining that level of incentives, as old grants expire or go out of the money. However, as far as we know, none of the studies that construct the portfolio measures address the potential effects of regulation on the trends in the usage of individual compensation instruments.

One important shortcoming of our data set is that it starts in 1993. Regulations on tax deductibility of CEO pay had just changed at the time (see IRC section 162(m)). Data on compensation practices prior to 1993 would be useful to the understanding of the distortions that 162(m), and other tax advantages introduced earlier, may have induced on pay practices.²² Detailed compensation data for a broad representative set of firms going further back in time is not available; however, Frydman and Saks (2010) provide a historical analysis of a limited set of firms.²³

As part of their analysis, Frydman and Saks (2010) plot the median of the partial sums of salary and bonus payments, successively adding the value of stock and option grants. They find that, even though the usage of options picks up considerably after taxation advantages are introduced in 1950, their relative importance in total compensation, as well as that of stock grants, does not become significant until the 1980s.²⁴ Since their sample of firms is necessarily limited (because of the long historical scope), and for comparison purposes, we replicate their graphical analysis for our sample in Figure 4.²⁵ For the

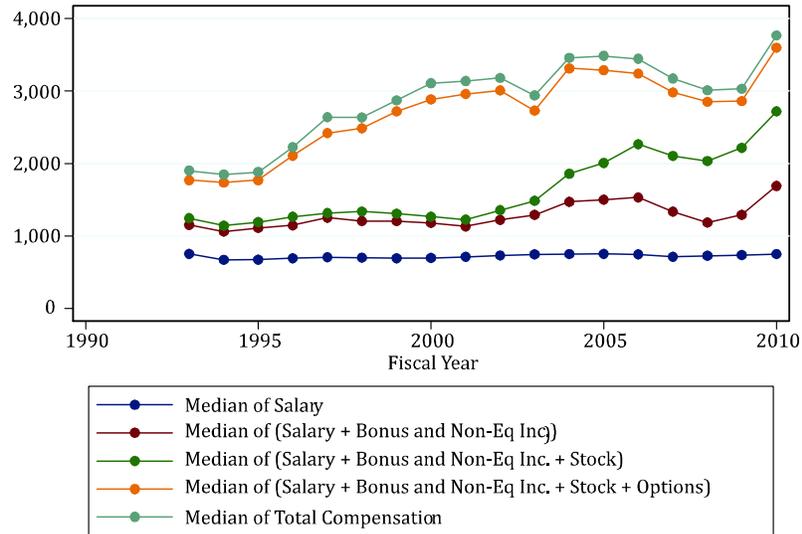
²² See Jarque (2008) for a review.

²³ See also Lewellen (1968).

²⁴ See Frydman and Saks (2010, Figure 2, p. 2,108).

²⁵ See Frydman and Saks (2010, Figure 1, p. 2,107).

Figure 4 Median Total Compensation and its Main Components

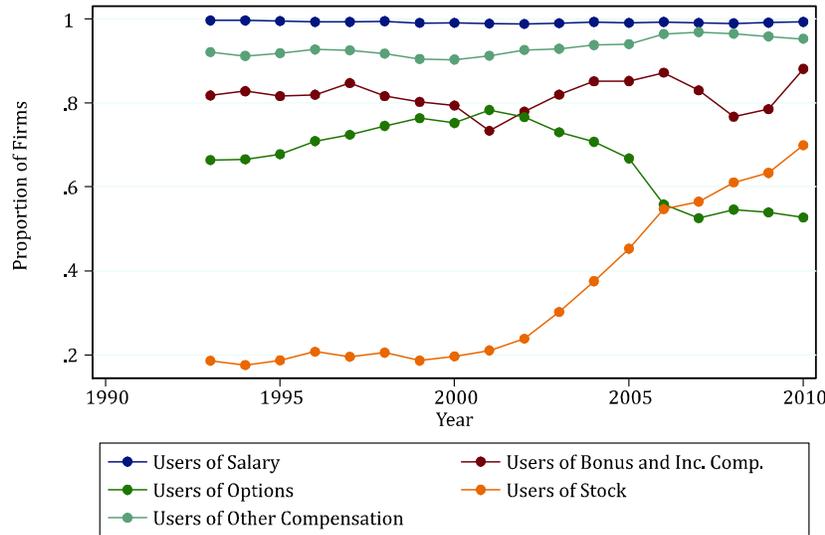


overlapping period from 1992 to 2005, and again with the caveat of not controlling for individual characteristics in this simple graphical analysis, we confirm their findings: Option grants have been an increasingly important component of the median pay of CEOs for the whole period, while the importance of stock awards started to pick up around 2002. With respect to the mean compensation that we plotted in Figure 1, we see that the importance of options was not as marked for median pay in the 1999–2001 period as it was for mean pay. Other than that, the main patterns seem to align between the two figures.

3. THE COMPOSITION OF PAY PACKAGES: THE EXTENSIVE MARGIN

We start this section by documenting the usage of the different compensation instruments over time. Then we proceed to analyze more formally which firm characteristics may be relevant for the choice of instruments of compensation. We find that variables like size and industry classification have some explanatory power over whether firms decide to include options or stock in their compensation packages.

Figure 5 Evolution of the Percentage of Firms that Use Each Instrument



Changes in regulation during the period we are studying exhibit the highest correlation with changes in usage patterns.

The Use of Different Compensation Instruments: A First Look

For all the firms in our sample, we check year by year which ones use each instrument (for example, a firm “uses” stock if it reports a positive stock grant to their CEO, regardless of the amount of the grant). This is plotted in Figure 5.

As is apparent from the graph, the use of both salary and other compensation is fairly universal and fairly constant over time (with a slight trend up for other compensation in the last five years). The use of bonus and incentive compensation is volatile around 85 percent, with no obvious trends. But the most striking feature in Figure 5 is the run-up in the use of restricted stock grants starting around 2003, which coincides with an important decrease in the use of option grants.

Given the strong variation over time in the usage of stock and options, it is worth thinking about the factors that could potentially be determining the decision of a firm to include either type of grant

in the compensation package to its CEO. Here we point to three main factors: (1) differences in tax advantages and accounting standards, (2) differences in sensitivity to firm performance, and (3) fixed costs of adoption of each instrument.

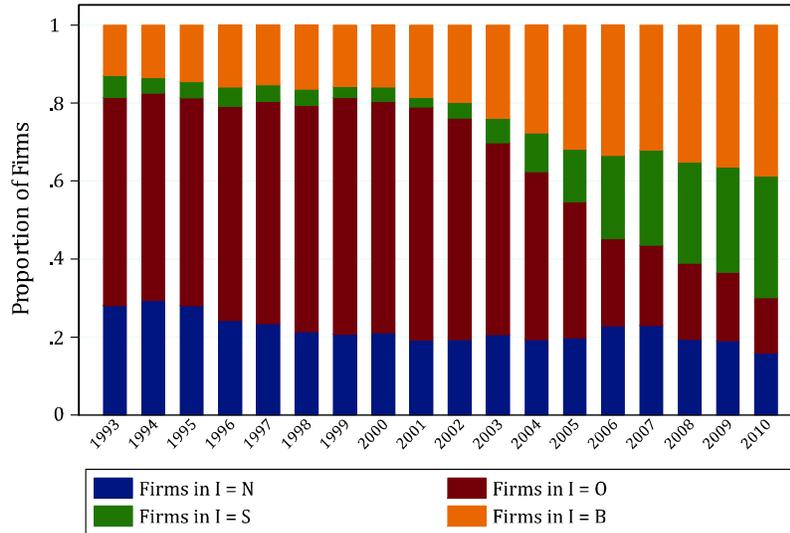
First we turn to tax and expensing differences. As we discuss at length in the introduction, both the passage of the Sarbanes-Oxley Act in 2002, and, especially, the change in expensing requirements and valuation of options in SFAS 123R approved in 2006 (and expected and voluntarily adopted by many firms as early as 2002), seem to have decreased the relative attractiveness of option grants over stock grants. We can summarize the comparison between the two instruments as follows. Restricted stock grants do not qualify for a tax deduction, they have to be accounted for as compensation expenses, and, before 2002, they had to be reported as insider transactions within 10 days of the grant. Options were more advantageous than stock before 2002 because they only had to be reported as insider transactions by the end of the fiscal year of the company; after Sarbanes-Oxley, both types of grants have to be reported within two business days of the transaction. Options were more advantageous than stock before 2006 because (i) they could be deducted for tax purposes, and (ii) they did not need to be expensed; after 2006, advantage (i) is still present, but (ii) is no longer there.

Second, stock and options may implement different incentives for the CEO. That is, in principle, without any accounting or tax differential treatment, stock and options could be substitutes in a compensation package: One could transfer a given amount of resources to the CEO either with a stock grant or with an option grant of equal expected value. However, the value of each of these two grants could change differently with changes in the value of the firm, i.e., the sensitivity of the compensation may be different depending on whether it includes only options or only stock (or both). Hence, idiosyncratic characteristics of the firm, like industry, size, or financial health, may determine the optimal sensitivity of pay to performance, and hence instrument choice.

Third, there may be a fixed cost of including an extra instrument in a compensation package (perhaps related to communication of new or more complex compensation practices to shareholders and creditors); this would imply that larger firms decide to include a different set of compensation instruments than their smaller counterparts.

To shed some light on these and other potential hypotheses, we will formally analyze the correlation of different firm characteristics on the choice of compensation instruments. We start our analysis of the data by classifying firms into four mutually exclusive groups, \mathcal{I} , according

Figure 6 Evolution Over Time of the Percentage of Firms in Each I Group



to which set of compensation instruments they use:

$$\mathcal{I} = \{S, O, B, N\},$$

with typical element I . That is,

- a firm with $I = S$ includes restricted stock grants (but no options) in its compensation package to the CEO,
- a firm with $I = O$ includes options (but no stock),
- a firm with $I = B$ includes both restricted stock and options, and
- a firm with $I = N$ includes none of the two.

Table 1 Descriptive Statistics of the Sample

	<i>S</i>		<i>O</i>		<i>B</i>		<i>N</i>		asset val	age	tenure	female
	Lev.	Diff.	Lev.	Diff.	Lev.	Diff.	Lev.	Diff.	(mil. \$)	(y)	(y)	(%)
Overall	.12		.43		.24		.22		14,924	55	7	.02
Per. I	.04	<i>base</i>	.57	<i>base</i>	.16	<i>base</i>	.24	<i>base</i>	11,926	56	7	.01
Per. II	.09	.05	.45	-.12	.26	.10	.20	-.04	16,893	56	7	.02
Per. III	.26	.17	.19	-.26	.35	.09	.20	.00	18,155	55	7	.03
Other	.11	<i>base</i>	.42	<i>base</i>	.21	<i>base</i>	.26	<i>base</i>	5,545	54	7	.03
Min/Man	.10	-.01	.47	.05	.25	.04	.19	-.07	6,030	56	7	.02
FIRE	.17	.06	.36	-.06	.27	.06	.20	-.06	61,872	56	8	.01
Utilities	.17	.06	.36	-.06	.23	.02	.25	-.01	13,833	56	6	.01
size Q ₁	.11	<i>base</i>	.48	<i>base</i>	.20	<i>base</i>	.21	<i>base</i>	324	54	8	.03
size Q ₂	.12	.01	.45	-.03	.27	.07	.17	-.04	1,150	55	8	.02
size Q ₃	.12	.01	.41	-.07	.32	.12	.15	-.06	3,675	56	7	.02
size Q ₄	.11	.00	.40	-.08	.37	.17	.12	-.09	54,604	57	7	.01
age Q ₁	.12	<i>base</i>	.45	<i>base</i>	.23	<i>base</i>	.21	<i>base</i>	8,460	47	5	.03
age Q ₂	.11	-.01	.45	.00	.26	.03	.17	-.04	17,100	54	6	.02
age Q ₃	.12	.01	.42	-.03	.28	.05	.18	-.03	17,769	58	7	.01
age Q ₄	.12	.01	.39	-.06	.20	-.03	.28	.07	16,583	65	12	.00
tenure Q ₁	.11	<i>base</i>	.41	<i>base</i>	.28	<i>base</i>	.19	<i>base</i>	18,103	53	1	.03
tenure Q ₂	.12	.01	.42	.01	.27	-.01	.18	-.01	17,516	54	4	.02
tenure Q ₃	.13	.02	.43	.02	.24	-.04	.20	.01	14,912	55	7	.02
tenure Q ₄	.12	.01	.44	.03	.18	-.08	.27	.08	11,481	59	17	.01
Male	.12	<i>base</i>	.43	<i>base</i>	.24	<i>base</i>	.22	<i>base</i>	15,095	56	7	0
Female	.18	.06	.34	-.09	.27	.03	.20	-.02	5,594	52	5	1

Figure 6 presents the evolution of the proportion of firms in each of these four groups over our sample period. The evidence is consistent with the changes in regulation prompting firms to switch from using options only (O) to using stock only (S); but it is also apparent that a higher portion of firms use both instruments (B), suggesting that some firms may have chosen to add stock to the use of options, rather than completely substituting options with stock. Next, we formally evaluate the role of changes in regulation in these variations in usage patterns, after also considering other potential determinant factors for these patterns, such as the size of the firm and the industry to which it belongs.

The Determinants of the Composition of Pay Packages

Table 1 presents the breakup of firms in the compensation groups in \mathcal{I} according to the regulatory periods, the industry group, and several firm and CEO characteristics available in Execucomp. It also includes statistics that describe the cross relations between these variables. We have established the existence of three different subperiods in our sample determined by important changes in regulation. Table 1 reports, in its first three rows, the fraction of firms that choose each instrument in the sample, and the changes in these fractions after the two changes in regulation. We denote as period I observations those from 1993 to 2001, as period II those from 2002 to 2005, and as period III those from 2006 to 2010. We see in the table that the fraction of firms in S increases in both subperiods, but especially in the later one. The fraction of firms in O options decreases, again more sharply in the later subperiod. The fraction of firms in N remains fairly constant over time around its overall mean of 22 percent. As we explained in the previous subsection, both regulations had the effect of decreasing the relative attractiveness of options over stock grants. Given this, it is useful to trace the changes in the fraction of firms that use options at all, i.e., $O \cup B$. In period I, we see that 73 percent of firms had options in their compensation packages. In period II this fraction remains fairly constant, at 71 percent: The decrease in O is almost exactly offset by the increase in B . That is, in the second subperiod firms were more likely to use options together with stock, rather than alone, but still as likely as before to use options at all. However, in the last subperiod the fraction drastically decreases to 54 percent: Although the fraction of firms in B increases, the decrease in O is three times as large. This is consistent with the annual evidence presented in Figures 5 and 6, which

show that the main adjustment in the usage of options was gradual and took place mainly over the course of period II.

For industry classification, we use the simple four groups of firms proposed by Murphy (1999). Firms are classified into: 1) mining and manufacturing, 2) finance and real estate (FIRE), 3) utilities, and 4) a mixed group containing any other firm. Table 1 reports that the choice of S is relatively more likely in FIRE and utilities, while that of O is more likely in mining and manufacturing and other. The choice of B is relatively more likely in FIRE, and less in other. Finally, the proportion of firms choosing N is much lower in mining and manufacturing and FIRE.

Next we report the breakout into compensation groups according to size. The literature has established that size is an important factor in the determination of pay levels. We use total assets as a measure of size.²⁶ Year by year, we classify the firms in our sample according to which of the four quantiles of the distribution of asset value they belong. Table 1 reports the fraction of firms that choose each instrument in the sample, and the differences from the fractions for the control group, which is the quantile of smallest firms. The patterns of usage of S seem to be independent of size, while O and N are relatively more popular in smaller firms. On the other hand, using stock and options together (B) is more frequent in larger firms.

Other potentially important characteristics are the tenure, age, and gender of the CEO. We briefly discuss each of these in turn.

Younger executives may have different career concerns than older ones, less experience, or different attitudes toward risk. More tenured executives may be more vested in the firm by means of historical grants, or firm-specific human capital. In Table 1 we see that the choice of S seems to be fairly independent of both age and tenure. The choice of O , instead, is more frequent for younger executives, while, interestingly, given the natural correlation of these two variables, it is less frequent for shorter tenured ones. The frequencies of choice of B are hump-shaped with respect to age, and decreasing for tenure. Firms seem more likely to use none of the instruments more frequently for long-tenured executives, and less frequently for middle-aged ones.

Some have argued that women are more risk averse than men (see Schubert et al. [1999] for a discussion of the evidence); this could influence the choice of compensation instrument. We only have 548

²⁶ For recent estimates, see Gabaix and Landier (2008, Table I, p. 66). Other size measures used in the literature are the number of employees and sales value. We confirm that in our data set the size measure with the highest R^2 for the level of pay is asset value. Details are available upon request.

firm-executive-year observations that correspond to a female CEO, versus 30,032 for males. Despite this, we report the average use of instruments by gender in Table 1 to point to one apparently significant difference: Female CEOs are about 10 percent less likely than male CEOs to receive options exclusively.

We now proceed to validate these raw statistics by performing a formal check on the effect of firm characteristics on the choice of instrument. We model the value to a given firm i of choosing a set of instruments I at time t as

$$V(I)_{it} = \alpha + \sum_{k=1}^3 \beta_k \text{industry}_i + \sum_{k=1}^2 \beta_{3+k} \text{period}_t + \beta_6 \text{tenure}_{it} \\ + \beta_7 \text{age}_{it} + \beta_8 \ln(\text{assets})_{it} + \beta_9 \text{female}_{it} + \varepsilon_{it}. \quad (1)$$

In words, the value $V(I)_{it}$ is assumed to depend linearly on a constant, α , dummy variables for the three distinct regulatory periods in the sample, an indicator variable for the industry group to which that firm i belongs, and the characteristics of firm i in year t that we selected based on our sample analysis. We do not observe directly the value $V(I)_{it}$, but rather the discrete choice of firms for $I \in \{S, O, B, N\}$. Hence, our statistical model is

$$\Pr(V(I)_{it} > V(I')_{it}) \quad \forall I' \neq I,$$

where the probability of observing the choice of a given I depends on whether $V(I)_{it}$, the value derived by a firm i from using instrument I at time t , is higher than the value of the other instruments. We assume the noise term ε_t has a type I extreme value distribution, so our discrete choice regression is a multinomial logit.

One concern with the interpretation of the results of the regression is the potential for colinearity. Table 1 reports, starting in the column labeled "Period," the averages of each variable in the subgroups defined in the different rows of the table. When analyzing those, the most salient fact is the uneven average size across periods (average size is increasing) and across industry groups (FIRE contains firms that are, on average, 10 times the size of firms in mining and manufacturing). However, when we plot the actual size distribution across periods it is not significantly different, thanks to the high variation in the size of firms within periods that we get by using the cross section. The difference in the distribution of size across industry groups is more

Table 2 Regression Results

Instrument Regression spec.	S			O			B			N		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
<i>Mean</i> (Pr(I X))	0.122*** (0.002)	0.123*** (0.002)	0.122*** (0.002)	0.427*** (0.003)	0.427*** (0.003)	0.427*** (0.003)	0.248*** (0.003)	0.248*** (0.002)	0.248*** (0.003)	0.203*** (0.002)	0.202*** (0.002)	0.203*** (0.002)
<i>sic4</i> : <i>ALE</i> ("Other")	0.110*** (0.004)	0.108*** (0.003)	0.110*** (0.004)	0.416*** (0.006)	0.408*** (0.005)	0.411*** (0.006)	0.245*** (0.005)	0.222*** (0.005)	0.233*** (0.005)	0.229*** (0.005)	0.263*** (0.004)	0.245*** (0.005)
<i>AME</i> (Min/Man)	-0.006 (0.004)	-0.006 (0.004)	-0.006 (0.004)	0.037*** (0.007)	0.032*** (0.007)	0.041*** (0.007)	0.036*** (0.006)	0.038*** (0.006)	0.040*** (0.006)	-0.067*** (0.006)	-0.064*** (0.005)	-0.075*** (0.006)
<i>AME</i> (FIRE)	0.038*** (0.007)	0.049*** (0.007)	0.038*** (0.007)	-0.015 (0.010)	0.034*** (0.010)	0.006 (0.011)	-0.049*** (0.008)	0.034*** (0.009)	-0.013 (0.009)	0.026** (0.010)	-0.117*** (0.007)	-0.031*** (0.009)
<i>AME</i> (Utilities)	0.080*** (0.008)	0.089*** (0.008)	0.082*** (0.008)	-0.076*** (0.010)	-0.018 (0.010)	-0.053*** (0.010)	-0.043*** (0.009)	0.025** (0.009)	-0.013 (0.009)	0.039*** (0.010)	-0.095*** (0.007)	-0.016 (0.009)
<i>Period</i> : <i>ALE</i> (I)	0.039*** (0.002)	0.040*** (0.002)	0.039*** (0.002)	0.579*** (0.005)	0.596*** (0.004)	0.577*** (0.004)	0.165*** (0.003)	0.173*** (0.003)	0.164*** (0.003)	0.216*** (0.004)	0.192*** (0.003)	0.220*** (0.004)
<i>AME</i> (II)	0.049*** (0.004)	0.048*** (0.004)	0.049*** (0.004)	-0.121*** (0.008)	-0.145*** (0.007)	-0.119*** (0.008)	0.096*** (0.006)	0.079*** (0.006)	0.097*** (0.006)	-0.025*** (0.006)	0.019*** (0.005)	-0.027*** (0.006)
<i>AME</i> (III)	0.227*** (0.005)	0.222*** (0.005)	0.227*** (0.005)	-0.390*** (0.006)	-0.410*** (0.006)	-0.387*** (0.006)	0.186*** (0.006)	0.169*** (0.006)	0.192*** (0.006)	-0.023*** (0.006)	0.020*** (0.005)	-0.031*** (0.005)
<i>AME</i> (ln(<i>assets</i>)) (mean = 7.71, SD = 1.78)	-0.001 (0.001)	-0.005** (0.002)	-0.000 (0.002)	-0.008*** (0.002)	-0.048*** (0.002)	-0.020*** (0.002)	0.051*** (0.002)	0.001 (0.002)	0.030*** (0.002)	-0.042*** (0.002)	0.052*** (0.002)	-0.009*** (0.002)
<i>AME</i> (ln(TDC1)) (mean = 8.00, SD = 1.16)		0.010*** (0.002)			0.102*** (0.003)			0.117*** (0.003)			-0.228*** (0.003)	
<i>AME</i> (<i>Ave</i> (ln(TDC1))) (mean = 8.21, SD = 0.91)			0.000 (0.003)			0.034*** (0.004)			0.054*** (0.004)			-0.088*** (0.003)
N	26736	26686	26736	26736	26686	26736	26736	26686	26736	26736	26686	26736
<i>Pseudo R</i> ²	0.109	0.215	0.120	0.109	0.215	0.120	0.109	0.215	0.120	0.109	0.215	0.120

Standard errors in parentheses. *AME*s are not reported for variables that were not statistically significant.**p* < 0.05, ***p* < 0.01, ****p* < 0.001

apparent. However, we perform a robustness check of the main qualitative features of our regression results by running our regression in four different samples according to industry group, and we confirm them all.²⁷

The results of the regression using the benchmark specification for $V(I)$ in (1) are reported in Table 2, under the column labeled as regression specification (1). In the first row of Table 2, we report the regression sample averages, or, equivalently, the average predicted probability in the model. The numbers differ slightly from those reported in Table 1 because some of the observations have missing values for some of the regressors, and hence they are dropped from the sample.

In order to provide an intuitive sense of the estimated relative importance of each regressor, the rest of the rows in the table report the average of the partial derivatives of the probability of usage, or the average marginal effect of each explanatory variable x_{ij} in the vector of all explanatory variables X_i , defined as

$$AME(j) \equiv Mean_i \left(\frac{\partial \Pr(I|X_i)}{\delta x_{ij}} \right).$$

That is, using the estimated coefficients, we calculate how much the probability of using each instrument changes for each of the firms in the sample when we marginally increase the value of a given explanatory variable x_{ij} , evaluated at the true value of the vector of regressors X_i for firm i ; then we take the average of those marginal changes over i . Note that the marginal effects are calculated in a slightly different way for discrete variables. The marginal effects with respect to the variable “Period,” for example, represent the average change induced by hypothetically switching a firm from the base period, I, to each of the remaining periods.²⁸ Formally,

$$AME(j) \equiv Mean_i \left[\Pr \left(I | X_i^{-x_{ij}}, x_{ij} = n \right) - \Pr \left(I | X_i^{-x_{ij}}, x_{ij} = base \right) \right],$$

for all n different than $base$, where $base$ denotes the value of the regressor x_j , in this case period I, and $n \neq base$ represents period II and period III. The notation $X_i^{-x_{ij}}$ represents the vector of regressors X_i excluding regressor x_j . In order to provide a benchmark to evaluate these discrete changes in probability, we also report, for discrete

²⁷ Details are available upon request.

²⁸ To calculate the marginal effect for “2002–2005,” Stata calculates the predicted probabilities by setting to 1 the dummy for the baseline period of “1993–2001” into each observation while leaving all other regressors at their true sample values. Then it calculates this predicted probability again by substituting “2002–2005” instead. The average of this difference is the reported marginal effect.

regressors, the level of the average predicted probability in the sample when setting $x_{ij} = base$ (we denote this by $ALE(j)$ in the table).

Some of the strongest economic effects are associated with the three regulatory subperiods. Size and industry are also statistically significant for all groups. Despite the differences reported in Table 1, and possibly due to high standard deviations, our controls for gender, age, and tenure of the CEO often do not have a statistically significant effect on the choice of compensation instruments, so we choose to not report the *AMEs* for these variables.²⁹ We now summarize the findings regarding period, industry, and size.

Regulatory periods *Firms move away from compensation packages that include only options after both regulatory changes, either to use only stock or to use options together with stock. They mainly add stock to their compensation packages during period II, and they mainly substitute stock for options in period III.*

We find that if the same firm went from living in period I, before Sarbanes-Oxley, to period II, the probability of it choosing *O* would decrease by a substantial 12 percentage points (pp), while that of choosing *S* would increase 5 pp and that of choosing *B* would increase by almost 10 pp. In period III, after FAS 123R went into effect, the probability of using options would be 39 pp lower than in the initial period, leaving it at about 20 percent. The most favored category in that switch would be stock, with a 23 pp increase, followed by both, with a 19 pp increase. The use of none decreases at a modest 2–3 pp in each of the two periods.

Industry classification *Firms in FIRE and utilities favor packages that include stock exclusively, or no grants at all, more frequently than the average firm. Both these industries make less use of packages that include options exclusively, or stock and options together. Firms in mining and manufacturing, in contrast, use options exclusively, or together with stock, slightly more than the average firm, and they are less likely to compensate without using any grants at all.*

The control industry, “other,” aligns with the average usage probabilities in the overall sample. We see that switching from “other” to “mining and manufacturing” is associated with a shift away from using *N* into *O*, or *B*, in comparable magnitude. Switching to FIRE is associated with an important shift away from *O* and *B* (by 2 and 5 pp,

²⁹ Details are available upon request.

respectively), into S and N (by 4 and 3 pp, respectively). Switching to utilities, which includes transportation, communications, electric, gas, and sanitary services, presents, perhaps surprisingly, a similar pattern than FIRE. For utilities, however, the effects are even stronger: The decrease in B and O is by 4 and 8 pp, respectively, and the increase in N and S is by 4 and 8 pp, respectively.

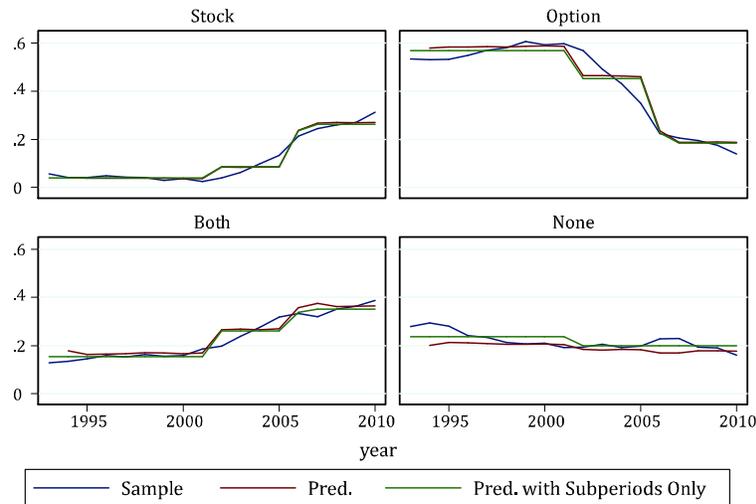
Size *Larger firms are less likely to use compensation packages that include options exclusively, and more likely to use those that include both stock and options. They are also less likely to compensate without using any type of grants at all.*

Firm size is a continuous variable (the log of the value of assets measured in thousands of 2010\$). An increase in firm size significantly increases the probability of choosing B , to the detriment of choosing O or N . It is difficult to compare the economic importance of size with respect to the discrete variables that we just commented on, since the numbers in the table represent the effect of infinitesimal increases in size, not changes in industry or period as above. For comparison, we can calculate the implied average change in the probabilities for an increase in log size equal to one standard deviation. This back-of-the-envelope calculation implies that the probability of choosing B increases by approximately 9 pp, while that of O decreases by 1 and that of N by 7. This suggests that the magnitude of changes associated with size is similar to that of changes in the industry classification, but smaller than that of the regulatory subperiods.

The Role of Individual Firm Characteristics

In Figure 7 we provide a simple graphical evaluation of the fit of the model in equation (1). We have plotted the sample percentage of users of each instrument by year in each of the subplots (blue line), the predicted probability of usage by the full model in (1) (red line), and a limited model that includes as regressors only the period dummy indicators (green line).³⁰ The main result that emerges from this comparison is that the explanatory power of the variables that indicate the different subperiods is high. Although the econometric model is not able to fit the smooth decline in the use of options, we have already pointed out earlier in this article that the new standards in FAS 123R

³⁰ The SEC adopted FAS 123R reporting rules for firms filing their proxy statements after December 15, 2006. Hence, we classify firms as being in period III if the month in which their fiscal year ends falls after November 2006. This means that in the year 2006, the firms in the sample are split across two subperiods. This explains the extra kink in the predictions using the restricted model.

Figure 7 Fit of the Model Over Time

were already recommended by the FASB in 1993, and some companies started adopting them earlier than 2006. This may explain most of the discrepancies between the green line and the true data.

In contrast with the good fit of the model over time, a low pseudo- R^2 seems to suggest that, even including the individual characteristics, our model does not do a good job in explaining individual cross-sectional variation in the use of instruments. As is apparent from the figure, the individual characteristics of the firm that we include in the full regression do not add much to the explanatory power of the model over time. In fact, the pseudo- R^2 of the regression using only the regulatory subperiods as explanatory variables is already 7 percent, compared to the 11 percent of the full model in specification (1). More work is needed to understand which individual characteristics of firms determine their choice of compensation instruments.

The Role of the Level of Pay

One potential explanatory variable of the choice of instrument that we left out of our analysis in regression specification (1) is the level of pay itself. It may be the case that tax advantages, or transparency concerns of the firm, make it more convenient for the firm to pay large sums to its CEO in the form of stock or options, rather than through

Table 3 Average Total Compensation and Average Mean Compensation by Firm

Period		<i>S</i>	<i>O</i>	<i>B</i>	<i>N</i>	Across Groups
I	$TDC1_{it}$	5,683 (37,035)	6,211 (16,821)	8,809 (15,811)	1,668 (3,360)	5,519 (16,176)
	$\text{mean}(TDC1)_i$	5,752 (7,804)	5,707 (7,707)	7,457 (7,707)	4,227 (6,211)	5,630 (7,002)
II	$TDC1_{it}$	5,672 (8,023)	5,677 (7,399)	8,854 (10,304)	2,185 (3,627)	5,808 (8,106)
	$\text{mean}(TDC1)_i$	5,422 (6,447)	5,380 (5,876)	7,206 (7,510)	4,380 (7,398)	5,660 (6,771)
III	$TDC1_{it}$	5,539 (7,861)	4,754 (6,645)	7,621 (7,827)	1,762 (2,866)	5,372 (7,218)
	$\text{mean}(TDC1)_i$	5,236 (6,163)	4,616 (5,196)	6,688 (6,421)	3,625 (5,605)	5,310 (6,090)
Across Periods	$TDC1_{it}$	5,585 (16,417)	5,893 (14,020)	8,299 (11,448)	1,804 (3,299)	5,542 (12,412)
	$\text{mean}(TDC1)_i$	5,350 (6,499)	5,486 (6,409)	7,056 (7,125)	4,095 (6,335)	5,542 (6,691)

Notes: Standard deviations in parenthesis.

a salary or a bonus program. To explore this possibility, we present in Table 3 the average level of total compensation (first row of all periods, labeled as $TDC1$) by group and subperiod. We can see that firms in N have a remarkably lower level of compensation, across all subperiods. Moreover, firms choosing B have the highest average compensation in all subperiods. The statistics for the group using only stock or only options are interesting: The level of pay is higher for O in period I, when options were more likely to be used on their own than stock (see Table 1). During period II, average pay is equal across the two groups of users. As we discussed in Section 3, this is a period when Sarbanes-Oxley had just been passed, making the choice of options more costly—at least in terms of opportunities for backdating and maybe in terms of public image. In period III, after the new accounting standards that made the valuation of options less arbitrary became compulsory, the ranking of average pay reverses: CEOs of firms that are users of stock are paid, on average, more than those that are users of options.

In order to explore formally the explanatory power of the level of pay after controlling for firm characteristics, we replicate the regression in equation (1), but add the level of total compensation (the log of the variable $TDC1$ in Execucomp) as a regressor. The results are reported in Table 2, under the column labeled as regression specification (2).

The meaningfulness of these estimated effects needs to be evaluated in the context of the mechanics of compensation, since there is an obvious relation between the level of compensation and the use of grants. This mechanical relationship exists unless we think that sometimes firms issue grants that are very small in value. In the sample, the minimum value for stock and option grants is in the order of \$3; the 1st percentile value is \$18,667 for stock grants and \$38,355 for options; the 10th percentiles values are about \$200,000 and \$250,000, respectively. Given that the 10th percentile of salary payments in the sample is \$364,000, and that of total compensation is in the order of \$750,000, these statistics suggest that fairly low values of the grants are possible and not that uncommon.

Another concern with the results of regression specification (2) is endogeneity: Since stock and options are risky assets, CEOs receiving their compensation in the form of grants (as opposed to salary or other less risky instruments, such as bonuses) may need to be compensated for their risk aversion with higher levels of pay. See Hall and Murphy (2002) for a formal explanation and quantification of the effect of risk aversion on the value of grants to executives, and the comparison of that value to the cost for the firm.

Total compensation is indeed a significant variable according to the results of the multinomial logit. Also, the pseudo- R^2 doubles with respect to specification (1). We find that a marginal increase in the level of pay leaves the probability of using stock almost unchanged, while it increases the one for choosing O by 10 percent and of B by 12 percent; it decreases the probability of using N by 23 percent.

The effects of size change significantly in specification (2). An increase in size now has a small but significant positive effect on the probability of S . The effects on O remain negative but increase significantly in magnitude. Moreover, the positive relation between size and choosing B becomes negligible (and insignificant) when controlling for the level of pay. Finally, the negative effect of size on the probability of choosing N changes to positive when controlling for the level of pay, suggesting that if a given firm is granting a relatively high level of total compensation, the fact that it is a larger firm actually makes it less likely to include stock or options in its compensation package.

We can also consider the changes in the marginal effects of the rest of the regressors with respect to those reported in specification (1). As for the period variable, the new model has similar implications both for period II and III when it regards the choice of S . However, the negative effect on the probability of choosing O is even stronger, while that of choosing B is still positive but weaker. The effect on choosing N changes sign in both periods with respect to specification (1): Although

magnitudes are small, firms are more likely to choose N in later periods than in the initial one. Note in Table 3 that there is no clear trend of average compensation over time across groups; however, compensation is lower in later periods for firms that include options in their packages.

As for the industry dummy, while the results for mining and manufacturing are very robust, for FIRE and utilities we observe some important changes. While firms in FIRE were more likely to choose S or N in specification (1), when including the level of total compensation as a control they are more likely to choose S , and O or B are now favored (in similar magnitudes), while N is now less likely to be chosen. Utilities is also more likely to choose S or N in specification (1); in specification (2) it becomes a likely user of B and a less likely user of N . A possible explanation of these reversals in the sign of the coefficients is that, given their size, firms in FIRE and utilities tend to have lower levels of pay, which are associated with a lower probability of choosing B and a higher probability of choosing N ; when the level of pay is not a control, that effect is assigned by the model to the industry dummy.

One may suspect that the covariance of the regressors with the level of pay is a potential cause of these changes in the estimated coefficients. However, the covariance is not perfect, and both size and pay remain significant in the robustness check, suggesting that specification (1) may have an omitted variable problem. Numbers need to be taken with caution.

As a final robustness check, we replicate the regression in equation (1) but add as a regressor the average level of total compensation (the log of the average of the variable $TDC1$ in Execucomp) of a firm across the years that it stays in the sample, rather than the actual level of $TDC1$ in each year. The results are reported in Table 2, under the column labeled as regression specification (3). Table 3 reports the average and standard deviation of this measure of pay in the sample (labeled $\text{mean}(TDC1)_i$). The most striking feature is the much higher pay for firms choosing N when compared to the average of contemporary level of pay. This reflects the fact that many of the firms in N are in one of the other compensation groups in some of the years.

The hope in including average pay as a regressor is that this may break slightly the mechanical link between the level of pay and the presence of grants, and rather pick up some firm characteristics that are correlated with, for example, the outside opportunity of the CEO, or any other characteristic that determines his average pay across the years but not necessarily the timing of the grants. We see in Table 3 that the pseudo- R^2 is higher than in specification (1), but much lower than in specification (2). The average level of pay is a significant

explanatory variable for O , B , and N , and the sign of the coefficients is aligned with that of the contemporaneous level of pay, but its economic importance is much smaller, confirming that some of the effects of the level of pay on the choice of instruments are purely mechanical.

Choosing N and the Timing of Grants

There is some anecdotal evidence that companies tend to have fixed timing rules when it comes to giving stock or option grants to their executives. Hence, when we observe a firm choosing N in our sample it may just mean that the firm is in a “non-granting” year, but that it will grant again the following year, or in a couple of years, depending on its timing rule. For example, taking the compensation of Steve Jobs over his tenure as Apple CEO (see Section 1), according to our classification, the company chose N in 10 of the years, O in 3, and S in 1. Why firms may not want to smooth out grants is, to our knowledge, an open question, and beyond the scope of this article. However, the common practice of having the selling restrictions of both stock and option grants vest progressively over time does provide some smoothing. Unfortunately, there is no good data readily available on these vesting periods. Nonetheless, we should keep in mind that if the practice of timing grants on a regular basis is really prevalent, then the statistics about usage presented here should be understood as informative about the timing of grants, and changes in usage patterns would be informative about changes in this timing.

A thorough analysis of the recidence patterns in the usage of instruments is beyond the scope of this article, and is left for future research. However, in order to provide a sense of how much of the variation in instrument choice in the data is not coming from timing of grants, we now report on a measure of the frequency of instrument use at the individual firm level: We calculate the fraction of years that a given firm is in each of the groups, or the “firm’s time share of I .” Denoting by t_{iI} the number of years that a firm i is in compensation group I , and by T_i the total number of years that firm i is in the sample, firm i ’s time share of I is defined

$$\tau_{iI} \equiv \frac{t_{iI}}{T_i}$$

for each I in \mathcal{I} . To give more meaning to the extreme values $\tau_{iI} = 0$ and $\tau_{iI} = 1$, we construct a balanced subset of the sample that includes only firms that we observe for at least six years ($T_i \geq 6$, a total of 489 firms out of the original full sample of 3,248 firms). From the fact that there are mass points at 0 (and, to a lesser extent, at 1) in the frequencies for this subsample, we conclude that, provided compensation cycles

Table 4 Percentage of Firms that are Never in a Given Compensation Group

Percent of Firms with $\tau_{iI} = 0$	<i>S</i>		<i>O</i>		<i>B</i>		<i>N</i>	
	Bal.	Full	Bal.	Full	Bal.	Full	Bal.	Full
Period								
I	.82	.86	.08	.12	.49	.61	.40	.42
II	.80	.80	.26	.29	.39	.53	.65	.60
III	.59	.55	.65	.64	.29	.42	.72	.59

are shorter than six years, not all the firms are following alternating times for the inclusion of options or stock grants in their compensation packages. This means that at least some of the variation that we see in the data comes from meaningful choices about the usage of the different compensation instruments.

Because the timing choices themselves may be influenced by the regulation period, in Tables 4 and 5 we report statistics of τ_{iI} by regulatory subperiod. We report this for the balanced subsample (denoted “Bal.”), as well as for our original full sample (denoted “Full”).

Table 4 reports the fraction of firms with $\tau_{iI} = 0$, i.e., they are never in compensation group *I*. It shows a pattern consistent with the evidence in our previous regression results: The fraction of firms that never were in group *S* or *B* decreases over time, while that of firms never choosing *O* increases. Interestingly, the increase in the fraction of firms with $\tau_{iN} = 0$ over time suggests that, if anything, timing decisions have changed toward using grants more frequently.

Table 5 reports the average value of τ_{iI} contingent on it being positive; that is, the average time share τ_{iI} for firms that are in compensation group *I* for at least one year. To report both the averages and their significances, we run an ordinary least squares regression of τ_{iI} on period dummies, for each *I*. The first column under each *I* reports, for the balanced sample, the coefficients for the constant (the level in the control period, I) and the included dummies (the change in the average τ_{iI} in each subsequent period with respect to period I), while the second column reports the same coefficients for the larger sample of firms that are in the data for at least six years. The patterns and significances are remarkably similar across the two samples of firms. There is no evidence of a significant decrease in the fraction of years that firms choose to grant options only (τ_{iO}) in period II, while it is significant both statistically and economically in period III. There is an important upward trend for both τ_{iS} and τ_{iB} , and a lot less markedly

Table 5 Mean Time Shares

Mean ($\tau_{iI} \tau_{iI} > 0$)	<i>S</i>		<i>O</i>		<i>B</i>		<i>N</i>	
	Bal.	Full	Bal.	Full	Bal.	Full	Bal.	Full
Period								
I: Level	.26 (.03)	.28 (.02)	.62 (.01)	.65 (.01)	.36 (.02)	.40 (.01)	.34 (.01)	.41 (.01)
II: Change from Period I	.15 (.04)	.17 (.02)	[-.02] (.02)	[-.02] (.01)	.19 (.02)	.15 (.01)	.07 (.02)	.10 (.01)
III: Change from Period I	.31 (.04)	.27 (.02)	-.14 (.03)	-.14 (.02)	.33 (.02)	.24 (.01)	[.05] (.03)	.08 (.01)
Adjusted R^2	.17	.13	.03	.03	.18	.10	.01	.03
<i>N</i>	387	1,345	979	3,532	894	2,531	602	2,489

Notes: Time share τ_{iI} represents the fraction of years that firm i belongs to group I , out of the total number of years that firm i is in the sample. This table reports mean time shares for each I , for firms with positive τ_{iI} . Square brackets indicate insignificance at the 5 percent confidence level.

for τ_{iN} .³¹ That is, (1) firms that choose O do so less frequently in period III, (2) firms that choose S , or B , do so more frequently in the later periods than in the initial one, and (3) firms that choose N do so only slightly more often in the last two periods than in the first one. Since these changes in grant timing patterns align with the trends in the usage of instruments that we have reported in Table 2, we conclude that our results could be due, at least partly, to a change in the frequency of usage of stock and options, rather than a change in the number of different firms that use them.

It is important to keep in mind that the evidence on the timing of grants that we have provided in this section is partial, since it does not control for the amount of past grants and it only exploits the panel aspect of the data in a limited way. It would be interesting to perform the analysis of usage that we do here with a comprehensive measure of the wealth of the CEO vested in the firm at each point in time (as in Clementi and Cooley [2010]), as a way of controlling for outstanding incentives. This is left for future research.

³¹ Note that Table 5 is providing evidence for firms that have $\tau_{iI} > 0$, and these firms differ across I s; hence, the percentages across rows do not typically sum up to 1.

Table 6 Shares of Total Compensation, by Instrument

	Salary (%)	BIC (%)	Stock (%)	Option (%)	Other (%)
All	.32	.23	.11	.28	.06
<i>S</i>	.27	.23	.45	0	.05
<i>O</i>	.27	.20	0	.49	.04
<i>B</i>	.19	.20	.26	.31	.04
<i>N</i>	.59	.30	0	0	.10

4. THE IMPORTANCE OF DIFFERENT COMPENSATION INSTRUMENTS: THE INTENSIVE MARGIN

In the previous section we asked what determines the choice of compensation instruments. A natural complementary question to that is what is the relative importance of each instrument in the total compensation of the CEO. In this section, we provide some simple statistics about the share of total compensation that salary, bonus and incentive compensation (BIC), stock grants, option grants, and “other compensation” represent.

Table 6 documents the average of these shares in our sample, disaggregated by groups of users. The most salient feature of those statistics is the difference in the shares of grants across firms in *S*, *O*, and *B*: Firms in *B* have a combined share of grants of 57 percent, higher than the shares of grants for firms using stock exclusively (45 percent) or options exclusively (49 percent). The share of BIC is similar for firms in *S*, *O*, and *B*, around 20 percent. In contrast, firms in *N*, who do not use stock or options, use both BIC and “other compensation” more intensely than the rest of firms, but the share of the only incentive instrument, BIC, is 30 percent, well below the combined shares of incentive instruments (BIC + stock + option) of the rest of the firms. In other words, BIC, stock, and options do not appear to be perfect substitutes for each other. This evidence complements what we presented in Table 3 about the relationship between the level of compensation and the usage choices, suggesting that the relative importance of different instruments may be related to the choice of instruments through the level of pay. We saw in Table 3 that firms in *N* have levels of total compensation between one-third and one-fourth of the rest of firms. In spite of this, the relative importance of the salary is much higher for them. Hence, there seems to be a fixed component in the determinant of the salary, or a “cap,” which is somewhat independent of whether the firms choose to also award grants or not. The most obvious

Table 7 Shares of Total Compensation, by Instrument

Freq.		Salary (%)	BIC (%)	Stock (%)	Option (%)	Other (%)
	Period I	.35	.22	.04	.33	.05
.04	<i>S</i>	.34	.26	.33	0	.07
.57	<i>O</i>	.28	.20	0	.48	.04
.16	<i>B</i>	.22	.18	.20	.36	.04
.24	<i>N</i>	.62	.28	0	0	.09
	Period II	.29	.23	.10	.32	.05
.09	<i>S</i>	.27	.26	.42	0	.06
.45	<i>O</i>	.25	.20	0	.51	.04
.26	<i>B</i>	.18	.20	.25	.33	.04
.20	<i>N</i>	.56	.32	0	0	.11
	Period III	.29	.23	.24	.18	.06
.26	<i>S</i>	.25	.22	.49	0	.05
.19	<i>O</i>	.27	.21	0	.48	.04
.35	<i>B</i>	.18	.21	.31	.27	.04
.20	<i>N</i>	.55	.32	0	0	.12

explanation is the limits to tax deductions for salaries above a certain level.³² However, other factors may be important, like the need to provide incentives through variable pay. This also possibly plays a role in explaining the difference in the shares of salary across the firms in *S*, *O*, and *B*. The share of salary is the lowest (19 percent) for firms in *B*, which are the ones that have the highest total compensation according to Table 3. However, the share of salary is equal for firms in *S* than for firms in *O*, in spite of the average total compensation in *S* being 90 percent of that in *O*.

Our previous analysis has shown that the use of instruments differs importantly across subperiods, and to some extent across industry groups. Hence, we now look at the average shares controlling for these two variables.

Table 7 presents evidence on the changes in the relative importance of the instruments over the three different regulation subsamples. For convenience, we replicate the sample frequencies of each group of compensation, within a period, that we already discussed following Table 1.

³² The Omnibus Budget Reconciliation Act Resolution 162(m) of 1992 imposed a \$1 million cap on the amount of the CEO's non-performance-based compensation that qualifies for a tax deduction. See Jarque (2008) for a review of the academic literature that studied the effects of that change of regulation on pay practices.

When we look at the shares for all the users together, we see that while the shares of BIC and “other compensation” remained fairly constant at about 23 percent and 5 percent, respectively, the share of salary was higher in period I (35 percent as opposed to 29 percent post-2002). The share granted in the form of options also experienced a sharp decline, but only in period III, when it went from 32 percent to 18 percent. The share of compensation that is no longer granted through salary after 2002 and no longer granted through options after 2006 is granted through stock: There is an increase in the share of stock of 6 pp in period II, and then of 14 extra pp in period III. These changes in the share of stock over time (intensive margin) are in line with the changes in the choice of S reported in Table 1 (extensive margin), where we saw that firms tended to “add” stock to their compensation package in period II, rather than completely substitute options for stock. Note, however, that these numbers for the share of total compensation that are given in the form of stock are representative both of firms in S and B in Table 1. We discuss the data in each compensation group next.

When we look at the statistics disaggregated by user groups, we see slightly different changes over the regulatory periods for each of them. The most striking fact may be the increase in the share of stock, which happens both for firms that are in S and in B . For firms in S , the share of stock increases by 9 pp in period II (compensated mainly by a decrease in the share of salary of 7 pp), and then by 7 pp in period III (compensated mainly by a decrease in the share of BIC by 4 pp). For firms in B , the share of stock increases by about 5 pp each period, while the share of options decreases (3 pp in period II, 6 extra pp in period III). In addition, for firms in O the share of options stays constant overall (and it even increases by 3 pp in period II). In other words, for the firms that continue to rely exclusively on option grants in spite of the regulatory hurdles, the relative importance of options with respect to salary, BIC, and “other compensation” does not decrease. That is, if what we observe is a response to the regulatory changes, it seems to have taken place through the extensive margin (with firms in O going from 57 percent of the sample to 19 percent), rather than the intensive one. This suggests that there might be some fixed cost to adopting a new instrument of compensation, maybe related to accounting costs or perhaps to communication to shareholders.

Table 8 presents the shares of each compensation instrument by industry group. The variation in the shares across instrument users, within a given industry group, is fairly in line with the patterns by users that we described in Table 6, so we do not report the

Table 8 Shares of Total Compensation, by Industry Group

	Salary (%)	BIC (%)	Stock (%)	Option (%)	Other (%)
Min/Man	.32	.22	.11	.31	.05
FIRE	.29	.27	.15	.23	.06
Utilities	.34	.25	.14	.21	.06
Other	.33	.20	.11	.30	.06

disaggregated numbers here.³³ One main conclusion stands out from Table 8—mining and manufacturing and other use options and salary more intensely than do FIRE and utilities, which rely more on BIC and stock. FIRE, which includes financial firms, has in fact the lowest share for salary. It is important to keep in mind that, as reported in Table 1, the proportion of firms in each user group is not constant across industry groups; this, together with the (omitted) evidence that shares for user groups within industry align with those reported in Table 6, implies that most of the variation across industries is due to composition effects, without important industry-specific patterns for the shares of each compensation instrument.

5. CONCLUSION

In the last decade several regulatory changes took place in the United States regarding the reporting and expensing of stock option grants. This article provides an empirical analysis of the impact of these changes in the composition of pay packages for CEOs at the largest U.S. firms from 1993 to 2010. Both the passage of the Sarbanes-Oxley Act in 2002 and the changes in accounting standards in SFAS 123R mandated by the SEC in 2006 erased some advantage of granting options versus stock as part of the compensation of CEOs. We find evidence indicating that firms may have responded to this by shifting away from options and into stock. Even though, after the two regulatory changes, there is still a significant portion of firms in the sample that choose to grant options to their CEO (about 55 percent of firms in the 2006–2010 period, compared to 67 percent before 2002), alone or combined with stock, the fraction of firms that are awarding options but not stock in a given year decreases (from 57 percent before 2002 to 19 percent after 2006).

³³ A more detailed table with shares across industries and compensation groups is available upon request.

However, while only 4 percent of firms used exclusively stock grants before 2002, this percentage increases over the period we analyze to reach 26 percent after 2006.

How firms decide whether to include options, stock, both, or none of the two types of grants in their pay packages remains to be understood, but we find some regularities. Firms in finance and in utilities are more likely to use stock or neither, while firms in mining and manufacturing are more likely to use options, or stock and options together. Larger firms tend to use stock and options together, although this effect disappears if we control for the level of pay, which is higher at larger firms. A higher level of pay is associated with a higher probability of using stock and options together, or only options.

We also find that different compensation instruments do not appear to be perfect substitutes within compensation packages. The relative importance of bonuses in overall compensation has not decreased over time, while that of the salary has, in favor of stock and option grants. Perhaps surprisingly given the decrease in the popularity of option grants starting in the early 2000s, the relative importance of options in relation to the total amount of compensation has not decreased over time for firms that still include options in their compensation packages.

REFERENCES

- Baker, George P., and Brian J. Hall. 2004. "CEO Incentives and Firm Size." *Journal of Labor Economics* 22 (October): 767–98.
- Bickley, James M. 2012. "Employee Stock Options: Tax Treatment and Tax Issues." Congressional Research Service Report for Congress (June 15).
- Brown, Lawrence D., and Yen-Jung Lee. 2011. "Changes in Option-Based Compensation Around the Issuance of SFAS 123R." *Journal of Business Finance & Accounting* 38 (November/December): 1,053–95.
- Cheng, Ing-Haw, Harrison G. Hong, and Jose A. Scheinkman. 2012. "Yesterday's Heroes: Compensation and Creative Risk-Taking." ECGI - Finance Working Paper No. 285/2010; AFA 2011 Denver Meetings Paper (June 24).

- Clementi, Gian Luca, and Thomas F. Cooley. 2010. "Executive Compensation: Facts." Fondazione Eni Enrico Mattei Working Paper 2010.89.
- Core, John, and Wayne Guay. 1999. "The Use of Equity Grants to Manage Optimal Equity Incentive Levels." *Journal of Accounting and Economics* 28 (December): 151–84.
- Frydman, Carola, and Raven E. Saks. 2010. "Executive Compensation: A New View from a Long-Term Perspective, 1936–2005." *Review of Financial Studies* 23 (May): 2,099–138.
- Gabaix, Xavier, and Augustin Landier. 2008. "Why Has CEO Pay Increased So Much?" *The Quarterly Journal of Economics* 123 (1): 49–100.
- Guay, Wayne, David Larcker, and John Core. 2005. "Equity Incentives." In *Top Pay and Performance: International and Strategic Approach*, edited by Shaun Tyson and Frank Bournois. Burlington, Mass.: Elsevier Limited, Chapter 8.
- Guay, Wayne, S. P. Kothari, and Richard Sloan. 2003. "Accounting for Employee Stock Options." *American Economic Review* 93 (May): 405–9.
- Hall, Brian J., and Jeffrey B. Liebman. 1998. "Are CEOs Really Paid Like Bureaucrats?" *The Quarterly Journal of Economics* 113 (August): 653–91.
- Hall, Brian J., and Jeffrey B. Liebman. 2000. "The Taxation of Executive Compensation." In *Tax Policy and the Economy*, Volume 14, edited by James M. Poterba. Cambridge, Mass.: MIT Press, 1–44.
- Hall, Brian J., and Kevin J. Murphy. 2002. "Stock Options for Undiversified Executives." *Journal of Accounting and Economics* 33 (February): 3–42.
- Heron, Randall A., and Erik Lie. 2007. "Does Backdating Explain the Stock Price Pattern around Executive Stock Option Grants?" *Journal of Financial Economics* 83 (February): 271–95.
- Jarque, Arantxa. 2008. "CEO Compensation: Recent Trends and Regulation." Federal Reserve Bank of Richmond *Economic Quarterly* 94 (Summer): 265–300.
- Jensen, Michael C., and Kevin J. Murphy. 1990. "Performance Pay and Top-Management Incentives." *Journal of Political Economy* 98 (April): 225–64.

- Lewellen, Wilbur G. 1968. *Executive Compensation in Large Industrial Corporations*. New York: National Bureau of Economic Research.
- Meyers, Arthur S. 2012. "Preserving Your Company's Tax Deduction for Stock Awards." *StockSense: Quarterly Newsletter from Fidelity Stock Plan Services* (December).
- Murphy, Kevin J. 1999. "Executive Compensation." In *Handbook of Labor Economics*, Vol. 3b, edited by Orley C. Ashenfelter and David Card. Amsterdam: Elsevier Science North Holland; 2,485-563.
- Murphy, Kevin J. 2003. "Stock-Based Pay in New Economy Firms." *Journal of Accounting and Economics* 34 (January): 129-47.
- Rosen, Sherwin. 1992. "Contracts and the Market for Executives." In *Contract Economics*, edited by Lars Werin and Hans Wijkander. Cambridge, Mass.: Blackwell Publishers.
- Schaefer, Scott. 1998. "The Dependence of Pay-Performance Sensitivity on the Size of the Firm." *The Review of Economics and Statistics* 80 (August): 436-43.
- Schubert, Renate, Martin Brown, Matthias Gysler, and Hans Wolfgang Brachinger. 1999. "Financial Decision-Making: Are Women Really More Risk-Averse?" *The American Economic Review* 89 (May): 381-5.
- Securities and Exchange Commission. 2006. "Executive Compensation and Related Person Disclosure." Available at www.sec.gov/rules/final/2006/33-8732a.pdf.