

October 2005

**Can Heterogeneity, Undiversified Risk, and Trading Frictions  
Solve the Equity Premium Puzzle?**

John Heaton\*  
Deborah Lucas\*\*

\* University of Chicago and the NBER

\*\* Northwestern University and the NBER

---

Prepared for the Handbook of Investments -- Equity Risk Premium.

## Can Heterogeneity, Undiversified Risk, and Trading Frictions Solve the Equity Premium Puzzle?

### Abstract

Can the historical equity premium be explained as a rational equilibrium outcome when risk-averse agents with conventional preferences are faced with non-diversifiable sources of risk (e.g., from labor or entrepreneurial income), and when trading frictions prevent them from using financial assets to effectively self-insure transitory shocks? Our research suggests that it is difficult to generate the historical equity premium in realistically parameterized models of this sort. Nevertheless, investigations of these factors clearly reveal the ingredients necessary for any consumption-based model to match returns data. Using simplified versions of some of our earlier models and other models in the literature, in this paper we illustrate the promise and limitations of incomplete risk diversification and trading frictions as explanations for the equity premium puzzle. We also present new results on the likely importance of entrepreneurial income risk.

## 1. Introduction

The equity premium puzzle is the inability of a standard representative consumer asset pricing model, calibrated with aggregate data, to match the historical differential between average stock returns and the risk-free rate. In pursuing a solution to the puzzle, our research strategy has been to weaken one or more of the maintained assumptions in the standard model. In particular, we have focused on deviations from complete markets arising from a variety of factors. These include undiversifiable background risks such as labor income and private business income, trading frictions, and limited portfolio diversification, all of which can reduce tolerance for investment risk. We have also considered whether these deviations from the standard model become more important when agents' preferences deviate from the constant relative risk aversion (CRRA) specification assumed by Mehra and Prescott (1985). The results of these investigations suggest the robustness of the puzzle to these important deviations from market completeness. Further, they suggest the ingredients necessary for any consumption-based model to match the data, which may be helpful in evaluating proposed solutions. Using calibrated models similar to our earlier work and others found in the literature, in this paper we illustrate the promise and limitations of incomplete risk diversification and trading costs as explanations for the equity premium puzzle.

In a consumption-based model, the critical determinants of the predicted equity premium are: the variance of consumption risk, its correlation with stock returns, and the curvature of the assumed utility function. An increase in any of these factors increases investors' distaste for stock market risk, and tends to increase the predicted equity premium. Aggregate consumption growth exhibits low volatility, and is only weakly correlated with stock returns. Hence a moderately risk-averse individual faced with this consumption process will require only a small premium to assume stock market risk.

By contrast, many individuals' consumption appears to be much more volatile than in aggregate, due to factors such as incomplete diversification of income risk or poor portfolio diversification. In some cases individual consumption may also be more highly correlated with stock returns than is aggregate consumption. It is therefore natural to ask whether differences between aggregate and individual consumption processes can help explain the puzzle. It should be emphasized that this is primarily an empirical question, not a theoretical one. As shown in Section 2, it is fairly easy to construct a risky consumption process that allows the standard model to match the equity premium (although with CRRA preferences it is difficult to simultaneously match the low volatility of the risk-free rate). The challenge is to reconcile the equity premium, and asset returns more generally, with consumption processes that are suggested by data.

To directly test whether an otherwise standard consumption-based model could explain asset returns when calibrated with individual data, it would be convenient to study the statistical properties of individual consumption. But individual consumption, when it is measured at all, is measured with considerable error. Expenditures do not equate to consumption, for instance, because of the lumpiness of durable purchases. Aggregating across different categories of consumption also can be problematic. A more insurmountable difficulty is the absence of a U.S. data set that includes a long time series of data on household consumption.<sup>1</sup>

Time series data is available, by contrast, on individual income. A popular source is the Panel Survey of Income Dynamics (PSID), which tracks the various components of income for a representative sample of households over many decades. The data can be used to look for differences in income risk across households with different characteristics, such as between

---

<sup>1</sup> Food consumption in the PSID is an exception, but food may not be a reliable proxy for a broader consumption measure.

stockholders and non-stockholders. It also can be used to test the standard model, since the model is essentially a non-linear filter that maps income to consumption by solving consumers' constrained optimization problems. Most of the research discussed in this chapter relies on this approach.

Several complications arise when the exogenous driving process is income rather than consumption, as model predictions are quite sensitive to the assumed market structure and income process. The most critical assumptions are: (1) the form of the budget and wealth constraints, particularly the severity of borrowing and short sales restrictions; (2) the persistence of the income process; (3) the minimum income realization; and (4) the correlation between individual income growth and the stock market. These considerations suggest that empirical evidence on household income processes, their correlation with asset returns, and credit market frictions, will determine whether explanations of the equity premium puzzle based on incomplete risk-sharing will ultimately be persuasive.

To briefly elaborate on these sensitivities: The structure of trading frictions and the persistence of income shocks determine the extent to which individuals can smooth income shocks through capital markets. Borrowing and short sales restrictions, and any trading costs, limit how much idiosyncratic income risk can be effectively smoothed by saving. More persistent income shocks similarly limit peoples' ability to self-insure by borrowing and lending, since they represent a permanent change in wealth that can only be absorbed by consumption changes. The minimum income realization puts a floor on consumption. It thereby serves as a risk-free bond that increases tolerance for investment risk. The higher is the floor on income and hence on consumption, the lower is the predicted equity premium. Finally, the predicted equity premium is very sensitive to the correlation between individual income shocks and stock returns. Without a

significant positive correlation between the two, even a very risky individual income process does little to increase the predicted premium.

The seemingly conflicting results in the literature on the success of a heterogeneous agent approach can be largely understood in terms of differences in these critical assumptions. Very persistent individual income processes, or very limited opportunities to smooth transitory shocks by transacting in financial securities, combined with a significant correlation between non-investment income and investment income, and a low floor rate on overall income, can generate a sizable equity premium (e.g., Mankiw (1986), Constantinides and Duffie (1996), Storesletten, Telmer and Yaron (2001), Constantinides, Donaldson, Mehra (2003)). On the other hand, with moderate persistence of income shocks, some trading frictions, a non-zero floor on income, and weak correlation between investment and non-investment income, it is difficult to explain the equity premium in this class of models (e.g., Heaton and Lucas(1992, 1995, 1996), Lucas (1994), Telmer(1993)).

When taking heterogeneity into account, a salient feature of the data is that stocks are concentrated in the portfolios of high-wealth households. As first emphasized by Mankiw and Zeldes (1991), it is then the consumption process of stockholders that should be most relevant for asset prices. In fact, the sources of income risk for high-wealth households look quite different than for the typical household, with labor income playing a smaller role and income from privately held businesses becoming significant. In section 3 we discuss our findings on the implications of private business income as a background source of risk, and for the first time examine these effects in a calibrated model.

An alternative to non-investment income as the cause of higher individual consumption risk is undiversified investment risk. That is, if individual stock portfolios are less diversified than in

aggregate, it is the statistical properties of those portfolios that affect the returns investors demand. Lack of diversification can arise, for instance, if having managers or business owners hold high equity stakes in their businesses overcomes agency problems. Such undiversified investment income is particularly important to the wealthy households likely to be the marginal participants in equity markets. As with non-investment income, the critical consideration is the statistical properties of the resulting consumption process of stockholders. Consumption financed primarily with income from poorly diversified portfolios has properties consistent with a high equity premium – a non-negligible probability of low consumption realizations, high variance, and high correlation with the stock market. As with other explanations based on consumption volatility arising from market imperfections, the deeper question is whether there is evidence of frictions large enough to induce such a costly lack of risk sharing.

In remainder of this paper we elaborate on these themes, using simple calibrated models to illustrate some of the theoretical points, and summarizing the related empirical evidence. In section 2 we discuss the literature on labor income as a source of uninsurable risk affecting asset returns, with and without trading costs. In section 3 we consider income risk from closely held businesses; and in Section 4 the risk arising from incomplete portfolio diversification. Section 5 concludes.

## 2. Labor Income as Background Risk

The first papers to consider undiversifiable income risk as an explanation of the equity premium puzzle focused on labor income. Labor income is a natural candidate because of its size (it accounts for about 75 percent of aggregate income) and the difficulty, due to problems of moral hazard and legal limits on selling it forward, of insuring it.

Mankiw (1986) first illustrated the potential for labor income risk to resolve the equity premium puzzle with a stylized two-period model. He shows that if some fraction of the population is forced to bear a disproportionate share of aggregate shocks, the predicted equity premium can be made arbitrarily large. A basic question explored in two closely related papers (Lucas (1994) and Telmer (1993)) is whether Mankiw's results continue to hold in an infinite horizon model similar in structure to Mehra and Prescott (1985), but with the addition of exogenous idiosyncratic labor income shocks.

The model that we will use to illustrate some general conclusions that can be drawn from these analyses is the following. Assume that investors maximize expected utility over a horizon  $T$ , generally taken to be infinite:

$$U_{t,i} = E_t \left[ \sum_{x=0}^T \beta^x (c_{t+x,i}^{1-\gamma} - 1) / (1-\gamma) \right] \quad (1)$$

Agent  $i$  invests  $s_{t+1,i}$  in stocks,  $b_{t+1,i}$  in bonds and consumes  $c_{t,i}$  at time  $t$ . The  $I$  types of agents are distinguished by their income processes and possibly their access to capital markets.

The consumption and saving choice is subject to the flow wealth constraint:

$$c_{t,i} + s_{t+1,i} + b_{t+1,i} \leq s_{t,i}(1 + r_t^s) + b_{t,i}(1 + r_t^b) + y_{t,i} \quad (2)$$

where  $r_t^s$  is the return on stocks at time  $t$ , and  $r_t^b$  is the return on bonds at time  $t$ , and  $y_{t,i}$  is risky, exogenously specified, non-tradable income.

The resulting Euler equation for asset of type  $j, j = s, b$  is:

$$E_t \left\{ (c_{t+1,i} / c_{t,i})^{-\gamma} (1 + r_{t+1}^j) \leq 1 / \beta \right\} \quad (3)$$

Equation (3) holds with equality for any unconstrained investor  $i$ . In that case, rearranging gives an expression for the equity premium:

$$E_t \left\{ \left( c_{t+1,i} / c_{t,i} \right)^{-\gamma} \left( r_{t+1}^s - r_{t+1}^b \right) = 0 \right\} \quad (4)$$

The ability to self-insure by saving is limited by exogenous borrowing and short sales constraints, which are used to control access to capital markets. By assumption investors can borrow only a limited amount in the bond market, and can short only a limited amount of stocks. When a borrowing or short sales constraint is binding, (3) becomes an inequality, and (4) only holds for unconstrained agents. In fact, asset prices are always determined by the common marginal rate of substitution of the unconstrained agents, since in equilibrium they must be content with their choices. The presence of credit-constrained agents implies that the identity of the marginal investor changes over time, so in general the aggregate consumption process does not price assets. The changing identity of the marginal investor also precludes closed form solutions to the model, and complicates its numerical solution.

Assuming an aggregate endowment of one share of stock paying a stochastic aggregate dividend,  $d_t$ , risk-free bonds in zero net supply, and aggregate labor income  $Y_t$ , the market clearing conditions at each time  $t$  are:

$$\sum_{i=1}^I c_{t,i} = d_t + Y_t \quad (5a)$$

$$\sum_{i=1}^I b_{t,i} = 0 \quad (5b)$$

$$\sum_{i=1}^I s_{t,i} = 1 \quad (5c)$$

An equilibrium for this model is a sequence of stock and bond prices, and consumption and portfolio allocations for each agent, such that equations (1) to (3) hold and the market clearing conditions (5) hold in all states and at all times.

Notice that background income process  $y_t$  affects the asset prices defined by (3) only indirectly, through its affect on the variability of consumption and its correlation with financial returns. While this background risk could come from a variety of sources, including wages, restricted pension holdings, housing rents, and private businesses, in this section we emphasize labor income as the risk source.

Under the assumption that consumption growth and returns are log-normally distributed conditional on information at time  $t$ , (4) can be written as:

$$E_t \{r_{t+1}^s - r_{t+1}^b\} = -\frac{1}{2} \text{var}_t(r_{t+1}^s) + \gamma \text{cov}_t \left[ \ln(c_{t+1,i} / c_{t,i}), r_{t+1}^s \right] \quad (6)$$

The equity premium increases in the covariance between stock returns and consumption growth, and the coefficient of relative risk aversion. Rough calibrations of (6) illustrate the original equity premium puzzle, and also its potential solution by appealing to higher consumption risk induced by market imperfections.

Using equation (6) and historical statistics on stock returns and consumption growth, we can ask what value of  $\gamma$  implies an equity premium of 7 percent? Consistent with historical data, assume a standard deviation of stock returns of 18 percent, a standard deviation of aggregate consumption growth of 3.7 percent, and a correlation between aggregate consumption growth and stock returns of 0.3. This implies a risk aversion coefficient of 43, well outside the range considered plausible. Conversely, one can ask about the implied equity premium for a given level of risk aversion. At

$\gamma = 10$ , the top of the range considered by Mehra and Prescott, the implied equity premium for these parameters is only 0.38 percent.<sup>2</sup>

Similar experiments, assuming higher levels of consumption growth variability, suggest a much better match with the data. For example, setting consumption volatility to 12 percent and holding other parameters as above, (6) implies that a risk aversion coefficient of 13.3 produces the target risk premium of 7 percent. At  $\gamma = 10$ , the implied equity premium is a respectable 4.9 percent.

Equation (6) can also be used to demonstrate the sensitivity of the predicted premium to the correlation between consumption growth and stock returns. In the Mehra and Prescott's analysis, dividend volatility and consumption volatility are equated, inducing a fairly high correlation between returns and income. Under the extreme assumption of perfect correlation between returns and consumption growth, and holding other parameters as in the low consumption risk example, the implied equity premium increases from .38 percent to 5 percent for  $\gamma = 10$ . With consumption volatility of 12 percent, the predicted premium rises from 4.9 to 20 percent. On the other hand, if the correlation between aggregate consumption and stock returns is set close to zero, as empirical evidence suggests may be the case, the equity premium is negligible even if consumption variability is high.

These calculations illustrate an important conclusion from these earlier papers -- high consumption variability, in combination with a moderate to high correlation between consumption growth and stock returns, has the potential to resolve the equity premium puzzle.

---

<sup>2</sup> By assuming volatile stock returns, rather deriving the volatility of stock returns from a dividend process calibrated with historical data, these examples actually understate the severity of the equity premium puzzle.

Before turning to more complicated calibration results, it is useful to demonstrate another key finding of the earlier investigations: In calibrated models in which labor income has a positive minimum realization, the capitalized value of that minimum serves as a bond that can significantly increase tolerance for stock market risk. Because of this, adding risky labor income can actually make it more difficult to resolve the equity premium puzzle.

The effect of floor labor income can be illustrated most clearly in a decision theoretic portfolio choice model, rather than in an equilibrium asset pricing model. In Heaton and Lucas (2000a) we analyze a model that is the portfolio choice analogue of the model described by equations (1) – (5), calibrated to U.S. data. We find that including a realistic amount of risky labor income results in portfolios with 100 percent stock investments, or when borrowing is permitted, a levered position in stocks. To see why this occurs, it is convenient to consider Merton’s (1971) closed form solution for the share of wealth invested in stocks,  $w^*$ :

$$w^*W = \frac{(E(r_s) - r_f)(W + y)}{\gamma \text{var}(r_s)} \quad (7)$$

In this model  $W$  is financial wealth,  $y$  is the constant flow of labor income, and  $r_f$  is the risk-free rate. In the absence of labor income this model generates reasonable portfolio shares for moderate levels of risk aversion. For instance, assuming an equity premium of 6 percent and a 15 percent annual standard deviation of stock returns, a portfolio with 50 percent in stocks implies a coefficient of relative risk aversion of 2.67. Now consider labor income of the same magnitude as financial wealth, which roughly matches the mean ratio of non-housing, non-business wealth to permanent income in the PSID (Carroll and Samwick (1995)). Holding the risk aversion coefficient at 2.67, predicted stockholdings rise to 100 percent of financial wealth. For households with even less financial wealth relative to income, the implication is that they would

borrow to buy stocks. Intuitively, the certain labor income is analogous to the fixed income stream from a bond, so it substitutes for bonds in the agent's utility function.

While the simple calculations based on equation (6) and (7) are suggestive, they abstract from the effect of more carefully modeling the joint statistical processes governing dividends and labor income, and the possibility that the effects of borrowing and short sales constraints will alter the relation between income and consumption in a way that has a significant effect on required returns. Table 1, reproduced from Lucas (1994), confirms that the conclusions drawn from equation (6) can carry over to a more fully specified model of labor income and dividends. The Table is generated from a calibrated version of the model described above, where two infinitely-lived agents receive equal and offsetting idiosyncratic labor income shocks when aggregate output is low. The idiosyncratic shocks are structured so that the employed worker receives income 1.22 times that of the unemployed worker when aggregate income growth is low. The aggregate income process is identical to that in Mehra and Prescott (1985), with dividends equal to a constant 30 percent of aggregate income, and labor income making up the balance. Table 1 shows that the resulting equity premium matches the historical average premium, even with risk aversion as low as 2.5, but only under the strong assumption of no access to capital markets so that consumption equals income.

(Table 1 here)

Table 1 also suggests that the risk-free rate puzzle -- the observation that the standard model also predicts too high a risk-free rate -- is resolved in this model. The reason is that with heterogeneous agents, the risk-free rate clears the market for agents who are not constrained. The price setters are then the subset of the population whose current income and desire to save are relatively high. This implies that the market clearing interest rate is lower than with a

representative agent. A general finding in this literature is that binding borrowing constraints make it relatively easy to solve the risk-free rate puzzle, whereas the equity premium puzzle is much more robust to the addition of financial frictions.

The results of Table 1 are reversed with the opening of capital markets. When agents are assumed to have access to trading in either the stock or bond market, a similarly calibrated model (in terms of the income process and preference parameters) fails to reproduce the historical equity premium. Even with quite limited market access, the predicted equity premium falls to levels comparable to those found by Mehra and Prescott (1985), and equilibrium individual consumption volatility can be indistinguishable from aggregate consumption volatility (Lucas (1994)). For instance, the predicted equity premium is close to zero when agents are allowed to trade but not short-sell stocks, and borrowing is prohibited. The reason is that agents effectively smooth transitory idiosyncratic shocks by accumulating savings (here in the form of stocks) when income is high and reducing savings when income is low. The floor on labor income, as discussed above, also increases risk tolerance.

The general conclusions that emerge from these calibrations is that if income shocks cannot be insured and individual consumption is sufficiently volatile as a result, this could be an explanation for the equity premium puzzle. Making a persuasive case for this explanation, however, requires finding evidence that income shocks are fairly permanent, as assumed by Mankiw (1986) and Constantinides and Duffie (1996), or establishing that the barriers to trading in financial markets are sufficient to discourage self-insuring transitory income shocks. Another important observation is that in these types of models, the only way to prevent self insurance is to restrict access to all financial markets, since agents readily substitute to trading in the low-cost market. These quantitative issues are addressed in Heaton and Lucas (1996), where we calibrate a related model that incorporates transaction costs and an empirically estimated income process.

## 2.1 Calibrating the Income Process

The persistence and volatility of individual income processes are critical determinants of the predicted equity premium in this class of models. Hence model predictions are sensitive to the calibration of these processes. A robust finding of the literature is that more permanent income shocks cannot be self-insured by trading in financial securities, and hence can lead to a high predicted equity premium in heterogeneous agent models. This is implicit in Mankiw (1986), and established very generally in Constantinides and Duffie (1996). In consumption-based asset pricing models, the reason that permanent income shocks imply much higher consumption risk than transitory shocks is that the former represent a much larger wealth shock.

The connection between the persistence of shocks and their wealth effects can be demonstrated most easily using the standard consumption-savings model. Assuming log preferences and complete markets, consumption is a constant fraction of wealth each period, equal to the risk-free rate multiplied by wealth. Hence the volatility of wealth translates into the volatility of consumption. Permanent shocks to income induce large innovations in wealth, and influence consumption one for one. For example, consider a permanent increase in income of \$100 per year, and a risk-free rate of 3 percent. For an infinitely lived agent wealth increases by  $\$100/.03$ . Each year, consumption increases by  $.03(\$100/.03) = \$100$ ; the permanent income shock translates to a permanent consumption shock. By contrast, a transitory shock of the same magnitude adds only \$100 to wealth, and consumption increases by only 3 percent of this, or \$3 per year.

For this reason, it seems preferable to focus on measuring and realistically calibrating wealth shocks, rather than trying to resolve the empirically complicated question of whether income

shocks are predominantly permanent or transitory. A further reason to emphasize the size of wealth shocks, rather than the permanent-transitory distinction, is that with finitely lived individuals the difference between the two is not well-defined. In fact, agents with very short horizons (e.g., just another period or two to live) effectively have permanent shocks regardless of the statistical properties of the income process.

Aside from the unconditional volatility of wealth shocks, the implied premium is affected by the conditional volatility in individual shocks, as emphasized by Constantinides and Duffie (1996). Income specifications exhibiting higher individual conditional volatility in the low aggregate state imply a higher equity premium than processes with the same unconditional volatility but no state dependency. A related observation is that income specifications with the possibility of a catastrophic state also are more successful in reproducing the historical equity premium (Freeman (2004), Rietz (1988)). Mechanically, the reason that higher conditional volatility or a low probability catastrophic state increase the implied equity premium in this class of models is that they allow a coincidence between very bad individual and aggregate shocks, without causing a counterfactually high correlation between individual income growth and stock returns. That is, if the distribution of individual shocks simply shifted down in recessions and up in recoveries by a constant amount, it would induce counterfactually high correlation between labor income and dividends. Increasing dispersion in the low aggregate state increases the predicted equity premium because it creates a very low income realization that is correlated with the low dividend state. In the case of symmetrically distributed conditional volatility, the higher dispersion also creates a very high individual income realization in the low aggregate state. Because of the non-linearity of marginal utility, however, the high individual income realizations have much less influence on the predicted premium than the low realizations.

There are several practical considerations in calibrating the income process for heterogeneous agent models. Model solutions are complicated by the much higher dimensional state space needed to track the cross-sectional wealth distribution. To make the problem numerically tractable, it is imperative to keep the exogenous state space small enough to be manageable. A first order autoregressive representation of the income growth process is conveniently parsimonious, and fits the data reasonably well. It allows a very high degree of persistence, although it precludes permanent shocks. For example, in Heaton and Lucas (1996) we estimate the aggregate process as:

$$X_t^a = \begin{bmatrix} .1487 & .0557 \\ -.5016 & .9168 \end{bmatrix} X_{t-1}^a + \begin{bmatrix} .0278 & 0 \\ .0121 & .0536 \end{bmatrix} \varepsilon_t^a + \begin{bmatrix} .1961 \\ -.2607 \end{bmatrix} \quad (8)$$

where  $\gamma_t^a = Y_t^a / Y_{t-1}^a$  is aggregate income growth,  $\delta_t = D_t^a / Y_t^a$  is the dividend share, and  $X_t^a = [\log(\gamma_t^a), \log(\delta_t)]$ .

In the calibrations reported below, several specifications are considered for the individual income dynamics. In the base case, the individual income dynamics, based on data from the PSID, are also represented by an AR(1) process:

$$\log(\eta_t^i) = -3.35 + .529 \log(\eta_{t-1}^i) + \varepsilon_t^i \quad (9)$$

where  $\eta_t^i = Y_t^i / Y_{t-1}^i$ , and the coefficients are the sample means from individual household estimates.  $\sigma(\varepsilon_t^i)$  is the sample mean of the household standard deviation of  $\varepsilon_t^i$ . To calibrate the model, equations (8) and (9) were discretized and represented as a first-order Markov chain with eight possible states. Recent work by Guvenen (2005) that revisits the decomposition of income risk between permanent and transitory shocks appears to support this specification.

Using a variant of (9) that includes aggregate income growth on the right-hand side, we find little evidence of countercyclical volatility in individual income growth. Nevertheless, in order to test the conjecture that the countercyclical component exists but is not reflected in the relatively short time series used in these estimates, we also calibrated the heterogeneous agent model with twice the conditional variance in the low aggregate growth state as in the high aggregate growth state, preserving the unconditional variance. That calibration is consistent with the imputed conditional volatility estimates of Storlesletten, Telmer and Yaron (2004). The calibration results suggest that conditional volatility of this magnitude is not sufficient in itself to generate a significant equity premium.

In section 3 below, we modify these income processes to represent an economy where entrepreneurs are the dominant participants in financial markets. The purpose is to see whether the greater income volatility and its higher correlation with the market imply a significantly larger equity premium than in previous calibrations based on labor income.

## 2.2 Adding Trading Frictions

Since the risk associated with measured labor income processes appears insufficient in itself to generate a significant equity premium, we turn to the question of whether individual risk becomes more important in the presence of trading frictions. Trading frictions that can be easily analyzed in this class of models fall into two broad categories: transactions costs associated with trading financial securities, and borrowing and short sales constraints. Both types of frictions influence equilibrium returns, through several distinct channels that are discussed in this subsection. While predicted returns are affected in the expected direction, frictions must be severe to match the historical equity premium.

### 2.2.1 Transactions Costs

Trading costs in the form of quadratic or pseudo-proportional transactions costs<sup>3</sup> have the potential to increase the predicted equity premium through both a direct effect of changing relative prices, and an indirect effect of reducing risk sharing. Borrowing costs paid exclusively by the borrower increase the predicted equity premium through a direct effect, which is that for a given level of borrowing, higher trading costs imply a lower market-clearing interest rate. The indirect effect is that the reduction in trading volume due to trading costs increases equilibrium consumption volatility, which in principle can also increase the predicted equity premium. In Heaton and Lucas (1996) we decompose these two effects, and conclude that the direct effect is likely to be quantitatively more important than the indirect effect. The relatively small indirect effect is related to the finding that trade shifts to the market with lower costs. This allows most labor income risk to be shared, unless trading costs are assumed to be large in all asset markets.

The direct effect of trading costs depends critically on which counterparty is assumed to pay them. To maintain a given level of borrowing demand, a one percent increase in trading costs requires a one percent lower interest rate. Similarly, to maintain a given supply of loans requires an increase in interest rates that offsets any increase in trading costs. When transaction costs are evenly split between borrowers and lenders, these effects are largely offsetting and the net effect on the equilibrium interest rate is small. Instead, the equilibrium is restored by a reduction in the amount transacted. If, however, borrowers directly pay trading costs but lenders do not, the equilibrium interest rate will be lower than if the transactions costs are split evenly across the two counterparties. The rate adjustment makes up for the wedge between the all-in cost of borrowing and lending, and restores equilibrium (in this case there is both a rate and quantity effect). Since

---

<sup>3</sup> Pseudo-proportional costs are defined as quadratic in a region close to zero, and linear outside that region. This structure, as well as quadratic costs, have the advantage of differentiability, and so are easier to work with than strictly proportional or fixed costs, which involve discontinuities.

the risk-free rate is taken to be a lending rate from the perspective of households (generally the Treasury bill rate) and obtaining consumer credit involves various fees, such a cost asymmetry is plausible. Similar logic applies to the stock market – only asymmetric costs have a significant direct effect on equilibrium stock returns. These observations are consistent with the finding from calibration exercises that the only cost configuration that yields a sizable equity premium is with all trading costs paid by borrowers in the bond market, but trading costs split evenly between buyers and sellers in the stock market.

The indirect effect reflects that for a given interest rate, a borrower will take out a smaller loan the higher are trading costs. Similarly, for a given interest rate a lender will extend less credit if he has to bear a trading cost. Consequently, trading costs reduce the total volume of trade. Reduced trading volume implies reduced risk sharing, which can decrease the tolerance for investment risk and increase the equity premium. With reasonable trading costs, however, our calibrations suggest that the indirect effect is likely to be small, accounting for less than one percent of the equity premium. This is because with CRRA utility, when effective risk aversion is assumed to be high enough to generate a non-negligible equity premium, agents are willing to bear fairly large transactions costs to smooth consumption. The resulting consumption process, although less smooth than in the same model without trading frictions, is still very similar to that in the representative agent case.

Fixed costs also can contribute to the indirect effect of a higher equity premium due to reduced risk sharing. Explicitly incorporating fixed costs is difficult because they create non-convexities that preclude the solution methods generally used to solve these models. One reason that fixed costs can affect returns is that they discourage small investors from participating in the stock market, which concentrates market risk on a subset of the population. The non-participation channel is explored in Section 4.

The model described by equations (1) - (5), when augmented with quadratic or pseudo-proportional trading costs and calibrated with the income process described in section 2.1, generates predictions consistent with the above descriptions. For example, in Heaton and Lucas (1996) we consider trading costs in the stock market,  $\kappa(s_{t+1,i}, s_{t,i}; Z_t)$ , where  $\kappa$  is assumed to be differentiable in its two arguments, and  $Z_t$  is the state. Both sellers and buyers pay a symmetric cost. Trading costs in the market for one-period bonds are denoted by  $\omega(b_{t+1,i}; Z_t)$ , where  $\omega$  is also differentiable. In some specifications costs are symmetric between borrowers and lenders, and in others only the borrower pays the cost. For unconstrained agents, the first order conditions (3) and (4) are replaced by:

$$(c_{t,i})^{-\gamma} \left( p_t^s + \kappa_1(s_{t+1,i}, s_{t,i}; Z_t) \right) = \beta E_t \left\{ (c_{t+1,i})^{-\gamma} \left( p_{t+1}^s + d_{t+1} - \kappa_2(s_{t+2,i}, s_{t+1,i}; Z_{t+1}) \right) \right\} \quad (10)$$

and

$$(c_{t,i})^{-\gamma} \left( p_t^b + \omega_1(b_{t+1,i}; Z_t) \right) = \beta E_t \left\{ (c_{t+1,i})^{-\gamma} \right\} \quad (11)$$

The market clearing conditions and budget constraints are also adjusted to include trading costs. As discussed above, the results are sensitive to the size and incidence of trading costs, and the severity of the borrowing constraint.

Figure 1, (Figure 1 from Heaton and Lucas (1996)) illustrates the main result, that the direct effects are much larger than the indirect effect, for the case of quadratic costs and one-sided borrowing costs. In the figure, the parameter  $\omega$  determines the level of transactions costs in the stock market. Borrowing costs are assumed to also increase with  $\omega$ . The line labeled “Net

Premium” is the portion of the premium attributable to consumption risk, which is a measure of the indirect effect.

(Figure 1 here)

Very large transactions costs are needed to generate a significant equity premium. To see this, figure 2 (Figure 2 from Heaton and Lucas (1996)) presents the average stock and bond market transactions costs as functions of  $\omega$ . At the high end of the range for the equity premium, the average trading costs in the stock market are 5 percent, and the marginal trading cost is 10 percent. These types of results lead us to the conclusion that moderate trading costs and realistic labor income risk are not sufficient to resolve the equity premium puzzle.

(Figure 2 here)

### 2.2.2 Borrowing and Short Sales Constraints

While proportional or quadratic transactions costs alter the volume and terms of trade, they do not change asset market participation rates; everyone still participates in securities markets, but trading volume falls. If such costs were the only frictions, the aggregate consumption growth process would still determine asset prices according to equations (10) and (11). In contrast, with intermittently binding borrowing and short sales constraints, market participants’ identity can change over time. It is these constraints that preclude any simple aggregation over individual consumption processes.

Like the indirect effect of transactions costs, borrowing and short sales constraints can increase consumption volatility and the predicted equity premium by reducing opportunities for risk

sharing. For constrained agents, consumption equals income. For unconstrained agents, there are fewer trading partners.

Borrowing and short sales constraints can be imposed exogenously, or determined endogenously. The earlier literature on trading frictions imposed exogenous constraints, set to levels motivated by observed borrowing and stock trading behavior. For instance, it is commonly assumed that short sales of stock are prohibitively costly, and that borrowing is limited to a modest fraction (e.g., 10 to 20 percent) of annual average labor income. Two main results emerge: (1) borrowing constraints can dramatically reduce the risk-free rate; and (2) if income shocks are persistent but not permanent, only very severe borrowing and short sales constraints stop consumption smoothing via asset markets. The first result, as explained earlier, is due to the fact that the constrained agents would be willing to pay a high interest rate to borrow, but the unconstrained agents whose preferences determine asset prices want to save. The dominance of savers in pricing drives down the risk-free rate. The second result is less general. It arises because the total value of financial wealth is high relative to the estimated size of uninsurable shocks, and the shocks are not too persistent. This allows a large portion of individual income shocks to be buffered by drawing down financial assets, even if borrowing is restricted to only a small fraction of average annual income.

Recently there has been a growing interest in endogenous borrowing constraints, where state contingent limits on debt are derived from assumptions about the enforcement mechanism for debt repayment. An important question is whether endogenous constraints tend to be more restrictive, or to generate more correlated risk, and hence whether they might explain a higher equity premium. The results of these exercises have been mixed. Zhang (1997) assumes that the punishment for default is a permanent forfeiture of the right to participate in asset markets. Due to the severity of the punishment, the resulting borrowing constraint is effectively more lenient

than assumed in most analyses with exogenously imposed constraints. Thus Zhang does not find that these endogenous constraints resolve the equity premium puzzle. Lustig (2003) proposes a much more stringent rule for accessing capital markets. He assumes that agents can only borrow to the extent that they have enough financial collateral to ensure the promised debt repayment is made in full. This has the effect of precluding any borrowing by non-stockholders. In the model the identity of non-stockholders, and hence those facing the constraint, changes over time. This induces another source of aggregate risk in the model.

### 3. Entrepreneurial Income as Background Risk

The analyses discussed in the last section strongly suggest that although labor income is an important source of largely undiversifiable risk for most households, it is unlikely to explain the equity premium puzzle in long horizon models, even in the presence of trading frictions. Labor income shocks have a large transitory component that can be self-insured by saving. There is low correlation with stock market returns, at least at annual or shorter frequencies. Also problematic is that labor income, as well as other income sources such as unemployment insurance, welfare, and intra-family transfers, tend to put a floor on income that serves as a safe bond, and thereby increases the propensity to take risk in the stock market. These observations motivate a search for other sources of background risk that might have more promising statistical properties.

Entrepreneurial income from closely held businesses has properties that suggest it could be an important source of background risk that is more correlated with stock returns than is labor income. We present evidence for this in Heaton and Lucas (2000c), and augment and update it in Curcuru et. al. (2004). The main considerations include: (1) its higher correlation with stock returns; (2) its higher volatility; (3) the much higher average financial wealth of business owners

than wage earners; and (4) the large portion of stock market value held by households that also have entrepreneurial income.

Using aggregate data from the National Income and Product Accounts, and stock returns from CRSP for the period from 1947 to 2003, we find a correlation between aggregate wage income growth and stock returns of only .06, but a higher correlation between aggregate business income growth and stock returns of .11. The volatility of aggregate business income growth is also more than twice as high as for wage income growth; 4.52 percent vs. 2.06 percent annually. Using the Panel of Individual Tax Return Data for the 1979 to 1990 period, we estimate that individual non-farm proprietary income has a volatility of 64 percent annually, almost twice as high as the growth rate of real wage income of 35 percent annually.

Entrepreneurs not only face riskier income streams, but they are more likely to influence asset market returns due to their much higher wealth and higher rates of stock market participation. Table 2 reports estimates from the 2001 Survey of Consumer Finances on the portfolio characteristics of business owners versus non-owners. Business owners have an average net worth that is four times higher than non-business owners. The typical share of assets invested in their own businesses is 32.5 percent, suggesting significant undiversified risk exposure. Business owners also hold substantial wealth in liquid assets including stocks. We estimate that households with businesses worth in excess of \$10,000 account for about 1/3 of total stockholding.

(Table 2 here)

In Heaton and Lucas (2000c), we use this evidence to motivate the estimation of a linear asset pricing model with proprietary income as a risk factor. We find that a model that includes

aggregate proprietary income, a value weighted stock index, and a credit spread outperforms a similar model with labor income in place of proprietary income. When the Fama French pricing factors HML and SMB are also included, proprietary income remains significant, while HML becomes insignificant. These results suggest that entrepreneurial income is a risk factor that influences asset prices. The analysis, however, does not directly address the equity premium puzzle.

It has not been established, however, whether explicitly modeling the income process of entrepreneurs in a heterogeneous agent model such as the one described in section 2.2 can generate a larger predicted premium than one calibrated with labor income. The only work we are aware of that addresses this question is Polkovnichenko (1999), who considers a model in which entrepreneurs and wage earners trade in financial securities to smooth consumption. He finds that the model can explain the larger holdings of stocks by entrepreneurs due to a higher precautionary demand for savings resulting from their more volatile income. The equilibrium risk premium, however, is similar to that in the representative agent case. The small price effect, as in the models discussed above, is because there is sufficient trading in financial securities to buffer individual stocks. His analysis, however, abstracts both from trading costs and idiosyncratic risk.

To further explore whether entrepreneurial income risk might resolve the equity premium puzzle, we recalibrate the model described in Section 2 above. The calibration is designed to give entrepreneurial income risk its best chance to succeed in this class of models. For tractability we are limited to two agents, both of whom are taken to be entrepreneurs. Restricting attention to the portion of income going to entrepreneurs can be interpreted as reflecting their much higher asset market participation rate, and large share of aggregate financial wealth. We assume an aggregate income process split equally between the dividends from traded stocks and from aggregate

entrepreneurial income. Consistent with the data, aggregate entrepreneurial income has twice the volatility of aggregate labor income, idiosyncratic entrepreneurial income shocks have twice the volatility of idiosyncratic labor income shocks. To maintain high correlation between entrepreneurial income and dividend income, we continue to assume that there is little variation in the share of dividend income in aggregate income. Since proprietary income is the only other source of income in the model, this assumption has the effect of increasing dividend volatility. Lettau, Ludvigson and Wachter (2005) present evidence in support of this specification of dividend volatility.

We do not have direct evidence on the persistence of entrepreneurial income, but consider two cases for the AR(1) coefficient. In the first, as estimated for labor income, the coefficient is 0.53. We call this the “low persistence” case. In the second, reflecting the possibility of more persistent shocks, the autoregressive parameter is increased to 0.8. We call this the “high persistence” case. We also consider two cases for the conditional volatility of entrepreneurial income. In the first, called the “constant shock distribution case”, the volatility of idiosyncratic shocks is independent of the aggregate state. In the second, called the “cyclical shock distribution case,” the conditional volatility of entrepreneurial income is 25% higher in the low aggregate growth state than in the high aggregate growth state.

Table 3 reports the predictions of the model for stock returns, bond returns and the equity premium. We continue to assume that  $\beta = 0.95$  and  $\gamma = 1.5$ . As in the results calibrated to labor income, the predicted equity premium is not large. Even with the large shocks of these examples, the model predicts that agents can readily trade to self insure.

#### 4. Limited Participation and Limited Diversification

Limited stock market participation can increase the required equity premium by concentrating stock market risk on a subset of the population (e.g., Basak and Cuocco (1998), Constantinides, Donaldson and Mehra (1998), Saito (1995), Polkovnichenko (1998) and Vissing-Jorgensen(2002)). The poor diversification of many individual portfolios is also a source of incomplete risk sharing. Both have the theoretical potential to explain the equity premium puzzle. As with other explanations based on market incompleteness, however, the unresolved question is whether these effects are large enough to make a quantitatively significant difference. In Heaton and Lucas (2000b), we consider both of these mechanisms in an overlapping generations (OG) model. We conclude that the increases in participation of the magnitude witnessed in the past two decades are unlikely to cause a significant reduction in the predicted equity premium going forward, but that improved portfolio diversification might explain a fall in the equity premium of several percentage points.

To understand why limited participation may have little quantitative significance for the equity premium, it is useful to review basic facts about the distribution of wealth, and its dynamics over time. Calculations from the SCF suggest that despite increases in stock market participation in the last 15 years, wealth and stock holdings in the U.S. remain highly concentrated in dollar terms. For example, in 1989 the top 10 percent of the wealth distribution held 84 percent of the stock. This dropped slightly to 83 percent in 1995 and further to 76.6 percent in 2001. Although these high concentrations might suggest the potential for greater risk-sharing, the low wealth levels of households not already holding stock means that risk-sharing could be little improved by low-wealth households taking pure long positions in the market. Presumably risk could be better shared with levered or derivative positions, but such behavior is not observed, presumably

because of trading costs or market imperfections. Gomes and Michaelides (2005) also reach the conclusion that participation rates do not explain the equity premium, in a richer calibrated life cycle model with a small fixed cost for stock market participation, and an endogenous participation decision.

Although raw changes in participation are unlikely to explain changes in expected returns, the way in which the wealthy participate, beyond the distinction between entrepreneurs and non-entrepreneurs, could be important. Tallies of stock market participation rates do not distinguish between diversified and undiversified investment strategies. In Heaton and Lucas (2000c) we show theoretically that changes in the level of diversification of the wealthy (substitution from individual stocks to diversified pension and mutual funds) is an important change in portfolio choice behavior that might explain some reduction in expected stock returns. The technical mechanism is that less diversified portfolios have fatter tails and in particular: higher volatility, a non-negligible probability of a catastrophic outcome, and greater skewness. The question remains, however, if greater diversification is valuable enough to justify a high equity premium, why greater diversification is not achieved.

### Conclusion

We conclude that there is no simple resolution of the equity premium puzzle based on heterogeneity, undiversified risk, and trading frictions in models that are otherwise structurally similar to that in Mehra and Prescott (1986). Nevertheless, these factors influence economic behavior, and are likely to play a significant role in explaining asset prices. Perhaps most importantly, the continuing concentration of asset market participation by wealthy and older households suggests that better information on the preferences, incentives and constraints for this group could improve our understanding of the equity premium, and asset prices more generally.

Table 1: Model predictions of asset returns and consumption growth with idiosyncratic shocks and no trading in stocks or bonds			
$\gamma$	1.0	1.5	2.5
$E(r^s)$	.071	.076	.098
$\sigma(r^s)$	.035	.055	.100
$E(r^b)$	.034	.020	-.002
$\sigma(r^b)$	.029	.015	.046
$E(r^s-r^b)$	.037	.056	.099
$\sigma(r^s-r^b)$	.007	.080	.081
$E(C'/C)$	.021	.019	.021
$\sigma(C'/C)$	.080	.080	.081

Table 2: Mean Portfolio Characteristics of Business Owners vs. Non-Owners		
	<u>Owners</u>	<u>Non-Owners</u>
Liquid Fin. Assets / Total Assets	24.9	37.9
Stocks / Liquid Fin. Assets	55.8	47.8
Bonds / Liquid Fin. Assets	18.2	20.3
Cash / Liquid Fin. Assets	26.1	31.9
Owner-Occupied Housing / Total Assets	34.3	54.2
Other Real Estate / Total Assets	6.8	5.7
Business / Total Assets	32.5	-
Age	49.1	52.0
Education (Years)	14.4	13.5
Income	\$ 169,693	\$ 69,533
Net Worth	\$ 1,298,065	\$ 323,255

Tabulations are from the 2001 SCF, and based on survey weights. Source: Curcuru et. al.

Table 3: Expected Stock Return, Bond Return and Equity Premium with Risks Calibrated Using Entrepreneurial Income

<i>A. Constant Shock Distribution</i>		
Persistence of Idiosyncratic Shocks		
	<u>Low</u>	<u>High</u>
Expected Stock Return	8.0%	8.0%
Expected Bond Return	7.9%	7.9%
Equity Premium	0.1%	0.1%
<i>B. Cyclical Shock Distribution</i>		
Persistence of Idiosyncratic Shocks		
	<u>Low</u>	<u>High</u>
Expected Stock Return	8.1%	8.1%
Expected Bond Return	7.9%	7.9%
Equity Premium	0.2%	0.2%

Figure 1: Effect of Transactions Costs on Returns

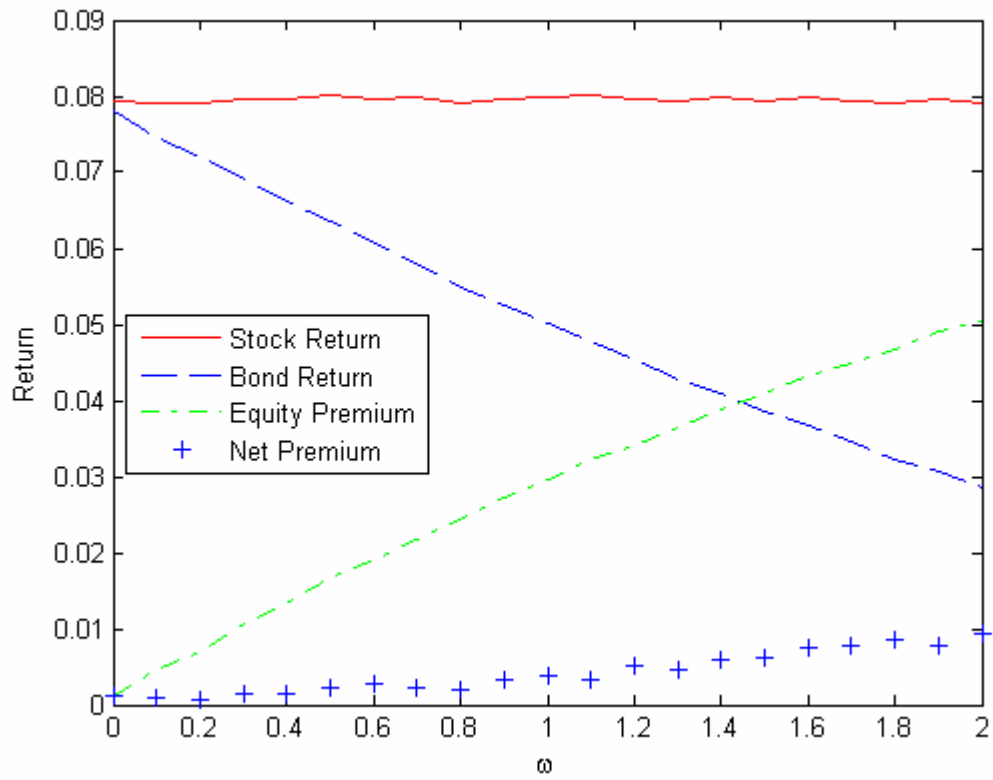
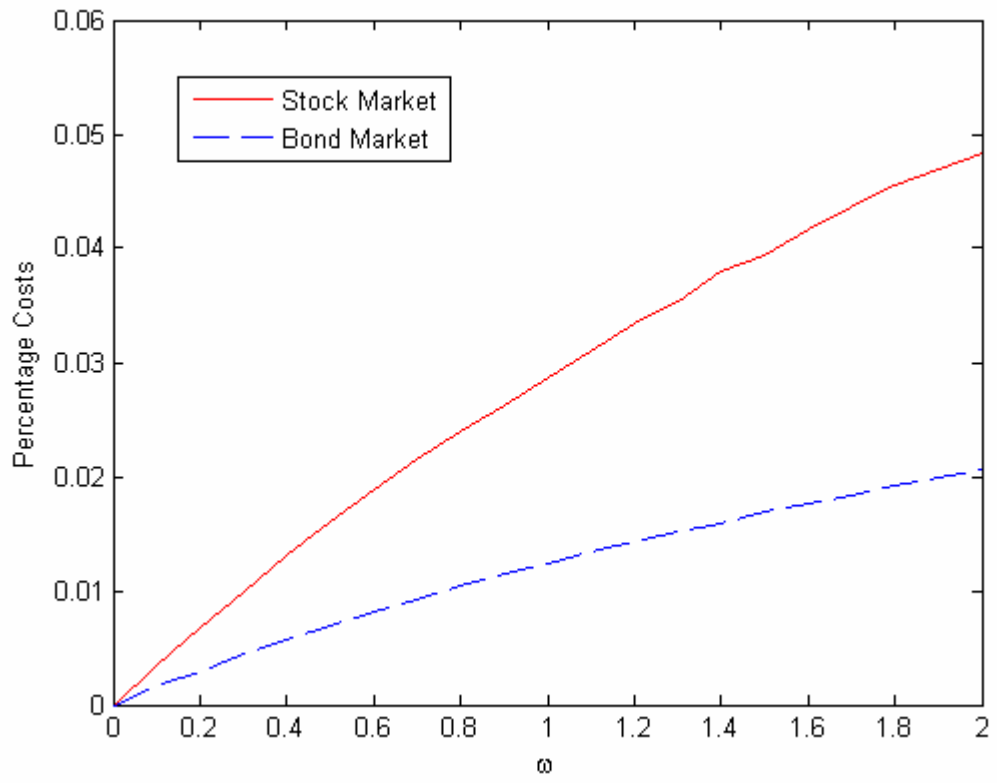


Figure 2: Average Transactions Costs



## References

- Constantinides, George and Darrell Duffie (1996), "Asset Pricing with Heterogeneous Consumers," *Journal of Political Economy*, Vol. 104, pp. 219-40.
- Constantinides, George M., John B. Donaldson and Rajnish Mehra (2002), "Junior Can't Borrow: A New Perspective on the Equity Premium Puzzle." *Quarterly Journal of Economics*, 117 pp 269-96.
- Curcuru, Stephanie, Heaton, J., Lucas, D., Moore, D. (2004), "Heterogeneity and Portfolio Choice: Theory and Evidence," Handbook of Financial Econometrics, forthcoming
- Freeman, Mark C. (2004), "Can Market Incompleteness Resolve Asset Pricing Puzzles?" *Journal of Business Finance and Accounting*, 31(7)&(8), pp 928-949.
- Gomes, Francisco and Alexander Michaelides (2005), "Asset Pricing with Limited Risk Sharing and Heterogeneous Agents," manuscript, London School of Economics
- Guvenen, Fatih (2005), "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?" manuscript, University of Rochester
- Heaton, John and Deborah Lucas (1992), "The Effects of Incomplete Insurance Markets and Trading Costs in a Consumption-Based Asset Pricing Model," *Journal of Economic Dynamics and Control*.
- \_\_\_\_\_ (1995), "The Importance of Investor Heterogeneity and Financial Market Imperfections for the Behavior of Asset Prices," *Carnegie Rochester Papers*
- \_\_\_\_\_ (1996), "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing," *Journal of Political Economy*
- \_\_\_\_\_. (2000a). "Portfolio Choice in the Presence of Background Risk," *Economic Journal*
- \_\_\_\_\_ (2000b), "Stock Prices and Fundamentals," *NBER Macroeconomics Annual*
- \_\_\_\_\_ (2000c) "Asset Pricing and Portfolio Choice: The Role of Entrepreneurial Risk" (with John Heaton), *Journal of Finance*
- \_\_\_\_\_ (2004), "Capital Structure, Hurdle Rates and Portfolio Choice – Interactions in an Entrepreneurial Firm," working paper, Northwestern University.
- Lettau, M., S. Ludvigson and J. Wachter (2005), "The Declining Equity Premium: What Role Does Macroeconomic Risk Play?" manuscript, New York University.
- Lucas, D. (1994), "Asset Pricing with Undiversifiable Income Risk and Short Sales Constraints: Deepening the Equity Premium Puzzle," *Journal of Monetary Economics*.
- Lustig, H. (2004), "The Market Price of Aggregate Risk and the Wealth Distribution," manuscript, UCLA.

Mankiw, N. Gregory (1986), "The Equity Premium and the Concentration of Aggregate Shocks," *Journal of Financial Economics* 17, pp 211-219.

Mankiw, N. G. and S. Zeldes (1991), "The Consumption of Stockholders and Non-Stockholders," *Journal of Financial Economics*, vol. 29, pp. 97-112

Mehra, Rajnish and Edward Prescott (1985), "The Equity Premium – A Puzzle," *Journal of Monetary Economics* 15, 145-161.

Polkovnichenko, Valerie (1999), "Heterogeneity and Proprietary Income Risk: Implications for Stock Market Participation and Asset Prices," manuscript, University of Minnesota

Reitz, T.A. (1988), "The Equity Premium: A Solution," *Journal of Monetary Economics*, Vol. 22, pp. 117-31.

Saito, Makoto (1995), "Limited Market Participation and Asset Pricing," manuscript, University of British Columbia

Storesletten, K. C. Telmer and A. Yaron (2001), "Asset Pricing with Idiosyncratic Risk and Overlapping Generations," manuscript, Carnegie Mellon University

\_\_\_\_\_ (2004), "Cyclical Dynamics in Idiosyncratic Labor-Market Risk," *Journal of Political Economy*,

Telmer, Chris I. (1993), "Asset Pricing Puzzles and Incomplete Markets," *Journal of Finance*, 48, 1803-1832.

Vising Jorgensen, Annette (2002), "Limited Asset Market Participation and the Elasticity of Intertemporal Substitution," *Journal of Political Economy* 110, 835-53.

Zhang, Harold (1997), "Endogenous Borrowing Constraints with Incomplete Markets," *The Journal of Finance*, Vol. 52, No. 5. (Dec., 1997), pp. 2187-2209.