On the Excess Returns to Illiquidity*

Robert Novy-Marx†
University of Chicago

This Draft: April 19, 2005

Abstract

This paper argues that the high expected returns observed on illiquid assets should be expected theoretically, but are not actually a premium for illiquidity, *per se*. Instead, illiquidity, like size, is a proxy for *any* unobserved risk. Liquidity should therefore have explanatory power in any asset pricing model that is not perfectly specified, with low measured liquidity forecasting high expected returns. The magnitude of the expected premium can be similar to that associated with the omitted risk factor.

Keywords: Liquidity, Asset Pricing, Factor Models, Omitted Variable Bias.

JEL Classification: G12.

*I would like to thank Jonathan Berk, George Constantinides, Greg Duffee, Madhur Duggar, John Heaton, Tom Knox, Milena Novy-Marx, Ľuboš Pástor, Monika Piazzesi, and an anonymous referee, for discussions and comments. All errors are mine alone.

†University of Chicago Graduate School of Business, 5807 S Woodlawn Ave, Chicago, IL 60637. Email: rnm@gsb.uchicago.edu.

1 Introduction

A large empirical literature now documents the impact of illiquidity on returns. Illiquid assets provide, on average, significantly higher returns than liquid assets, *ceteris paribus*. That is, literature documents the existence of a significant "illiquidity premium."[1] This holds true both in studies that consider the levels of assets' liquidities, such as Amihud and Mendelson (1986) or Brennan and Subrahmanyam (1996), and in studies that consider assets' exposures to changes in market liquidity, such as Pástor and Stambaugh (2003).

At the same time, the bulk of the theoretical work in this area predicts that any illiquidity premium should be quite small.[2] Constantinides (1986) and Vayanos (1998) both argue that proportional transaction costs should significantly impact trading frequency but have only a minimal effect on prices. Heaton and Lucas (1996) find that transaction costs do not generate significant premia in an economy in which agents trade to share labor-income risk. Lo, Mamaysky and Wang (2001) generate moderate price discounts due to illiquidity, but the resulting return premium is still quite low. Huang (2003) finds that illiquidity premia should be inconsequential, barring other significant constraints.

These theoretical predictions for a low illiquidity premium are unsurprising once one understands the intuition driving the results. The low illiquidity premium in all of these models is driven by one or both of just two factors: endogenous portfolio selection and endogenous timing of trades. The intuition as to why endogenizing portfolio selection reduces the illiquidity premium depends only on the existence of at least some long horizon investors. Short horizon investors will choose only to hold the most liquid assets, leaving those least impacted by transaction costs, the long horizon investors, to hold the less liquid assets, a phenomenon Amihud and Mendelson (1986) term "clientele effects." Long horizons investors are able to amortize any "tollgate" charges. Even a transaction cost of five percent represents only a sixteen basis point drag on annual returns to an investor with a thirty-year horizon. If long horizon investors only need compensation, in the form of higher returns, sufficient to make them indifferent between holding liquid and illiquid

[1] The literature somewhat confusingly uses the terms "liquidity premium" and "illiquidity premium" interchangeably. Strictly speaking, the liquidity premium refers to the fact that liquid assets command a price premium, trading higher than otherwise comparable illiquid assets, whereas the illiquidity premium refers to the fact that illiquid assets provide a return premium, delivering higher expected yields than comparable liquid assets. We prefer illiquidity premium because of the analogous usage of the more standard "risk premium." That is, we will use illiquidity premium to refer to the excess expected returns generated by illiquid assets in the same way that risk premium refers to the extra returns generated by risky assets.

[2] Amihud and Mendelson (1986) provides a notable exception. Their model, however, explicitly forbids long horizon investors from undertaking liquidity "arbitrage." Allowing for competition across investor type would reduce the magnitude of the model's predicted illiquidity premium.

assets, *ceteris paribus*, then the return premium provided by illiquid assets must be quite low. The intuition as to why endogenizing trade timing reduces the illiquidity premium is even more straightforward. Illiquid assets are more expensive to trade but this, *ipso facto*, leads investors to choose to trade illiquid assets less frequently. Trading these assets less frequently mitigates the costs imposed by illiquidity. Trading these assets optimally dramatically reduces the premium required to make an investor indifferent between holding these assets and comparable liquid assets.

These theoretical predictions notwithstanding, the existence of a significant illiquidity premium is now fairly well established. Pástor and Stambaugh (2003) suggest one plausible explanation. They propose that some assets, illiquid in a dimension they term "sensitivity to liquidity risk," tend to have low returns when aggregate market liquidity dries up, which are also times when investors value returns highly. If a security's lowest returns concur with unfavorable shifts in investor welfare it should trade at a discount. If illiquid assets' returns are negatively correlated with state prices then investors will demand high compensatory average returns. Acharya and Pedersen (2003) have considered this type of correlation theoretically. They explicitly consider the correlation between trading costs, which impact net returns, and market returns, which proxy for state prices, in an asset pricing framework. Within this framework they derive a significant illiquidity premium.Their model assumes, however, that each investor liquidates their entire portfolio every period. Incorporating investors with longer horizons into the model, or endogenizing trading, would again mitigate or even abrogate the results, leaving us still without a theoretical reason to expect an illiquidity premium.[3]

Given that investors should not, on theoretical grounds, demand significant compensation to hold illiquid assets, how can we make sense of the overwhelming empirical evidence that illiquid assets provide significantly higher returns? The central issue here is causality. While there is no reason to believe that investors should demand high returns for holding illiquid assets, there is every reason to believe that illiquid assets should yield high returns. That is, while there is not a causal relationship between illiquidity and high excess returns, the coincidence of the two is quite expected.

A related argument is found in Berk's (1995) "A Critique of Size-Related Anomalies." Berk (1995) argues that small size is not a *cause* of high expected returns, but rather that assets exposed to greater risks trade at deeper discounts (*i.e.*, are smaller, by market capitalization) and, *ipso facto*, generate higher expected returns. That is, while small size is not

[3]Lustig (2003) and Vayanos (2004) also considers the illiquidity premium generated by covariance between returns and state prices. In these equilibrium models the covariance arises as a result of binding solvency constraints (Lustig) or asset sales forced by mutual fund withdrawals (Vayanos).

the *proximate* cause of high expected returns, the coincidence is expected because exposure to risk leads to both small size and high expected returns.

That the observed "illiquidity premium" is really an expected coincidence between illiquidity and high expected returns ultimately follows from a similar argument: exposure to underlying risk leads to both high return premia and low liquidity. However, the line of reasoning is necessarily somewhat less direct than the argument pertaining to size. Following this less direct argument is helped by conceptualizing the market along lines similar to those found in Campbell, Grossman and Wang (1993), the paper that motivates Pástor and Stambaugh's (2003) return reversal measure of illiquidity.

Consider the market as composed of two parts. The first part consists of "liquidity traders," investors that hold and trade stocks for exogenous reasons. To avoid confusion arising from this standard but unrelated use of the term "liquidity," we will refer to these investors henceforth as "noninformational traders," or simply "NTs." The second part of the market consists of risk-averse utility maximizers. These "attentive traders," or "ATs," hold optimal price contingent quantities of each asset, actively making portfolio decisions endogenous to risk-reward trade-offs in the economy. They constitute a risk bearing capacity, and are willing to accommodate the NTs' trading demands if appropriately compensated. This compensation takes the form of high expected returns, which result from the assets trading at a discount.

That is, the prices of securities are effectively set by what the ATs are willing to pay; the NTs are not directly involved in price determination. NTs do influence prices indirectly, however, as liquidity trades affect prices through the supply channel. Net noninformational trading in any given security effectively constitutes a change in the supply of the security, from the point of view of the attentive traders, and this impacts the market clearing price as the ATs have downward sloping asset demand.

While we do not usually think of downward-sloping demand for financial assets explicitly, in part because they are generally in fixed aggregate net supply, the idea is implicit in much of asset pricing theory. For example, the idea of downward-sloping asset demand underpins the Capital Asset Pricing Model. Given quite general investor preferences, the price of an asset is just its expected payoff minus a correction due to higher moments of the payoff distribution. Because the total expected payoff is linear in the supply of the asset, whereas the correction is dependent on higher-order powers of supply, the price of a risk goes to zero as the supply of the risk goes to zero. That is, because idiosyncratic risk is, by definition, at the zero quantity point on the supply curve it is not priced, while market risk is in positive net supply and therefore lower on the demand curve.

Conceptualizing the market as consisting of these two groups, exogenously motivated NTs and risk-averse, utility-maximizing ATs with their consequently downward-sloping asset demand, helps clarify other phenomena observed in the market. For example, analysts' statements regarding price drops due to "selling pressure" can seem nonsensical, given that there are always buyers on the other side, but in this context the statements make perfect sense. Selling pressure is an excess of selling demand. Selling pressure can only refer to the NTs' trading demands, as ATs do not demand trade at all, only acting as counterparties when adequately compensated. It is the order flow of the NTs that consequently determines whether pressure is on the buy side or the sell side.

Price drops at the end of lock-outs, or with secondary offerings, provide another example of phenomena that may be illuminated by analysis in the context of our market concept. These price drops make perfect sense, even in the absence of asymmetric information, if the corresponding events increase the ATs' claims to the underlying asset. In this case they effectively amount to an outward shift in the supply curve, which leads to lower prices. [4]

Before moving on to a discussion of how this view of the market helps us to understand the illiquidity premium, it is first useful to discuss liquidity in general more carefully. Pástor and Stambaugh (2003) capture both the spirit of the concept, and the difficulty of pinning it down precisely, when they state that "liquidity is a broad and elusive concept that generally denotes the ability to trade large quantities quickly, at low cost, and without moving the price." While not precise, this statement captures the idea that it "costs" less to trade liquid assets, where costs here are both direct and indirect. Direct costs include fees and the bid-ask spread. Indirect costs arise from the fact that an investor's current trades compete against her future trading needs: when liquidating a large position, today's trading has a detrimental effect on the price at which tomorrow's trades will be executed. Our concept of liquidity captures both direct and indirect costs in what we believe is the most economically sensible way.

Liquid assets, to us, are those for which the "liquidity cost" of trading is low, where the "liquidity cost" is simply the difference between the "value" at current prices and the actual economic value. More precisely, an asset's liquidity will be defined as the inverse of the unit liquidity cost of trading the asset, where the liquidity cost is the difference between the "mark-to-market value" of holdings and the actual amount one can get for those holdings assuming optimal liquidation.

Returning to our discussion of the illiquidity premium, it is clear that the liquidity cost

[4] Asset prices are essentially impacted here through the "portfolio balance channel." This is the same channel that central banks depend on whenever they "sterilize" their interventions.

of trading an asset, or risk factor, is related to the slope of the demand curve for the factor. Order flow from the NTs results in a change in the supply of the factor (from the point of view of the ATs), resulting in a change in the factor's price. The magnitude of this "nontransitory" change in price corresponds perfectly to common microstructure notions of liquidity relating to "the permanent price impact of a trade."[5] The size of the price impact is directly related to the cost of liquidating a position. The steeper the demand curve, the greater the price impact, and the more expensive it is to trade out of a position. But the slope of the demand curve is due to the uncertainty in the factor's payoff: for quite general preferences the slope of the demand curve for the factor is increasing in the uncertainty of the factor's returns. Uncertainty is, of course, also the source of the factor's excess returns, and these returns are also generally increasing in uncertainty. This coincidence between illiquidity and high excess returns will also be observed in individual assets, to the extent that trading in these individual assets is correlated with trading in the risk factors to which they are exposed.[6]

That is, uncertainty generates the slope of the demand curve for an asset, and this slope, which is the source of illiquidity costs, largely determines the magnitude of the price discount at which the asset trades. This price discount is in turn the source of higher expected returns. The greater the variance, the steeper the demand curve and the greater the price discount, and thus the lower the liquidity and the greater the expected excess returns. In other words, the discount at which an asset trades and the slope of the demand curve for the asset, *i.e.*, expected excess return and illiquidity, are both expressions of the same underlying uncertainty.[7]

Liquidity measures' ability to explain excess returns should remain after controlling for other risks. In fact, reasoning similar to that found in Berk (1995) shows that liquidity

[5]The connection between liquidity and the slope of asset demand is quite clear, even if implicit, in the microstructure literature. In the Kyle (1985) model the illiquidity parameter λ is exactly the slope of the market maker's demand curve. Much of the empirical microstructure literature also uses the notion of the permanent price impact of a trade as a measure of illiquidity (see, for example, Hasbrouck (1988)).

[6]While individual assets are small relative to the market, and thus agents' demand for any given asset should be almost flat *in a partial equilibrium analysis*, if trade in an asset is correlated with trade in the risk factors to which it is exposed we would expect to observe downward sloping demand for the asset, in the sense that trade in the asset is systematically correlated with changes in the asset's price. Empirically, liquidity is not measured using a controlled experiment in which the researcher generates purely asset specific idiosyncratic trades, but by taking the order flow for the asset as given, and this order flow is correlated with trading in assets exposed to similar risks.

[7]The results of the "information risk" models of Easley et al. (2002) and O'Hara (2003) are driven by the same coincidence. In these models some agents have inside (or "private") information regarding an asset's pay-out. This steepens the demand curve of the asset's residual buyer, who does not posses the inside information, because the quantity they must hold is negatively correlated with the asset's pay-out. As a result the asset exposed to this "information risk" yields higher expected returns, but is more expensive to trade.

measures should add explanatory power to any asset pricing model that does not completely explain returns. The slope of the demand for an asset is a measure of the asset's risk, so is positively correlated with any risk factor. Liquidity, which is inversely related to the slope of demand for the asset, is consequently negatively correlated with all risk factors. This fact alone guarantees that illiquidity will proxy, to some degree, for any omitted risk.

Alternatively, the argument may be structured along the lines of Ball (1978), in terms of omitted-variable bias: estimates of the impact of illiquidity on asset returns overstate the true impact as a result of omitted variables. To see this, consider the result of adding illiquidity to any regression of returns onto an incomplete set of risk factors. Adding illiquidity, which is correlated with every priced risk factor and thus outside the span of any incomplete set, increases the dimension of the space onto which returns are projected. The estimated factor betas will be biased as a result of omitted variables. The estimated illiquidity beta will be biased toward overstating illiquidity's true impact because illiquidity is positively correlated with *any* risk factor. Illiquidity will have explanatory power, therefore, in any regression that omits risk factors, even if illiquidity is truly unpriced.

We should also expect to see higher returns associated with assets that become especially illiquid when market liquidity dries up, *i.e.*, assets that are sensitive to liquidity risk. To see that these assets will have high average returns, consider the following thought experiment. Suppose some market factor infrequently has periods of high conditional uncertainty, for example an "emerging market" factor, the returns to which occasionally become highly uncertain. Assets associated with this factor are sensitive to liquidity risk. During periods of high uncertainty the slope of the demand curve for this factor steepens, steepening the demand curve for the market as a whole, but disproportionately steepening the demand curve for assets heavily exposed to the factor. That is, periods of high uncertainty regarding this factor are associated with a market-wide liquidity crunch, but one that especially impacts assets that are heavily exposed to the factor. But these assets also generate high excess returns. They are exposed to a "peso problem" type volatility. Assets exposed to this factor have the possibility of high conditional variance in the future, so trade at a discount today, resulting in high expected returns on these assets even in normal market conditions. These assets also generate higher unconditional expected returns, due to risk-aversion at the market level, than do similar assets with the same unconditional variance but constant volatility.

In light of our argument that illiquid assets must provide higher returns, *ceteris paribus*, than liquid assets, how should we approach the extant literature on the illiquidity pre-

mium?[8] Illiquidity is associated with higher returns, but it is not illiquidity, *per se*, that causes these high returns. Rather, illiquidity proxies for unobserved risk factors. That *ad hoc* regression variables can proxy for unobserved risks is not novel, and care should be taken in general not to assign *causation* to right hand side variables when the source of their explanatory power is not well understood. We would, for example, find a high R^2 if we were to regress left shoe sales on right shoe sales, though of course right shoe sales are not the cause of left shoe sales. Right shoe sales do tell you a great deal, however, about the demand for shoes.

In the same manner, understanding that illiquidity does not drive, but is merely correlated with, high excess returns does not mitigate the importance of these earlier results. Far from diminishing these earlier results' value, this understanding increases the importance of accurately estimating the liquidity characteristics of individual assets. Just as size is an important dimension in portfolio selection, liquidity composition should be a consideration in optimal asset allocation. But whereas size is directly observable, liquidity, and sensitivity to liquidity risk, are not. Accurate estimation of assets' liquidity characteristics can aid portfolio selection, improving investor welfare.

2 Illiquidity and Excess Returns

This section considers more formally the predictive power of liquidity, and sensitivity to liquidity risk, for forecasting returns. The market is composed, in this simple model, of two types of investors. The first type are "liquidity" or "noninformational" traders ("NTs"), who are assumed to trade as a result of exogenous shocks. The second type are risk-averse utility maximizers. These attentive traders ("ATs") hold optimal price-contingent quantities of each asset. This segment of the economy admits a representative investor with constant absolute risk aversion preferences with risk aversion parameter α.[9] The systematic risks in the economy are comprised of a set of n orthogonal risk factors, or "factor portfolios," v_i for $i \in \{1, 2, ..., n\}$, which we will assume are normally distributed. That is, the payoff to the i^{th} factor, which is realized at $t = 2$, is distributed $N(\mu_i, \sigma_i^2)$. At $t = 1$ the trading demands of the NTs, who initially hold a fraction (or "quantity") $1 - q_i^0$ of each factor, are realized. These trading demands, $-dq_i$ for $i \in \{1, 2, ..., n\}$, are assumed to be distributed $N(0, \sigma_{q_i}^2)$. We will assume for simplicity that dq_i is uncorrelated with other random variables in the

[8] The argument in this paper is, of course, that differing liquidities indicate assets are actually *ceteris non paribus*; here *ceteris paribus* is used to mean "other things *apparently* equal."

[9] The CARA utility formulation is particularly tractable because the demand for risky assets is independent of wealth. Additionally, the ATs' aggregate asset demands are linear in prices.

economy. The risk free rate is r. Markets must clear, so prices adjust such that ATs are willing to satisfy the NTs' trading demand. Both types of traders, NTs and ATs, live only two periods and consume only at $t = 2$, when the assets pay off.

2.1 The Returns to Illiquidity

The first order condition for the maximum of attentive traders' expected utility implies that the ATs' demand for each factor is linear in the factor's price, $\frac{\mu_i - (1+r)P_i}{\alpha\sigma_i^2}$. Market clearing then demands that each factor is priced at $P_i = (1+r)^{-1}(\mu_i - q_i\alpha\sigma_i^2)$, where $q_i = q_i^0 + dq_i$ is the quantity of the i^{th} factor held by the ATs at $t = 1$.

In this economy a sort of linear factor pricing model holds, where the expected net payoffs are linear in the factor weights. Any asset θ, which essentially consists of a portfolio of the factors plus idiosyncratic variance, $\theta = w_1 v_1 + w_2 v_2 + \cdots + w_n v_n + \varepsilon_\theta$, has an expected net payoff

$$\mathbf{E}[\theta] - P_\theta \;=\; \mathbf{E}\left[\sum_{i=1}^{n} w_i(v_i - P_i)\right] \;=\; rP_\theta + \sum_{i=1}^{n} w_i q_i \alpha\sigma_i^2, \tag{1}$$

which is linear in the weights w_i for $i \in \{1, 2, ..., n\}$. The expected net excess return to the i^{th} factor is given by

$$\mathbf{E}\left[R_i^e\right] \;=\; \frac{\mu_i - P_i}{P_i} - r \;=\; \frac{q_i\alpha\sigma_i^2}{P_i}. \tag{2}$$

Scaling by the standard deviation in the factor's payoff yields the market price of exposure to factor volatility

$$\xi_i \;=\; \frac{q_i\alpha\sigma_i}{P_k}. \tag{3}$$

Now suppose we were to test a misspecified asset pricing model, which omits the k^{th} risk factor. That is, suppose we run the regression

$$y_i \;=\; a_i + \sum_{\substack{j=1 \\ j\neq k}}^{n} \beta_{ij} x_j + \varepsilon_i, \tag{4}$$

where y_i denotes the return on asset i and x_j denotes the return on the j^{th} factor.

While the k^{th} factor is missing from the regression, liquidity measures will contain information about an asset's exposure to the missing factor. In general, the liquidity of the k^{th} factor is just the inverse of the associated price impact parameter, defined here as

the price impact per dollar traded, which we will denote λ_k. This parameter, essentially the Kyle-λ (1985), measures how expensive it is to liquidate a position. Liquid assets are ones for which the associated impact parameters are low on average, and for which it is consequently cheap to trade out of a position. Now this price impact parameter, given by

$$\lambda_k \equiv \frac{-dP_k}{P_k dq_k} = \frac{\alpha\sigma_k^2}{P_k}, \tag{5}$$

is really just the slope of the ATs' demand. Substituting into equation (2) yields

$$\mathbf{E}\left[R_k^e\right] = q_k\lambda_k. \tag{6}$$

That is, the expected excess return on the k^{th} factor is proportional to λ_k, the slope of the demand for the factor. Alternatively, we have that factor illiquidity is proportional to the market price of exposure to factor risk,

$$\lambda_k = \frac{\sigma_k}{q_k}\xi_k. \tag{7}$$

The expected excess return to any factor, and the market price of exposure to factor volatility, are therefore negatively correlated with the factor's liquidity.

Now suppose that in our regression we have two distinct assets for which we estimate the exact same factor loadings, but for which we have differing liquidity estimates. The less liquid asset *must* have higher expected returns going forward, despite the identical factor load estimates. The less liquid asset is likely to have a higher (respectively, lower) loading on the omitted factor if the omitted factor is "less liquid" (respectively, "more liquid") than the other factors. Because of the assets' probable relative exposures to the omitted risk factor, and the fact that factor liquidity is negatively correlated with expected returns, we should expect the less liquid asset to provide higher returns than the more liquid asset. That is, liquidity has explanatory power for returns. The explanatory power of the illiquidity estimate depends only on some misspecification in the asset pricing model, and not on which factor, or set of factors, was omitted.

2.2 A Note on Intermediation

For the sake of simplicity we have assumed that the NTs trade directly with the representative AT. It is important to note, however, that the price impact of trading, while still proportional to the slope of the ATs' demand, is magnified by intermediation. Order flow

seen by any given market maker is informative regarding probable net market-wide trading demand, and market makers use this information when setting prices.

We can illustrate the role of intermediation by reconsidering our previous example, but assume that the NTs cannot trade directly with the ATs, and instead trade with intermediaries who then trade the orders on to the ATs at market clearing prices. Suppose these intermediaries consist of N competitive risk-neutral market makers, and that net trading in the i^{th} factor handled by the j^{th} market maker is distributed $\tau_i^j = \nu_i + \xi_i^j$, where $\nu_i \sim N(0, \sigma_{\nu_i}^2)$ is a common component of the NTs' trading seen by each market maker and $\xi_i^j \overset{i.i.d.}{\sim} N(0, \sigma_{\xi_i}^2)$ is an idiosyncratic component seen only by the j^{th} market maker.

The *ex-post* realized change in a factor's price, after the market makers pass the NTs' trades on to the ATs, is given by $\frac{1}{P_i}\frac{dP_i}{dq_i}dq_i = \lambda_i dq_i$. This is not, however, the price at which the market makers trade with the NTs. The market makers, whom are competitive and risk neutral, adjust prices so as to trade at the *expected* price conditional on the information they have: the individual order flow they receive. The impact that an order τ_i^j has on the price at which the j^{th} market maker is willing to trade the i^{th} factor is therefore given by

$$\begin{aligned} \frac{\Delta P_i^j}{P_i} &= \lambda_i \mathbf{E}\left[dq_i \,\middle|\, \tau_i^j\right] \\ &= \lambda_i \left(\tau_i^j + (N-1)\mathbf{E}\left[\nu_i \,\middle|\, \tau_i^j\right]\right) \\ &= \left(1 + (N-1)\frac{\sigma_{\nu_i}^2}{\sigma_{\nu_i}^2 + \sigma_{\xi_i}^2}\right)\tau_i^j \lambda_i \end{aligned} \tag{8}$$

That is, the price impact of trading in the intermediated economy is greater, by a factor $1 + (N-1)\sigma_{\nu_i}^2/(\sigma_{\nu_i}^2 + \sigma_{\xi_i}^2)$, because market makers condition on order flow information when setting prices. The price impact of trading is, however, still proportional to λ_i, the slope of the ATs' demand for the factor.

2.3 The Returns to Sensitivity to Liquidity Risk

To study the impact of sensitivity to liquidity risk we will now return to the simple model of the previous section, but relax the assumption that the payoffs of each of the n risk factors are normally distributed. In particular, we will assume that some factor has periods (*i.e.*, some probability) of higher than normal conditional variance. In particular, suppose this factor, the n^{th} without loss of generality, has high variance in the "high uncertainty" state of the world, which is realized with probability p, and low variance otherwise. Assuming the distribution is still conditionally normal, and letting $\mathbb{1}_h$ denote the indicator variable for

the high uncertainty state of the world, we have that v_n is distributed $N(\mu_n, \text{Var}\left[v_n \mid \mathbb{1}_h\right])$ where $\text{Var}\left[v_n \mid \mathbb{1}_h = 1\right] = \sigma_h^2 > \sigma_l^2 = \text{Var}\left[v_n \mid \mathbb{1}_h = 0\right]$. We will assume for simplicity that $\mathbb{1}_h$, which is realized prior to the NTs' trading demands, is uncorrelated with other random variables in the economy.

The factor, which is exposed to what we will call "uncertainty risk," will be sensitive to liquidity risk. When the high uncertainty state of the world is realized this factor becomes less liquid, because the slope of the demand curve steepens. Because this factor is in the market portfolio market liquidity also drops, so the factor liquidity covaries positively with market liquidity. The factor is consequently, by definition, sensitive to liquidity risk. But this factor will also generate higher unconditional returns than predicted by an econometrician who estimates factor moments. The ATs value this factor less than they would a normally distributed factor with the same mean and variance, simply because they are risk-averse. The ATs will only hold the factor, therefore, if it is priced lower than a normally distributed factor with the same expected payoff and volatility. That is, $P_n < P_{\hat{n}}$, where $P_{\hat{n}}$ denotes what the price of the n^{th} factor would have been if the payoff were actually distributed $N(\mu_n, \text{Var}[v_n])$. The fact that it is priced lower but has the same expected payoff means it has higher expected returns. The extent to which an asset's liquidity covaries with the market's and the magnitude of an asset's expected supernormal returns are both correlated with the asset's loading on the factor, so sensitivity to liquidity risk predicts returns.

Conditional on a realization of the good state of the world the return will be even more surprising. At these times the ATs are compensated for an unrealized risk. These exceptionally high returns are earned essentially for insuring against rare events, *i.e.*, for exposure to what is often termed "peso problem"-type volatility.

Conditional on a realization of the uncertain state of the world we will observe the return reversals, predicted by Campbell, Grossman and Wang (1993) and exploited by Pástor and Stambaugh (2003), associated with exposure to liquidity risk. When conditional uncertainty increases, the prices of assets exposed to the factor with high conditional uncertainty will drop. These assets will then, as a consequence, generate high expected returns going forward.

2.4 The General Omitted-Variable Bias Argument

Now suppose, more generally, that assets are correctly priced by some linear factor model with n priced risk factors, where n may be unknown, and illiquidity is not independently

priced. That is, suppose the expected excess return to an asset is

$$\mathbf{E}\left[R^e\right] = \mathbf{X}\boldsymbol{\beta} \tag{9}$$

where $\mathbf{X} = (X_1, ..., X_n)$ is a row vector of the returns to the n risk factors, $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)'$ is a column vector of the asset's exposure to these factors, and illiquidity is not a risk factor.

Now suppose one were to test a misspecified factor model which omits one or more risk factors. In particular, suppose one ran the "short regression" of excess returns onto an incomplete set of $k < n$ risk factors and an illiquidity factor

$$R^e = \mathbf{X}_1\mathbf{b}_1^* + X_{il}b_{il}^* + \varepsilon \tag{10}$$

where $\mathbf{X}_1$ is the vector of returns to the factors under consideration, $\mathbf{b}_1^*$ is the vector of the asset's estimated exposures to these factors, X_{il} is the return to an illiquidity factor, and b_{il}^* is the asset's estimated exposure to the illiquidity factor.[10] In this short regression, as a result of omitted-variable bias, the expected estimated exposures are not the true exposures. That is, the regression yields biased estimates, with $\mathbf{E}[\mathbf{b}_1^*] \neq \boldsymbol{\beta}_1$ and $\mathbf{E}[b_{il}^*] \neq 0$.

More can be said about these biases, and about the bias in the estimate of the illiquidity beta in particular, using the omitted-variable bias formula. Let $\mathbf{X}_2$ be a vector of returns to $n-k$ risk factors that, in conjunction with the short-regression factors, span the risk-space, which may be chosen orthogonal to the factors in $\mathbf{X}_1$ without loss of generality. Including these factors in the regression yields the unbiased "long regression"

$$R^e = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + X_{il}b_{il} + \varepsilon \tag{11}$$

in which the expected estimated sensitivities to the priced risk factors are the true sensitivities, $\mathbf{E}\left[\binom{\mathbf{b}_1}{\mathbf{b}_2}\right] = \binom{\boldsymbol{\beta}_1}{\boldsymbol{\beta}_2}$, and in which $\mathbf{E}[b_{il}] = \beta_{il} = 0$, because illiquidity is truly unpriced. The omitted-variable bias formula then gives the relation between the expected estimated sensitivities in the short regression and the true sensitivities

$$\mathbf{E}\left[\binom{\mathbf{b}_1^*}{b_{il}^*}\right] = \binom{\boldsymbol{\beta}_1}{\beta_{il}} + \mathbf{F}\boldsymbol{\beta}_2 \tag{12}$$

where $\mathbf{F} = [(\mathbf{X}_1, X_{il})'(\mathbf{X}_1, X_{il})]^{-1}(\mathbf{X}_1, X_{il})'\mathbf{X}_2$ is the projection of the omitted factors' returns onto the returns of the factors in the short regression. The factors in $\mathbf{X}_2$ were chosen orthogonal to those in $\mathbf{X}_1$ so only the omitted factors' projection onto illiquidity is

[10] We have implicitly assumed here that the factors considered in the short regression were those corresponding to the first k elements of $\mathbf{X}$. This is without loss of generality, as the enumeration was arbitrary.

nontrivial, *i.e.*, every row of **F** is zeros except the last. In conjunction with the assumption that illiquidity is unpriced, *i.e.*, that $\beta_{il} = 0$, this gives the asset's expected estimated illiquidity beta in the short regression,

$$\mathbf{E}\left[b_{il}^{*}\right] = \mathbf{F}_{il}\boldsymbol{\beta}_2 \tag{13}$$

where $\mathbf{F}_{il}$ is the last row of **F**, the projection of the omitted variables' returns onto the illiquidity factor's return. Now because illiquidity covaries positively with every risk factor, each element of $\mathbf{F}_{il}$ is positive. This is enough, because the expected excess return to each risk factor is positive, to ensure that $\mathbf{E}\left[b_{il}^{*}\right]$ is strictly positive, and increasing in the asset's exposure to each of the omitted risks. Because expected returns are also increasing in the exposure to each of the priced risk factors, expected returns are positively correlated with assets' estimated illiquidity betas. That is, illiquidity will have explanatory power for returns in the misspecified asset pricing model, even though illiquidity is not, by assumption, independently priced.

3 Magnitude of the Asset Pricing Implications

The magnitude of the empirical manifestation of the omitted variable bias, *i.e.*, the size of the predicted illiquidity premium, may be addressed in two ways. One is to simulate the economy, to determine if an economically significant illiquidity premium emerges in misspecified regressions even when the data generating process contains no such premium. The other is to calibrate a factor model to determine the magnitude of the omitted risk factor required to generate the illiquidity premium observed in the data. We will consider both of these methods below.

3.1 The Illiquidity Premium in Simulations

Simulations of the economy presented in the previous section reveal a significant illiquidity premium. This premium is present even though the agents have no preference for liquidity, *per se.*

Consider the simplest non-trivial example of the economy presented in section 2. In this economy an asset θ^j consists of an exposure w^j to the single, unobserved systematic risk-factor v, the payoff of which is distributed $N(\mu, \sigma^2)$, plus a mean-zero idiosyncratic payoff

ε^j.[11] We will assume there is a continuum of assets, with mass normalized to one, and that the exposures to the omitted risk factor are normally distributed. The assets' loadings on the risk factor are then distributed $N(1, \sigma_\beta^2)$, where σ_β is the dispersion in the factor loadings.

The attentive traders initially hold q^0 of the factor, and hold $q = q^0 + dq$ after satisfying the noninformational traders' trading demand, where dq is distributed $N(0, \sigma_q^2)$ and is uncorrelated with other variables in the economy. Order flow for asset j consists of systematic order flow, proportional to the order flow for the systematic risk factor dq, and idiosyncratic order flow dq_i^j, distributed $N(0, \sigma_{q_i}^2)$. The order flow on asset j is therefore $dq^j = dq + dq_i^j$.

The change in the price of asset j results from the systematic component of the order flow, because this is the component correlated with the change in the ATs' aggregate exposure to the risk factor. The observed price impact from trading is

$$\hat{\lambda}^j = \frac{dP^j}{dq^j} = \frac{w^j \lambda}{1 + \frac{dq_i^j}{dq}} \tag{14}$$

where $\lambda = \alpha\sigma^2$ is the slope of the ATs' demand for the risk factor.[12] We can see, from the previous equation, that $\hat{\lambda}^j$ is a noisy measure of the asset's true loading on the omitted risk factor, w^j.

Performing a quintile sort on assets based on the observed $\hat{\lambda}^i$s in simulations, we see...

SIMULATION RESULTS TO GO HERE

Johnson (2004) also observes a significant illiquidity premium in simulations of several economies in which agents have no explicit demand for liquidity. He considers some standard representative-agent endowment economies, including those of Santos and Veronesi (2004), in which industries have stationary consumption shares, and Cochrane, Longstaff, and Santa-Clara (2003), in which there are two independent Lucas trees. Johnson finds that the degree of illiquidity in these frictionless economies can be similar to that observed empirically.

[11] Simulation results in the one-factor economy are closely related to those in the multi-factor economy performed on assets first sorted, as is common in the literature, by loadings on observed factors.

[12] In the CARA setting, where it is payoffs, and not returns, that are normally distributed, it is natural to measure price impact using dP^j/dq^j, instead of using $dP^j/P^j dq^j$ as we did in our theoretical discussion.

3.2 The Required Magnitude of the Omitted Risk Factor

A simple calibration can also determine the magnitude of the omitted risk factor required to produce the empirically observed illiquidity premium.[13] The calibration results will depend on both how well the empirically employed illiquidity measure proxies for the slope of the relevant price setting agents' demands for the omitted factor, and the degree of variation in assets' true loadings on the omitted factor.

The size of the illiquidity premium that arises as a result of an omitted risk factor (or factors) depends on two things: the magnitude of the omitted risk factor, and the spread in factor loadings between portfolios of liquid and illiquid assets. That is, we should expect to see an illiquidity premium of

$$\mathbf{E}[r_{il}^{e}] \;=\; \left(\beta_{il} - \beta_{liq}\right) \mathbf{E}[R_{n}^{e}], \tag{15}$$

where $E[r_{il}^{e}]$ is the difference in expected returns between portfolios of otherwise identical illiquid and liquid assets, β_{il} and β_{liq} are the portfolios' factor loadings on the omitted risk factor, and $\mathbf{E}[R_{n}^{e}]$ is the expected excess return on the omitted factor.

The spread between factor betas on portfolios of illiquid and liquid assets, $\beta_{il} - \beta_{liq}$, itself depends not only on the variation in assets' true loadings on the omitted factor, but also on the extent to which a sort on illiquidity translates into a sort on the factor loadings (*i.e.*, the magnitude of the errors-in-variables problem). Sorting on illiquidity is a noisy sort on assets' true loadings on the omitted factor, because the empirical measures of illiquidity employed in the literature (for example, volume, or price change scaled by signed volume), are imperfect proxies for the theoretically relevant price impact parameter, which measures the slope of the market's asset demand. The quality of the proxy can be expressed as a "signal-to-noise" ratio. A low noise-to-signal ratio implies illiquidity is a good proxy for the unobserved true weighting on the omitted risk factor. Conversely, with a noise-to-signal ratio greater than two, in fewer than one in ten thousand cases are we able to infer with statistical certainty at the 95 percent level that an asset's true loading on the omitted factor is different from assets' mean loading on the factor.

The quality of the sort also depends on the quality of the loading estimates of the included risk factors, because misestimation of these loadings introduces variation into the included factor loadings of "otherwise identical" assets. This can diminishes the quality of the illiquidity measure as a signal regarding the loading on the omitted factor, but the mea-

[13] The argument presented here addresses the impact of the errors-in-variables problem on the magnitude of the omitted variable bias quite generally, and is not limited to our discussion of illiquidity.

sure then also conveys information regarding the misestimated loadings on the included risk factors. Through this channel illiquidity will actually have predictive power forecasting asset returns even in a fully specified model. That is, even if our model includes every risk factor we should still expect a portfolio of illiquid assets to generate higher returns than a portfolio of otherwise seemingly identical liquid assets. After initially sorting assets based on estimated risk factor loadings, a second sort on liquidity will produce an illiquid portfolio that contains assets with true average factor loadings higher than the estimated loadings and a liquid portfolio that contains assets with true average factor loadings lower than the estimated loadings. These biases result in a spread between the expected returns to the liquid and illiquid portfolios.

It is straight forward to calculate, under some normality assumptions, combinations of factor loading dispersions and noise-to-signal ratios that produce the illiquidity premia observed in the data (see appendix). Table 1 shows the magnitude of the omitted risk-factor required to produce an observed illiquidity premium of four percent, for various combinations of standard deviations in asset factor loadings and noise-to-signal ratios. The cross-sectional dispersion in assets' exposures to the omitted risk factor, σ, increases across the columns from left to right. The magnitude of the omitted risk factor required to generate the four percent illiquidity premium is decreasing in σ, because the high-low spread between sorted portfolios' exposure to the risk factor is increasing in the variance of individual assets' loadings on the factor. The noise-to-signal ratio, σ_ϵ/σ, increases as one moves down the rows. The magnitude of the omitted risk factor required to generate the four percent illiquidity premium is increasing in σ_ϵ/σ, because with more noise the liquidity sort generates less of a high-low spread between sorted portfolios' exposures to the omitted risk factor.

Table 1
Implied Magnitude of the Omitted Risk Factor, percent per year

Noise-to-signal ratio	Factor load standard deviation (σ)				
σ_ϵ/σ	0.2	0.5	1	2	5
0.2	7.27	2.91	1.45	0.73	0.29
0.5	7.97	3.19	1.59	0.80	0.32
1	10.09	4.03	2.02	1.01	0.40
2	15.95	6.38	3.19	1.59	0.64
5	36.37	14.55	7.27	3.64	1.45

In the upper right corner of the table an extremely modest factor premium of 0.29 percent per year is sufficient to generate the observed illiquidity premium, because the sort on illiquidity generates high variation in the portfolios' loadings on the omitted risk factor. Conversely, in the lower left corner of the table an implausibly high factor risk premium of 36.37 percent per year is required to generate the observed illiquidity premium because, with little dispersion in assets' exposure to the omitted risk factor and illiquidity a poor proxy for this exposure, the sort produces almost no variation in loadings on the omitted risk factor between portfolios of liquid and illiquid assets. In the middle of the table, with a moderate standard deviation in assets' loading on the omitted risk factor in the range of 0.5 to 1.0, and a moderate noise-to-signal ratio in the range of 1.0 to 2.0, the magnitude of the omitted risk factor required to generate the observed illiquidity premium is in the range of 2.02 to 6.38 percent per year. That is, in this range the observed illiquidity premium will be of roughly the same magnitude as the expected excess return to the omitted factor.

4 Conclusion

The "illiquidity premium puzzle" is only puzzling because the illiquidity premium has been interpreted as a compensatory premium for holding illiquidity. Illiquid assets do not pay significantly higher returns than otherwise comparable liquid assets because investors demand compensation for exposure to illiquidity. Rather, illiquidity may indicate exposures to other risks for which investors do demand compensation. The existence of the illiquidity premium does not require that investors value liquidity, *per se*, but only that the slopes of investors' demands for risks are correlated with the corresponding factors' expected returns. This is all but guaranteed by the fact that the magnitude of each factor's expected return, *i.e.,* the magnitude of the price discount at which it trades relative to its expected payoff, is itself largely determined by the slope of investors' demand for the factor. That is, both the market price of exposure to any given risk factor and the sensitivity of this price of risk to trading are proportional to the slope of the market's demand. As a result, even if investors do not have a preference for liquidity, illiquidity must still be positively correlated with returns in the cross-section.

A Appendix

A.1 The magnitude of the errors-in-variables problem under the normality assumption

Assuming assets' true loadings on the missing factor are normally distributed with variance σ^2, we can calculate the difference in factor betas between the upper and lower "f-quantiles," *i.e.*, between portfolios formed by grouping the fraction f of assets with the highest and lowest factor loadings. These portfolios consist of assets with factor loadings more than $-N^{-1}(f)$ standard deviations form the mean, so the difference in average factor loadings is

$$\begin{aligned}\beta_H - \beta_L &= \frac{1}{f}\int_{\sigma N^{-1}(1-f)}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} x dx - \frac{1}{f}\int_{-\infty}^{\sigma N^{-1}(f)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} x dx \\ &= \frac{2f^{-1}}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(N^{-1}(f)\right)^2} \sigma. \end{aligned} \tag{16}$$

The dispersion in factor loadings achieved through a sort on liquidity is less than what could be achieved if one could sort on the loadings directly, as a result of an errors-in-variables problem. Suppose illiquidity measures any monotonic transformation of a noisy measure of an asset's loading on the omitted risk factor. Then, assuming normally distributed noise with variance σ_ϵ^2, the high-low spread in factor betas between portfolios sorted on liquidity, *i.e.*, formed by grouping fractions f of the least and most liquid assets, is given by

$$\beta_{il} - \beta_{liq} = 2f^{-1}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} N\left(\frac{x-\theta}{\sigma_\epsilon}\right) x dx \tag{17}$$

where $\theta = -\sqrt{\sigma^2 + \sigma_\epsilon^2}\, N^{-1}(f)$ is the minimum distance from mean measured liquidity for inclusion in the liquid/illiquid portfolios. Integrating this equation by parts and dividing by equation (16) gives the relationship between the variation in factor loading achieved through the illiquidity sort to the true variation,

$$\beta_{il} - \beta_{liq} = \frac{\beta_H - \beta_L}{\sqrt{1 + (\sigma_\epsilon/\sigma)^2}}. \tag{18}$$

Equations (15), (16), and (18) taken together imply that

$$\mathbf{E}[R_n^e] \;=\; \sqrt{\tfrac{\pi}{2}} f e^{-\frac{1}{2}\left(N^{-1}(f)\right)^2} \frac{\sqrt{1+(\sigma_\epsilon/\sigma)^2}}{\sigma} \mathbf{E}[r_{il}^e]. \tag{19}$$

Using a quintile sort for portfolio formation ($f = 0.2$) and assuming an illiquidity premium of four percent per year ($\mathbf{E}[r_{il}^e] = 4\%$), a magnitude common in the literature, yields

$$\mathbf{E}[R_n^e] \;=\; \frac{\sqrt{1+(\sigma_\epsilon/\sigma)^2}}{\sigma} \times 1.43\%. \tag{20}$$

This is the magnitude of the omitted risk factor that would be necessary to generate the entire observed illiquidity premium. It depends on the variation in assets' loadings in the omitted factor (σ) and noise-to-signal ratio (σ_ϵ/σ).

References

[1] Acharya, Viral V., and Lasse Heje Pedersen, 2003, Asset Pricing and Liquidity Risk, Working Paper.

[2] Amihud, Yakov, and Haim Mendelson, 1986, Asset Pricing and the Bid-Ask Spread, *Journal of Financial Economics* 17, pp. 223-249.

[3] Amihud, Yakov, 2002, Illiquidity and Stock Returns: Cross-Section and Time-Series Effects, *Journal of Financial Markets* 5, pp. 31-56.

[4] Ball, Ray, 1978, Anomalies in The Relationship Between Securities' Yields and Yield-Surrogates, *Journal of Financial Economics* 6, pp. 103-126.

[5] Berk, Jonathan, 1995, A Critique of Size Relate Anomalies, *Review of Financial Studies* 8, pp. 275-286.

[6] Brennan, Michael J., and Avanidhar Subrahmanyam, 1996, Market Microstructure and Asset Pricing: On the Compensation for Illiquidity in Stock Returns, *Journal of Financial Economics* 41, pp. 441-464.

[7] Campbell, John Y., Sanford J. Grossman, and Jiang Wang, 1993, Trading Volume and Serial Correlation in Stock Returns, *Quarterly Journal of Economics* 108, pp. 905-939.

[8] Cochrane, John H., 2001, Asset Pricing (Princeton University Press, Princeton, New Jersey).

[9] Cochrane, John H., Francis A. Longstaff, and Pedro Santa-Clara, 2003, Two Trees: Asset Price Dynamics Induced by Market Clearing, University of Chicago Working Paper.

[10] Constantinides, George M., 1986, Capital Market Equilibrium with Transaction Costs, *Journal of Political Economy* 94, pp. 842-862.

[11] Easley, David, Soeren Hvidkjaer, and Maureen O'Hara, 2002, Is Information Risk a Determinant of Asset Returns?, *Journal of Finance* 57, pp. 2185-2221.

[12] Goldberger, Arthur S., 1991, A Course in Econometrics (Harvard University Press, Cambridge, Massachusetts).

[13] Hasbrouck, Joel, 1988, Trades, Quotes, Inventories, and Information, *Journal of Financial Economics* 22, pp. 229-252.

[14] Heaton, John, and Deborah J. Lucas, 1996, Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing, *Journal of Political Economy* 104, pp. 443-487.

[15] Huang, Ming, 2003, Liquidity Shocks and Equilibrium Liquidity Premia, *Journal of Economic Theory* 109, pp. 104-121.

[16] Johnson, Timothy C., 2004, Dynamic Liquidity in Endowment Economies, Working Paper (London Business School).

[17] Kyle, Albert S., 1985, Continuous Auctions and Insider Trading, *Econometrica* 53, pp. 1315-1336.

[18] Lo, Andrew W., Harry Mamaysky and Jiang Wang, 2001, Asset Prices and Trading Volume Under Fixed Transaction Costs, NBER Working Paper.

[19] Lustig, Hanno, 2001, The Market Price of Aggregate Risk and the Wealth Distribution, Working Paper.

[20] O'Hara, Maureen, 2003, Presidential Address: Liquidity and Price Discovery, *Journal of Finance* 63, pp. 1335-1354.

[21] Pástor, Ľuboš, and Robert F. Stambaugh, 2003, Liquidity Risk and Expected Stock Returns, *Journal of Political Economy* 111, pp. 642-685.

[22] Santos, Tano, and Pietro Veronesi, 2004, Labor Income and Predictable Stock Returns, University of Chicago Working Paper.

[23] Vayanos, Dimitri, 1998, Transaction Costs and Asset prices: A Dynamic Equilibrium Model, *Review of Financial Studies* 11, pp. 1-58.

[24] Vayanos, Dimitri, and Jean-Luc Vila 1999, Equilibrium Interest Rates and Liquidity Premium with Transaction Costs, *Economic Theory* 13, pp. 509-539.

[25] Vayanos, Dimitri, 2004, Flight to Quality, Flight to Liquidity, and the Pricing of Risk, Working Paper.