Asset Price Dynamics in Partially Segmented Markets *

Robin Greenwood
Harvard University and NBER

Samuel G. Hanson
Harvard University and NBER

Gordon Y. Liao
Harvard University

May 2016

Abstract

How do large supply shocks in one financial market affect asset prices in other markets? We develop a model in which capital moves quickly within an asset class, but slowly between asset classes. While most investors specialize in a single asset class, a handful of generalists can gradually re-allocate capital across markets. Upon arrival of a supply shock, prices of risk in the impacted asset class become disconnected from those in others. Over the long-run, capital flows between markets and prices of risk become more closely aligned. While prices in the impacted market initially overreact to shocks, under plausible conditions, prices in related asset classes underreact.

*We thank Daniel Bergstresser, Yueran Ma, Jeremy Stein, Adi Sunderam, and seminar participants at Berkeley Haas, Brandeis, MIT Sloan, NYU Stern, University of Minnesota Carlson, University of North Carolina Kenan-Flager, and University of Texas McCombs for useful feedback. We thank David Biery for research assistance. Greenwood and Hanson gratefully acknowledge funding from the Division of Research at Harvard Business School.

1 Introduction

How do large supply shocks in one financial market affect the pricing of assets in other markets? For example, suppose that the Federal Reserve announces that it will sell a large portfolio of long-term U.S. Treasury bonds. How would such an announcement impact yields in the Treasury market? How should we expect the yields on corporate bonds and mortgage-backed securities, which are also exposed to interest rate risk, to react? And how should the resulting price dynamics play out over time?

If markets for different asset classes are tightly integrated, then a shock that affects the pricing of a risk factor in one asset class will have a similar effect on other asset classes exposed to the same risk. When markets are more segmented, however, prices of risk in one market may be disconnected from those in other markets. Segmentation arises because institutional and informational frictions lead investors to specialize in a particular asset class or a narrow set of assets (Merton [1987], Grossman and Miller [1988], Shleifer and Vishny [1997]). Although specialization can facilitate arbitrage across securities within an asset class, it can impede arbitrage across asset classes. Following a large supply shock, specialists' limited willingness to trade across markets may lead the pricing of risk to become disconnected across markets.

The degree of segmentation between different financial markets depends on time horizon. Over the long run, the forces of arbitrage ensure that capital will flow from underpriced markets to overpriced markets. However, the process of market integration can be slow, because investors with the flexibility to trade across asset classes do not do so immediately. For example, investment committees at pension funds and endowments—who have the flexibility to allocate capital across asset classes—typically only reallocate capital annually or biannually.

In this paper, we develop a dynamic model of financial markets in which capital moves quickly between securities within a given asset class, but more slowly between different asset classes. Our key contribution is to show how supply shocks in one market are reflected in prices, investor behavior, and flows into neighboring markets. In particular, we show how the reaction of neighboring markets depends on time horizon. We develop the model using a stylized depiction of fixed-income assets that trade in partially segmented markets. As we explain below, fixed income markets are a natural setting for this analysis, because bond yields naturally encode information about future risk premia. And, from a practical perspective, the fault lines between different segments of the fixed-income markets can be quite stark.

Consider two long-term risky assets trading in partially segmented markets, such as corporate bonds and Treasury securities. Both assets are exposed to a common fundamental risk factor, making them partial substitutes. This means that, absent frictions, their prices would be tightly linked by cross-market arbitrage. To introduce market segmentation, we assume that there are two sets of risk-averse market specialists, each of whom can flexibly trade one of the risky assets as well as a short-term risk-free asset. Specialists are unable to allocate capital across the two markets. However, markets are partially integrated by risk-averse generalist investors who periodically reevaluate their portfolios and shift between the two risky assets. This setup is similar to Gromb and Vayanos (2002), except that the cross-market arbitrageurs are slow-moving, much like in Duffie (2010). Because of the gradual nature of cross-market arbitrage, markets are more integrated in the long run than the short run.

In the setting we have just described, what happens when there is an unanticipated supply shock in one market? Suppose, for concreteness, that the Federal Reserve announces that it will sell a large

portfolio of long-term U.S. Treasury bonds, permanently expanding the amount of interest rate risk that investors need to bear in equilibrium. Treasury market specialists react immediately to the shock, absorbing the increased supply into their inventories. The risk premium on long-term Treasury bonds will rise, lifting their yields. However, Treasury yields will overreact—the short-run price impact will exceed the long-run impact—because the amount of capital that can initially accommodate the shock is limited to specialists and a handful generalists. Over the long-run, generalist investors will allocate additional capital to the Treasury market, muting the price impact of the supply shock at longer horizons. Price dynamics of this sort are similar to those described in Duffie (2010).

Our key contribution is to characterize how prices evolve in related markets that are not directly impacted by the supply shock. Consider the question of how corporate bond prices (or stock prices for that matter) will react to a shock to the supply of long-term Treasuries. Although this supply shock does not directly impact the corporate bond market, this market is indirectly affected because generalist investors will respond by increasing their holdings of long-term Treasuries and reducing their holdings of long-term corporate bonds. These cross-market capital flows drive down the prices of corporate bonds and push up corporate bond yields. In this way, the trading of generalist asset allocators transmits supply shocks across markets, serving to increase market integration. While yields in the Treasury market initially overreact to the supply shock, under plausible parameter values, we show that corporate bond yields will underreact: the short-run price impact is less than the long-run impact. The overreaction of Treasury yields and the underreaction of corporate yields are both driven by the fact that generalists only reallocate capital slowly. As a result, it takes time for financial markets to fully digest large supply shocks.

If all investors were generalists, the two markets would be fully integrated in the sense that exposures to common risk factors would always have the same prices in the two markets. However, in the more realistic case when markets are partially segmented, risk prices can differ across markets. This occurs because risks cannot be easily unbundled from assets and because markets receive periodic supply shocks, making cross-market arbitrage risky for generalists, as in the model developed by Gromb and Vayanos (2002). For example, interest rate risk may not be priced identically in the corporate bond market and the Treasury market. Following a supply shock, the premia associated with similar risk exposures can differ significantly between the two asset markets. As generalists react to pricing discrepancies across markets, differences in risk premia will gradually narrow. However, the differences will not vanish in the long run because of the permanent risks associated with cross-market arbitrage. Put differently, partial segmentation creates a form of noise trader risk.

The price dynamics in our model depend critically on the fractions of specialists in each market, the number of time periods it takes generalists to fully rebalance their portfolios, and the degree of substitutability between the two asset markets. The fraction of specialists and generalist investors play an especially important role. When there are a small number of slow-moving generalists, the Treasury market overreacts while the corporate market underreacts to the shock to Treasury supply. However, if there are many slow-moving generalists, markets are well-integrated and supply shocks can result in short-run overreaction in both markets.

We also use the model to explore the impact of *anticipated* future supply shocks. For example, suppose that the Fed announces it will sell long-term bonds starting in two years. How will the Treasury and corporate bond yields react to the announcement, and how should prices adjust when the

Fed actually starts selling? Although yields in both markets react immediately, both markets exhibit significant underreaction: the short-run price impact is significantly less than the long-run impact. Immediately upon the announcement, generalists begin to gradually adjust in the direction of the anticipated shock, buying Treasuries and selling corporate bonds. Treasury specialists provide temporary liquidity to these generalists, by selling their Treasury holdings, planning to replenish their inventories once the supply shock actually lands in two years. The result is a protracted adjustment process starting from the announcement and continuing until well after the supply shock lands. Longer delays between the announcement and the arrival of the supply shock result in a more gradual adjustment process.

In describing our model, we have made no distinction between a risky "asset" and the "market" in which it trades. This distinction arises when we introduce multiple risky assets into each market. Individual assets differ in their degree of exposure to common risk factors. For example, the "market" for U.S. Treasury securities contains bonds of many different maturities, which have different exposures to interest rate risk. Extending our model to allow for multiple securities per asset market, we show that a conditional CAPM prices all assets in the first market and that another conditional CAPM—with different prices of risk—prices all assets in the second market. Critically, these two market-specific pricing models are linked over time by the cross-market arbitrage activities of slow-moving asset allocators. For example, the pricing of interest rate risk for 2-year Treasuries is always perfectly consistent with the pricing of interest rate risk for 10-year Treasuries. However, the pricing of interest rate risk in the Treasury market may differ somewhat from that in the corporate bond market. And these cross-market differences will be most pronounced following the arrival of major shocks that take time for slow-moving generalists to digest.

The question of how asset prices adjust across partially segmented markets is of enormous practical importance. Consider the recent large-scale purchases of long-term government bonds by central banks in the United States, United Kingdom, Japan, and Europe, often referred to as "quantitative easing." A key question about these policies is whether they impact the prices of financial assets outside of the market for government bonds. The favored methodology for answering this question has been to use event studies of intraday or one-day price changes following central bank policy announcements. A number of these studies have concluded that the effects of quantitative easing are most pronounced in the market in which the central bank is transacting, with only modest spillovers to other related markets (Woodford [2012] and Krishnamurthy and Vissing-Jorgensen [2013]). Others have suggested that at longer horizons, the spillovers are more significant. Mamaysky (2014) suggests that if one expands the measurement window by a few days or weeks, the effects in other markets may be much larger. Feunou et al (2015) suggest that U.S. quantitative easing was transmitted to Canadian bond markets over time via portfolio flows.

Our model suggests that the short-run price impact of a supply shock on different markets may not accurately reveal the long-run impact, which is often of greater interest to policymakers. We illustrate this idea by analyzing the statistical power of short-run event studies within our model. We show that the horizon at which statistical power is maximized is often much shorter than the horizon at which the long-run price impact is achieved. In summary, event study methodology is not well suited to analyzing the impact of market supply shocks.

Our model is closely related to two strands of research in financial economics. The idea that front-

line arbitrageurs in financial markets are highly specialized traces back to Merton (1987) and Grossman and Miller (1988), and is a central tenet of the theory of limited arbitrage (De Long et al [1990], Shleifer and Vishny [1997], and Gromb and Vayanos [2002]). A small literature in finance describes asset prices and returns in segmented markets (Stapleton and Subrahmanyam [1977], Errunza and Losq [1985], Merton [1987]). More recently, a number of researchers have demonstrated downward-sloping demand curves for individual financial asset classes, which would be puzzling if markets were fully integrated (Gabaix, Krishnamurthy, and Vigneron [2007], Gârleanu, Pedersen, and Poteshman [2009], Greenwood and Vayanos [2014], and Hanson [2014]). These researchers have often motivated their analysis by positing an extreme form of market segmentation in which a different pricing kernel is used to price the securities in each distinct asset class. Our paper emphasizes how the actions of slow-moving asset allocators serve to link these market-specific pricing kernels together, thereby offering a middle ground between these models positing extreme segmentation and traditional models featuring perfect integration.

Second, our paper is related to research on "slow-moving capital," which is the idea that capital does not flow as quickly towards attractive investment opportunities as textbook theories might suggest (Mitchell, Pedersen, Pulvino [2007], Duffie [2010], Acharya, Shin, and Yorulmazer [2013]). Here, our model draws most heavily from Duffie (2010), who studies the implications of slow moving capital for price dynamics in a single asset market. Duffie and Strulovici (2012) present a model of the movement of capital across two partially segmented markets, but their focus is on the endogenous speed of capital mobility, which we take as exogenous. Our key contribution here is to characterize the dynamics of prices across related asset markets and to describe the patterns of cross-market arbitrage in response to large supply or demand shocks.

2 Model

We develop the model in two steps. We first develop a tractable, benchmark model for pricing long-term fixed-income assets that are exposed to *both* interest rate risk and default risk. The model builds on the default-free term structure models in Vayanos and Vila (2009) and Greenwood and Vayanos (2014) in which interest rate risk is priced by a set of specialized, risk-averse bond arbitrageurs, leading to a downward-sloping aggregate demand curve for bond risk factors. In this first step, we develop a simple way to incorporate default risk into this class of models. In the second step, we introduce a second asset class and a richer institutional trading environment that contains both generalists and specialists. In this richer environment, we describe how prices and investor positions in both markets evolve following a supply shock that directly impacts only one market.

2.1 Single asset model

2.1.1 Defaultable perpetuities

Consider a homogenous portfolio of perpetual, defaultable bonds each of which promises to pay a coupon of C each period. Let $P_{L,t}$ denote the the price of each long-term bond at time t. Suppose that a random fraction h_{t+1} of the bonds default at $t+1$ and are worth $(1-L_{t+1})(P_{L,t+1}+C)$ where $0 \leq L_{t+1} < 1$ is the (possibly random) loss-given-default as a fraction of market value. The remaining

fraction $(1-h_{t+1})$ of the bonds do not default and are worth $(P_{L,t+1}+C)$. Thus, the return on the bond portfolio is

$$1+R_{L,t+1}=\frac{(1-Z_{t+1})(P_{L,t+1}+C)}{P_{L,t}}, \tag{1}$$

where $Z_{t+1}=h_{t+1}L_{t+1}$, satisfying $0\leq Z_{t+1}<1$, is the portfolio default realization at time $t+1$. If $Z_{t+1}\equiv 0$, the bonds are default-free. If Z_{t+1} is stochastic, the bonds are defaultable with high realizations of Z_{t+1} corresponding to larger default losses at time $t+1$. This formulation of default risk follows Duffie and Singleton's (1999) "recovery of market value" assumption which has become standard in the credit risk literature.

To generate a tractable linear model, we use a Campbell-Shiller (1988) log-linear approximation to the return on this portfolio of defaultable perpetuities. Specifically, defining $\theta\equiv 1/(1+C)<1$, the one-period log return on the bonds is

$$r_{L,t+1}\equiv \ln\left(1+R^{L}_{t+1}\right)\approx \overbrace{\frac{1}{1-\theta}}^{D} y_{L,t}-\overbrace{\frac{\theta}{1-\theta}}^{D-1} y_{L,t+1}-z_{t+1}, \tag{2}$$

where $y_{L,t}$ is the log yield-to-maturity at time t,

$$D=\frac{1}{1-\theta}=\frac{C+1}{C} \tag{3}$$

is the Macaulay duration when the bonds are trading at par, and $z_t=-\ln(1-Z_t)$ is the log default loss at time t.[1]

To derive this approximation note that the Campbell-Shiller (1988) approximation of the 1-period log return is

$$\begin{aligned} r_{L,t+1} &= \ln(P_{L,t+1}+C)-p_{L.t}-z_{t+1} \\ &\approx \kappa+\theta p_{L.t+1}+(1-\theta)c-p_{L.t}-z_{t+1} \end{aligned} \tag{4}$$

where $\theta=1/(1+\exp(c-\overline{p}_L))$ and $\kappa=-\log(\theta)-(1-\theta)\log\left(\theta^{-1}-1\right)$ are parameters of the log-linearization. Iterating equation (4) forward, we find that the log bond price is

$$p_{L,t}=(1-\theta)^{-1}\kappa+c-\textstyle\sum_{i=0}^{\infty}\theta^{i}E_t\left[r_{L,t+i+1}+z_{t+i+1}\right]. \tag{5}$$

Applying this approximation to *promised cashflows* (i.e., $z_{t+i+1}\equiv 0$ for all $i\geq 0$) and the *yield-to-maturity*, defined as the *constant return* that equates bond price and the discounted value of *promised* cashflows, we obtain

$$p_{L,t}=(1-\theta)^{-1}\kappa+c-(1-\theta)^{-1}y_{L,t}. \tag{6}$$

Equation (2) then follows by substituting the expression for $p_{L,t}$ in equation (6) into the Campbell-Shiller return approximation in equation (4). Assuming the steady-state price of the bonds is par ($\overline{p}_L=0$), we have $\theta=1/(1+C)$. Thus, bond duration is $D=-\partial p_{L,t}/\partial y_{L,t}=(1-\theta)^{-1}=(1+C)/C$.

[1] This log-linear approximation for default-free coupon-bearing bonds appears in Chapter 10 of Campbell, Lo, and MacKinlay (1997). Our approximation for defaultable bonds then follows trivially given the assumption that default losses are a (random) fraction of market value.

Since $-\partial p_{L,t}/\partial y_{L,t} = -\left(\partial P_{L,t}/\partial Y_{L,t}\right)\left(\left(1+Y_{L,t}\right)/P_{L,t}\right) = \left(Y^{\vdots}_{L,t}+1\right)/Y_{L,t}$ this corresponds to Macaulay duration when the bonds are trading at par ($Y_{L,t} = C$).

2.1.2 Risk factors

Investors in defaultable long-term bonds are exposed to three different types of risk: interest rate risk, default risk, and supply risk. First, investors are exposed to *interest rate risk*. In our model, investors face an exogenous short-term interest rate that evolves randomly over time and will suffer a capital loss on their bond holdings if short-term rates rise unexpectedly. Second, investors face *default risk*: the future period-by-period default realization is unknown and evolves randomly over time. Finally, investors are exposed to *supply risk*: there are random supply shocks which impact the prices and yields on long-term bonds, holding fixed the expected future path of short-term interest rates and expected future defaults. Thus, using Campbell's (1991) terminology, interest rate risk and default risk are forms of fundamental "cash flow" risk, whereas supply risk is a form of "discount rate" risk.

We make the following concrete assumptions:

- **Short-term interest rates:** The log short-term riskless rate available to investors between time t and $t+1$, denoted r_t, is known at time t. We assume that r_t also follows an exogenous AR(1) process

$$r_{t+1} = \overline{r} + \rho_r\left(r_t - \overline{r}\right) + \varepsilon_{r,t+1}, \tag{7}$$

  where $Var_t\left[\varepsilon_{r,t+1}\right] = \sigma_r^2$. One can think of the short-term rate as being determined outside the model either by monetary policy or by a stochastic short-term storage technology that is available in perfectly elastic supply.

- **Default losses:** We assume that the default process z_t follows

$$z_{t+1} = \overline{z} + \rho_z\left(z_t - \overline{z}\right) + \varepsilon_{z,t+1} \tag{8}$$

  where $Var_t\left[\varepsilon_{z,t+1}\right] = \sigma_z^2$.

- **Supply:** We assume that the perpetuity is available in an exogenous, time-varying supply s_t. We assume that supply follows an AR(1) process

$$s_{t+1} = \overline{s} + \rho_s\left(s_t - \overline{s}\right) + \varepsilon_{s,t+1}, \tag{9}$$

  where $Var_t\left[\varepsilon_{s,t+1}\right] = \sigma_s^2$.

For simplicity, we will assume that $\varepsilon_{s,t+1}$, $\varepsilon_{r,t+1}$, and $\varepsilon_{z,t+1}$ are mutually orthogonal. However, it is straightforward to relax this assumption.

2.1.3 Specialist demand and market clearing

There is a unit mass of specialized bond arbitrageurs, each with risk tolerance τ. Specialist arbitrageurs can earn an uncertain future return of $r_{L,t+1}$ from to t to $t+1$ by investing in the defaultable long-term bond. Alternatively, they can earn a certain return of r_t by investing at the short-term interest rate. Specialist arbitrageurs are concerned with their interim wealth.

Formally, we assume that at date t specialist arbitrageurs have mean-variance preferences over their wealth at $t+1$. This means that arbitrageurs choose their holdings of the perpetuity to solve

$$\max_{b_t} \left\{ b_t E_t \left[rx_{L,t+1} \right] - (2\tau)^{-1} (b_t)^2 \, Var_t \left[rx_{L,t+1} \right] \right\}, \tag{10}$$

where $rx_{L,t+1} \equiv r_{L,t+1} - r_t$ is the log excess returns on the defaultable long–term bond over the short-term interest rate between t and $t+1$. Thus, arbitrageur demand for the risky bond is

$$b_t = \tau \frac{E_t \left[rx_{L,t+1} \right]}{Var_t \left[rx_{L,t+1} \right]}. \tag{11}$$

Equation (11) says that arbitrageurs borrow at the short-term rate and invest in risky long-term bonds when the expected return on perpetuities exceeds that the short rate ($E_t \left[rx_{L,t+1} \right] > 0$). Conversely, arbitrageurs sell short bonds and invest at the short rate when $E_t \left[rx_{L,t+1} \right] < 0$. And they respond more aggressively to these movements in risk premia when they are more risk tolerant and when the variance of excess bond returns is low.

Market clearing ($b_t^* = s_t$) implies that the bond risk premium, $E_t \left[rx_{L,t+1} \right]$ is given by

$$E_t \left[rx_{L,t+1} \right] = \tau^{-1} V_L^{(1)} s_t, \tag{12}$$

where $V_L^{(1)} = Var_t \left[(D-1) \, y_{L,t+1} + z_{t+1} \right]$ is the equilibrium variance of 1-period excess returns.

Thus, bond risk-premia are increasing in bond supply, s_t. When a positive supply shock arrives, bond risk premia jump instantaneously. If the shock is almost permanent ($\rho_s \approx 1$), the impact on the risk premium will be long lived. If the shock is transient ($0 < \rho_s \ll 1$), supply will quickly revert to steady-state ($\overline{s}$) and risk premia will revert to their steady-state level, $\tau^{-1} V_L^{(1)} \overline{s}$.

2.1.4 Solution and equilibrium yields

To solve the model, we conjecture that equilibrium bond yields take the linear form

$$y_{L,t} = \alpha_0 + \alpha_r \left(r_t - \overline{r} \right) + \alpha_z \left(z_t - \overline{z} \right) + \alpha_s \left(s_t - \overline{s} \right). \tag{13}$$

Using this conjecture, in the Internet Appendix we show that a linear equilibrium of this form exists so long as arbitrageurs are sufficiently risk tolerant (i.e., if τ is large enough). We show that the equilibrium variance of 1-period excess bond returns, $V_L^{(1)}$, must satisfy the following quadratic equation

$$V_L^{(1)} = \left(\frac{\theta}{1 - \rho_r \theta} \sigma_r \right)^2 + \left(\frac{1}{1 - \rho_z \theta} \sigma_z \right)^2 + \left(\tau^{-1} \frac{\theta}{1 - \rho_s \theta} \sigma_s \right)^2 \left(V_L^{(1)} \right)^2. \tag{14}$$

The total risk premium can be decomposed into compensation for bearing interest rate risk, compensation for bearing credit risk, and compensation for bearing supply risk:

$$E_t\left[rx_{L,t+1}\right] = \overbrace{\tau^{-1}\left(\frac{\theta}{1-\rho_r\theta}\sigma_r\right)^2 s_t}^{\text{Interest rate risk premium}} + \overbrace{\tau^{-1}\left(\frac{1}{1-\rho_z\theta}\sigma_z\right)^2 s_t}^{\text{Credit risk premium}} + \overbrace{\tau^{-1}\left(\tau^{-1}\frac{\theta}{1-\rho_s\theta}\sigma_s\right)^2\left(V_L^{(1)}\right)^2 s_t}^{\text{Supply risk premium}}. \tag{15}$$

The level of supply (s_t) appears three times on the right hand side of equation (15) because all three components of the total risk premium move in lock in our single asset model.

As in Greenwood and Vayanos (2014), when there is supply risk ($\sigma_s^2 > 0$) a linear equilibrium only exists if bond arbitrageurs are sufficiently risk tolerant.[2] If arbitrageurs are sufficiently risk tolerant, there are two possible solutions to (14): one in which yields are highly sensitive to supply shocks and one in which yields are less sensitive. What is the intuition for the multiplicity of equilibria? If yields are highly sensitive to supply shocks, then bonds become highly risky for arbitrageurs. Hence, arbitrageurs absorb supply shocks only if they are compensated by large changes in yields, making the high sensitivity of yields to shocks self-fulfilling. Conversely, if yields are less sensitive to supply shocks, then bonds become less risky for arbitrageurs and arbitrageurs are willingly absorb supply shocks even if they are only compensated by modest changes in yields. Equilibrium multiplicity of this sort is common in overlapping generations models such as ours where arbitrageurs with short investment horizons hold a long-lived asset that is subject to supply shocks (see e.g., DeLong, Shleifer, Summers, and Waldmann [1990]).

Following Greenwood and Vayanos (2014), we focus on the well-behaved and economically relevant equilibrium in which yields are less sensitive to supply shocks, which corresponds to the smaller root of equation (14).[3] It is then straightforward to show that $V_L^{(1)}$ is increasing in σ_r^2, σ_z^2, σ_s^2, ρ_r, ρ_z, ρ_s, and D $[= (1-\theta)^{-1}]$ and decreasing in τ. Thus, for a given level of bond supply, the total risk premium is larger when short-term rates are more volatile, when there is greater uncertainty about future defaults, and when supply shocks are more volatile. Furthermore, the risk premium is larger when each of these three processes is more persistent. Finally, the risk premium is increasing in the duration of the perpetuity and is decreasing in arbitrageur risk tolerance.

Rewriting equation (2) as $y_{L,t} = E_t\left[(1-\theta)\left(r_t + rx_{L,t+1} + z_{t+1}\right) + \theta y_{L,t+1}\right]$ and iterating forward, we see that the equilibrium yield on the defaultable perpetuity is a weighted average of expected future short rates, future default losses, and future risk premia

$$y_{L,t} = (1-\theta)\sum_{i=0}^{\infty}\theta^i E_t[\overbrace{r_{t+i}}^{\text{Short rate}} + \overbrace{z_{t+i+1}}^{\text{Default loss}} + \overbrace{\tau^{-1}V_L^{(1)}s_{t+i}}^{\text{Risk premium}}]. \tag{16}$$

[2] If τ is too small and there are supply shocks ($\sigma_s^2 > 0$), no linear equilibrium exits because bonds become extremely risky for arbitrageurs and it is impossible to clear the market. See the Internet Appendix.

[3] As $\sigma_s^2 \to 0$, this smaller root converges to the solution for $V_L^{(1)}$ when $\sigma_s^2 = 0$ (i.e., to $\left((\theta\sigma)_r / (1-\rho_r\theta)\right)^2 + \left(\sigma_z/(1-\rho_z\theta)\right)^2$) whereas the larger root diverges to infinity as $\sigma_s^2 \to 0$. All of the relevant comparative statics on $V_L^{(1)}$ have the intuitive signs at the smaller root, but have the opposite signs at the larger root.

Because of the coupon-bearing nature of the long-term bond, equation (16) shows that expected short rates, default losses, and risk premia in the near future have a larger effect on bond yields than those in the distant future.[4] Making use of the assumed AR(1) dynamics for r_t, z_t, and s_t, we can express the equilibrium yield as

$$y_{L,t} = \overbrace{\left[\overline{r} + \frac{1-\theta}{1-\rho_r\theta}\left(r_t - \overline{r}\right)\right]}^{\text{Expected future short rates}} + \overbrace{\left[\overline{z} + \frac{1-\theta}{1-\rho_z\theta}\rho_z\left(z_t - \overline{z}\right)\right]}^{\text{Expected future default losses}} + \overbrace{\left[\tau^{-1}V_L^{(1)}\overline{s} + \tau^{-1}V_L^{(1)}\frac{1-\theta}{1-\rho_s\theta}\left(s_t - \overline{s}\right)\right]}^{\text{Risk premium}}. \tag{17}$$

Equation (17) shows that the perpetuity yield is more sensitive to movements in short rates when the short-rate process is more persistent and when bond duration is shorter (i.e., $\partial^2 y_{L,t}/\partial r_t \partial \rho_r > 0$ and $\partial^2 y_{L,t}/\partial r_t \partial D < 0$). Similarly, the yield is more sensitive to movements in current default losses (z_t) when the default process is more persistent and when bond duration is shorter. Yields are more sensitive to bond supply when short-rates are more volatile or more persistent or when defaults are more volatile or more persistent. Finally, yields are also more sensitive to supply shocks when risk tolerance is low, supply shocks are more volatile, or supply shocks are more persistent.[5]

2.2 Partially segmented markets

With this machinery in place, we now introduce a second risky asset and a richer trading environment, to capture the idea that the two assets trade in partially segmented markets. Our goal is to study how shocks to asset supply in one market are transmitted over time to the second market.

2.2.1 Asset markets

Suppose now that there are two portfolios of perpetual risky assets, A and B. A is default–free and exposed only to interest rate risk. Borrowing notation from above, portfolio A pays a coupon of C_A each period, so the gross return on A is $1 + R_{A,t+1} = \left(P_{A,t+1} + C_A\right)/P_{A,t}$. The log excess return on the A portfolio over the short-term interest rate from time t to $t+1$ is

$$rx_{A,t+1} \approx \frac{1}{1-\theta_A} y_{A,t} - \frac{\theta_A}{1-\theta_A} y_{A,t+1} - r_t, \tag{18}$$

where $\theta_A = 1/\left(1 + C_A\right)$.

The second portfolio, B, is subject to default risk which makes it an imperfect substitute for asset A. Specifically, the B portfolio carries a *promised* coupon payment of C_B each period. However, the gross return on the B portfolio from time t to $t+1$ is $1 + R_{B,t+1} = \left(1 - Z_{t+1}\right)\left(P_{B,t+1} + C_B\right)/P_{B,t}$

[4] This is similar to Campbell and Shiller's (1988) analysis of the price of a dividend-paying stock.

[5] The sign of $\partial^2 y_{L,t}/\partial s_t \partial D$ is ambiguous since $\partial V_L^{(1)}/\partial D > 0$, but $\partial\left[\left(1-\theta\right)/\left(1-\rho_s\theta\right)\right]/\partial D < 0$. This corresponds to the finding in Vayanos and Greenwood (2014) that, depending on the persistence of supply shocks, a current increase in bond supply can have a greater impact on the yields of intermediate or long-dated bonds. Specifically, highly persistent supply shocks have the greatest impact on long-dated yields, while transitory supply shocks have the greatest impact on intermediate-dated yields.

where $0 \leq Z_{t+1} \leq 1$ is the default realization at time $t+1$. Therefore, the log excess return on B from time t to $t+1$ is

$$rx_{B,t+1} \approx \frac{1}{1-\theta_B} y_{B,t} - \frac{\theta_B}{1-\theta_B} y_{B,t+1} - z_{t+1} - r_t, \tag{19}$$

where $\theta_B = 1/(1+C_B)$. The additional z_{t+1} term in equation (19) reflects the time $t+1$ default realization that is specific to the B asset. The variance of z_{t+1} determines, in part, the degree of substitutability between assets A and B.

We assume that the processes for the short rate r_t and for default losses z_t are as in equations (7) and (8) above. However, we assume that the two asset markets are subject to different supply shocks which also limits their substitutability for investors with shorter horizons. The net supply that investors must hold of asset A evolves according to

$$s_{A,t+1} = \overline{s}_A + \rho_{s_A}(s_{A,t} - \overline{s}_A) + \varepsilon_{s_A,t+1}, \tag{20}$$

where $Var_t[\varepsilon_{s_A,t+1}] = \sigma^2_{s_A}$. Similarly, the net supply that investors must hold of asset B evolves as

$$s_{B,t+1} = \overline{s}_B + \rho_{s_B}(s_{B,t} - \overline{s}_B) + \varepsilon_{s_B,t+1}, \tag{21}$$

where $Var_t[\varepsilon_{s_B,t+1}] = \sigma^2_{s_B}$.We continue to assume that $\varepsilon_{r,t+1}$, $\varepsilon_{z,t+1}$, $\varepsilon_{s_A,t+1}$, and $\varepsilon_{s_B,t+1}$ are mutually orthogonal.

2.2.2 Market participants

There are three types of investors, all with identical risk tolerance τ. Investors are distinguished by their ability to transact in different markets and by the frequency with which they can rebalance their portfolios. Fast-moving A-specialists are free to adjust their holdings of the A asset and the riskless short-term asset each period; however, A-specialists cannot hold the B asset. A-specialists are present in mass q_A and we denote their demand for A by $b_{A,t}$. Analogously, fast-moving B-specialists can freely adjust their holdings of the B asset and the riskless asset each period, but cannot hold the A asset. B-specialists are present in mass q_B and their demand for asset B is $b_{B,t}$.

The third group of investors is a set of slow-moving generalists who can adjust their holdings of A and B asset, as well as the riskless short-term asset, but can do so only every k periods. Generalists are present in mass $1-q_A-q_B$. Fraction $1/k$ of these generalists investors are active each period and can reallocate their portfolios between the A and B assets. However, they must then maintain this same portfolio allocation for the next k periods. As in Duffie (2010), this is a reduced form way to model the frictions that limit the speed of capital flows across markets.

The market structure we have described here is a natural way to capture the industrial organization of real world asset management. Due to agency and informational problems, savers are only willing to give delegated managers the discretion to adjust their portfolios quickly if the manager accepts a narrow, *specialized* mandate. These same agency and informational frictions also mean that savers are only willing to give managers the discretion to adjust quickly if the manager gives them an open-ended claim (e.g., Stein (2005)). As a result, fast-moving investors often have endogenously *short horizons*. By contrast, most institutions, such as endowments and pensions, that have *longer horizons* and possess *greater flexibility* to re-allocate capital across asset classes are subject to governance

mechanisms—themselves a response to informational and agency frictions—that limit the speed of any such capital movement. In combination, we believe that a model with fast-moving specialists and slow-moving generalists is a tractable, reduced-form way to capture real-world arbitrage frictions.

In this paper, we focus on a "medium-run" equilibrium in which the parameters governing market structure (q_A, q_B, and k) are regarded as fixed and exogenously given. However, one could extend the model to endogenize the market structure. In the resulting "very long-run" equilibrium, q_A, q_B, and k would adjust so that A specialists, B specialists, and generalists all have the same expected utility in the long-run.[6]

Fast-moving A-specialists and B-specialists have mean-variance preferences over 1-period portfolio log returns. Thus, their demands are given by

$$b_{A,t} = \tau \frac{E_t\left[rx_{A,t+1}\right]}{Var_t\left[rx_{A,t+1}\right]}, \tag{22}$$

and

$$b_{B,t} = \tau \frac{E_t\left[rx_{B,t+1}\right]}{Var_t\left[rx_{B,t+1}\right]}. \tag{23}$$

Since they only rebalance their portfolios every k periods, slow-moving generalist investors have mean-variance preferences over their k-period *cumulative* portfolio excess return. Defining $rx_{A,t\to t+k} \equiv \sum_{i=1}^{k} rx_{A,t+i}$ and $rx_{B,t\to t+k} \equiv \sum_{i=1}^{k} rx_{B,t+i}$ as the cumulative k-period returns from t to $t+k$ on A and B, the k-period portfolio excess return of generalists who are active at t is[7]

$$rx_{d_t,t\to t+k} = d_{A,t} \times rx_{A,t\to t+k} + d_{B,t} \times rx_{B,t\to t+k}. \tag{24}$$

Thus, generalist investors who are active at time t choose their holdings of asset A and B, denoted $d_{A,t}$ and $d_{B,t}$, to solve

$$\max_{d_{A,t},d_{B,t}} \left\{ E_t\left[rx_{d_t,t\to t+k}\right] - (2\tau)^{-1}\left(Var_t\left[rx_{d_t,t\to t+k}\right]\right) \right\}. \tag{25}$$

This implies that

$$\begin{bmatrix} d_{A,t} \\ d_{B,t} \end{bmatrix} = \frac{\tau}{1 - R^{2(k)}_{AB}} \begin{bmatrix} \frac{E_t\left[rx_{A,t\to t+k}\right]}{Var_t[rx_{A,t\to t+k}]} - \beta^{(k)}_{B|A} \frac{E_t\left[rx_{B,t\to t+k}\right]}{Var_t[rx_{B,t\to t+k}]} \\ \frac{E_t\left[rx_{B,t\to t+k}\right]}{Var_t[rx_{B,t\to t+k}]} - \beta^{(k)}_{A|B} \frac{E_t\left[rx_{A,t\to t+k}\right]}{Var_t[rx_{A,t\to t+k}]} \end{bmatrix}, \tag{26}$$

where, for example, $\beta^{(k)}_{B|A}$ is the coefficient from a linear regression of $rx_{B,t\to t+k}$ on $rx_{A,t\to t+k}$ and $R^{2(k)}_{AB}$ is the goodness of fit from this regression.[8]

[6] For instance, one could assume that A and B specialists must pay a cost to set up a specialized, fast-moving fund and that generalists must pay a cost in order to adjust more quickly. q_A, q_B, and k would then need to adjust so that (i) investors expect to earn the same long-run Sharpe ratio, net of costs, from all three structures and (ii) generalists' marginal benefit from adjusting their portfolios more frequently equals the marginal cost of more frequent adjustment.

[7] Formally, this means we assume that slow-moving generalists re-invest all capital initially allocated to the A market (B market) in the A market (B market) over their k-period investment horizon. Also, our implicit log-linearization of the portfolio return omits the second-order Jensen's inequality adjustments familiar from Campbell and Viceira (2002). However, in the case of low-volatility fixed-income instruments, these adjustments are quantitatively small and do not alter the core economic intuition of the model.

[8] We obtain similar results if we alter equation (25) to reflect the fact that the cumulative return from rolling over an

Equation (26) says that, all else equal, generalist investors allocate more capital to market A when asset A becomes more attractive from a narrow risk-reward standpoint (i.e., $d_{A,t}$ is increasing in $E_t[\sum_{i=1}^{k} rx_{A,t+i}]/Var_t[rx_{A,t\to t+k}]$). Further, assuming B and A co-move positively ($\beta^{(k)}_{B|A} > 0$), generalists allocate less capital to market A when asset B becomes more attractive from a risk-reward standpoint (i.e., $d_{A,t}$ is decreasing in $E_t[\sum_{i=1}^{k} rx_{B,t+i}]/Var_t[rx_{B,t\to t+k}]$). In this way, the response of generalist investors transmits supply shocks in the B market to the A market, promoting cross-market integration over time. Cross-market capital flows become more responsive to differences in risk-reward between markets when the two assets are closer substitutes (i.e., when $R^{2(k)}_{AB}$ is higher). In the limit as the two assets become perfect substitutes, $R^{2(k)}_{AB}$ approaches 1 and generalist investors become extremely aggressive in exploiting any cross-market pricing differences.

2.2.3 Equilibrium yields

In market A at time t, there is a mass q_A of fast-moving specialists, each with demand $b_{A,t}$, and a mass $(1 - q_A - q_B)\, k^{-1}$ of active slow-moving generalists, each with demand $d_{A,t}$. These investors must accommodate the *active supply*, which is the total supply of $s_{A,t}$ less any supply held off the market by inactive generalist investors, $(1 - q_A - q_B)\, k^{-1} \sum_{j=1}^{k-1} d_{A,t-j}$. Thus, the market-clearing condition for asset A is

$$\overbrace{q_A b_{A,t}}^{\substack{\text{Specialist}\\ \text{demand}}} + \overbrace{(1 - q_A - q_B) k^{-1} d_{A,t}}^{\substack{\text{Active generalist}\\ \text{demand}}} = \overbrace{s_{A,t}}^{\substack{\text{Total bond}\\ \text{supply}}} - \overbrace{(1 - q_A - q_B)(k^{-1} \sum_{i=1}^{k-1} d_{A,t-i})}^{\substack{\text{Inactive generalist}\\ \text{holdings}}}. \tag{27}$$

The market-clearing condition for asset B is analogous.

We conjecture that equilibrium yields and generalist demands are linear functions of a state vector, $\mathbf{x}_t$, that includes the steady-state deviations of the short-term interest rate, the default realization, the supply of asset A, the supply of asset B, inactive generalist holdings of asset A, and inactive generalist holdings of asset B. Formally, we conjecture that long-term yields in market A and B are

$$y_{A,t} = \alpha_{A0} + \boldsymbol{\alpha}'_{A1} \mathbf{x}_t, \tag{28}$$

$$y_{B,t} = \alpha_{B0} + \boldsymbol{\alpha}'_{B1} \mathbf{x}_t, \tag{29}$$

and that the demands of slow-moving generalists are

$$d_{A,t} = \delta_{A0} + \boldsymbol{\delta}'_{A1} \mathbf{x}_t, \tag{30}$$

$$d_{B,t} = \delta_{B0} + \boldsymbol{\delta}'_{B1} \mathbf{x}_t, \tag{31}$$

investment at the short-rate for k periods, $\sum_{i=0}^{k-1} r_{t+i}$, is unknown at time t. As in Campbell and Viceira (2001), this adds an I-CAPM-like hedging motive for holding long-duration assets that have high excess returns when $\sum_{i=0}^{k-1} r_{t+i}$ turns out to be lower than expected. Formally, this means that generalists solve $\max_{d_{A,t}, d_{B,t}} \left\{ E_t \left[r_{P,t\to t+k} \right] - \frac{1}{2\tau} \left(Var_t \left[r_{P,t\to t+k} \right] \right) \right\}$ where $r_{P,t\to t+k} = (\sum_{i=0}^{k-1} r_{t+i}) + d_{A,t} \times (\sum_{i=1}^{k} rx_{A,t+i}) + d_{B,t} \times (\sum_{i=1}^{k} rx_{B,t+i})$. The solution takes the same form as (26), replacing $E_t[\sum_{i=1}^{k} rx_{A,t+i}]$ with $E_t[\sum_{i=1}^{k} rx_{A,t+i}] - \tau^{-1}\, Cov_t[\sum_{i=1}^{k} rx^{A}_{t+i}, \sum_{i=0}^{k-1} r_{t+i}]$ and simiilarly for asset B.

where the $2(1+k)\times 1$ dimensional state vector, $\mathbf{x}_t$, is given by

$$\mathbf{x}_t = [r_t - \overline{r}, z_t - \overline{z}, s_{A,t} - \overline{s}_A, s_{B,t} - \overline{s}_B, d_{A,t-1} - \delta_{A0}, \cdots, d_{A,t-(k-1)} - \delta_{A0}, d_{B,t-1} - \delta_{B0}, \cdots, d_{B,t-(k-1)} - \delta_{B0}]'. \tag{32}$$

These assumptions imply that the state vector follows an AR(1) process

$$\mathbf{x}_{t+1} = \mathbf{\Gamma}\mathbf{x}_t + \boldsymbol{\epsilon}_{t+1}, \tag{33}$$

where the transition matrix $\mathbf{\Gamma}$ depends on generalist demands.

As we show in the Internet Appendix, equilibrium yields take the same basic form as in (17) with only specialist investors. For market A, the yield is given by

$$\begin{aligned} y_{A,t} = & \overbrace{\left\{\overline{r} + \left(\frac{1-\theta_A}{1-\rho_r\theta_A}\right)(r_t - \overline{r})\right\}}^{\text{Expected future short rates}} \\ & + \overbrace{\left[(q_A\tau)^{-1} V_A^{(1)} \left(\overline{s}_A - (1-q_A-q_B)\,\delta_{A0}\right)\right]}^{\text{Unconditional term premia}} \\ & + \overbrace{\left[(q_A\tau)^{-1} V_A^{(1)} \begin{pmatrix} \frac{1-\theta_A}{1-\theta_A\rho_{s_A}}(s_{A,t} - \overline{s}_A) \\ -(1-\theta_A)(1-q_A-q_B)\,k^{-1}\sum_{i=0}^{\infty}\theta_A^i E_t[\sum_{j=0}^{k-1}(d_{A,t+i-j} - \delta_{A0})] \end{pmatrix}\right]}^{\text{Conditional term premia}}, \end{aligned} \tag{34}$$

where $V_A^{(1)} = Var_t\left[rx_{A,t+1}\right]$ is the equilibrium variance of 1-period excess returns on asset A. The yield for asset B has an extra term relating to expected future defaults, but is otherwise similar

$$\begin{aligned} y_{B,t} = & \overbrace{\left\{\overline{r} + \left(\frac{1-\theta_B}{1-\rho_r\theta_B}\right)(r_t - \overline{r})\right\}}^{\text{Expected future short rates}} + \overbrace{\left\{\overline{z} + \frac{1-\theta}{1-\rho_z\theta}\rho_z\,(z_t - \overline{z})\right\}}^{\text{Expected future default losses}} \\ & + \overbrace{\left[(q_B\tau)^{-1} V_B^{(1)} \left(\overline{s}_B - (1-q_A-q_B)\,\delta_{B0}\right)\right]}^{\text{Unconditional term/credit premia}} \\ & + \overbrace{(q_B\tau)^{-1} V_B^{(1)} \begin{bmatrix} \frac{1-\theta_B}{1-\theta_B\rho_{s_B}}(s_{B,t} - \overline{s}_B) \\ -(1-\theta_B)(1-q_A-q_B)\,k^{-1}\sum_{i=0}^{\infty}\theta_B^i E_t[\sum_{j=0}^{k-1}(d_{B,t+i-j} - \delta_{B0})] \end{bmatrix}}^{\text{Conditional term/credit premia}}. \end{aligned} \tag{35}$$

Although equations (34) and (35) show that yields in markets A and B take a similar algebraic form, the risk premia in the two markets will not be the same because of the different risks that market specialists must bear in equilibrium.

As explained further in the Internet Appendix, solving the model involves finding a solution to a system of $8k$ polynomial equations in $8k$ unknowns. Specifically, we need to determine the way that equilibrium yields and active generalist demand in markets A and B respond to shifts in asset supply in A and B: this generates 8 unknowns. We also need to determine how equilibrium yields and active generalist demand in A and B respond to the holdings of inactive generalists: this generates $8(k-1)$ unknowns.

As in the single-asset case, in the presence of supply shocks, a solution only exists if investors are sufficiently risk tolerant (i.e., for τ sufficiently large). And, there can be a multiplicity of equilibrium solutions. However, as above, there is a unique solution that has well-behaved limiting behavior. And, as we show below, this solution delivers comparative statics that accord with common sense.

What is the economic intuition for the multiplicity of equilibria? In our two-asset model with slow-moving investors, there are three separate forces that give rise to equilibrium multiplicity:

1. Since specialists have short-horizons, a steeply-downward sloping demand curve creates a self-fulfilling form of discount rate risk for specialists, just as in the single-asset model. However, the relevant and well-behaved solution features a smaller equilibrium response of A yields to A supply shocks, and similarly for asset B. As above, the solutions featuring a larger response to supply shocks explode in the limiting case where supply risk vanishes.

2. Although generalists have longer investment horizons than specialists, their investment horizons are still shorter than the maturity (perpetual) of the A and B assets. Since generalists are concerned about the supply risk associated with cross-market arbitrage, the degree of equilibrium segmentation between the A and B can be self-fulfilling. For instance, if yields in market B are insensitive to shocks to the supply of A (and vice versa), cross-market arbitrage becomes very risky for generalists. Hence, generalists will not aggressively integrate markets, making the low sensitivity of B yields to A supply shocks self-fulfilling. Conversely, if generalists behave as if markets are highly integrated, then cross-market arbitrage becomes less risky and, yields in B will be more sensitive to A supply shocks (and vice versa). However, the relevant and well-behaved solution always features more aggressive cross-market arbitrage and, thus, tighter cross-market integration. The solutions with weak cross-market arbitrage explode in the limit where supply risk vanishes: to induce generalists to absorb a A supply shock, the yields in B must drop massively in response to a tiny rise in the supply of A.

3. The final source of multiplicity stems from the way that active generalists and, therefore, bond yields react to the holdings of inactive generalists. In the unique, well-behaved equilibrium, active generalists reduce their holdings less than one-for-one in response to abnormally large holdings of inactive generalists. As a result, large holdings of inactive generalists reduce equilibrium yields. However, there are also solutions in which active generalists "overreact" to the holdings of inactive generalists, reducing their holding more than one-for-one. This can lead to situations where large holdings of inactive generalists actually raises equilibrium yields. This solution behaves oddly in the limit where the number of generalists grows vanishingly small, with a tiny number of active generalists taking extremely large bets.

We solve this system of polynomial equations numerically in Python using the Powell hybrid algorithm. This algorithm performs a quasi-Newton search to find roots of a system of nonlinear equations starting from an initial guess vector. To find all of the roots, we apply this algorithm by sampling over $10,000$ different initial conditions. As discussed above, we restrict attention to those solutions where active generalists reduce their holdings less than one-for-one in response to abnormally large holdings of inactive generalists. Of these, we focus on the single solution where the price of A

(B) is less sensitive to shocks to the supply of A (B) and is more sensitive to shocks to the supply of B (A).[9]

2.3 Defining market integration

What do we mean by "market integration"? We define markets as being *integrated in the short-run* if, at each date, *conditional risk premia* in both markets reflect the same conditional prices of factor risk. For example, the pricing of interest rate risk is conditionally integrated across markets if, at each date, the expected return per unit of exposure to short-rate shocks is the same in markets A and B. Similarly, we will say that markets are *integrated in the long-run* when average, or *unconditional risk premia* in both markets reflect the same unconditional prices of risk. Unconditional integration is therefore a weaker form of market integration than conditional integration.

Note that, in our model, market integration has nothing to do with the speed by which *fundamental cash flow news* is reflected in asset prices. In our model, fundamental cash flow news is reflected instantaneously in both markets. To see this, consider the terms in curly brackets in equations (34) and (35) above. Both A and B share exposure to news about changes in future short rates and this news is reflected *identically* in their yields.

In our model, the degree of market integration depends on which investors can bear risk at different horizons and is driven by two parameters: $(1 - q_A - q_B)$ and k. The first parameter, $(1 - q_A - q_B)$, is the population share of generalists. This parameter determines the degree of long-run integration between markets. For instance, if $(1 - q_A - q_B) \approx 1$, markets will be well integrated in the long-run even if k is large. The second parameter, k, indexes the speed with which generalist capital can flow between markets. Thus, k determines the degree of short-run integration. Markets are perfectly segmented if $(1 - q_A - q_B) = 0$ or $k \to \infty$. If either of these conditions holds, the two markets operate independently of each other.

Formally, collect all of the 1-period returns in a vector $\mathbf{rx}_{t+1}$ and the asset supplies in a vector $\mathbf{s}_t$. Letting $rx_{M_t,t+1} = \mathbf{s}_t'\mathbf{rx}_{t+1}$, markets are *integrated in the short-run* if

$$\begin{aligned} E_t\left[\mathbf{rx}_{t+1}\right] &= \tau^{-1} Var_t\left[\mathbf{rx}_{t+1}\right]\mathbf{s}_t \\ &= \boldsymbol{\beta}_t\left[\mathbf{rx}_{t+1}, rx_{M_t,t+1}\right] E_t\left[rx_{M_t,t+1}\right] \end{aligned} \tag{36}$$

where $\beta_t\left[\mathbf{rx}_{t+1}, rx_{M_t,t+1}\right] = Var_t\left[\mathbf{rx}_{t+1}\right]\mathbf{s}_t / \left(\mathbf{s}_t' Var_t\left[\mathbf{rx}_{t\to t+j}\right]\mathbf{s}_t\right)$ and $E_t\left[rx_{M_t,t+1}\right] = \mathbf{s}_t' E_t\left[\mathbf{rx}_{t+1}\right]$. In other words, markets are integrated in the short-run if, at each date, a conditional-CAPM based on the current market portfolio ($rx_{M_t,t+1} = \mathbf{s}_t'\mathbf{rx}_{t+1}$) prices both the A and B assets. In our model, markets are integrated in the short-run if and only if $(1 - q_A - q_B) = 1$ and $k = 1$.[10]

[9] Specifically, we select solution vectors that satisfy the restrictions $-1 < \sum_{i=1}^{k-1} \delta_{A1}[d_{A,t-i}] < 0$ and $-1 < \sum_{i=1}^{k-1} \delta_{B1}[d_{B,t-i}] < 0$, where $\delta_{A1}[d_{A,t-i}]$ denotes the element of the $\boldsymbol{\delta}_{A1}$ solution vector that captures the way that active generalists' demands for A responds to inactive generalists' holdings of A in period $t-i$. We then pick the single solution among the remaining with the smallest value of $\alpha_{A1}[s_A]$ and $\alpha_{B1}[s_B]$.

[10] In our setting, a conditional-CAPM holds if and only if the conditional prices of factor risk are the same in both markets at each date. To see this, write $rx_{A,t+1} - E_t\left[rx_{A,t+1}\right] = \boldsymbol{\phi}_A'\boldsymbol{\varepsilon}_{t+1}$ where $\boldsymbol{\varepsilon}_{t+1}$ are the (four) factor innovations and $\boldsymbol{\phi}_A$ are the factor loadings for asset A. Proceeding similarly for market B and stacking these equations, we have $\mathbf{rx}_{t+1} - E_t\left[\mathbf{rx}_{t+1}\right] = \boldsymbol{\Phi}\boldsymbol{\varepsilon}_{t+1}$ where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_A\ \boldsymbol{\phi}_B]'$. Therefore, when $(1 - p_A - p_B) = 1$ and $k = 1$, we have $E_t\left[\mathbf{rx}_{t+1}\right] = \tau^{-1} Var_t\left[\mathbf{rx}_{t+1}\right]\mathbf{s}_t = \boldsymbol{\Phi}\left(\tau^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}'\mathbf{s}_t\right) = \boldsymbol{\Phi}\boldsymbol{\lambda}_t$ where $\boldsymbol{\lambda}_t = \left(\tau^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}'\mathbf{s}_t\right)$ are the (four) conditional prices of factor risk at time t. By contrast, when $(1 - p_A - p_B) \neq 1$ and $k \neq 1$, there is no conditional CAPM that will price the 1-period returns on the A and B assets.

Similarly, markets are *integrated in the long-run* if

$$\begin{aligned} E\left[\mathbf{rx}_{t\to t+k}\right] &= \tau^{-1} Var_t\left[\mathbf{rx}_{t\to t+k}\right] E\left[\mathbf{s}_t\right] \\ &= \boldsymbol{\beta}[\mathbf{rx}_{t\to t+k}, rx_{\overline{M},t\to t+k}] E[rx_{\overline{M},t\to t+k}], \end{aligned} \tag{37}$$

where $\beta[\mathbf{rx}_{t\to t+k}, rx_{\overline{M},t\to t+k}] = Var_t\left[\mathbf{rx}_{t\to t+k}\right] E\left[\mathbf{s}_t\right] / \left(E\left[\mathbf{s}_t'\right] Var_t\left[\mathbf{rx}_{t\to t+k}\right] E\left[\mathbf{s}_t\right]\right)$ and $E[rx_{\overline{M},t\to t+k}] = E\left[\mathbf{s}_t'\right] E\left[\mathbf{rx}_{t\to t+k}\right]$. In other words, markets are integrated in the long-run if the same unconditional-CAPM based on the average market portfolio ($rx_{\overline{M},t\to t+k} = E\left[\mathbf{s}_t'\right] \mathbf{rx}_{t\to t+k}$) prices both the A and B assets on average. In our model, markets are integrated in the long-run if and only if $(1 - q_A - q_B) = 1$, irrespective of k.

Economically, the reason markets are not integrated is because cross-market arbitrage is risky for generalists, much like in Gromb and Vayanos (2002). Unless $(1 - q_A - q_B) = 1$ and $k = 1$, short-run integration fails because generalists demand compensation for the risk associated with the short-run trades they place to exploit the cross-market pricing differences that arise following supply shocks. Similarly, unless $(1 - q_A - q_B) = 1$, long-run integration fails because generalists are engaged in a risky "cross-market arbitrage" trade even in the long run and must be compensated for its risks. Specifically, when $(1 - q_A - q_B) < 1$, generalists will not hold the market portfolio in the steady-state (i.e., $E\left[d_{A,t}\right] \neq E\left[s_{A,t}\right]$ and $E\left[d_{B,t}\right] \neq E\left[s_{B,t}\right]$). Relative to the market portfolio, generalists' portfolio will incorporate a tilt that reflects cross-market pricing differences. And, generalists will demand compensation for bearing the risks stemming from this portfolio tilt.[11] Thus, in the general case where $(1 - q_A - q_B) < 1$ and $k > 1$, we obtain neither short-run nor long-run market integration.

How should one think about the relevant values for $(1 - q_A - q_B)$ and k empirically? Clearly, the relevant values of $(1 - q_A - q_B)$ and k depend crucially on the two markets being considered. For instance, U.S. Treasury and U.S. Agency bonds are often overseen by the same portfolio manager within a large institution. As a result, we would expect $(1 - q_A - q_B)$ to be near 1 and k to be low, so the two markets would be tightly integrated even in the short-run: Treasury supply shocks would be rapidly transmitted to Agency debt markets and vice versa. However, in other cases, such as U.S. Treasury bonds and corporate bonds, or the fixed-income market and the equity market, it is natural to think that $(1 - q_A - q_B)$ is well below 1 and that $k > 1$. Although different asset classes are often held by the same generalists—e.g., pension funds or endowments, most of these investors are quite slow to reallocate capital.

3 Market integration following large supply shocks

How do prices adjust across different asset markets following large supply shocks? Here we use our model to explore asset price dynamics following shocks. We are interested in understanding how these dynamics depend on the model's underlying parameters, especially $(1 - q_A - q_B)$ and k.

Table 1 lists the illustrative set of parameter values that we use in these numerical exercises. For the purposes of these illustrations, it may be helpful to think of market A as the U.S. Treasury market

[11] In the symetric case where $q_A = q_B$ and $\overline{s}_A = \overline{s}_B$, we have $\delta_{B0} > \delta_{A0}$. The reason is that the B asset is riskier than the A asset since the former is exposed to cash-flow risk. As a result, B-specialists will hold less of the B-asset than A-specialists hold of the A-asset. Relative to the market portfolio, this means that generalists will be overweight the B asset and underweight the A asset.

and market B as the corporate bond market. We use annualized values so that one period in our numerical exercises corresponds to one year. The total average supply of assets in each market is normalized to be one unit.

We begin our analysis by choosing $k = 4$ years and $q_A = q_B = 45\%$, but later show comparative statics for these parameters. Based on these values, our simulations assume that most of the capital in each market is operated by specialists, with 10% being controlled by flexible generalist investors, one-fourth of whom re-allocate their portfolios each year. Our choice of $k = 4$ is somewhat arbitrary, but we think of this as capturing the empirically relevant case of pension funds or endowments who typically review their asset allocations on an annual or biannual basis and, even then, only sluggishly adjust their portfolios towards some evolving target.

3.1 Unanticipated supply shocks

Baseline example We first consider the impact of an unanticipated supply shock that increases the supply of asset A (Treasuries) by 50% in period 10. To make the intuition as stark as possible, we focus on the case of a near-permanent supply shock and set $\rho_{s_A} = 0.999$. Specifically, Figure 1 illustrates the price impact of this shock, plotting the evolution of expected annual returns and bond yields in market A (Treasuries) and market B (corporate bonds). Figure 2 shows how specialists and generalist investors adjust their holdings in response to the shock. Finally, Figure 3 plots the yield spread between the B asset (corporate bonds) and the A asset (Treasuries).

Prior to the supply shock in period 10, Figure 1 shows that the risk premium in market B (corporate bonds) is 0.64% per annum versus a risk premium of 0.48% in market A (Treasuries). The additional risk premium of 0.16% obtains because market B (corporate bonds) is subject to default risk, which exposes investors to an additional source of cash flow risk and amplifies the supply risk facing corporate bond holders. The initial yield in market B is 4.84% per annum versus a yield of 4.48% in market A. The steady-state yield in market A equals the average short-term riskfree rate of $\overline{r} = 4.00\%$ plus the steady-state risk premia of 0.48%. The 0.36% steady-state yield spread between the B and A markets equals the difference in steady-state risk premia of 0.16% plus the market B's expected default losses of $\overline{z} = 0.20\%$ per annum.

When the supply shock hits the market A in period 10, expected returns and yields in both markets react immediately. Figure 1.A shows that expected returns in market A *overreact* and reach a peak of 0.70% before ultimately falling back to a long-run level of 0.64%. The overreaction of expected returns for asset A illustrates a general property of models that feature slow-moving capital such as Duffie (2010), namely, the relative *steepness* of short-run demand curves and relative *flatness* of long-run demand curves.

In contrast, Figure 1.A shows the key novel implication of our model: expected returns in market B actually *underreact* to the shock to the supply of A, rising slowly from 0.64% to a new long-run level of 0.72%. Why does market A overreact to the supply shock while market B underreacts? Figure 2.A shows how the positions of different market participants evolve over time. Following the initial supply shock in market A, both specialist demand in A ($b_{A,t}$) and active generalist demand in A ($d_{A,t}$) spike upwards. As a partial hedge against their increased holdings of A, active generalists reduce their holdings in market B ($d_{B,t}$). This reduction in generalists' B holdings is motivated by a need to reduce the common short-rate risk (and supply risk) across their holdings in both markets. To fill

the void left by the generalists, specialists in market B must hold more of the B asset ($b_{B,t}$). As time passes and more generalists reallocate their portfolios in response to the shock, the active demands for A ($b_{A,t}$ and $d_{A,t}$) decline slowly towards their new long-run levels.

In our model, the dynamics of bond risk premia are tied to the dynamics of the "active supply" of A and B that must be absorbed by active market participants each period. By active supply we mean the total supply less the assets that are being held off the market by inactive generalists, corresponding to the right-hand side of equation (27). The evolution of the active supplies is shown in Figure 2.B. The dynamics of active supply mirror those for bond risk premia shown in Figure 1.A. The initial supply shock to A in period 10 immediately increases the active supply in A but has no immediate effect on the active supply of B. This is because slow-moving generalists have yet to reduce their holdings in market B. Over the ensuing periods, generalists gradually increase their holdings of A and reduce their holdings of B. Therefore, the active supply in A gradually declines while the active supply in B gradually rises.

Recall that $k = 4$ in this example, so by period 13 all generalist investors have re-allocated their portfolios in response to the supply shock in period 10. However, the gradual adjustment of generalists gives rise to modest echo effects after period 13, generating a series of damping oscillations that converge to the new long-run equilibrium. As in Duffie (2010), these oscillations arise because generalists who reallocate soon after the supply shock hits take large opportunistic positions. These large positions temporarily reduce the active supply of A (and increase the supply of B) and then need to be absorbed in later periods.

In market A, conditional risk premia are the sum of risk premia related to interest rate risk, the supply of A, and the supply of B. Changes in total risk premia are driven primarily by the pricing of interest rate risk. Conditional risk premia in market B can be similarly decomposed into its components, which also include a premium for cash flow risk. Following the supply shock, the premia associated with interest rate risk differs significantly between the two asset markets. As generalists react to this pricing discrepancy, the difference in interest rate risk premia between the two markets gradually narrows. However, the difference does will not vanish in the long run because of the permanent risks associated with cross-market arbitrage.

Because markets are partially segmented, large supply shocks can have surprising effects on seemingly unrelated risk premia in our model. For example, because it triggers significant cross-market capital flows, the shock to the supply of asset A (Treasuries) actually raises the risk premium that corporate bond investors earn for bearing default risk in market B (corporate bonds), even though Treasury bonds themselves have no exposure to default risk. In this way, our model may shed light on the otherwise puzzling finding that central bank purchases of long-term government bonds appear to have reduced credit risk premia (Krishnamurthy and Vissing-Jorgensen [2011]).

Figure 1.B shows the reactions of bond yields in both markets. The overreaction of the A market and the underreaction of the B market is more muted in yield space than in risk premium space. This is natural since bond yields reflect weighted averages of future bond risk premia.[12] Market A yield overreacts by 11% of the total long-run impact and market B yield underreacts by 19% of the total long-run impact.

[12] Specifically, generalizing (16) we have $y_{A,t} = (1-\theta_A)\sum_{i=0}^{\infty}\theta_A^i E_t[r_{t+i} + \tau^{-1}V_A^{(1)}b_{A,t+i}]$ and $y_{B,t} = (1-\theta_B)\sum_{i=0}^{\infty}\theta_B^i E_t[r_{t+i} + z_{t+i+1} + \tau^{-1}V_B^{(1)}b_{B,t+i}]$.

We can assess the evolution of market segmentation over time by examining the dynamics of yield spreads. Figure 3 shows that the yield spread between the two markets, $y_B - y_A$, compresses due to the increased supply of asset A. However, because yield A overreacts and yield B underreacts, the yield spread overreacts even more (31%) than the A market yield.

Comparative statics In Table 2, we perform a variety of comparative statics exercises to illustrate how the price dynamics following supply shocks depend on the parameters of our model. We focus on the parameters governing market structure: the population share of generalist investors $(1 - q_A - q_B)$ and the frequency at which generalists can rebalance (k).

For a given set of model parameters, we summarize the impact of the supply shock on both the A and B markets by listing the yields and expected annual returns in (i) the period before the shock arrives (labeled as "pre-shock level"), (ii) the period when the shock arrives (labeled as "short-run Δ"), and (iii) in $2k$ periods after the shock arrives (labeled as "long-run Δ").

We define the degree to which bond yields over- or underreact as the difference between the short-run change and the long-run change, expressed as a percentage of the long-run change[13]

$$\%Over\text{-}Reaction(y) \equiv \frac{(y_t - y_{t-1}) - (y_{t+2k} - y_{t-1})}{(y_{t+2k} - y_{t-1})}.$$

Our measure of over-reaction for risk premia, $\%Over\text{-}Reaction(E[rx])$, is defined analogously. According to this definition, using our baseline set of parameters, yields in market A overreact by approximately 11%, while yields in market B underreact by 19%.

The second row in Table 2 shows that, if market participants are more risk tolerant, this reduces the price impact of the supply shock on both market A and market B. Changing investor risk tolerance has a similar impact on the short- and long-run response of yields to shocks. Thus, the degree of overreaction or underreaction in each market is unchanged in percentage terms.

We next change the mix between generalist and specialist investors. q_A and q_B indicate the relative fraction of specialists in market A and B, respectively. In row 3, we set $q_A = q_B = 0.5$ so there are no generalists and the two markets are completely segmented: a supply shock in the market A is not transmitted to the market B and vice versa.

In contrast, in the case of many generalists and few specialists, the markets are well integrated, so that both the A and B markets overreact to a supply shock that directly hits only the A market. In this case, shown in row 4 which sets $q_A = q_B = 0.2$, the two markets behave as essentially one and the result is similar to the single-market case with slow-moving capital studied in Duffie (2010).

We next change the mix between market A specialists and market B specialists, holding fixed the overall mix between generalists and specialists. Row 5 of Table 2 shows that if we hold the total number of specialists the same at $q_A + q_B = 0.9$, then as we increase the proportion of specialists in B and decrease the proportion of specialists in A, we get more over-reaction in A. The B market is only modestly affected by this change because the supply shock is primarily being absorbed by generalists

[13] Since our supply shock is not quite permanent, we subtract off the constant $(1-\rho_{S_A}^{2k})/\rho_{S_A}^{2k}$ from $\%Over\text{-}Reaction$ to ensure that our measure is zero the case of perfectly conditionally-integrated ($1-p_A-p_B = k = 1$) or perfectly segmented markets ($1 - p_A - p_B = 0$) in which there is no "over-reaction" but only "reaction." This is because in these limiting cases we have $[(y_t - y_{t-1}) - (y_{t+2k} - y_{t-1})] / [y_{t+2k} - y_{t-1}] = \left[\alpha_{s_A} s_{A,t} - \alpha_{s_A} s_{A,t} \rho_{s_A}^{2k}\right] / \alpha_{s_A} s_{A,t} \rho_{s_A}^{2k} = (1-\rho_{s_A}^{2k})/\rho_{s_A}^{2k}$. For $k = 4$ and $\rho_{s_A} = 0.999$, this constant is 0.8%.

anyway.

Recall that k is the number of periods it takes for generalists to fully reallocate their portfolios and that $k = 4$ in our base case. In row 6 we instead set $k = 2$, so half the generalists reallocate their portfolio each period, and the other half reallocate in the next period. Naturally, this smaller value of k reduces the over-reaction in market A and the under-reaction in market B. Similarly, when we set $k = 6$ in row 7, there is more over-reaction in market A, and more under-reaction in market B.

Note that k also affects the unconditional risk premium that investors earn over the long run. As we increase k, there are two competing effects on unconditional risk premium. On the one hand, generalists with longer horizons worry less about a fixed amount of transitory discount rate risk, leading to a decline in the unconditional price of discount rate risk.[14] On the other hand, the steady state quantity of discount rate risk that investors must bear actually grows with generalist horizons k.[15] As shown in Table 2, the latter effect generally tends to dominate.[16] In summary, as we increase k, *supply shocks* have a larger impact on conditional risk premia and this increase in *supply risk* tends to raise unconditional risk premia.

In row 8, we ask how our results depend on the relative sizes of the A and B markets. Relative to our base case of two equally sized markets, we find that generalists are better able to integrate a small market with a larger market. We keep $\overline{s}_A + \overline{s}_B = 2$ and $q_A + q_B = 0.9$, but we now assume that $\overline{s}_A = 5/3$ and $\overline{s}_B = 1/3$ so average supply in market A is $5\times$ that in market B. We also assume that $q_A = 0.45 \times \overline{s}_A$ and $q_B = 0.45 \times \overline{s}_B$ as in the baseline, which implies $q_A = 0.75$ and $q_B = 0.15$. As in our baseline, we consider a shock that raises the supply of A by 0.5. Row 8 shows that A over-reacts less and that B under-react less to the shock than under our baseline. The explanation is that market B is now much smaller relative to total generalist risk tolerance. As a result, a cross-market arbitrage position of a given size is better able to keep prices in market B close to those in market B.

Finally, we ask whether the supply shock scenario we have considered has symmetric effects if it is delivered in market B. Row 9 of the table shows that the answer is yes: the price impact on market A when the supply shock hits market B is exactly the same as the price impact on market B when the supply shock hits A**.** This symmetry of cross-market price impact is natural and is a general property of our model under certain conditions.[17]

Row 10 in Table 2 shows that the degree of market integration depends on the amount of default

[14] Since mean-reverting supply shocks generate negative serial correlation in returns, the variance ratio $Var_t[rx_{A,t\to t+k}]/k$ will be decreasing in k holding fixed the endogenous parameters that govern the return generating process. Thus, as in Campbell and Viceira (2002), longer-horizon investors worry less about transitory supply (discount rate) risk, leading them to take larger positions in risky assets.

[15] Formally, as we increase k, the endogenous parameters that govern the return generating process are not held fixed. Since fewer long-horizon investors are active in a given period, the short-term price impact of supply shocks grows, leading to an rise in the quantity of discount rate risk.

[16] Formally, let $E[rx^A_{t+1}] = (p_A\tau)^{-1}\left(\overline{s}^A - (1-p_A-p_B)\,\delta_{A0}\right)V^{(1)}_A$ denote the unconditional risk premium. We have

$$\frac{\partial E[rx^A_{t+1}]}{\partial k} = (p_A\tau)^{-1}\left[\frac{\partial V^{(1)}_A}{\partial k}\left(\overline{s}^A - (1-p_A-p_B)\,\delta_{A0}\right) - V^{(1)}_A\,(1-p_A-p_B)\,\frac{\partial \delta_{A0}}{\partial k}\right]$$

When $\sigma^2_{s_A}$,$\sigma^2_{s_B} = 0$, $\partial V^{(1)}_A/\partial k = \partial\delta_{A0}/\partial k = 0$ and unconditional risk premia are independent of k. When $\sigma^2_{s_A}$,$\sigma^2_{s_B} > 0$, we have $\partial V^{(1)}_A/\partial k > 0$ and $\partial\delta_{A0}/\partial k > 0$. In general, we find that the former effect tends to dominate, so that unconditional risk premia are increasing in k.

[17] Specifically, the cross-market price impact is symmetric in expected return space so long as $\rho_{s_A} = \rho_{s_B}$ and is symmetric is yield space so long as $\rho_{s_A} = \rho_{s_B}$ and $\theta_A = \theta_B$.

risk in market B. As we increase σ_z^2 from row 8 to row 9, we have less long-run and short-run integration between the two markets, because A and B are more distant substitutes, consistent with the logic of Wurgler and Zhuravskaya (2002). If B-specific risk is large, then the market with the shock will have a larger peak because generalists are less willing to integrate the markets. However, there is not much effect on under-reaction in the market that does not receive the shock.

As we reduce the amount default risk in market B, shocks to the supply of B have a larger price impact on market A because the two assets are closer substitutes and cross-market arbitrage is less risky. In the limit when $\sigma_z^2 = 0$, the markets would be perfect substitutes and thus perfectly integrated: conditional risk premia would be identical in the two markets.

3.2 Anticipated supply shocks

We next study asset price dynamics following the *announcement* of a large *future change* in asset supply. As an example of a large pre-announced supply change, consider the Large Scale Asset Purchase programs initiated by the Federal Reserve between 2008 and 2013. The Fed's initial announcement of long-term bond purchases occurred on November 25, 2008, but asset purchases did not begin until January 2009 and continued in the months thereafter.

To mimic the announcement of a future increase in the supply of asset A, we assume that $s_{A,t}$ jumps up at some time t and we simultaneously increase the demands of inactive generalist investors for A such that the *active supply* of asset A does not change at time t. This means that, unlike the case of an unanticipated supply shock, it would be possible to clear the market at time t without any increase in the holdings of A specialists or active generalists. Thus, the only reason that prices change when a future supply change is announced is because the announcement leads active long-horizon generalists to opportunistically increase their current holdings of A.

Formally, letting $\varepsilon_t[X_t] = X_t - E_{t-1}[X_t]$ denote the time t *innovation* or *surprise* to some random process X_t, an anticipated supply shock is defined so that the innovation to the right-hand of equation (27) is zero when it is announced at time t:

$$\varepsilon_t[s_{A,t}] - (1 - q_A - q_B)\, k^{-1} \sum_{j=1}^{k-1} \varepsilon_t[d_{A,t-j}] = 0. \tag{38}$$

Furthermore, if we vary $\{\varepsilon_t[d_{A,t-j}]\}_{j=1}^{k-1}$ holding fixed $\sum_{j=1}^{k-1} \varepsilon_t[d_{A,t-j}] = (k/(1 - q_A - q_B))\, \varepsilon_t[s_{A,t}]$, we can alter the announced *timing* of a supply change holding fixed the announced size of the *cumulative* change. As we discuss shortly, exercises of this sort can be used to evaluate different asset purchase strategies for a central bank seeking to affect the level of long-term interest rates or for share repurchasing firm seeking to give boost to its stock price.

We begin with the simple case of the pre-announcement of a one-time, near-permanent jump in the supply of asset A. Specifically, Figure 4 shows the dynamics of risk premia and bond yields when $k = 4$ and market participants learn in period 5 that the supply of asset A will increase by 50% in period 8. Figure 4.A shows that annual risk premia in market A actually *decline* slightly after the announcement in period 5 and before the supply rises in period 8. And risk premia in market A still jump up when bond supply actually jumps to its new level in period 8. The risk premium in market B, which is only indirectly affected due to cross-market arbitrage by generalists, rises gradually over time.

What drives these dynamics? Upon the announcement, generalists begin to gradually adjust in the direction of the anticipated shock, buying asset A and selling asset B. Risk premia decline in market A because A specialists provide temporary liquidity to generalists, planning to replenish their inventories once the supply shock actually lands.

Figure 4.B shows the evolution of bond yields in response to the anticipated supply shock. Yields in both market A and market B rise *gradually* to their new steady-state levels. The yields in market A underreact by 51% and the yields in market B underreact by 45%. Why do A yields rise gradually when the supply shock is pre-announced but over-reacted in Figure 1 when the same shock was unanticipated? First, risk premia in the near future have a larger effect on bond yields than those in the distant future. Second, risk-premia in the A market are only expected to rise significantly once the supply actually increases at time 8. In combination, these two facts imply that yields in market A must rise gradually over time.

Figure 5 shows how the positions of market participants evolve over time in response to this announced supply increase. Figure 5.A shows that active generalists opportunistically increase their holdings of A (d_A) and decrease their holdings of B (d_A) when the shock is announced at time 5.[18] The gradual build up of generalist demand in A is responsible for the slight decline in annual A market risk premia from periods 5 to 7 before the upward jump in period 8. Similarly, the gradual reduction of generalists' demand for B results in a slow rise in B market risk premia. In contrast to the generalists, the specialists' demand in market A (b_A) decreases initially then increases. This is because specialists can adjust quickly, and thus they have the ability to front-run the anticipated change in supply—specialists reduce their portfolio holdings of A just before the positive supply shock and increase holding of A immediately after the shock.

Figure 6 compares the dynamics of prices in the case of anticipated versus unanticipated increase in supply. In both cases, the supply of asset A increases by 50% at time 8. In the anticipated case, this supply increase is pre-announced at time 5, while in the unanticipated case, the shock is not pre-announced and is a surprise at time 8. Whether or not a shock is anticipated, the *long-run* impact on yields and risk premia is the same. However, the short-run effects can be quite different. When the shock is a surprise, yields and risk premia in the A market overreact more strongly at time 8. Pre-announcing the supply shock mobilizes slow-moving generalists before the supply of A actually rises at time 8. This early mobilization reduces the active supply of asset A that must be absorbed when the shock lands at time 8, damping the overreaction of market A. While introducing a delay between announcement and the increase in supply limits overreaction, this also lengthens the amount of time it takes for the full impact to be reflected in prices. This tension may be relevant for central bankers designing asset purchase programs: allowing for a long period of time between a purchase announcement and implementation results in less profits for market specialists and, therefore, less short-term cost of policy implementation. Of course, pre-announcement necessarily delays the desired

[18] Why do generalists buy A bonds *in advance* of the supply increase? Generalists have long-horizons ($k = 4$ in this example), but can only adjust their portfolios slowly. Although generalists expect A bond yields to rise over the next 4 periods and, therefore, expect to suffer a capital loss on A bonds, they expect this capital loss to be more than offset by an increase in income from holding A bonds. As a result, the expected cumulative 4-period excess return from holding A bonds rises when a future supply increase is announced at t. Formally, since $E_t\left[rx_{A,t\to t+4}\right] = \{\sum_{i=0}^{3} E_t[y_{A,t+i} - r_{t+i}]\} - \{(\theta_A/(1-\theta_A))\, E_t[y_{A,t+4} - y_{A,t}]\}$, this means that change in the first term in curly braces outweighs the second term in curly braces. Using equation (26), the significant rise in $E_t\left[rx_{A,t\to t+4}\right]$ then translates to an increase in $d_{A,t}$ and decline in $d_{B,t}$.

impact on asset prices.

We can also use the model to describe more complex paths of supply shocks. Consider the effects of an announcement that asset purchases are going to be made over multiple periods, much like the Federal Reserve did when it announced in 2013 that it would purchase $40 billion of MBS each month in QE3. In Figure 7, asset sales are announced in period 5 and carried out from period 6 to 8. Figure 7.A shows that risk premia rise gradually over the period of policy implementation. Figure 7.B shows that yields also rise gradually over the implementation period from time 6 to 8. In summary, we observe substantial under-reactions in returns and yields in both markets at the time of the policy announcement in period 5. An event study during the time of the initial announcement would capture some of the market reaction, but it could significantly understate the long-run impact.

In Figure 8, we show how generalists and specialist holdings react following the announcement in period 5 that the supply of A will rise by 50% from period 6 to period 8. Generalists and specialists both increase their demands for A in order to accommodate the overall increase in supply. Generalists also reduce their demand for B to partially hedge against their purchase of A. However, this market integration occurs at a slow pace since only one-fourth of the generalists can reallocate between the two markets each period. Lastly, specialists in B increase their demand (b_B) to fill the gap left by generalists.

3.3 Temporary supply shocks

Thus far, we have only considered the price impact of permanent supply shocks. However, many supply shocks may be more temporary in nature. For example, Hanson (2014) shows that shocks to the effective duration of US mortgage-backed securities have a half-life of just six months, generating highly transient shocks to the supply of interest rate risk in US fixed income markets. Similarly, presumably investors did not interpret the Federal Reserve's initial announcement that it would purchase large quantities of long-term bonds as a permanent reduction in supply: investors expected Quantitative Easing to be a temporary policy, generating a temporary supply shock that was expected to revert over time (Greenwood, Hanson, and Vayanos [2015]).

In Figure 9, we show the impact on bond risk premia of unanticipated shocks to the supply of asset A which have varying degrees of persistence as governed by ρ_{s_A}. In Figure 9, the supply shocks arrives at time 1 and, as above, this experiment assumes that $k = 4$. Unsurprisingly, when the supply shock reverts more quickly, the effect on risk premia on both markets decays more quickly over time. This decaying effect on risk premia simply mirrors the decaying supply shock and would obtain even if markets were perfectly integrated period by period (i.e., if $q_A = q_B = 0$ and $k = 1$).

More interestingly, Figure 9 shows that slow-moving generalists play a less active role in integrating markets when the shock is expected to be more temporary. This can be seen most easily by comparing the ratio of the impact on $E_t\left[rx_{B,t+1}\right]$ to the impact on $E_t\left[rx_{A,t+1}\right]$. Slow-moving generalists trade less aggressively when shocks are expected to be short-lived because they cannot move quickly enough to capture the transient difference in expected return between the two markets. This more muted cross-market arbitrage response reduces the impact of the supply shock on the B market. As a result, in a world with slow-moving capital, temporary supply shocks have a more localized impact on prices than more persist shocks.

4 Multiple assets per market

We now explore how the main results of our model carry over to a more complex setting in which there are multiple risky assets trading in each market. Subject to some mild conditions which guarantee that cross-market arbitrage remains risky, we show that the intuitions from the two risky asset model carry over to this richer setting. Specifically, we show that a particular conditional CAPM prices all assets in the first market and that a different conditional CAPM prices all assets in the second market. These two market-specific pricing models are linked over time by the cross-market arbitrage activities of slow-moving asset allocators, who take steps to equalize the price of risk in the two markets. Much like before, the degree of market integration depends on the risks faced by cross-market arbitrageurs.

4.1 Markets and assets

Suppose there are N bonds in market A, denoted A_1, A_2, ..., A_N. As above, we assume that the A-market bonds are default–free and are only exposed to interest rate risk.[19] However, the bonds have different durations. Specifically, asset A_n has duration D_{A_n}. As in equation (2), the log excess return on bond A_n over the short-term interest rate from time t to $t+1$ is

$$rx_{A_n,t+1} \approx \overbrace{\frac{1}{1-\theta_{A_n}}}^{D_{A_n}} y_{A_n,t} - \overbrace{\frac{\theta_{A_n}}{1-\theta_{A_n}}}^{D_{A_n}-1} y_{A_n,t+1} - r_t, \tag{39}$$

where $\theta_{A_n} = 1 - 1/D_{A_n}$.

We also assume that there are N defaultable bonds in market B, denoted B_1, B_2, ..., B_N. The return on bond B_n from time t to $t+1$ takes the form

$$1 + R_{B_n,t+1} = (1 - Z_{t+1})^{\psi_{B_n}} (1 - U_{B_n,t+1}) \frac{(\delta_{B_N} P_{B_n,t+1} + C_{B_n})}{P_{B_n,t}}, \tag{40}$$

where Z_{t+1} is a default process common to all bonds in the B market, ψ_{B_n} is the exposure of perpetuity B_n to this systematic default factor, and $U_{B_n,t+1}$ is an idiosyncratic default process that is specific to bond B_n. Therefore, the log excess return on bond B_n from time t to $t+1$ is

$$rx_{B_n,t+1} \approx \overbrace{\frac{1}{1-\theta_{B_n}}}^{D_{B_n}} y_{B_n,t} - \overbrace{\frac{\theta_{B_n}}{1-\theta_{B_n}}}^{D_{B_n}-1} y_{B_n,t+1} - \psi_{B_n} z_{t+1} - u_{B_n,t+1} - r_t, \tag{41}$$

[19] In order to have perpetuities with different durations, we introduce a set of "geometrically decaying perpetuities." Specifically, consider a perpetuity that promises to pay a decaying stream C, δC, $\delta^2 C$, $\delta^3 C$, ... where $(1-\delta) \in [0,1]$ denotes the geometric decay rate. Thus, $\delta = 0$ corresponds to 1-period debt and $\delta = 1$ is a consol bond. Assuming a yield of $Y_{L,t}$, the price of this security is $P_{L,t} = \sum_{j=1}^{\infty} (1+Y_{L,t})^{-j} \delta^{j-1} C = C/(1+Y_{L,t}-\delta)$ which implies a Macaulay duration of $-\partial p_{L,t}/\partial y_{L,t} = (1+Y_{L,t})/(1+Y_{L,t}-\delta)$. Suppose the perpetuity's price is $\overline{P}_L = 1$ and its yield is $\overline{Y}_L$ in the steady-state. This implies a coupon of $C = 1 - \delta + \overline{Y}_L$ and a steady-state duration of $-\partial p_{L,t}/\partial y_{L,t} = (C+\delta)/C$ which is increasing in δ.

Using the same steps as above, the log return on the decaying perpetuity from t to $t+1$ is approximately $r_{L,t+1} = \log[(\delta P_{L,t+1} + C)/P_{L,t}] \approx (1-\theta)^{-1} y_{L,t} - \theta(1-\theta)^{-1} y_{L,t+1}$. where $\theta = \delta/(\delta + \exp(c - \overline{p}_L))$. Since the steady-state price is par, we have $\theta = \delta/(\delta + C)$. Thus, bond duration is $-\partial p_{L,t}/\partial y_{L,t} = (1-\theta)^{-1} = (\delta + C)/C$ which corresponds to the Macaulay duration when the perpetuity is trading at par. Thus, we assume that security A_n has a geometric decay rate if δ_{A_n} and a coupon of $C_{A_n} = 1 - \delta_{A_n} + \overline{Y}_{A_n}$, implying a duration of $D_{A_n} = (1-\theta_{A_n})^{-1} = \left(1+\overline{Y}_{A_n}\right)/\left(1+\overline{Y}_{A_n}-\delta_{A_n}\right)$.

where $\theta_{B_n} = 1 - 1/D_{B_n}$ and $u_{B_n,t+1} = -\ln(1 - U_{B_n,t+1})$. Given this formulation for default losses it is perhaps most natural to think of bond B_n as corresponding to a *portfolio* of defaultable bonds, albeit one that *imperfectly diversified* and is therefore exposed to *idiosyncratic* default losses. For instance, one could think of B_n as representing a portfolio of all bonds in a certain industry with some specified credit rating and some specified maturity.

We assume that the processes for the short rate r_t and for the common default process z_t are as in equations (7) and (8) earlier. We assume that idiosyncratic default process for bond B_n follows

$$u_{B_n,t+1} = \overline{u}_{B_n} + \rho_{u_{B_n}} (u_{B_n,t} - \overline{u}_{B_n}) + \varepsilon_{u_{B_n},t+1}. \tag{42}$$

We assume the net supplies that investors must hold in the A assets are

$$\mathbf{s}_{A,t} = \mathbf{s}_{A0} + \mathbf{s}_{A1} \times s_{A,t} \tag{43}$$

where $s_{A,t}$ follows

$$s_{A,t+1} = \rho_{s_A} s_{A,t} + \varepsilon_{s_A,t+1}. \tag{44}$$

Similarly, the net supplies that investors must hold in the B assets are

$$\mathbf{s}_{B,t} = \mathbf{s}_{B0} + \mathbf{s}_{B1} \times s_{B,t},$$

where where $s_{B,t}$ follows

$$s_{B,t+1} = \rho_{s_B} s_{B,t} + \varepsilon_{s_B,t+1}. \tag{45}$$

We assume that $\varepsilon_{r,t+1}$, $\varepsilon_{z,t+1}$, $\varepsilon_{s_A,t+1}$, $\varepsilon_{s_B,t+1}$, $\varepsilon_{u_{B_1},t+1}$, $\varepsilon_{u_{B_2},t+1}$, ..., $\varepsilon_{u_{B_N},t+1}$ are mutually orthogonal.

4.2 Market participants

As above, there are three-types of investors, each with risk tolerance τ. A-specialists are present in mass q_A, B-specialists are present in mass q_B, and generalists are present in mass $(1 - q_A - q_B)$.

Fast-moving A-specialists are free to adjust their holdings of all assets in the A market (and the riskless short-term asset) each period, but cannot hold the B assets. Let $b_{A_n,t}$ denote the demand of A specialists for asset A_n and let $\mathbf{b}_{A,t}$ denote the $N \times 1$ vector of their holdings of each of the N assets in market A. Collecting the excess returns on these N assets in a vector, the excess return on A-specialists portfolio is thus $rx_{A_t,t+1} = (\mathbf{b}_{A,t})' \mathbf{rx}_{A,t+1}$.

A-specialists have mean-variance preferences over 1-period portfolio returns and solve

$$\max_{\mathbf{b}_{A,t}} \left\{ E_t[rx_{A_t,t+1}] - (2\tau)^{-1} Var_t[rx_{A_t,t+1}] \right\} = \max_{\mathbf{b}_{A,t}} \left\{ \mathbf{b}'_{A,t} E_t[\mathbf{rx}_{A,t+1}] - (2\tau)^{-1} \mathbf{b}'_{A,t} Var_t[\mathbf{rx}_{A,t+1}] \mathbf{b}_{A,t} \right\}.$$

Thus, the demands of A-specialists are given by

$$\mathbf{b}_{A,t} = \tau (Var_t[\mathbf{rx}_{A,t+1}])^{-1} E_t[\mathbf{rx}_{A,t+1}].$$

Since this implies $Var_t[rx_{A_t,t+1}] = \tau E_t[rx_{A_t,t+1}]$, we have

$$E_t[\mathbf{rx}_{A,t+1}] = \tau^{-1} Var_t[\mathbf{rx}_{A,t+1}] \mathbf{b}_{A,t} = \boldsymbol{\beta}_t[\mathbf{rx}_{A,t+1}, rx_{A_t,t+1}] E_t[rx_{A_t,t+1}]$$

where $\boldsymbol{\beta}_t\left[\mathbf{rx}_{A,t+1}, rx_{A_t,t+1}\right] = Cov_t\left[\mathbf{rx}_{A,t+1}, rx_{A_t,t+1}\right]/Var_t\left[rx_{A_t,t+1}\right]$. Thus, the 1-period returns on all A-market assets will be priced by a local conditional-CAPM that is specific to the A-market—i.e., where the relevant "market portfolio" is the time t portfolio of A-market specialists, $rx_{A_t,t+1} = (\mathbf{b}_{A,t})'\mathbf{rx}_{A,t+1}$.

Since a symmetric analysis hold for the N assets in market B, we will have two conditional-CAPMs: one for the assets in market A and another for the assets in market B. The key question is how these two conditional-CAPMs will be linked together in equilibrium by the cross-market arbitrage activities of generalist. As above, slow-moving generalists are present in mass $1 - q_A - q_B$. Fraction $1/k$ of generalists investors are active each period and choose the portfolios of assets from the A and B markets that they will hold over the following k periods.

4.3 The risk of cross-market arbitrage

With multiple assets, the key question concerns the risks that generalists face when they undertake cross-market arbitrage. Note that assets in market A are exposed to exogenous shocks to three state variables: r_{t+1}, $s_{A,t+1}$, and $s_{B,t+1}$. Assets in market B are exposed to exogenous shocks to r_{t+1}, $s_{A,t+1}$, and $s_{B,t+1}$ as well as exogenous shocks to z_{t+1}. In addition, each asset B_n is potentially exposed to idiosyncratic shocks to $u_{B_n,t+1}$.

An interesting complication arises if generalists are able to freely choose their holdings of the N assets in market A and the N assets in market B. In this case, it may be possible to use A assets to construct a "factor-mimicking portfolio" that is only exposed to shocks to r_{t+1} and to construct a similar factor-mimicking portfolio using only B assets. If this is possible then, unless the risks associated with shocks to r_{t+1} are being priced the same in the A and B markets at each date, generalists will have a riskless arbitrage opportunity.

In general, it is *possible* to construct (nearly-perfectly) factor-mimicking portfolios if both A and B markets contain many redundant assets. But even if it is *possible* to construct factor-mimicking portfolios in both A and B markets, this may not be feasible for slow-moving generalists. For instance, generalists may lack the expertise to construct these complicated (long-short) mimicking portfolios or may face institutional frictions that make this infeasible. We can distinguish between at least three cases:

1. **Case 1:** It is possible to construct factor-mimicking portfolios in both markets A and B for each of the common risk factors.

2. **Case 2:** It is not possible to construct factor-mimicking portfolios in this way.

3. **Case 3:** It is possible to construct factor-mimicking portfolio in this way, but generalists are not capable of doing so: generalists function as coarse asset-allocators as opposed to granular cross-market arbitrageurs.

We discuss each of these three cases in greater detail.

Case 1: It is possible to construct factor-mimicking portfolios To begin, suppose that the B assets are not exposed to idiosyncratic shocks. As explained in the Internet Appendix, k-period returns in this case will satisfy a linear factor model with $2(1+k)$ independent factors (recall that the

model's state vector, $\mathbf{x}_t$, contains $2(1+k)$ elements). Specifically, there are 2 factors that correspond to innovations to fundamentals, r_t and z_t, and k factors corresponding to innovations to each of the supply factors at different horizons.[20]

Suppose generalists can freely choose positions in all $2N$ assets where $N \geq 2(1+k)$ and all assets are non-redundant (their factor loadings must be linearly independent). In this case, it will be possible to construct factor-mimicking portfolios for shocks to r_t, z_t, $s_{A,t}$, and shocks to $s_{B,t}$ using only A assets and using only B assets.[21] As a result, the active generalists will work to perfectly integrate factor pricing between the A and B markets in the short-run. Intuitively, because cross asset-class arbitrage is riskless in this case, the same risk factor prices will always prevail in all asset markets—i.e., the two markets will perfectly integrated in the short run (conditionally) and the long run (unconditionally). Of course, because generalists are slow-moving, the risk factor prices that prevail in both the A and B markets will be subject to slow-moving capital effects, e.g., risk factor prices will overreact to shocks to the supply of that risk factor.

Even if the B market assets are subject to idiosyncratic default shocks, this outcome will obtain in the limit where we hold constant the total supply of A and B assets but allow $N \to \infty$. In this case, investors' portfolios will become arbitrarily granular, implying that it will be easy for generalist investors to diversify away these idiosyncratic shocks when constructing factor-mimicking portfolios. This is essentially the intuition behind Ross's (1976) Arbitrage Pricing Theory. Thus, if generalists can freely choose positions in all $2N$ assets, Case 1 will be a good approximation in the case when N is large relative to the number of common risk factors in the A and B markets.

Case 2: It is not possible to construct factor-mimicking portfolios Next suppose generalists can freely choose positions in all $2N$ assets. However, suppose that N is not large, so it is not possible to construct accurate factor-mimicking portfolios. If the B assets are not exposed to idiosyncratic shocks, cross-market arbitrage remains risky for generalists so long as $N < 2(1+k)$. And, assuming that the B assets are exposed to idiosyncratic shocks, cross-market arbitrage will remain risky even when $N \geq 2(1+k)$. Specifically, the N assets in market B are exposed to $N + 2(k+1)$ risk factors—$2(k+1)$ that are common and N that are asset-specific. A factor-mimicking portfolio is a set of N unknown positions in the B assets that must satisfy $N + 2(k+1)$ linear equations. In general, there is no such solution. And, if N is small and idiosyncratic volatility is large, then any portfolio will do a poor job of mimicking each common factor. As a result, cross-market arbitrage will remain risky for generalists, so cross-market integration will be imperfect, both in the short-run and the long-run. This case will be empirically relevant, if the "idiosyncratic risks" are not asset specific, but are shared by a large subset of assets in a market (e.g., industry default factors). In this case, cross-market arbitrage will always entail some amount of "basis risk," rendering it risky for generalists.

Case 3: Generalists are unable to construct factor-mimicking portfolios Finally, suppose it is possible to construct accurate factor-mimicking portfolios. However, suppose that generalists

[20] Specifically, k-period returns for all securities in markets A and B will load linearly on $\sum_{i=1}^{k} \varepsilon_{r,t+i}$ and k-period returns on all B-market securities will load linearly on $\sum_{i=1}^{k} \varepsilon_{z,t+i}$. However, the k-period returns on securities with different durations will load differently on supply shocks in the near versus distant future. Thus, all k-period returns satisfy a linear factor model in $(\sum_{i=1}^{k} \varepsilon_{r,t+i}, \sum_{i=1}^{k} \varepsilon_{z,t+i}, \{\varepsilon_{s_A,t+i}\}_{i=1}^{k}, \{\varepsilon_{s_B,t+i}\}_{i=1}^{k})$.

[21] If $N = 2(k+1)$, there will be a unique set of factor-mimicking portfolios using B market securities. If $N > 2(k+1)$, there will be multiple possible factor-mimicking portfolios.

cannot freely choose positions in all $2N$ assets. For instance, generalists may lack the expertise to construct the complicated factor-mimicking portfolios or may face institutional frictions that make this infeasible. In this case, generalists function as coarse asset-allocators and not granular cross-market arbitrageurs: their degrees of freedom for within market asset allocation are less than the number of common risk factors ($2(1+k)$). Under these conditions, the markets for A and B will again by imperfectly integrated in both the short- and long-run, because cross-market arbitrage is risky for generalists with granular portfolios of this sort.

To give a concrete example, we might suppose that generalists only have one-degree of freedom within each market: how much to allocate to a baseline portfolio in each market and cannot vary their allocation at all within markets. Specifically, one could assume that $\mathbf{d}_{A,t} = \mathbf{s}_{A0} \times d_{A,t}$ and $\mathbf{d}_{B,t} = \mathbf{s}_{B0} \times d_{B,t}$, so that choosing different values of $d_{A,t}$ and $d_{B,t}$ only moves the baseline portfolios, $\mathbf{s}_{A,0}$ and $\mathbf{s}_{B,0}$, up and down.

The bottom line is that, under plausible conditions, cross-market arbitrage will expose generalists to risk. As a result, the key insights of our simpler two asset model will carry over to the more general case where multiple assets trade in two partially segmented markets. Below we develop and solve an example for two markets and two assets in each market to illustrate the basic intuition.

4.4 Example: Two assets in each market

We now illustrate the impact of a supply shock on two different markets that each contain multiple assets. Specifically, we numerically solve the model in the case where generalists re-allocate their portfolios every $k = 2$ periods and with $N = 2$ assets in both the A and B markets. The two assets in each market differ solely in their durations. The short-term bonds, denoted $A1$ and $B1$, have a duration of 2 years (i.e., $D_{A_1} = D_{B_1} = 2$). The long-term bonds, denoted $A2$ and $B2$, have duration of 10 years (i.e., $D_{A_2} = D_{B_2} = 10$). As before, the two bonds in market A are default free whereas the two bonds in market B are exposed to default risk.[22]

Figure 10 shows the evolution of risk premia following an unexpected shock at time 10 which permanently increases the supply of long-term default-free bonds ($A2$) and reduces the supply of short-term default-free bonds ($A1$) by an equal amount. This scenario corresponds to a "reverse Operation Twist" in which the Federal Reserve sells long-term Treasuries and reinvests the proceeds in short-term Treasuries.

Since this supply shock increases the total amount of interest rate risk than investors must bear, Figure 10 shows that the risk premia for all four assets rise after impact. Furthermore, this supply shock has a larger impact on the risk premium for long-maturity bonds in each market, leading both the A and B yield curves to steepen (since the shock is permanent). These patterns are consistent with those generated by existing models of bond supply shocks (e.g., Greenwood and Vayanos [2014] and Greenwood, Hanson, and Vayanos [2015]). However, since markets are partially segmented in our example, it takes time for the slow-moving generalists to integrate the A and B markets following the shock, leading the risk premia of the two A assets to initially over-react and the risk premia of the two B assets to initially underreact in Figure 10. Furthermore, because cross-market arbitrage is risky for

[22] For simplicity, we assume that $B1$ and $B2$ have the same exposure to the common default process, z_t—i.e., we assume that $\psi_{B_1} = \psi_{B_2}$ in equation (41). We also assume that there is no idiosyncratic default risk—i.e., $u_{B_1,t} = u_{B_2,t} \equiv 0$.

generalists, market integration remains imperfect even in the long run. Specifically, even many period after the supply shock, Figure 10 shows that the yield curve in market A has steepened more than the yield curve in market B.

Figure 11 shows the evolution of investor bond holdings in response to the supply shock. At each date, Figure 11 plots the difference between investor holdings and their pre-shock holdings. At time 10, active generalists take the other side of the supply shock that hits market A, increasing their holdings of $A2$ (long-maturity) and decreasing their holdings of $A1$ (short-maturity). Naturally, active generalists utilize the B market to hedge out their increased exposure to interest rate risk, thereby transmitting the supply shock from market A to market B. Specifically, active generalists reduce their holdings of $B2$ (long-maturity) and increase their holdings of $B1$ (short-maturity) at time 10. Although there are some minor oscillations after time 11, both markets have largely "digested" the shock by time 11 since all generalists have had the opportunity to rebalance their portfolios.

5 Discussion and Applications

5.1 Event studies and assessment of Quantitative Easing programs

In response to a rapidly evolving financial crisis and worldwide recession, in late 2008 and early 2009, central banks around the world announced their intention to aggressively purchase government bonds and other long-term debt securities. On November 25, 2008, the Federal Reserve announced its intention to buy $100 billion in GSE debt and $500 billion in mortgage-backed securities (MBS), and followed up four months later with a significantly expanded purchase program that also included Treasuries. The Bank of England followed in quick succession, announcing £50 billion of private asset purchases in January 2009 and expanding the purchase program to include government bonds in March 2009. The Bank of Japan, already engaged in an asset purchase program since earlier in the decade, announced in December 2008 that the quantity of its monthly purchases of JGB securities would increase. By December 2013, global central banks presided over massive portfolios of long-term securities.

A crucial question in assessing the effectiveness of central bank asset purchase programs is whether they impacted securities prices beyond government bonds. Suppose, for example, that the impact of asset purchase programs was limited to markets in which the purchases were being made (Treasury bonds and MBS), perhaps because these markets are highly segmented from other financial markets. Such a finding should dampen central bankers' enthusiasm for these programs, and cast doubt that asset purchases could affect broader economic activity.

Our model provides a natural framework for understanding how these asset purchase programs should spill across different financial markets over time. According to our model, the largest short-run effects of these programs should be in the securities being purchased. In the long run, however, changes in risk premia in the market being targeted should spill over to non-targeted markets. Differences between the short-run and long-run price impact should reflect the degree to which the programs were anticipated, the length of time between the announcement date and implementation, and the effective degree of segmentation between different financial markets.

Most empirical studies of these purchase programs have used an event study methodology, focusing on the 1-day or even intraday impact on bond yields following announcements of future asset purchases.

In one of the first of these event studies, Gagnon, Raskin, Remache, and Sack (2011) report interest rate changes around a set of Federal Reserve announcement days between November 2008 and January 2010. Cumulating over all announcement dates associated with the Fed's first round of quantitative easing (QE1), they report a 62 basis points decline in 10-year US Treasury yields, a 123 basis points decline in agency MBS yields, and a 74 basis points decline in Baa-rated corporate bond yields. Krishnamurthy and Vissing-Jorgensen (2011) extend this analysis to the Fed's second round of quantitative easing (QE2) and also discuss the impact on other assets, including high yield corporate bonds. After controlling for other factors, Krishnamurthy and Vissing-Jorgensen conclude that the effects of asset purchases were most pronounced among the assets being purchased (MBS and Treasuries in QE1 and Treasuries in QE2), suggesting a high degree of segmentation between different fixed income markets. Neeley (2013) shows that the average cumulative change in 10-year government bond yields in Australia, Canada, Germany, Japan, and the UK immediately following major Fed QE announcements was 53 basis points, compared to the 107 basis points change in the yield on US 10-year notes. Neeley finds no significant impact on major stock market indices outside the US. In summary, at short horizons, there is modest evidence that asset purchases impacted other fixed income markets, and almost no evidence that they had any impact on the equity markets.

At the same time, some researchers have recognized that short horizon announcement returns may not capture the full impact of these asset purchase programs. In their empirical assessment of the Bank of England's quantitative easing program, Joyce, Lasoasa, Stevens and Tong (2010) suggest that it may have impacted corporate bonds and equities, although "these effects might be expected to take time to feed through, as it will take time for investors and asset managers to rebalance their portfolios." They further suggest that the impact on QE may subsequently be reflected in corporate issuance. Fratzcher, Lo Duca, and Straub (2013) suggest that the Fed's QE programs triggered portfolio flows that ultimately impacted emerging market asset prices and foreign exchange rates. Feunou et al (2015) present evidence that QE stimulated portfolio flows into Canada and had a large ultimate impact on Canadian bond yields. Mamaysky (2014) suggests that QE might ultimately spill into the asset markets through portfolio allocation, but notes that "it is unlikely that such portfolio flows can take place quickly."

Researchers have used different approaches to measure the long-run effects of QE. Joyce, Lasoasa, Stevens and Tong (2010) report the cumulative change in asset prices for the longer period between March 4, 2009 and May 31, 2010 in addition to 1-day announcement returns. They show that corporate bond yields fall by a cumulative 70 basis points around asset purchase announcements, but by 400 basis points over the longer period. Mamaysky (2014) takes a more tailored approach to each asset market: he chooses an announcement window that maximizes the statistical power of the measured return. Using this approach, he shows that the impact of QE on both equity and high yield bond markets is much larger after 15 days than what one would measure using a 1-day window. But even this approach may significantly understate the long-run effect, because, as we have noted, the full impact of supply shocks may easily take quarters or years to be felt.

In Figure 12, we reproduce and extend the main results from Mamaysky's study. We start with the set of QE announcement dates identified by Fawley and Neely (2013). We then limit the analysis to major announcement dates on which the 10-year Treasury yield changed by more than 10 basis points in absolute value. After applying this filter, we are left with eight announcement dates spread

across QE1, QE2 and QE3. For these announcement days, we compute the cumulative response of (i) 10-year Treasury yields, (ii) returns on the S&P 500, (iii) the VIX, and (iv) and high-yield corporate bond yields. We compute the response up to 20 trading days after the announcement date.

To assess statistical significance of these long horizon responses, we compare the response following actual QE announcement dates with the counterfactual "response" following a set of eight randomly chosen dates (drawn without replacement) from November 2008 to September 2012. This comparison allows us to gauge whether the responses following actual QE announcement dates were statistically "unusual" in any sense. We adjust for overlapping event windows using the same methodology as Mamaysky (2014). This procedure is repeated 5,000 times to generate a benchmark distribution of the cumulative response following these counterfactual announcement dates. Finally, we compute a one-sided p-value, for each outcome variable and at every horizon, as the fraction of counterfactual "responses" that are more extreme than the response following the actual QE announcements. For instance, a p-value of 5% on the S&P 500 index 10 days after the announcement indicates that the average return following QE announcements is greater than 95% of the counterfactual responses with randomly chosen announcement dates.

Figure 12 shows the response of different financial market prices as measured by the p-values defined above. 10-year Treasury yields immediately react to QE announcements. Treasury yields continue to outperform for an extended period following the initial announcement, reflecting the direct impact of the shocks on this market. High yield bonds and the stock market (the S&P 500) do not react as strongly upon announcement: 1-day announcement returns on high yield bonds and equities are not statistically significant, only outperforming 80% of simulated draws on initial impact. However, equities and high yield bonds continue to rally for more than a week following the announcement. By day 11 following announcements, these markets outperform 99% of the simulated responses.

Figure 12 also shows the VIX, which measures the implied volatility on S&P 500 index options and is often seen as a broad barometer of the pricing of risk across many different markets. The VIX shows almost no abnormal reaction on announcement. The VIX slowly declines in the week following announcement and the result is highly significant by day 11 following announcements.[23]

Our model clarifies the broader issue at stake: event studies are a useful methodology for detecting short-run price changes, but often lack the statistical power to detect price impact at longer horizons. However, if markets are segmented, the true long-run effects of supply and demand shocks may take time to materialize.[24] For instance, consider the effects of σ_z—the volatility of fundamental cash-flow shocks in market B—on our ability to detect the impact on prices in market B stemming from a supply shock that hits market A . When σ_z is large relative to σ_r, our model suggests that the econometrician may have little power to detect this gradual shift in risk premia in market B: there is simply too much B-specific fundamental news to reliably detect cross-market spillovers using a handful of announcement events. The statistical power would increase with the number of events, but power is nonetheless decreasing in σ_z. Thus, our model suggests that short-run event studies may have a hard time detecting spillover effects on markets, such as equities and high yield bonds, where there can be significant news about cash flow fundamentals. More generally, our framework suggests that event studies are an inappropriate methodology for measuring cross-market price impact, at least in

[23] We have found a similar reaction in the interest rate volatility implied by swaption prices.

[24] The event study methodology was originally developed in the 1970s to tackle questions of *informational efficiency* of stock prices, not changes in risk premia.

the short run.

Figure 13 illustrates the potential inability of event studies to detect cross-market spillover effects from an unanticipated supply shock in market A. In an environment with low short-rate volatility, a supply shock to market A can have statistically significant impact on market A (Treasuries) but not on market B (corporate bonds).[25] Even though the shock to the supply of A impacts yields in the B market, the short-term effect is not statistically significant—e.g., the confidence interval for the 1-day change includes zero—and, thus, it would not be detected by conventional event study techniques. In addition, the long-term effect, while economically meaningful, is also statistically insignificant.

In summary, our model suggests that we should be extremely cautious in using event studies to assess the long-run impact of supply shocks on market prices and risk premia. However, measuring the long-run impact of supply shocks across markets is inherently difficult because the full economic impact may occur over such a long time that it is swamped by other factors.

5.2 Corporate Arbitrage

In our model, generalist "asset allocators" play a critical role in integrating prices across markets. Earlier we suggested that these investors were best thought of as pension funds or endowments, who adjust asset allocation at an annual frequency. But generalists could also be nonfinancial corporations who continually access different financial markets in order to raise capital. Corporations have flexibility over which securities to issue, and can thus execute a form of arbitrage between otherwise segmented securities markets.

Consider a firm with financing needs who has the choice of issuing in the equity or debt markets. Suppose that debt markets have just received a large positive demand shock. For example, perhaps bond mutual funds experienced large inflows, causing a reduction in corporate bond yields. How should a nonfinancial firm respond? While the firm may have preferences over its capital structure, attractive pricing in the debt market would lead the firm to satisfy a greater fraction of its total financing needs by issuing debt that it would not otherwise.

A growing literature in behavioral corporate finance has suggested that nonfinancial corporations may have advantages vis a vis professional arbitrageurs in conducting cross-market arbitrage at low frequencies (Stein [2005] and Ma [2015]). The advantage of nonfinancial firms arises because arbitrageurs use investment vehicles—mutual funds or hedge funds—where capital can easily be withdrawn in response to poor temporary performance. This "open-ended" structure limits arbitrageurs' willingness to place the kinds of big, slow-to-converge bets required when trying to profit from asset-class level dislocations. By contrast, nonfinancial firms—with access to patient capital that cannot be withdrawn—can respond aggressively to dislocations at the asset-class level.

Is it reasonable to think of corporations as slow-moving generalists in the sense implied by our model? Several recent papers in behavioral corporate finance speak to the slow-moving nature of corporate arbitrage. Greenwood, Hanson, and Stein (2010) suggest, for example, that corporations adjust the maturity of their debt issuance in response to shifts in government debt maturity, a phenomenon they dub "gap filling." They show that when corporate issuance is measured annually, there is only a

[25] The parameter values used to generate Figure 13 reflect the low interest rate volatility during QE when short rates were pinned down at zero and supply risk was low. In addition, arbitrageur risk tolerance (τ) was small during this period.

modest correlation between government supply shocks and the maturity of corporate issuance. However, at horizons of two years or longer, or when the data are measured in levels rather than issuance, the evidence is much stronger. Graham, Leary, and Roberts (2014) show similar results in their study of corporate leverage between 1920 and 2010. Corporate leverage tends to rise when government debt falls, suggesting that corporations respond to supply shocks. But this result is stronger when the authors study the level of corporate leverage, rather than annual changes. Ma (2015) studies cross-market arbitrage by corporations directly. She shows that, at low frequencies, firms actively substitute between equity and debt markets in an attempt to exploit valuation differences across these markets. Liao (2016) provides evidence that cross-market arbitrage by corporations plays an important role in integrating credit markets in different currencies—e.g., the dollar-denominated credit market and the Euros-denominated credit market.

6 Conclusion

Modern financial markets are highly specialized. While specialization brings many benefits, the boundaries of securities markets are tested when there are large shocks to the supply of an entire asset class. In this paper, we develop a model to describe securities prices when shocks must draw in arbitrageurs from other related asset markets. We use the model to study the process by which capital flows across markets, and how quickly and by what magnitude prices adjust in different markets. Unlike textbook theories in which asset prices are determined solely by the *stock* of risky assets supplied, our approach suggests that *supply flows*—i.e., the rate at which the supply stock is changing—also matter in the short run. Even when a large amount of capital is mobile in the long run, different asset markets need not be fully integrated because market segmentation creates risks for arbitrageurs.

Our model explores the consequences of specialization when markets are hit with large shocks. However, we have taken the existence of specialists as given. But why are some asset classes dominated by specialists while others are widely held in the portfolios of generalists? And what determines the boundaries of specialists' expertise and, hence, the fault lines between different asset classes? Answering these questions remains an important task for future research.

References

Acharya, V., H. S. Shin, and T. Yorulmazer, 2013, "Fire-sale FDI," Korean Economic Review 27, 163-202.

Campbell, J. Y., 1991, "A Variance Decomposition for Stock Returns", Economic Journal 101:157–179.

Campbell, J. Y. and R. J. Shiller, 1988, "Stock Prices, Earnings, and Expected Dividends," Journal of Finance 43, 661-76.

Campbell, J. Y. and L. Viceira, 2002, Strategic Asset Allocation: Portfolio Choice for Long-Term Investors, Clarendon Lectures in Economics, Oxford University Press.

DeLong, J. B., A. Shleifer, L. H. Summers, and R. J.Waldmann, 1990, "Noise Trader Risk in Financial Markets," Journal of Political Economy 98, 703-738.

Duffie, D., "Asset Price Dynamics with Slow-Moving Capital", Journal of Finance 2010, 65: 1238-1268.

Duffie, D., and B. Strulovici, 2012, "Capital Mobility and Asset Pricing", Econometrica 80: 2469-2509.

Duffie, D., and K. Singleton, 1999, "Modeling Term Structures of Defaultable Bonds" , Review of Financial Studies 12: 687-720.

Errunza, V., and E. Losq, 1985, "International Asset Pricing under Mild Segmentation: Theory and Test" Journal of Finance 40, 105-124.

Fawley, B. and C. Neely, 2013, "Four Stories of Quantitative Easing," Federal Reserve Bank of St. Louis Review, 95 (1), 51–88.

Feunou, B., J. Fontaine, J. Kyeong, and J. Sierra, 2015 "Foreign Flows and Their Effects on Government of Canada Yields," Bank of Canada Staff Analytical Note.

Fratzscher, M., M. Lo Duca, and R. Straub, 2013, "On the International Spillovers of US Quantitative Easing," European Central Bank Working Paper.

Gagnon, J., M. Raskin, J. Remache, and B. Sack, 2011, "The Financial Market Effects of the Federal Reserve's Large-scale Asset Purchases," International Journal of Central Banking 7 , 3–43.

Garleanu, N., L. H. Pedersen, and A.M. Poteshman, 2009, "Demand-Based Option Pricing," The Review of Financial Studies 22, 4259-4299.

Graham, J. R., M. Leary, and M. Roberts, 2014, "A Century of Capital Structure: The Leveraging of Corporate America," Journal of Financial Economics forthcoming.

Greenwood, R., S.G. Hanson and J.C. Stein, 2010, "A Gap-Filling Theory of Corporate Debt Maturity" Journal of Finance 65, 993-1028.

Greenwood, R., S. G. Hanson, J. Rudoplh, and L. H. Summers, 2015, "The Optimal Maturity of Government Debt," in The $13 Trillion Question: How America Manages Its Debt, Brookings Institution Press.

Greenwood, R., S. G. Hanson, and D. Vayanos, 2015, "Forward Guidance in the Yield Curve: Short Rates versus Bond Supply," Working Paper.

Greenwood, R. and D. Vayanos, 2014, "Bond Supply and Excess Bond Returns," Review of Financial Studies 27, 663-713.

Gromb, D. and D. Vayanos, 2002, "Equilibrium and Welfare in Markets with Financially Constrained Arbitrageurs,"Journal of Financial Economics 66, 361-407.

Gromb, Denis and Dimitri Vayanos, 2015, "The Dynamics of Financially Constrained Arbitrage," Working paper.

Grossman, Sanford J & Miller, Merton H, 1988. "Liquidity and Market Structure," Journal of Finance 43, 617-37.

Hanson, Samuel G., 2014, "Mortgage Convexity," Journal of Financial Economics 113(2), 270-299.

Joyce, M. A. S., A. Lasaosa, I. Stevens, and M. Tong, 2011, "The Financial Market Impact of Quantitative Easing in the United Kingdom," International Journal of Central Banking 7, 113-161.

Krishnamurthy, A., O. Vigneron and X. Gabaix, 2007, "Limits of Arbitrage: Theory and Evidence from the Mortgage-Backed Securities Market", Journal of Finance, 62(2), 557-595.

Krishnamurthy, A. and A. Vissing-Jorgensen, 2011, "The Effects of Quantitative Easing on Interest Rates: Channels and Implications for Policy," Brookings Papers on Economic Activity, Fall 2011, 215-265.

Krishnamurthy, A. and A. Vissing-Jorgensen, 2012, "The Aggregate Demand for Treasury Debt," Journal of Political Economy, 120, 233–267.

Liao, G., 2016, "Credit Migration and Covered Interest Rate Parity", Working paper.

Ma, Y., 2015, "Non-financial Firms as Arbitrageurs in Their Own Securities", Working paper.

Mamaysky, Harry, 2014, "The Time Horizon of Price Responses to Quantitative Easing", Working Paper.

Merton, Robert C, 1987. "A Simple Model of Capital Market Equilibrium with Incomplete Information", Journal of Finance 42, 483-510.

Neely, C., 2012, "The Large-scale Asset Purchases had Large International Effects," Federal Reserve Bank of Saint Louis Working Paper.

Ross, S. A., 1976, "The Arbitrage Theory of Capital Asset Pricing," Journal of Economic Theory 13, 341-360.

Pedersen, L.H., M. Mitchell , and T. Pulvino, 2007. "Slow Moving Capital," American Economic Review 97, 215-220.

Shleifer, A., and R. W. Vishny, 1997. "The Limits of Arbitrage," Journal of Finance 52, 35-55.

Stapleton, R.C., and M. G.. Subrahmanyam, 1977, "Market Imperfections, Capital Market Equilibrium and Corporation Finance," Journal of Finance 32, 307-319.

Stein, J. C., 2005, "Why Are Most Funds Open-End? Competition and the Limits of Arbitrage," Quarterly Journal of Economics 120, 247-272.

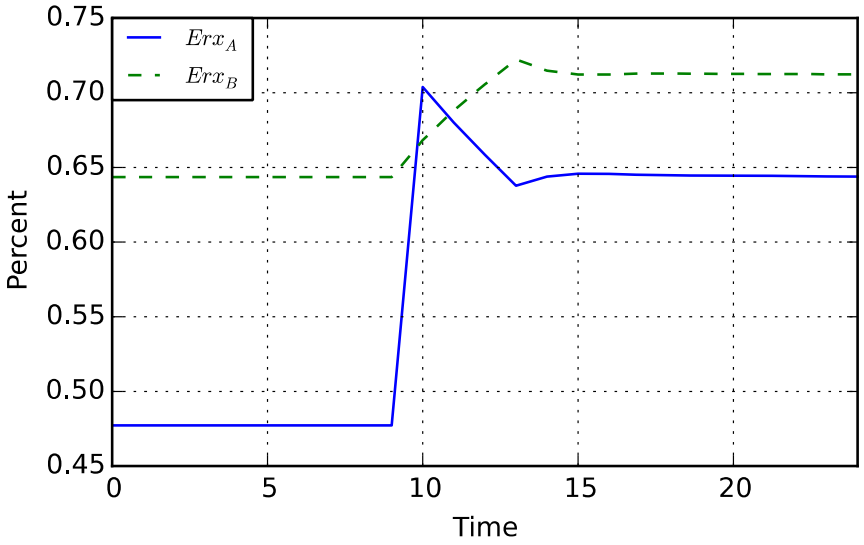
Woodford, M., 2012, "Methods of Policy Accommodation at the Interest-Rate Lower Bound," in The Changing Policy Landscape, Federal Reserve Bank of Kansas City.

Wurgler, J. and E. Zhuravskaya, 2002, "Does Arbitrage Flatten Demand Curves for Stocks?" The Journal of Business 75, 583–608.

Vayanos, D. and J. Vila, 2009, "A Preferred-Habitat Model of the Term Structure of Interest Rates," NBER Working Paper No. 15487.

Figure 1: Price impact of an unanticipated shock to the supply of asset A. This figure shows the impact on annual bond risk premia and bond yields of an unanticipated shock that increases the supply of asset A by 50% in period 10. Panel A shows the evolution of annual bond risk premia in market A, $E_t[rx_{A,t+1}]$, and market B, $E_t[rx_{B,t+1}]$, over time. Panel B shows the evolution of bond yields in market A, $y_{A,t}$, and market B, $y_{B,t}$, over time.

Panel A: Annual bond risk premia



Panel B: Bond yields

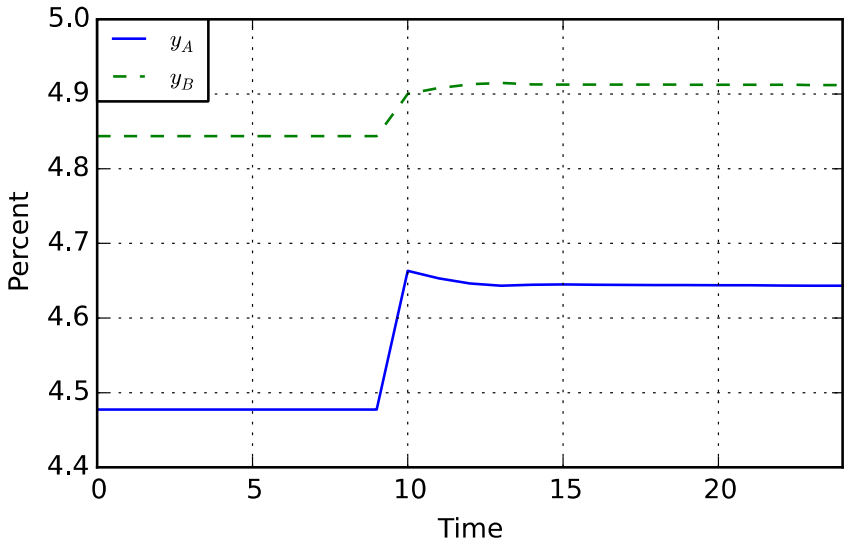
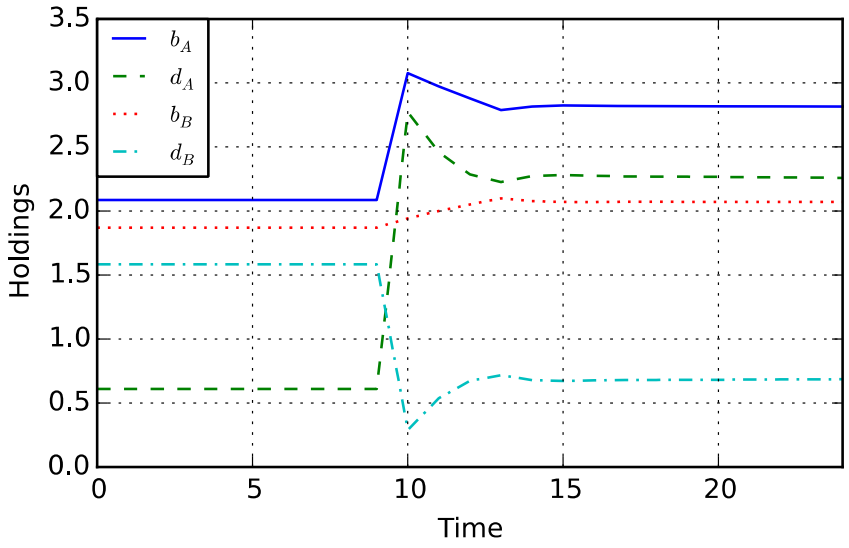


Figure 2: Portfolio adjustments in response to an unanticipated shock to the supply of asset A. This figure shows the impact on investor positions and active asset supplies of an unexpected shock that increases the supply of asset A by 50% in period 10. Panel A shows the evolution of specialists holdings in markets A and B ($b_{A,t}$ and $b_{B,t}$) as well as the positions of active generalists ($d_{A,t}$ and $d_{B,t}$). Panel B shows the evolution of the "active supplies" of assets A and B. The active supply of A is $s_{A,t} - (1 - q_A - q_B)k^{-1}\sum_{i=1}^{k-1} d_{A,t-i}$ and the active supply of B is defined analogously.

Panel A: Specialist holdings and positions of active generalists



Panel B: Active asset supply

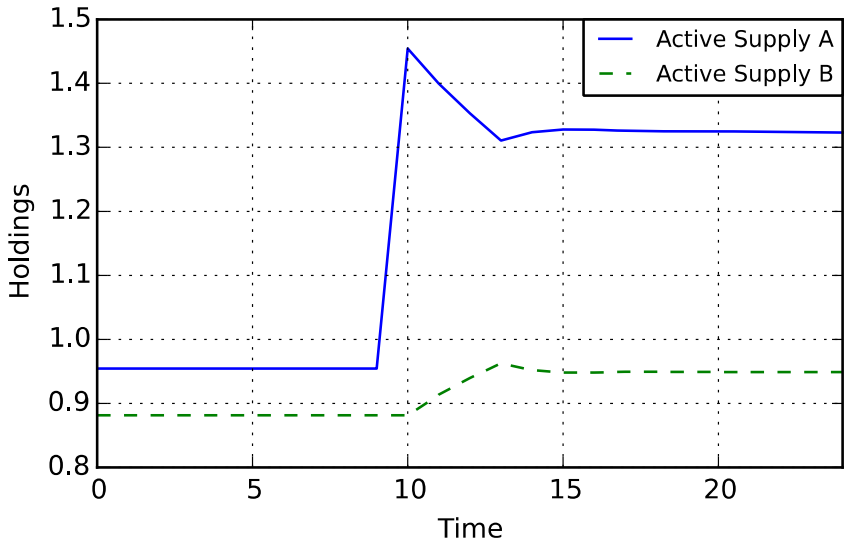


Figure 3: Yield spread impact of an unanticipated shock to the supply of asset A. This figure shows the impact on the yield spread between asset B and asset A, $y_{B,t}$ -$y_{A,t}$, of an unanticipated shock that increases the supply of asset A by 50% in period 10.

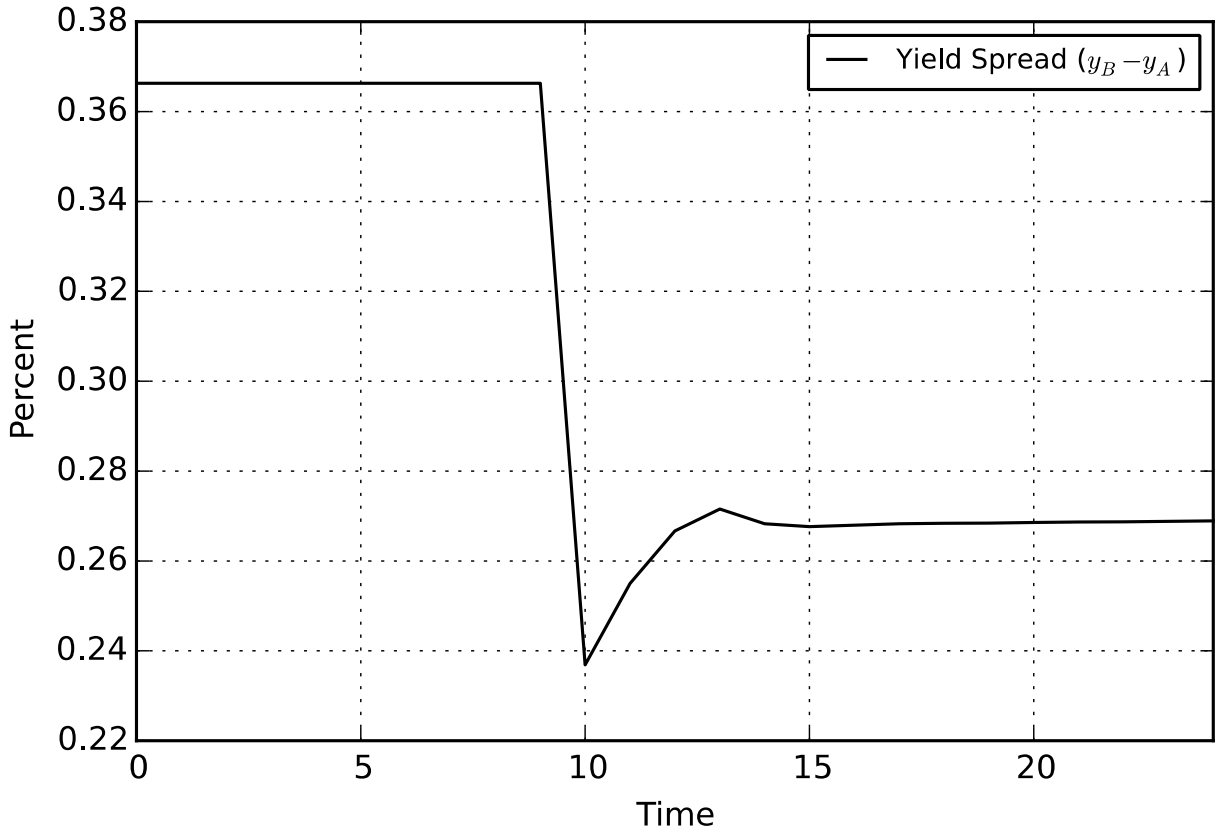
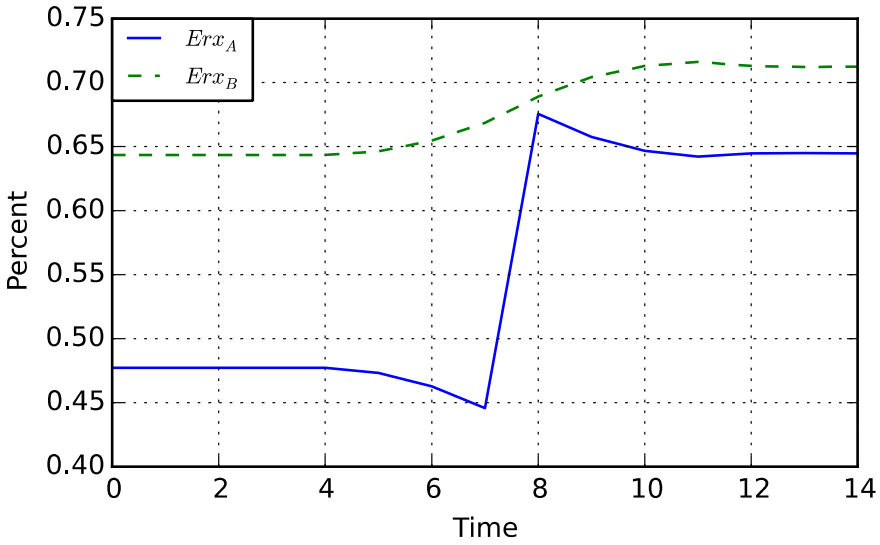


Figure 4: Price impact of an anticipated shock to the supply of asset A. This figure shows the impact on annual bond risk premia and bond yields of an anticipated supply shock: there is an announcemement in period 5 that the supply of asset A will jump by 50% in period 8. Panel A shows the evolution of annual bond risk premia in market A, $E_t[rx_{A,t+1}]$, and market B, $E_t[rx_{B,t+1}]$, over time. Panel B shows the evolution of bond yields in market A, $y_{A,t}$, and market B, $y_{B,t}$, over time.

Panel A: Annual bond risk premia



Panel B: Bond yields

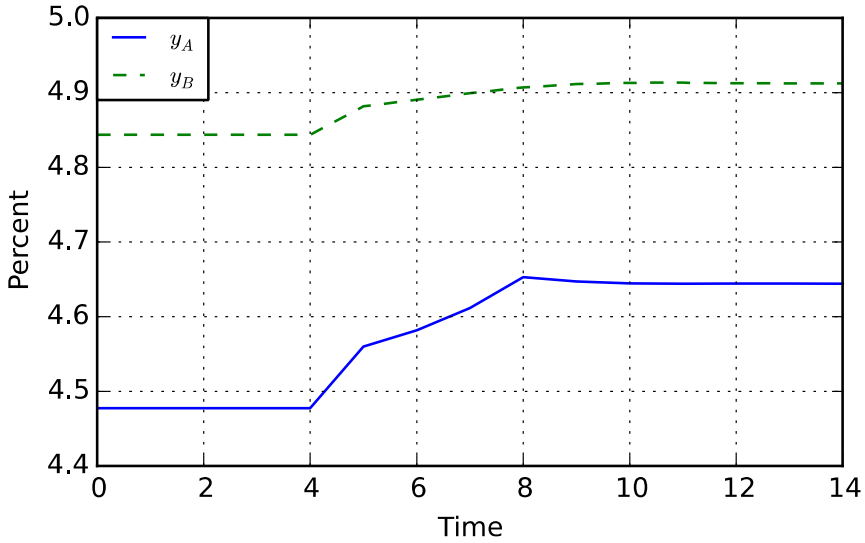
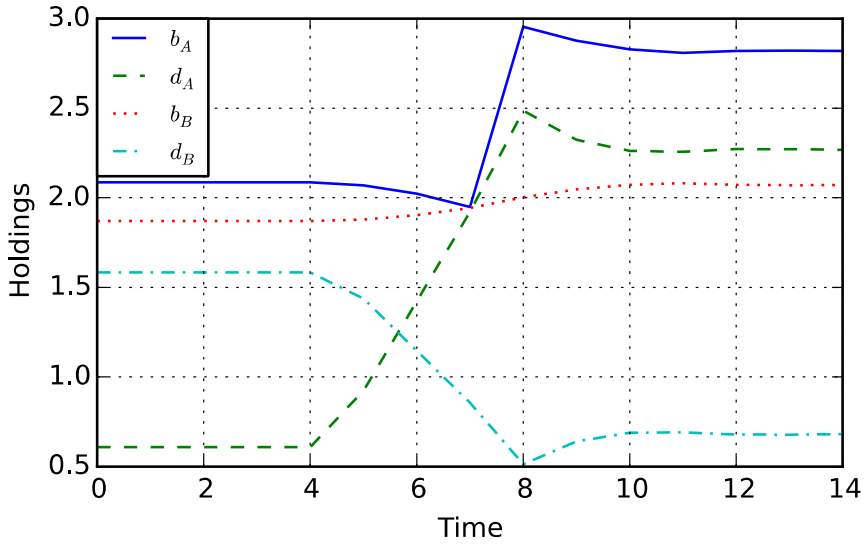


Figure 5: Portfolio adjustments in response to an anticipated shock to the supply of asset A. This figure shows the impact on investor positions and active asset supplies of an anticipated supply shock: there is an announcemement in period 5 that the supply of asset A will jump by 50% in period 8. Panel A shows the evolution of specialists holdings in markets A and B ($b_{A,t}$ and $b_{B,t}$) as well as the positions of active generalists ($d_{A,t}$ and $d_{B,t}$). Panel B shows the evolution of the "active supplies" of assets A and B. The active supply of A is $s_{A,t} - (1 - q_A - q_B)k^{-1}\sum_{i=1}^{k-1} d_{A,t-i}$ and the active supply of B is defined analogously.

Panel A: Specialist holdings and positions of active generalists



Panel B: Active asset supply

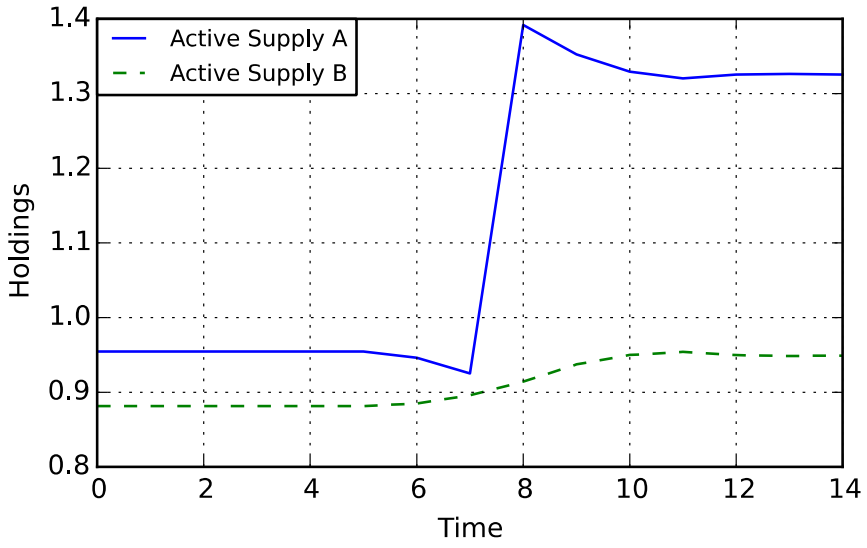
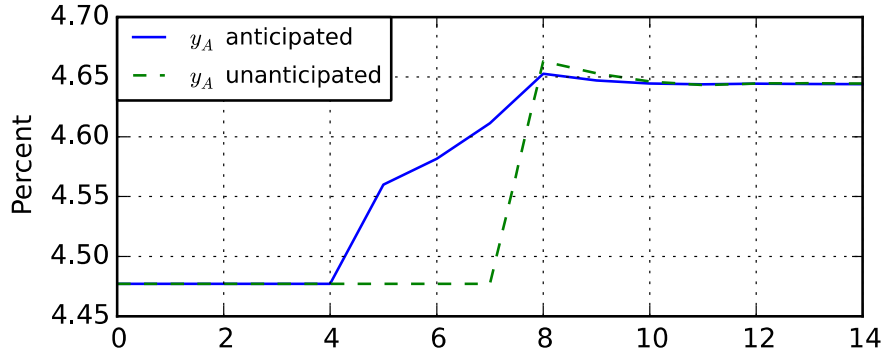
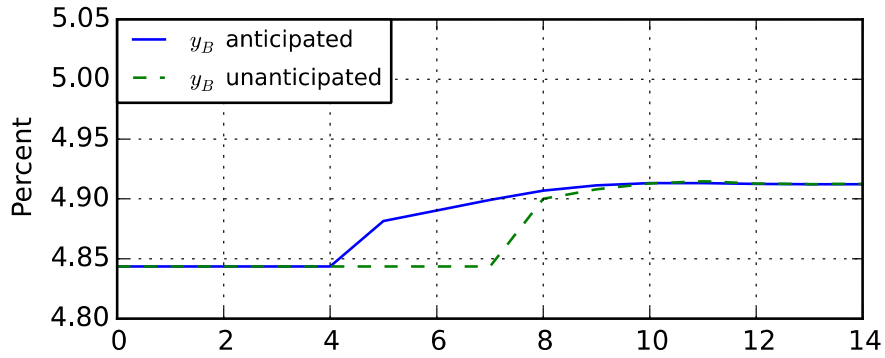


Figure 6: Comparison of anticipated and unanticipated shocks to the supply of asset A. This figure compares the price impact of anticipated and unanticipated supply shocks. The yield dynamics depicted in Figure 1.B and Figure 4.B and are presented here side by side for ease of comparison. In both cases, the supply of A asset increases by 50%. The solid lines show the yield dynamics when the supply shock is pre-announced in period 5 and arrives in period 8. The dashed lines show the yield dynamics when the supply shock unexpectedly arrives in period 8 without any prior announcement. Panel A shows yields in market A, Panel B shows yields in market B, and Panel C shows the yield spread, y_B - y_A.

Panel A: Yield of bond A



Panel B: Yield of bond B



Panel C: Yield spread between bonds B and A

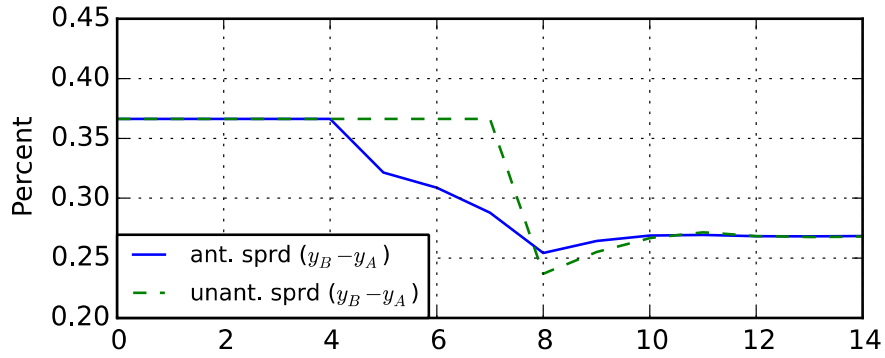
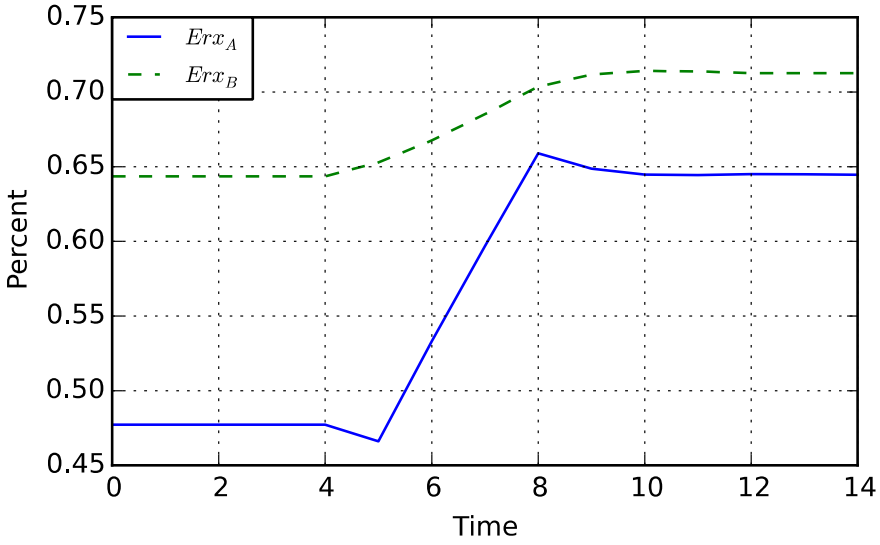


Figure 7: Price impact of an anticipated gradual rise in the supply of asset A. This figure shows the impact on annual bond risk premia and bond yields of an anticipated gradual supply shock: there is an announcemement in period 5 that the supply of asset A will increase by 50% with the increase spread out equallly between time 6 and time 8. Panel A shows the evolution of annual bond risk premia in market A, $E_t[rx_{A,t+1}]$, and market B, $E_t[rx_{B,t+1}]$, over time. Panel B shows the evolution of bond yields in market A, $y_{A,t}$, and market B, $y_{B,t}$, over time.

Panel A: Annual bond risk premia



Panel B: Bond yields

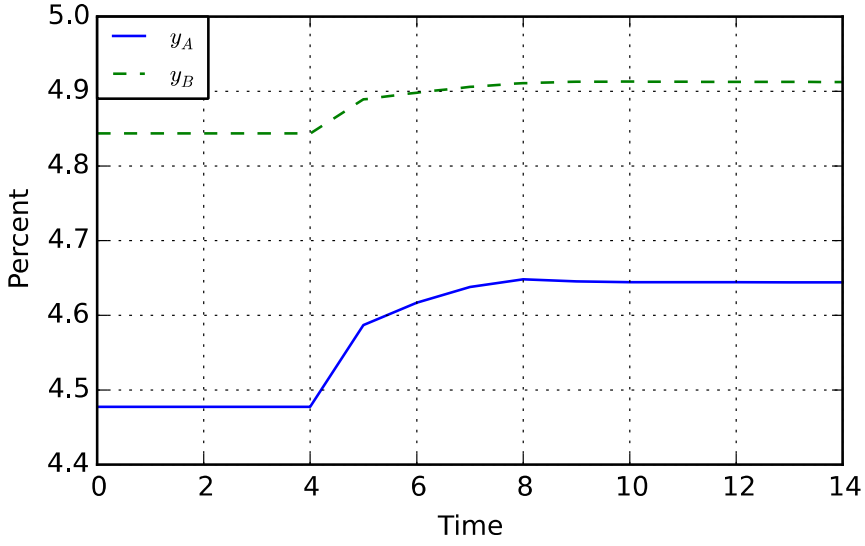
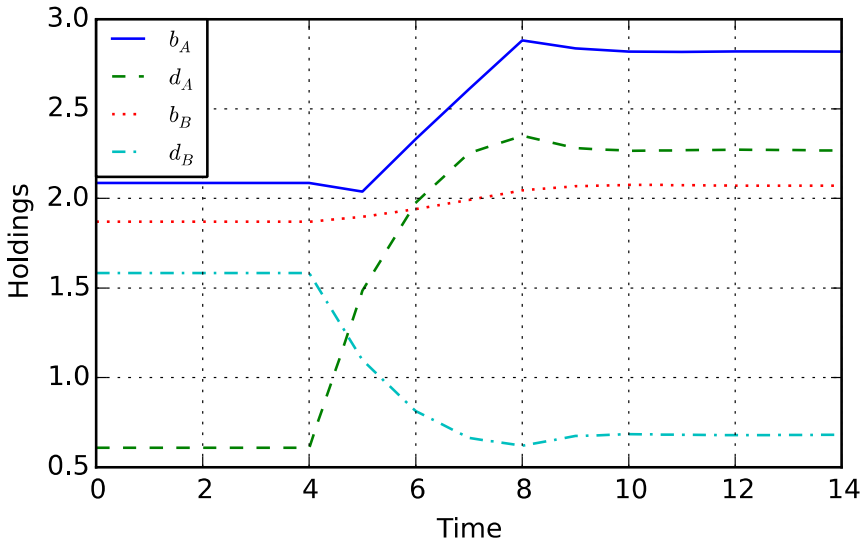


Figure 8: Portfolio adjustments in response to an anticipated gradual rise in the supply of A. This figure shows the impact on investor positions and active asset supplies of an anticipated gradual supply shock: there is an announcemement in period 5 that the supply of asset A will increase by 50% with the increase spread out equallly between time 6 and time 8. Panel A shows the evolution of specialists holdings in markets A and B ($b_{A,t}$ and $b_{B,t}$) as well as the positions of active generalists ($d_{A,t}$ and $d_{B,t}$). Panel B shows the evolution of the "active supplies" of assets A and B. The active supply of A is $s_{A,t} - (1 - q_A - q_B)k^{-1}\sum_{i=1}^{k-1} d_{A,t-i}$ and the active supply of B is defined analogously.

Panel A: Specialist holdings and positions of active generalists



Panel B: Active asset supply

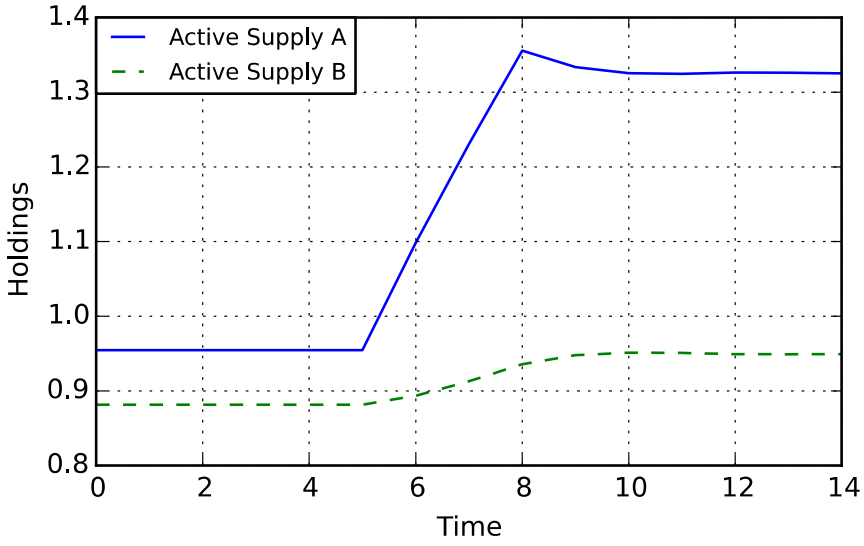
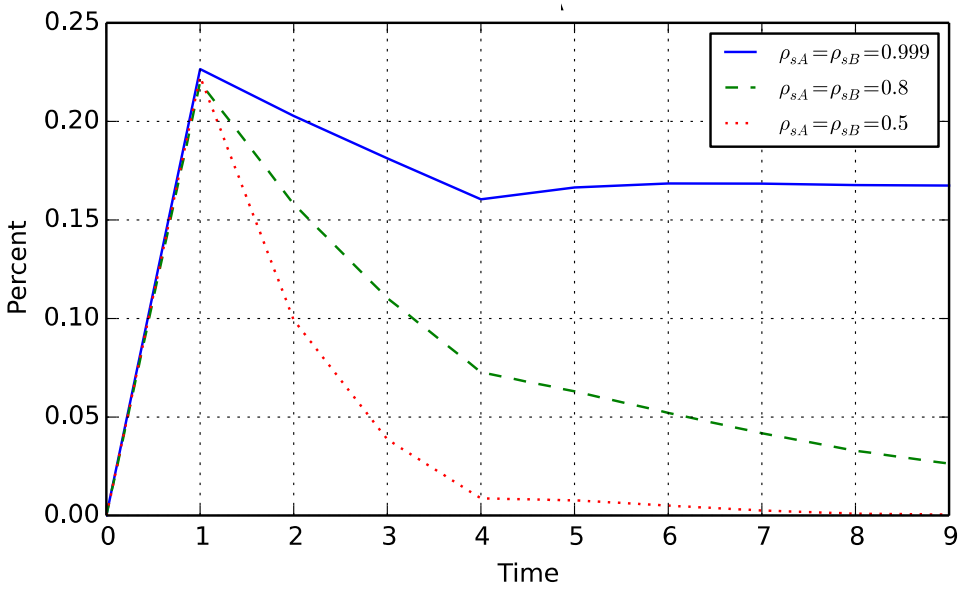


Figure 9: Price impact of a temporary shock to the supply of asset A. This figure shows the impact on bond risk premia of an unexpected supply shock with varying degrees of persistence. The supply of asset A increases by 50% in period 1. Panel A shows the evolution of conditional risk premia in market A, $E_t[rx_{A,t+1}]$ - $E_0[rx_{A,1}]$, and Panel B shows the evolution of conditional risk premia in market B, $E_t[rx_{B,t+1}]$ - $E_0[rx_{B,1}]$, over time.

Panel A: Changes in market A bond risk premia



Panel B: Changes in market B bond risk premia

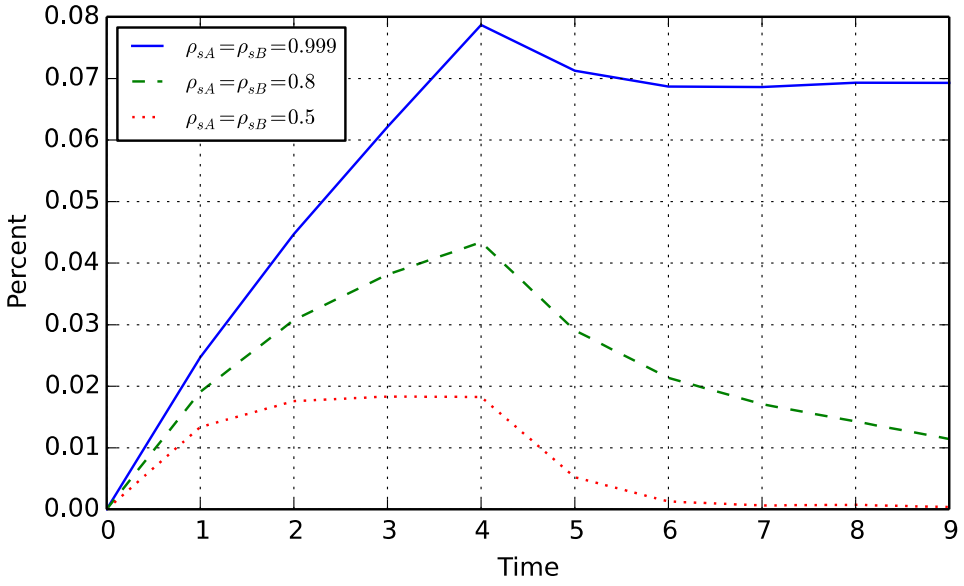
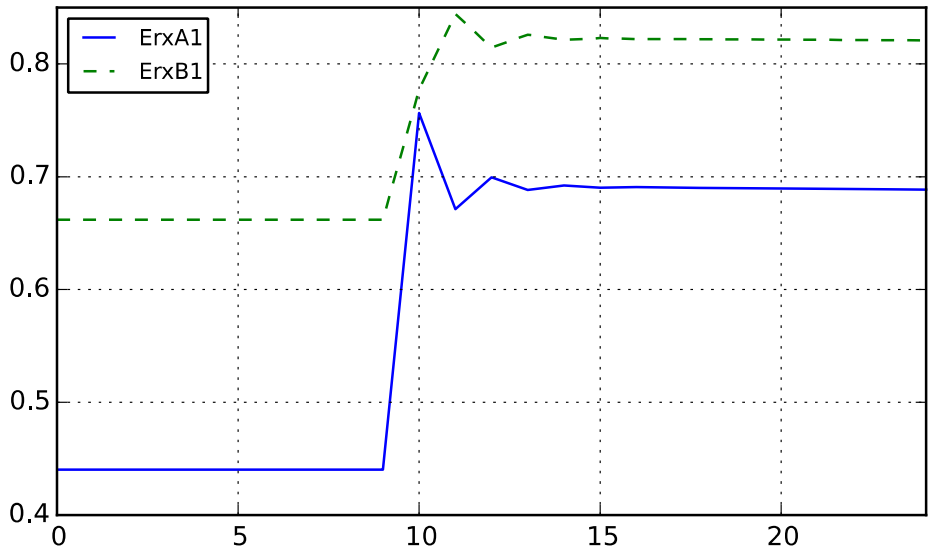


Figure 10: Price impact with multiple securities in each market. This figure shows the impact on bond risk premia of an unexpected supply shock that permanently increases the supply of long-term default-free bond ($A2$) and decreases the supply of short-term bond ($A1$) by an equal amount at time 10. Panel A shows the evolution of risk premia for short-term securities in each market ($A1$ and $B1$). Panel B shows the evolution of risk premia for long-term securities in each market ($A2$ and $B2$).

Panel A: Risk premia for short-maturity bonds in each market



Panel B: Risk premia for long-maturity bonds in each market

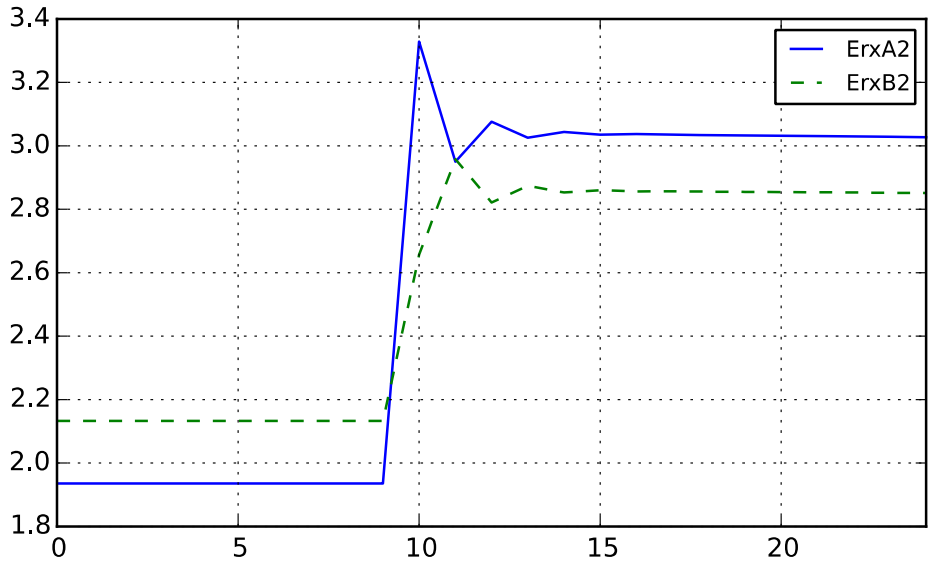
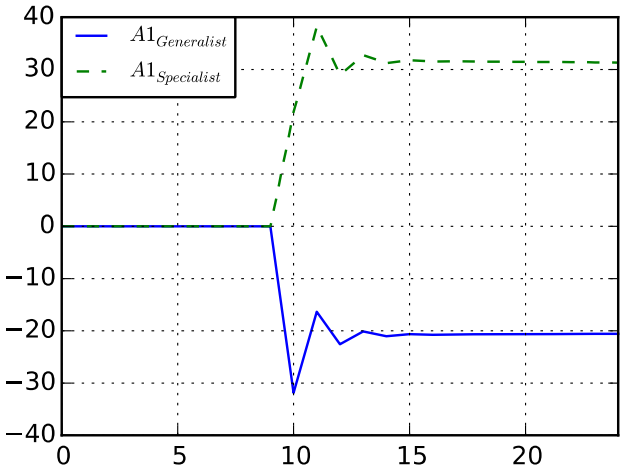
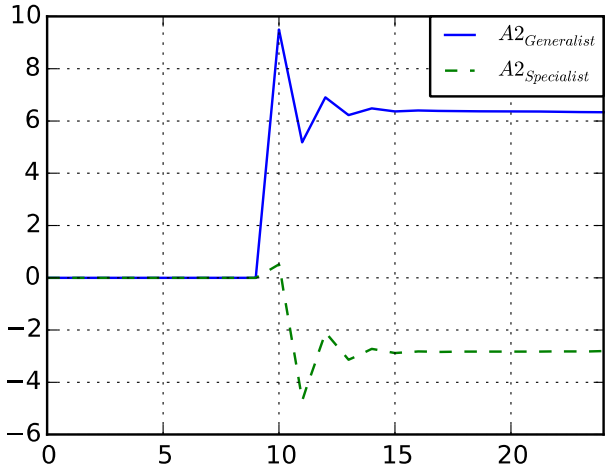


Figure 11: Portfolio adjustments with multiple securities in each market. This figure shows the impact on investor positions of an unexpected supply shock that permanently increases the supply of long-term default-free bonds ($A2$) and reduces the supply of short-term default-free bonds ($A1$) by an equal amount at time 10. The four panels shows the holdings of specialists and active generalists for short-maturity bonds in market A ($A1$), long-maturity bonds in market A ($A2$), short-maturity bonds in market B ($B1$), and long-maturity bonds in market B ($B2$). For simplicity, at each date, we plot the difference between investor holdings and their pre-shock holdings.

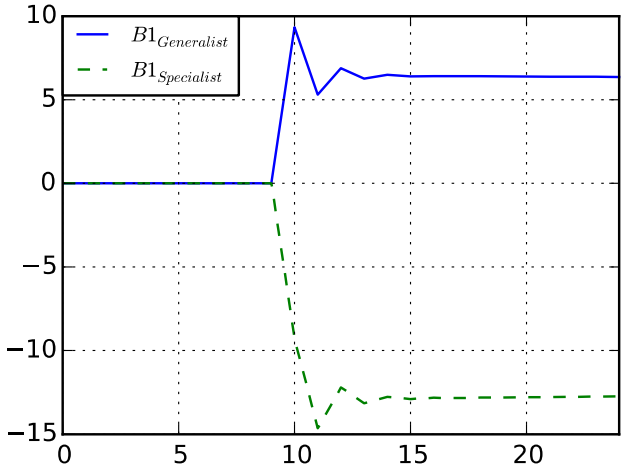
$A1$: Short-maturity bonds in market A



$A2$: Long-maturity bonds in market A



$B1$: Short-maturity bonds in market B



$B2$: Long-maturity bonds in market B

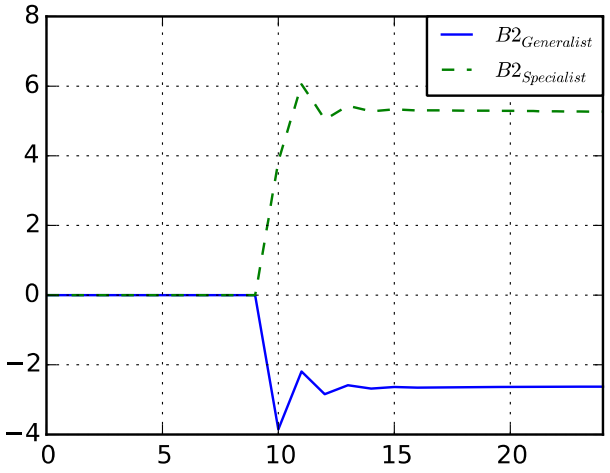


Figure 12: Responses to Federal Reserve Quantitative Easing announcements. This figure shows the response in the days following major Federal Reserve Quantitative Easing (QE) announcements of (i) the 10-year US Treasury yield, (ii) the return on the S&P 500, (iii) the VIX, and (iv) the yield on the Bank of America Merrill Lynch US High-Yield Index. For each financial market price, we calculate the average response following eight major QE announcements: 11/25/2008 (QE1), 12/1/2008 (QE1), 12/16/2008 (QE1), 1/28/2009 (QE1), 3/18/2009 (QE1), 8/27/2010 (QE2), 9/21/2010 (QE2), and 8/22/2012 (QE3). We calculate this response at a 1-day horizon (labeled as day 0 below), a 2-day horizon (labeled as day 1 below), and so on up to a 21-day horizon (labeled as day 20). The average response following these eight dates is compared against a benchmark distribution obtained by randomly drawing eight dates from November 2008 to September 2012 as counterfactual announcement dates. We use 5,000 simulated draws to generate this benchmark distribution. The p-value is then the fraction of simulated counterfactual responses that are more extreme than the response to the actual Federal Reserve QE announcement dates. For ease of comparison, the p-values are signed such that a low p-value for a fixed income instrument or the VIX indicates a *decline* in yield or volatility that is statistically significant based on the benchmark distribution. By contrast, a low p-value for the S&P 500 indicates a positive return that is statistically significant relative to the benchmark distribution.

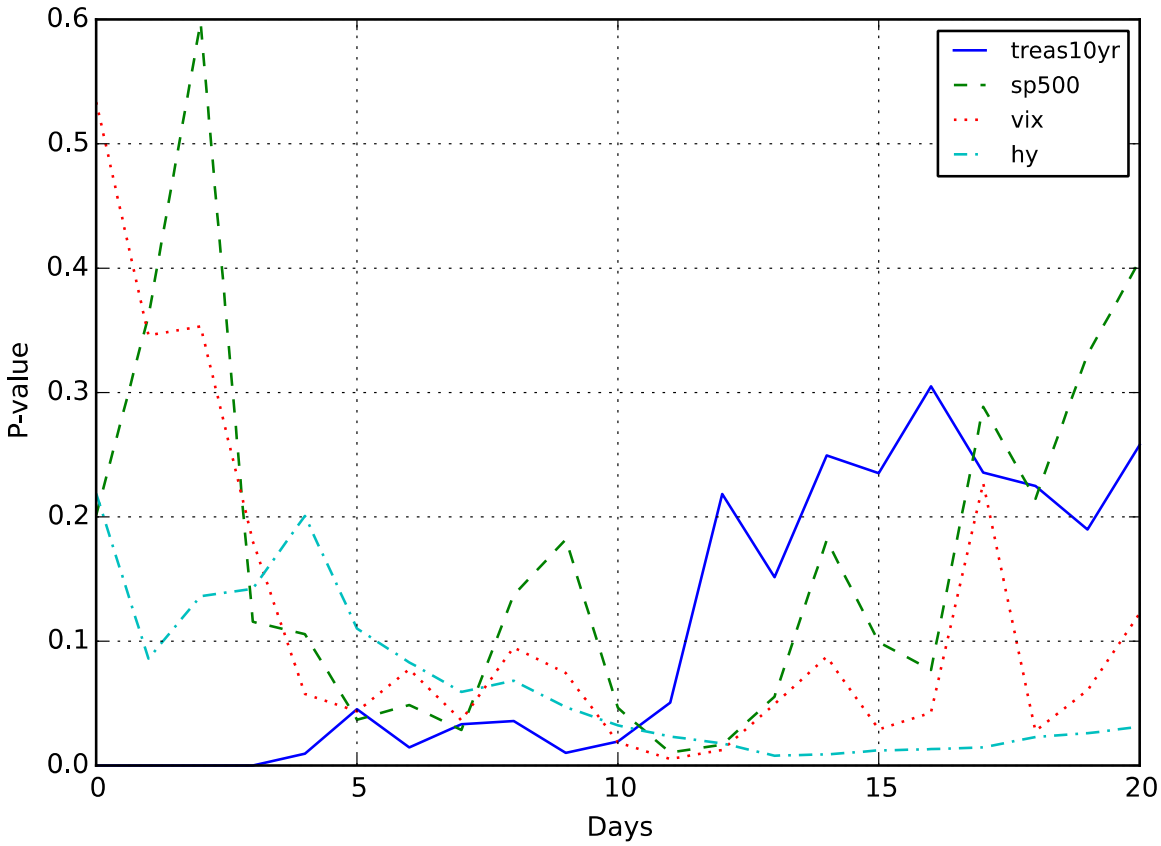


Figure 13: Event study confidence interval following an unanticipated shock to the supply of asset A. The yields of markets A and B and their respective 95% confidence intervals are shown. An unanticipated shock that doubles the supply of asset A is delievered in period 10. The following parameters are used: $\tau = 0.5, \sigma_{s_A} = \sigma_{s_B} = 0, \sigma_r = \sigma_z = 0.2\%$. All other parameters are the same as those listed in Table 1. For period $t > 9$, we compute the model-implied confidence interval for the cumulative changes in yields for market A and B from period 9, $y_{A,t}$ - $y_{A,9}$ and $y_{B,t}$ - $y_{B,9}$, assuming that all shocks are normally distributed. These confidence intervals are shaded in gray.

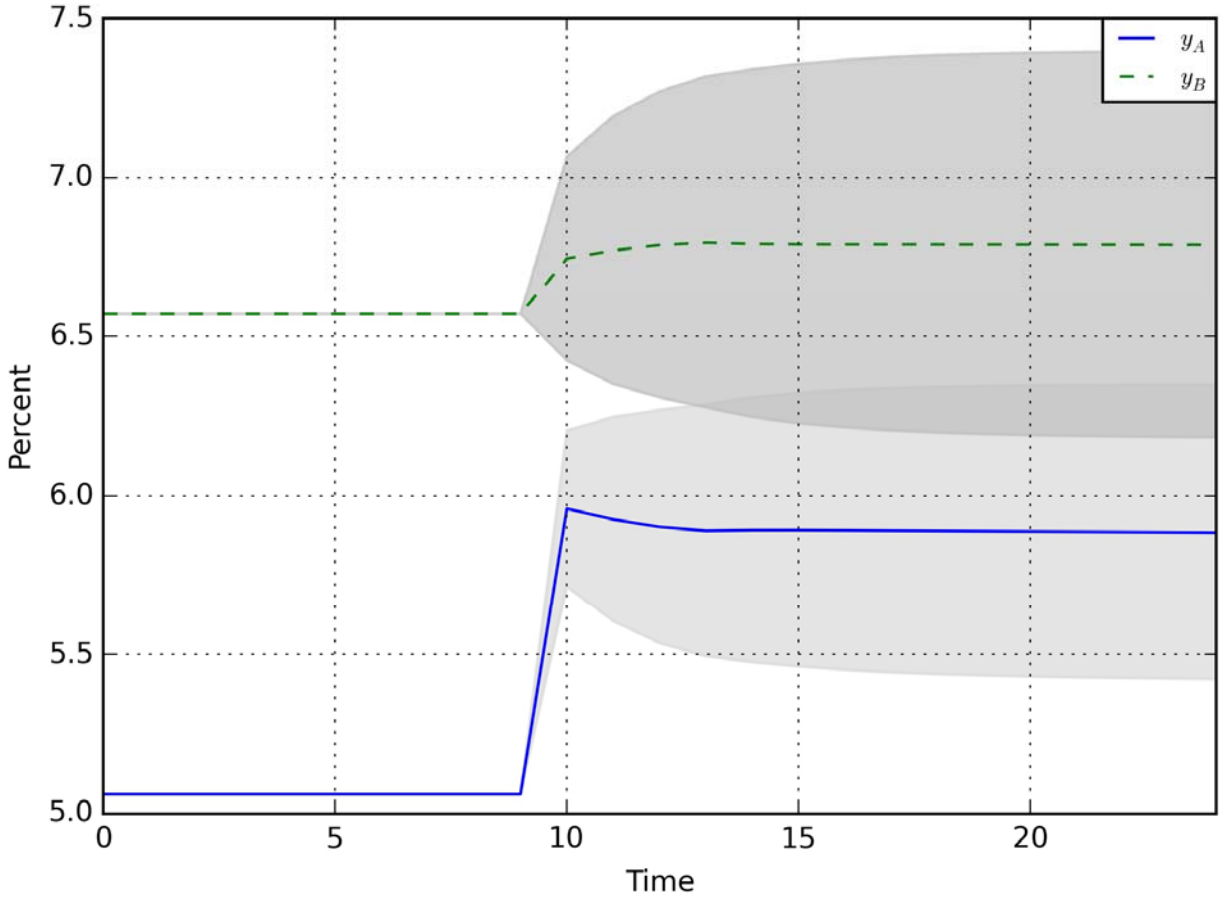


Table 1: Illustrative model parameters. This table presents the illustrative model parameters that we use throughout our numerical exercises. We use annualized values so that one period corresponds to one year.

Parameter	Description	Value
q_A, q_B	Percentage of investors that are specialists in A and B	45%
k	Number of periods between generalist portfolio rebalancing	4
$\bar{r}$	Average short-term riskless rate	4%
σ_r	Volatility of annual shocks to short-term riskless rate	1.3%
ρ_r	Annual persistence of short-term riskless rate	0.85
$\bar{z}$	Expected default losses per annum on asset B	0.2%
σ_z	Volatility of annual shocks to default losses on asset B	0.7%
ρ_z	Annual persistence of default losses on asset B	0.85
$\bar{s}_A, \bar{s}_B$	Average asset supplies	1
$\sigma_{s_A}, \sigma_{s_B}$	Volatility of annual supply shocks	0.6
ρ_{s_A}, ρ_{s_B}	Annual persistence of supply shocks	0.999
D_A, D_B	Macaulay duration in years (implies $\theta_A = \theta_B$=0.8)	5 years
τ	Investor risk tolerance	50

Table 2: Model comparative statics. This table shows how the price impact of the same supply shock — an unanticipated shock that increases asset supply by 50% — varies as we change key model parameters one at a time. All other parameters are held constant at the values listed in Table 1. For a given set of model parameters, we summarize the impact of the supply shock on both the A and B markets by listing (i) the yields and expected annual returns in the period before the shock arrives (labeled as "pre-shock level"), (ii) the changes in yields and expected annual returns in the period when the shock arrives (labeled as "short-run Δ"), and (iii) in 2k periods after the shock arrives (labeled as "long-run Δ"). Finally, we report the degree to which bond yields over- or underreact as the difference between the short-run change and the long-run change, expressed as a percentage of the long-run change

$$\%\textit{Over-Reaction}(y)=\frac{(y_t-y_{t-1})-(y_{t+2k}-y_{t-1})}{(y_{t+2k}-y_{t-1})}.$$

Our measure of over-reaction for risk premia is defined analogously.

		Market A								Market B							
		Risk premia, $E_t[rx_{A,t+1}]$				Yields, $y_{A,t}$				Risk premia, $E_t[rx_{B,t+1}]$				Yields, $y_{B,t}$			
		Pre shock level	Short run Δ	Long run Δ	Over-react	Pre shock level	Short run Δ	Long run Δ	Over-react	Pre shock level	Short run Δ	Long run Δ	Over-react	Pre shock level	Short run Δ	Long run Δ	Over-react
Supply shock hits market A																	
(1) Base case		0.48	0.23	0.17	35%	4.48	0.19	0.17	11%	0.64	0.02	0.07	-65%	4.84	0.06	0.07	-19%
(2) More risk tolerant	$\tau = 60$	0.39	0.18	0.14	35%	4.39	0.15	0.14	11%	0.52	0.02	0.06	-65%	4.72	0.05	0.06	-19%
(3) No Generalists	$q_A = q_B = 0.5$	0.47	0.24	0.23	0%	4.47	0.23	0.23	0%	0.73	0.00	0.00	NA	4.93	0.00	0.00	NA
(4) More Generalists	$q_A = q_B = 0.2$	0.46	0.35	0.12	180%	4.46	0.18	0.12	47%	0.58	0.13	0.11	20%	4.78	0.12	0.11	14%
(5) More B specialists	$q_A = 0.3,\ q_B = 0.6$	0.59	0.34	0.22	56%	4.59	0.26	0.22	17%	0.56	0.03	0.07	-64%	4.76	0.06	0.07	-18%
(6) Fast-adjusting Generalists	$k = 2$	0.47	0.20	0.17	21%	4.47	0.17	0.17	4%	0.64	0.04	0.07	-39%	4.84	0.06	0.07	-6%
(7) Slow-adjusting Generalists	$k = 6$	0.48	0.24	0.17	40%	4.48	0.20	0.17	17%	0.65	0.02	0.07	-76%	4.85	0.05	0.07	-30%
(8) Larger A, smaller B market	$\bar{s}_A = 5/3, \bar{s}_B = 1/3,$ $q_A = 0.75, q_B = 0.15$	0.38	0.11	0.10	14%	4.38	0.10	0.10	4%	0.55	0.03	0.07	-60%	4.75	0.06	0.07	-16%
Supply shock hits market B																	
(9) Base case		0.48	0.02	0.07	-65%	4.48	0.06	0.07	-19%	0.64	0.34	0.25	35%	4.84	0.28	0.25	11%
(10) More default risk	$\sigma_z = 1.4\%$	0.51	0.02	0.06	-69%	4.51	0.04	0.05	-21%	1.52	0.88	0.70	26%	5.72	0.76	0.69	8%