Andreas Fuster, Laurie Goodman, David Lucca, Laurel Madar, Linsey Molloy, and Paul Willen

The Rising Gap between Primary and Secondary Mortgage Rates

- While the primary-secondary mortgage rate spread is a closely tracked series, it is an imperfect measure of the pass-through between secondary-market valuations and primary-market borrowing costs.
- This study tracks cash flows during and after the mortgage origination and securitization process to determine how many dollars (per $100 loan) are absorbed by originators, either to cover costs or as originator profits.
- The authors calculate a series of originator profits and unmeasured costs (OPUCs) for the period 1994-2012, and show that these OPUCs increased significantly between 2008 and 2012.
- Although some mortgage origination costs may have risen, a large component of the rise in OPUCs remains unexplained by cost increases alone, pointing to increased profitability of originators.

1. Introduction

The vast majority of mortgage loans in the United States are securitized in the form of agency mortgage-backed securities (MBS). Principal and interest payments on these securities are passed through to investors and are guaranteed by the government-sponsored enterprises (GSEs) Fannie Mae or Freddie Mac or by the government organization Ginnie Mae.[1] Thus, investors in these securities are not subject to loan-specific credit risk; they face only interest rate and prepayment risk—the risk that borrowers may refinance the loan when rates are low.[2]

In the primary mortgage market, lenders make loans to borrowers at a certain interest rate, whereas in the secondary market, lenders securitize these loans into MBS and sell them to investors. When thinking about the relationship between these two markets, policymakers and market commentators usually pay close attention to the "primary-secondary spread." This spread is calculated as the difference between an average

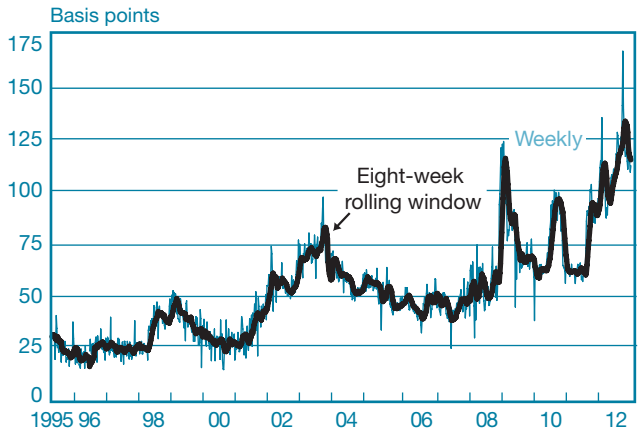
[1] Fannie Mae is the Federal National Mortgage Association (or FNMA); Freddie Mac is the Federal Home Loan Mortgage Corporation (FHLMC; also FGLMC); Ginnie Mae is the Government National Mortgage Association (GNMA).

[2] They also face the risk that borrowers prepay at lower-than-expected speeds when interest rates rise.

Andreas Fuster and David Lucca are senior economists in the Federal Reserve Bank of New York's Research and Statistics Group; Laurie Goodman is the center director of the Housing Finance Policy Center at the Urban Institute; Laurel Madar and Linsey Molloy are associates in the Bank's Markets Group; Paul Willen is a senior economist and policy advisor in the Federal Reserve Bank of Boston's Research Department.
Corresponding authors: andreas.fuster@ny.frb.org; david.lucca@ny.frb.org

This article is a revised version of a white paper originally prepared as background material for the workshop "The Spread between Primary and Secondary Mortgage Rates: Recent Trends and Prospects," held at the Federal Reserve Bank of New York on December 3, 2012. The authors thank Adam Ashcraft, Alan Boyce, James Egelhof, David Finkelstein, Kenneth Garbade, Brian Landy, Jamie McAndrews, Joseph Tracy, and Nate Wuerffel for helpful comments, and Shumin Li for help with the data. The views expressed are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York, the Federal Reserve Bank of Boston, or the Federal Reserve System.

Chart 1
The Primary-Secondary Spread



Sources: Bloomberg L.P.; Freddie Mac.

Chart 2
Back-of-the-Envelope Calculation of the Net Market Value of a Thirty-Year Fixed-Rate Mortgage Securitized in an Agency MBS

Sources: JPMorgan Chase; Freddie Mac; Fannie Mae; authors' calculations.

Notes: The chart shows the interpolated value of a mortgage-backed security (MBS) with coupon ($r_{primary}$ – g-fee) minus 100. The line reflects an eight-week rolling window average; the calculation uses back-month MBS prices.

mortgage interest rate (usually coming from the Freddie Mac Primary Mortgage Market Survey) and a representative yield on newly issued agency MBS—the "current-coupon rate."

Chart 1 shows a time series of the primary-secondary spread through the end of 2012. The spread was relatively stable from 1995 to 2000, at about 30 basis points; it subsequently widened to about 50 basis points through early 2008, but then reached more than 100 basis points in early 2009 and during 2012. Following the September 2012 Federal Open Market Committee announcement of additional MBS purchases, the spread temporarily rose to more than 150 basis points—a historical high that attracted much attention from policymakers and commentators at the time.

While the primary-secondary spread is a closely watched series, it is an imperfect proxy for the degree to which secondary-market movements are reflected in mortgage borrowing costs (the "pass-through") since, among other things, the secondary yield is not directly observed, but model-determined, and thus subject to model misspecification. Furthermore, mortgage market pass-through depends on the evolution of the GSEs' guarantee fees (or "g-fees," the price the GSEs charge for insuring the loan) as well as on mortgage originators' margins. To understand changes in the extent of pass-through over time, it is useful to track the two components separately. While g-fee changes are easily observable, we argue that originator margins are best studied by tracking the different cash flows during and after the origination process, rather than by looking at the primary-secondary spread (even after netting out g-fees). Indeed, since originators are selling the loans, their margin depends on the price at which they can sell them, rather than the interest rate on the security into which they sell the loans.

To get a sense of what lenders earn from selling loans, we first consider a simple "back-of-the-envelope" calculation. We track the secondary-market value of the typical offered mortgage loan (according to the Freddie Mac survey) over time, assuming that the lender securitizes and sells the loan as an agency MBS. To do so, we first deduct the g-fee from the loan's interest stream. We then compute the value of the remaining interest stream by interpolating MBS prices across coupons and subtracting the loan amount of $100.[3] Chart 2 shows that the approximate net market value of a mortgage grew from less than 100 basis points (or $1 per $100 loan) before 2009 to more than 350 basis points in the second half of 2012. Taken literally, the chart implies that lender costs (other than the g-fee), lender profits, or a combination of the two must have increased by 300 basis points, or a factor of four, in five years.

In this article, we first present a more detailed calculation of originator profits and costs, and then attempt to explain their rise by considering a number of possible factors

[3] For instance, assume that the mortgage note rate is 3.75 percent and the g-fee is 50 basis points, such that the remaining interest stream is 3.25 percent. Assuming that the 3.0 percent MBS trades at 102 and the 3.5 percent MBS trades at 104.5, the approximate market value of this mortgage in an MBS pool would then be simply the average of the two prices, 103.25, or 3.25 net of the loan principal.

affecting them. In section 2, we begin with a general discussion of the mortgage origination and securitization process, and how originator profits are determined. Here, we include a detailed discussion of the valuation of revenues from servicing and points as well as costs from g-fees, based on standard industry methods. Next, in section 3 we use these methods to derive a time series of average originator profits and unmeasured costs (OPUCs) for the period 1994-2012, which largely reflects the time-series pattern of Chart 2. We then compare OPUCs and the primary-secondary spread as measures of mortgage market pass-through. Finally, in section 4 we turn to possible explanations for the increase in OPUCs, including putback risk, changes in the valuation of mortgage servicing rights, pipeline hedging costs, capacity constraints, market concentration, and streamline refinancing programs. While some of the costs faced by originators may have risen over the period 2008-12, we conclude that a large component of the rise in OPUCs remains unexplained by cost increases alone, suggesting that originators' profits likely increased over this period. We then discuss possible sources of the rise in profitability. Capacity constraints likely played a significant role in enabling originator profits, especially during the early stages of refinancing waves. Pricing power coming from refinancing borrowers' switching costs could have been another factor sustaining originator profits.[4]

2. Measuring the Profitability of Mortgage Originations

2.1 The Origination and Securitization Process

The mortgage origination process begins when a borrower seeks a quote for a loan, either to purchase a home or to refinance an existing mortgage. Based on the borrower's credit score, stated income, loan amount, and expected loan-to-value (LTV) ratio, an originator offers the borrower a combination of an interest rate and an estimate of the amount of money the borrower will need to provide up front to close the loan.[5] For example, for a borrower who wants a $300,000, thirty-year fixed-rate mortgage, the originator may offer a 3.75 interest rate, known as the "note rate," with the borrower paying $3,000 (or 1.0 percent) in closing costs. If the borrower and originator agree on the terms, then the originator will typically guarantee these terms for a "lock-in period" of between thirty and ninety days, and the borrower will officially apply for the loan.

During the lock-in period, the originator processes the loan application, performing such steps as verifying the borrower's income and the home appraisal. Based on the results of this process, borrowers may ultimately not qualify for the loan, or for the rate that the originator initially offered. In addition, borrowers have the option to turn down the loan offer, for example, because another originator may have offered better loan terms. As a result, many loan applications do not result in closed loans. These "fall-outs" fluctuate over time and present a risk for originators, as we discuss in more detail in section 4.

Originators have a variety of alternatives to fund loans: they can securitize them in the private-label MBS market or in an agency MBS, sell them as whole loans, or keep them on their balance sheets. In the following discussion, we focus on loans that are "conforming" (meaning that they fulfill criteria based on loan amount and credit quality, so that they are eligible for securitization by the GSEs), and assume securitization in an agency MBS, meaning that this option either dominates or is equally profitable to the originator's alternatives.[6,7]

[4] Importantly, this article focuses on longer-term changes in the level of originator profits and costs, rather than on the high-frequency pass-through of changes in MBS valuations to the primary mortgage market.

[5] Throughout this article, we use the terms "lender" or "originator" somewhat imprecisely, as they lump together different origination channels that in practice operate quite differently. Currently, the most popular origination channel is the "retail channel" (for example, large commercial banks that lend directly), which accounts for about 60 percent of loan originations, up from around 40 percent over the period 2000-06 (source: *Inside Mortgage Finance*). The alternative "wholesale" channel consists of brokers and "correspondent" lenders. Brokers have relationships with different lenders that fund their loans, and account for about 10 percent of originations. Correspondent lenders account for 30 percent of originations, and are typically small independent mortgage banks that have credit lines from and sell loans (usually including servicing rights) to larger "aggregator" or "sponsor" banks. Our discussion in this section applies most directly to retail loans.

[6] The fraction of mortgages that are not securitized into agency MBS has steadily decreased in recent years, according to *Inside Mortgage Finance:* while the estimated securitization rate for conforming loans ranged from 74 to 82 percent over the period 2003-06, it has varied between 87 and 98 percent since then (the 2011 value was 93 percent). The private-label MBS market has effectively been shut down since mid-2007, with the exception of a few deals involving loans with amounts exceeding the agency conforming loan limits ("jumbo" loans).

[7] Our discussion throughout this article applies directly to conventional mortgages securitized by the GSEs Fannie Mae and Freddie Mac; the process of originating Federal Housing Administration (FHA) loans and securitizing them through Ginnie Mae is similar, but with some differences (such as insurance premia) that we do not cover here.

A key feature of an agency MBS is that principal and interest payments for these securities are guaranteed by the GSEs.[8] The GSEs charge a monthly flow payment, the g-fee, which is a fixed fraction of the loan balance. Flow g-fees do not depend on loan characteristics but may differ across loan originators. Until 2012, flow g-fees averaged approximately 20 basis points per year, but during 2012 they rose to about 40 basis points, reflecting a Congressionally mandated 10-basis-point increase to fund the 2012 payroll tax reduction and another 10-basis-point increase mandated by the Federal Housing Finance Agency (FHFA). As we discuss below, originators can convert all or part of the flow g-fee into an up-front premium by "buying down" the g-fee. Alternatively, they can increase the flow g-fee and receive an up-front transfer from the GSE by "buying up" the g-fee.

Since 2007, the GSEs have also been charging a separate up-front premium due upon delivery of the loan, known as the loan-level price adjustment (LLPA).[9] The LLPA contains a fixed charge for all loans (currently 25 basis points) known as an adverse-market delivery charge, as well as additional loan-specific charges that depend on loan characteristics such as the term of the loan, the LTV, and the borrower's FICO score. For instance, as of early 2013, the LLPA for a borrower with a FICO score of 730 and an LTV of 80 was 50 basis points (for a thirty-year fixed-rate loan; the charge is waived for loans with a term of fifteen or fewer years). Together with the 25-basis-point adverse-market delivery charge, this implies that the loan originator pays an up-front fee equal to 0.75 percent of the loan amount. Thus, the total up-front transfer between the originator and GSE consists of the LLPA plus or minus potential g-fee buy-ups or buy-downs, which can be either positive or negative. For simplicity, our discussion assumes that the transfer from the originator to the GSE is positive and refers to it as an "up-front insurance premium" (UIP).

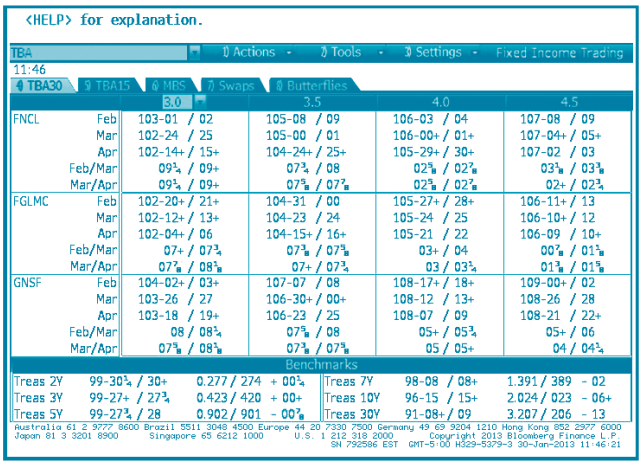
Once an originator chooses to securitize the loan in an agency MBS pool, it can select from different coupon rates, which typically vary by 50-basis-point increments. The note rate on the mortgage, for example, 3.75 percent, is always higher than the coupon rate on an agency MBS, for example, 3.0 percent. Who receives the residual 75-basis-point interest flow? Assuming the originator does not buy up or down the g-fee, approximately 40 basis points go to the GSEs (as of early 2013), leaving 35 basis points of "servicing income." The GSEs require the servicer to collect at least 25 basis points in servicing income, known as "base servicing." Base servicing is tied to the right

[8] If the loan is found to violate the representations and warranties made by the seller to the GSEs, the GSEs may put the loan back to the seller.

[9] LLPA is the official term used by Fannie Mae; Freddie Mac calls the corresponding premium "postsettlement delivery fee." The respective fee grids can be found at www.fanniemae.com/content/pricing/llpa-matrix.pdf and www.freddiemac.com/singlefamily/pdf/ex19.pdf.

Exhibit 1

Example of a TBA Price Screen



Source: Bloomberg L.P.

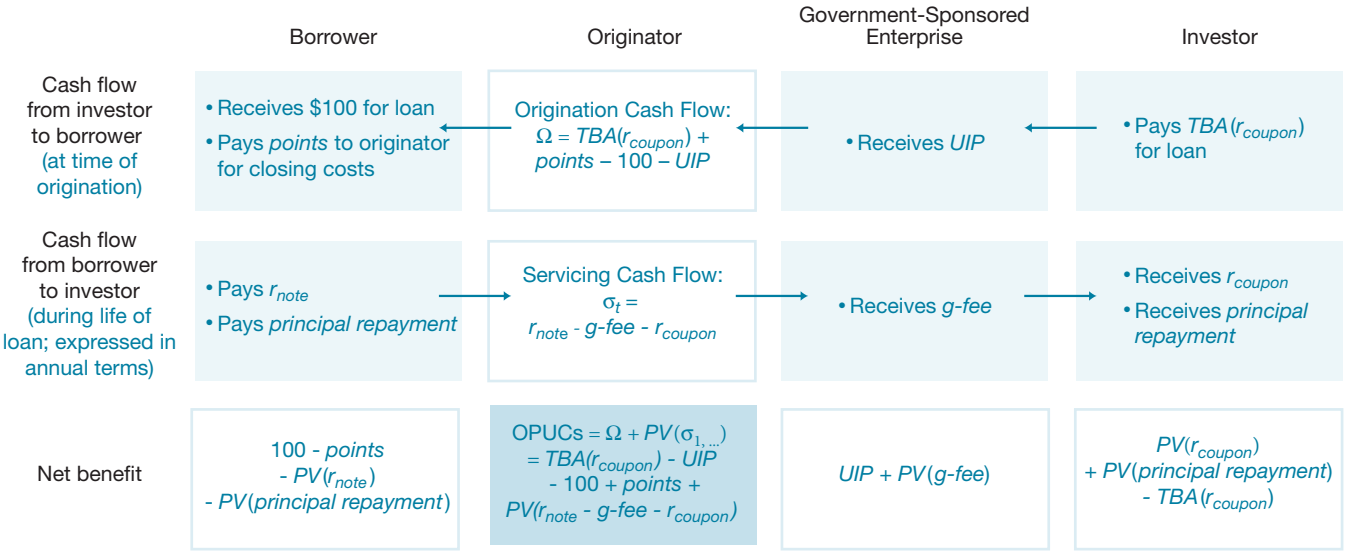
Notes: Prices are quoted in ticks, which represent 1/32nd of a dollar; for instance, 103-01 means 103 plus 1/32 = $103.03125 per $100 par value. The "+" sign represents half a tick (or 1/64). Quotes to the left of the "/" are bids, while those to the right are asks (or offers).

and obligation to service the loan (which involves, for instance, collecting payments from the borrower) and can be seized by the guaranteeing GSE if the servicer becomes insolvent. Servicing income in excess of 25 basis points—10 basis points in this example—is known as "excess servicing," and is a pure interest flow. One might surmise here that a loan in a 3.0 percent pool must have a rate of 3.65 percent or higher (3.0 plus 40 basis points for the g-fee plus 25 basis points for base servicing), but recall from above that the originator can buy down the g-fee so, in fact, the minimum note rate in a 3.0 percent pool is 3.25 percent. In practice, for a mortgage of a given note rate, originators compare the profitability of pooling it in different coupons, as described below.

Originators typically sell agency loans in the so-called TBA (to-be-announced) market. The TBA market is a forward market in which investors trade promises to deliver agency MBS at fixed dates one, two, or three calendar months in the future. For concreteness, Exhibit 1 displays TBA prices from Bloomberg at 11:45 a.m. on January 30, 2013. At this time, investors will pay 102 14+/32≈102.45 for a 3.0 percent Fannie Mae (here denoted FNCL) MBS for April settlement. To understand the role of the TBA market, suppose that Bank A expects to have $100 million of 3.5 percent note rate mortgages available for delivery in April. In order to hedge its interest rate risk, Bank A will then sell $100 million par of 3.0 percent pools "forward" in the TBA market at a price of $102.45 per $100 par, to be delivered on the standard settlement day in April. Over the following weeks,

Exhibit 2

Mortgage Loan Securitized in an Agency MBS and Sold in TBA Market: The Money Trail



Note: $TBA(r_{coupon})$ is the price of a mortgage-backed security (MBS) with coupon rate r_{coupon} in the "to-be-announced" market; *UIP* is up-front insurance premium (consisting of loan-level price adjustments plus or minus potential g-fee buy-ups or buy-downs); *PV* is present value.

Bank A assembles a pool of loans to be put in the security and delivers the loans to Fannie Mae, which then exchanges the loans for an MBS. This MBS is then delivered by Bank A on the contractual settlement day to the investor who currently owns the TBA forward contract in exchange for the promised $102.45 million. A key feature of a TBA trade is that at the time of trade, the seller does not specify which pools of loans it will deliver to the buyer—this information is "announced" only shortly before the trade settles. As a consequence, market participants generally price TBA contracts under the assumption that sellers will deliver the least valuable—or "cheapest-to-deliver"—pools at settlement.[10]

2.2 How Does an Originator Make Money on the Transaction?

A mortgage loan involves an initial cash flow at origination from investors to the borrower, and subsequent cash flows from the borrower to investors as the borrower repays the loan principal and interest. Exhibit 2 maps these cash flows for a mortgage loan securitized in a Fannie Mae MBS and sold in the TBA market. The top panel shows the origination cash flow, which involves the investor paying price $TBA(r_{coupon})$ to the originator in exchange for an MBS with coupon rate r_{coupon}.

[10] See Vickery and Wright (2013) for an overview of the TBA market.

From the investor's payment, an originator funds the loan and pays any *UIP* to Fannie Mae.[11] Together with points received from the borrower, the cash flow to the originator when the loan is made equals:

$$\Omega \equiv \text{Origination cash flow} = TBA(r_{coupon}) + points - 100 - UIP. \quad (1)$$

Through the life of the loan (middle panel of Exhibit 2), a borrower pays the note rate, r_{note}, from which Fannie Mae deducts the *g-fee* and the investor gets r_{coupon}, leaving servicing cash flow to the originator equal to:

$$\sigma_t \equiv \text{servicing cash flow}_t = r_{note} - g\text{-}fee - r_{coupon}. \quad (2)$$

Originator profits per loan are the sum of profits at origination (equation 1) and the present value (*PV*) of the servicing cash flow (equation 2) less all marginal costs (other than the *g-fee*) of originating and servicing the loan, which we call "unmeasured costs." Thus,

$$\text{originator profits} = \Omega + PV(\sigma_1, \sigma_2, \ldots) - \text{unmeasured costs}. \quad (3)$$

[11] Here and below, "originator" refers to all actors in the origination and servicing process, that is, if a loan is originated through a third-party mortgage broker, for instance, the broker will earn part of the value.

In our empirical exercise below, we study the sum of profits and unmeasured costs, which is what we can observe:

$$\text{originator profits and unmeasured costs (OPUCs)} = \Omega + PV(\sigma_1, \sigma_2, \ldots). \quad (4)$$

In later sections of the article, we attempt to assess to what extent changes in unmeasured costs can explain fluctuations in OPUCs.

We next consider a specific transaction to illustrate how the computations in Exhibit 2 are done in practice. Consider a loan of size $100 with a note rate of 3.75 percent locked in on January 30 for sixty days by a borrower with a FICO score of 730 and an LTV ratio of 80. The borrower agrees to pay 1 point to the originator for the closing, and the originator sells the loan into a TBA security with a 3.0 percent coupon for April settlement to allow sixty days for closing. Assuming the loan closes, how high are the OPUCs?

Computing the net revenue at origination, Ω, is relatively straightforward. According to Exhibit 1, investors pay $102.45 for every $100 of principal in a TBA security with a 3.0 percent coupon. As discussed earlier, the up-front insurance premium from the LLPA (and assuming no g-fee buy-up/-down) at the time was 0.75 percent of the loan (or 0.75 points). The originator collects 1 point from the borrower, remitting $100 for the loan, yielding Ω = 2.7 points.

Computing the net revenue at origination, Ω, is relatively straightforward.... Valuing the stream of servicing income after origination, (σ_1, σ_2, ...), is more complicated.

Valuing the stream of servicing income after origination, (σ_1, σ_2, ...), is more complicated. For now, we assume that the originator does not buy up or down the g-fee—a decision that we will revisit below. This means that from the borrower's interest flow of 3.75 percent, the GSEs collect 40 basis points, while the investors get 3.0 percent, leaving 35 basis points in flow servicing income, σ_t, decomposed into 25 basis points of base servicing and 10 basis points of excess servicing. There are a number of alternative ways to determine the present value of these flow payments:

IO Strip Prices or Coupon Swaps

Servicing income can be thought of as an interest-only (IO) strip, which is a security that pays a flow of interest payments, but no principal payments, to investors as long as a loan is active.[12] The main driver of the valuation of an IO strip is the duration of the loan—an IO strip is far more valuable if one expects the borrower to prepay in five years as opposed to one year; as in the latter case, interest payments accrue for a much shorter time period. One simple way to value IO strips is to construct them from TBA securities through coupon swaps. For example, going long on a 3.5 percent MBS and short on a 3.0 percent MBS generates interest cash flows of 50 basis points with prepayment properties that correspond roughly to loans in 3.0 and 3.5 pools. According to Exhibit 1, that 50-basis-point IO strip for April settlement would cost 2 11/32 (104 25+/32 minus 102 14+/32) ≈ 2.34. Since our originator has only 35 basis points of servicing, the coupon swap method would value servicing rights at 35/50 × 2.34 ≈ 1.6, resulting in OPUCs of 2.7 + 1.6 = 4.3 points.[13]

This method ignores the fact that base servicing generates other revenues, such as float income, in addition to the IO strip. To account for this additional value, it is often assumed that the base servicing is worth more than the present value of the IO strip. Assuming that base servicing is worth, for example, 25 percent more than excess servicing would yield a PV of servicing income of (25 × 1.25 + 10)/50 × 2.34 ≈ 1.9, so that OPUCs would equal 2.7 + 1.9 = 4.6 points.

Another shortcoming of the coupon swap method is that the coupon swap reflects differences in assumed loan characteristics across coupons. For example, TBA prices may reflect the fact that higher coupons are older securities having different prepayment characteristics. These differences will distort the valuations of interest streams from the coupon swaps.[14]

Constant Servicing Multiples

An alternative method for valuing servicing flows is to use fixed accounting multiples that reflect historical valuations of

[12] Another way to describe an IO strip is as an annuity with duration equal to the life of the loan.

[13] This is the method implicitly used in the back-of-the-envelope calculation in Chart 2, except that there we ignored points paid by the borrower.

[14] As an illustration, a 50-basis-point IO strip from a new 4.0 percent loan may not be worth as much as the price difference between the 3.5 and the 4.0 TBAs suggests, because the 4.0 TBAs may consist of loans that are older or credit impaired and thus prepay more slowly.

servicing. In the industry, the base servicing multiple is often assumed to be 5*x*, meaning that the present value of 25 basis points equals 1.25, while excess servicing is assumed to be valued at 4*x*, so that the value of the excess servicing in our example is 0.40. Using these servicing multiples, we see that the servicing income in our example is worth 1.65, meaning that OPUCs for this loan would be 2.7 + 1.65 = 4.35 points.

Buy-ups

As mentioned above, originators can convert the g-fee into an up-front premium, or vice versa, using buy-ups and buy-downs. A buy-up means that the flow g-fee increases, but to compensate, the GSE will reduce the UIP (or, in case it is negative, transfer money to the originator upon delivery of the loan). Thus, buying up the g-fee is a way to reduce the flow servicing income and increase income at the time of origination.

The GSEs offer a buy-up multiple, which is communicated to originators (but not otherwise publicly known), and varies over time, presumably with the level of the coupon swap. If, for example, the buy-up multiple is 3*x*, then a 10-basis-point increase in the g-fee reduces UIP by 30 basis points, lowering σ_t by 0.1 and raising Ω by 0.3. Note that only excess servicing, σ_t, -0.25, can be "monetized" this way, while 25 basis points of base servicing still need to be retained and valued by the originator. If we assume a base servicing multiple of 5*x*, as above, then buying up the g-fee by 10 basis points would lead to OPUCs of 3.0 + 1.25 = 4.25.

The buy-up multiple provides a lower bound on the valuation of excess servicing—the originator (or some other servicer) may value it at a higher multiple; but if it does not, it can sell its excess servicing to the GSEs. To what extent originators want to take advantage of this option depends on a number of factors. For example, as we discuss in section 4.1, the upcoming implementation of Basel III rules may require banks to hold additional capital against mortgage servicing assets, which may lower their effective valuation of servicing income. By buying up the g-fee, these banks can turn servicing cash flows that are subject to additional regulatory capital charges into cash. Another potential factor is the originator's beliefs about the prepayment properties of a pool of loans. For example, if a lender believes that the expected lifetime of a pool is shorter than average, it may choose to buy up the g-fee.

Market Prices of Servicing Rights

Finally, there is an active market for trading servicing rights, which can be sold by originators at origination or well afterward. One can use market prices to value servicing rights, but since not all servicing rights change hands, it is difficult to know whether the ones that trade are systematically more or less valuable than the ones that originators hold.

2.3 Best Execution

Lenders can decide to securitize a loan into securities having different coupons, which involves different origination and servicing cash flows. The strategy that maximizes OPUCs is known in the industry as "best (or optimal) execution."[15]

Thus far, we have assumed that the originator securitizes the loan in a 3.0 coupon. However, since the note rate is 3.75, the originator could alternatively sell it in a 3.5 coupon.[16] Given that the originator must retain 25-basis-point base servicing, such a choice would require buying down the entire 40-basis-point g-fee, meaning that instead of any flow payment to the GSE, the originator pays the full insurance premium up front. Exactly like the buy-up multiple discussed above, the GSEs also offer a (higher) buy-down multiple, which determines the cost of this up-front payment.

Using the prices in Exhibit 1, we note that the price of a 3.5 TBA coupon is 104 24+/32=104.77, meaning that changing coupons would increase loan sale revenues by 2.32 points. If we assume the buy-down multiple equals 7, then UIP would increase by 2.8 points relative to the 3.0 coupon case. Ω is thus equal to 2.22, or 0.48 less than it would be for the 3.0 coupon case. Meanwhile, servicing income is now simply σ_t = 0.25, as the flow g-fee has been bought down to zero, and with an assumed base servicing multiple of 5*x*, OPUCs for this execution would equal 2.22 + 1.25 = 3.47.

Comparing this OPUC value with the "constant servicing multiples" case above, we see that pooling into the 3.0 coupon would generate higher OPUCs than the 3.5 coupon and thus would be best execution for a mortgage with the 3.75 percent note rate.

However, this conclusion is sensitive to a number of assumptions—in particular, the valuation of excess servicing and the buy-down multiple.[17] As shown in Table 1, pooling in the higher coupon becomes more attractive as the buy-down multiple decreases or the excess servicing multiple decreases.

[15] See Bhattacharya, Berliner, and Fabozzi (2008) for an extensive discussion of pooling economics and mortgage pricing that also includes nonagency securitizations.

[16] The originator could also place the loan in a 2.5 percent or lower coupon—the only restriction is that the note rate cannot be more than 250 basis points above the coupon.

[17] As base servicing always needs to be retained, its valuation does not affect best execution—it shifts OPUCs up or down equally for all coupons.

Table 1

Dependence of Best Execution on Excess Servicing and Buy-Down Multiples

Excess Servicing Multiple	Buy-Down Multiple	OPUCs(3.0) (Points)	OPUCs(3.5) (Points)
4*x*	7*x*	4.35	3.47
4*x*	5*x*	4.35	4.27
3*x*	5*x*	4.25	4.27

Sources: Bloomberg L.P.; authors' calculations.
Note: OPUCs are originator profits and unmeasured costs.

2.4 Rate Sheets and Borrower Choice

Until now, we have taken the borrower choice as given—the borrower pays 1 point at origination and is offered a note rate of 3.75. However, from our OPUC calculations, it is clear that there are other combinations of note rate and points that would be equally profitable for the originator. For example, if the borrower paid a note rate of 4.0 instead, and the originator still pooled the loan into a 3.0 coupon, then excess servicing would increase by 25 basis points, leading to 1 point higher revenue under an excess servicing multiple of 4*x*. Therefore, the originator could maintain its profit margin by offering the borrower a combination of 0 points at closing and a note rate of 4.0.[18]

Indeed, originators offer borrowers precisely these sorts of alternatives between closing costs and rates. Table 2 shows part of a rate sheet provided by a bank to a loan officer on January 30, 2013.[19] The entries in the table are "discount points," which are points paid by the borrower at closing to lower the note rate on the loan. For example, assume that the total closing fees the originator would charge the borrower without any discount points would equal 1.58 points—sometimes referred to as "origination points." These fees include application processing costs, compensation for the loan officer, and also the LLPA (0.75 points in our example), which is usually charged directly to the borrower.

Our baseline borrower has a sixty-day lock-in period and a note rate of 3.75 percent; accordingly, based on the rate sheet, the borrower is contributing -0.581 discount points. This means that the bank is actually paying the borrower cash up front (often referred to as a "rebate"), which reduces closing costs from 1.58 points to the 1 point assumed throughout the example. If the borrower wanted a lower note rate, for example, 3.5 percent, then the closing costs would rise by 1.044 – (-0.581) = 1.625, or from 1 to 2.625 points. Alternatively, by choosing a rate of 4.125 percent, the borrower could get a rebate of 1.581 points and would pay nothing at closing.

As shown in the rate sheet, there is no single "mortgage rate." Rather, a large number of different note rates are available to borrowers on any given day, typically in increments of 0.125.[20] Originators simply change the number of discount points offered for the different note rates one or more times a day, reflecting secondary-market valuations (TBA prices), servicing valuations, and GSE buy-up/buy-down multiples.[21]

Table 2

Example of a Mortgage Rate Sheet

	Lock-in Period		
Note Rate	Fifteen Days	Thirty Days	Sixty Days
4.750	(3.956)	(3.831)	(3.706)
4.625	(3.831)	(3.706)	(3.581)
4.500	(3.706)	(3.581)	(3.456)
4.375	(3.331)	(3.206)	(3.081)
4.250	(3.081)	(2.956)	(2.831)
4.125	(1.831)	(1.706)	(1.581)
4.000	(1.456)	(1.331)	(1.206)
3.875	(1.081)	(0.956)	(0.831)
3.750	(0.831)	(0.706)	(0.581)
3.625	(0.081)	0.044	0.169
3.500	0.794	0.919	1.044
3.375	1.669	1.794	1.919
3.250	2.544	2.669	2.794
3.125	3.919	4.044	4.169

Source: www.53.com/wholesale-mortgage/wholesale-rate-sheets.html on January 30, 2013.

Notes: Figures are in percentage points of the loan amount. Loan type is a thirty-year fixed-rate loan. Column 1 shows the annual interest rate to be paid by the borrower over the life of the loan. Columns 2-4 show the points the borrower needs to pay up front to obtain the interest rate in column 1, for different lock-in periods. Parentheses denote negative figures.

[18] In fact, the 4.0 note rate might increase the profit margin, because it would potentially alter the best-execution coupon.

[19] Actual sample rate sheets can be found, for instance, at www.53.com/wholesale-mortgage/wholesale-rate-sheets.html. Most lenders do not make their rate sheets available to the public.

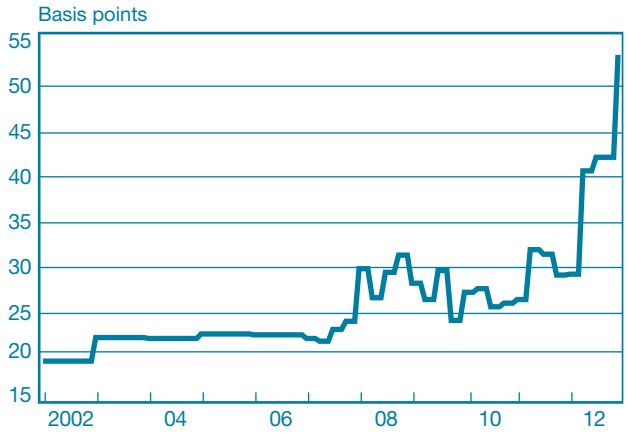
[20] That said, banks will often quote a headline mortgage rate, which is generally the lowest rate such that the number of discount points required from the borrower is "reasonable" (this rate is sometimes referred to as the "best-execution" rate for the borrower, not to be confused with the originator's best execution). In the example rate sheet, this rate would likely be 3.75 or 3.625, as going below 3.625 requires significant additional points from the borrower.

[21] The set of available note rates on a given day generally depends on which MBS coupons are actively traded in the secondary market.

2.5 Summary: Trade-offs, Trade-offs Everywhere

As shown in the preceding discussion, the different actors in the origination and securitization process have a number of trade-offs available to them. Borrowers can decide between paying more points up front and paying a higher interest rate later. Originators can choose between different coupons into which to pool a loan, which imply different origination and servicing cash flows; in addition, as part of this decision, originators can choose whether to pay the GSE insurance premium up front or as a flow. Finally, investors can choose to invest in securities with different coupons, with higher coupons requiring a larger initial outlay, but subsequently generating higher flow payments. Investor demand for different coupons, which reflects their prepayment and interest rate projections, ultimately affects originators' best-execution strategies and thus the point-rate grid offered to borrowers.

Chart 3
Average Effective Guarantee Fee



Source: Fannie Mae SEC Forms 10-K and 10-Q, various issues through 2012:Q4.

3. Measuring OPUCs over Time

Our goal in this section is to derive an empirical measure of average OPUCs (equation 4) for thirty-year fixed-rate mortgages for the period 1994 to 2012. To do so, we need to make a number of assumptions.

First, rather than valuing each possible loan note rate, we value a hypothetical mortgage having a note rate equal to the survey rate from Freddie Mac's Primary Mortgage Market Survey, at weekly frequency. We also use the weekly time series of average points paid from the same survey.

Second, rather than accounting separately for LLPAs and the flow g-fee, we use an "effective" g-fee, which assumes that LLPAs are paid over the life of the loan, as reported in Fannie Mae's Securities and Exchange Commission Form10-Q filings. The average size of the effective g-fee is shown in Chart 3. In our calculations, we incorporate anticipated changes in g-fees. In particular, the 10-basis-point increases that came into effect on April 1, 2012, and December 1, 2012, are assumed in our calculations to apply to loans originated January 1 and September 1, respectively, which is right after the increases were announced.

Third, as explained above, we need to value the servicing income flow. The coupon swap method has the advantage of being based on current market prices that reflect changes in the duration of the cash flows. But, as mentioned earlier, the coupon swap may also reflect differences in assumed loan characteristics across coupons; therefore, it may be a poor proxy for the value of an interest strip from a new loan.

To circumvent this issue, and also for the sake of simplicity, our baseline calculations use fixed multiples of 5*x* for base servicing, 4*x* for excess servicing, and 7*x* for buy-downs.[22] These are commonly assumed values in industry publications. Later, we explore the sensitivity of OPUCs to alternative assumptions.

Finally, we do a best-execution calculation, considering three different *TBA* coupons (using back-month prices) into which the mortgage could potentially be pooled.[23] The highest coupon is set such that it requires the originator to buy down some or all of the *g-fee* up front, while instead, for the other two possible coupon options, the originator retains positive excess servicing because the loan's interest payment is more than sufficient to cover the *g-fee* and base servicing.[24] The best execution among the three options determines our OPUC value for the week in question. Before turning to the weekly OPUC time series, we report in Table 3 a detailed OPUC calculation on a given day. We can infer, from the bottom of the table, that the mid-coupon execution is optimal in this example.

[22] We assume the buy-up multiple to be smaller than 4*x*, such that, in our calculations, buy-ups are never used.

[23] The use of back- rather than front-month TBA price contracts reflects the originators' desire to hedge price movements during the lock-in period, as discussed in more detail in section 4.

[24] Depending on the mortgage rate, pooling into the highest candidate coupon may not actually be a possibility—as explained, the mortgage rate needs to exceed the coupon rate by at least 25 basis points.

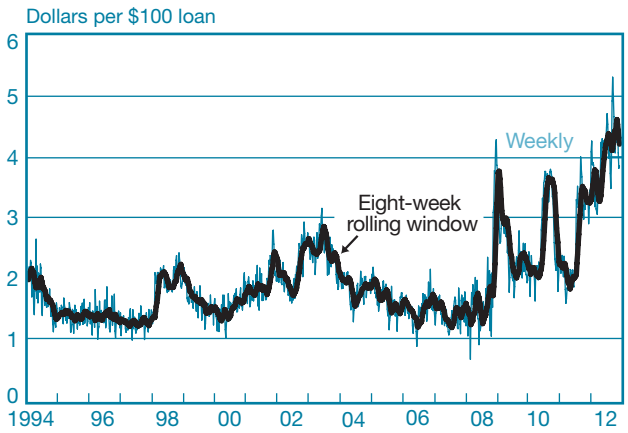
Table 3
Example of OPUCs Best-Execution Calculation

TBA Coupon (Percent)	3.5	4.0	4.5	(1)
Coupon-independent inputs (percent)				
Mortgage rate	4.78	4.78	4.78	**(2)**
Points	0.7	0.7	0.7	**(3)**
Effective g-fee	0.261	0.261	0.261	**(4)**
Base servicing	0.25	0.25	0.25	**(5)**
Excess servicing	0.769	0.269	-0.231	**(6)** = **(2)** − **(1)** − **(4)** − **(5)**
Coupon-specific inputs (dollars per par value)				
TBA price (back-month)	97.55	99.95	101.67	**(7)**
Value of base servicing	1.25	1.25	1.25	**(10)** = 5 × **(5)**
Value of excess servicing	3.08	1.08		**(11)** = 4 × **(6)** if **(6)** > 0
G-fee buy-down			-1.62	**(12)** = 7 × **(6)** if **(6)** < 0
Revenues from TBA sale less payout to borrower	-1.75	0.65	2.37	**(13)** = **(7)** − (100 − **(3)**)
Value of servicing net of g-fee	4.33	2.33	-0.37	**(14)** = **(10)** + **(11)** + **(12)**
OPUCs				
By coupon	2.58	2.98	2.00	**(15)** = **(13)** + **(14)**
Best-execution		2.98		**(16)** = max**(15)** if **(2)** − **(1)** > .25

Source: Authors' calculations.

Note: Calculation is for April 30, 2009. OPUCs are originator profits and unmeasured costs; TBA is "to-be-announced."

Chart 4
Originator Profits and Unmeasured Costs, 1994-2012



Sources: JPMorgan Chase; Freddie Mac; Fannie Mae; authors' calculations.

3.1 Results

The weekly OPUC series over the period 1994 to 2012 is shown in Chart 4. The series averaged about $1.50 between 1994 and 2001, then temporarily increased to the $2.00-$3.00 range over 2002-03, before declining again and remaining below $2.00 for most of the period 2005-08. The OPUC measure jumped dramatically to more than $3.50 in early 2009 and then again in mid-2010. Most notably, however, it increased further over 2012, and reached highs of more than $5 per $100 loan in the second half of the year, before declining again toward the end of 2012.

As shown in the back-of-the-envelope calculation in Chart 2, the higher valuation of loans in the MBS market is the main driver of the increase in OPUCs toward the end of our sample period. Relative to that figure, the increase in OPUCs over 2009-12 in Chart 4 is less dramatic; this is because the earlier calculation implicitly valued servicing through coupon swaps, which were very low in early 2009 but relatively high since 2010. In contrast, in Chart 4 we have

assumed constant multiples.[25] As we discuss in more detail below, servicing right valuations appear to have declined, rather than increased, over the past few years, supporting the use of fixed multiples rather than coupon swaps.

When interpreting the OPUC series, it is important to keep in mind a few notes. First, the measure uses data on thirty-year conventional fixed-rate mortgage loans only and therefore bears no direct information on other common types of loans, such as fifteen-year fixed-rate mortgages, adjustable-rate mortgages, Federal Housing Administration loans, or jumbos.

Second, since the measure uses survey rates/points and average g-fees, our OPUC series is an average industry measure rather than an originator-specific one. In addition, rates and points may be subject to measurement error that could distort the OPUC measure at high frequency, although this should not have much effect on low-frequency trends.

Third, the measure is a lower bound to the actual industry OPUCs, as it uses TBA prices to value loans, while originators may have more profitable options available. Indeed, as noted in section 2, about 10 percent of conforming loans are held on balance sheet, implying that originators find it more (or equally) profitable not to securitize these loans. In addition, a significant fraction of agency loans is securitized in specified MBS pools that trade at a premium, or pay-up, to TBAs. In fact, the fraction of mortgages sold into the non-TBA market appears to have increased substantially in 2012, relative to earlier years. Table 4 shows an estimate of pools that are being issued as specified ("spec") pools, rather than TBA pools.[26] Over the first ten months of 2012, only about 60 percent (value-weighted) of all pools were issued to be traded in the TBA market, while the rest were issued as spec pools. The increase in spec-pool issuance is due in part to Making House Affordable (MHA) loans originated under the Home Affordable Refinance Program (HARP), which account for about 20 percent of all issuance and typically trade at significant pay-ups to TBAs, owing to their lower expected prepayment speeds. For example, over the second half of 2012, Fannie 3.5 and 4 MHA pools with LTVs above 100 traded on average about 1 1/2 and 3 1/2 points higher than corresponding TBAs. Low-loan-balance pools, the second largest spec-pool type, received similarly high pay-ups.

The higher valuation of loans in the MBS market is the main driver of the increase in OPUCs toward the end of our sample period.

Table 4

Issuance of Various GSE Thirty-Year Fixed-Rate Pool Types, January–October 2012

Pool Type	Balance (Millions of Dollars)	Loan Count	Balance (Percent)	Count (Percent)
TBA	379,763	1,347,516	59	46
MHA[a]	124,779	559,180	20	19
Loan balance[b]	97,161	867,628	15	30
Other specified[c]	36,588	138,735	6	5
Total	638,292	2,913,059	100	100

Sources: Fannie Mae; Freddie Mac; 1010data; Amherst Securities.

Note: GSE is government-sponsored enterprise. TBA is "to-be-announced." MHA is the Making Home Affordable program.

[a]Includes pools that are 100 percent refi with 80<Orig LTV≤105, and pools with loans >105 LTV.
[b]Includes pools that contain only loans with balances less than or equal to $175,000.
[c]Includes 100 percent investor, NY, TX, PR, low FICO pools, and "mutt" pools (variety of specified loan types). Excludes GSE pool types that are jumbo, FH reinstated, co-op, FHA/VA, IO, relo, and assumable.

3.2 OPUCs, the Primary-Secondary Spread, and Pass-Through

In assessing the extent to which secondary-market movements pass through to mortgage loan rates, most commentators focus on the primary-secondary spread—the difference between primary mortgage rates and the yield on MBS securities implied by TBA prices. As shown in Chart 1, the spread reached record-high levels over the course of 2012, suggesting that declines in primary mortgage rates did not keep pace with those on secondary rates. For example, while the primary-secondary spread averaged 73 basis points in 2011, the corresponding number was 113 basis points in 2012.

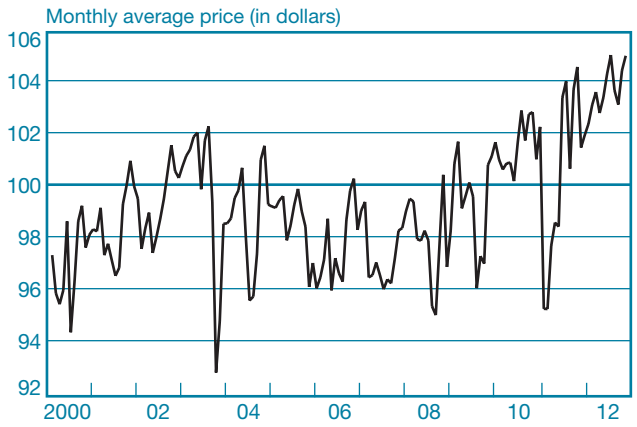
While the primary-secondary spread is a closely tracked series, it is an imperfect measure of the pass-through between secondary-market valuations and primary-market borrowing costs for several reasons.

[25] Another difference is that we take changes in points paid by borrowers into account, but this matters relatively little (the average amount of points paid by borrowers was relatively stable, between 0.4 and 0.8 over the period 2006-12).

[26] We do not know with certainty whether a pool is ultimately traded in the TBA market or as a specified pool; we simply assume that pools that strictly adhere to certain specified pool criteria are also subsequently traded as such.

Chart 5

Price of Lowest Fannie Mae TBA Thirty-Year Coupon with Sizable Issuance



Source: eMBS; JPMorgan Chase.

Notes: TBA is "to-be-announced." "Sizable issuance" means that the coupon accounts for at least 10 percent of total issuance in that month.

First, the yield on any MBS is not directly observable, because the timing of cash flows depends on prepayments. Therefore, the calculation of the yield is based on the MBS price and cash flow projections from a prepayment model, which itself uses as inputs projections of conditioning variables (for example, interest rates and house prices). In addition, for TBA contracts, the projected cash flows and the yield also depend on the characteristics of the assumed cheapest-to-deliver pool. The resulting yield is thus subject to errors due to model misspecification.

Second, the primary-secondary spread typically relies on the theoretical construct of a "current coupon MBS." The current coupon is a hypothetical TBA security that trades at par and has a yield meant to be representative of those on newly issued securities.[27] Historically, this par contract has usually fallen between two other actively traded TBA coupons; however, in recent times, even the lowest coupon with nontrivial issuance has generally traded significantly above par (Chart 5). As a result, the current coupon rate is obtained as an extrapolation from market prices, rather than a less error-prone interpolation between two traded points.[28] Importantly, the impact of potential prepayment model misspecification on yields is amplified when the security trades significantly above (or below) par because the yield on the security depends on the timing of the amortization of the bond premium.

A better way to think about pass-through is to look directly at what happens with the money paid by an investor in the secondary market—does it go to borrowers, originators, or the GSEs (either up front, or through equivalent flow payments)? The purpose of the OPUC measure is to track how many dollars (per $100 loan) get absorbed by originators, either to cover costs other than the g-fee, or as originator profits.[29] G-fees also contribute to the overall cost of mortgage credit intermediation—increasing these fees means that less money goes to borrowers (or equivalently, that they need to pay a higher rate). So, full pass-through of secondary-market movements to borrowers would require OPUCs and g-fees to remain constant (or, alternatively, a rise in g-fees would need to be offset by a decrease in OPUCs).

In panel A of Chart 6, we conduct a counterfactual exercise in which we compute a hypothetical survey note rate during 2012, assuming that either the OPUCs only (dark blue line), or both the OPUCs and the g-fee (light blue line), had stayed at their average levels in 2011:Q4.[30] The comparison of the light blue line with the black line, the actual realized mortgage rate, shows that had the cost of mortgage intermediation stayed constant relative to 2011:Q4, mortgage rates during 2012 would at times have been substantially lower, with a maximum gap between the two rates of 55 basis points in early October 2012.

Comparing the black line with the dark blue line (holding only OPUCs fixed but letting g-fees increase), we note that over most of 2012, much of the gap between the actual and counterfactual rate derives from the rise in OPUCs.

[27] An alternative is to calculate the yield on a particular security, which may trade at a pay-up to the cheapest-to-deliver security. However, such a calculation is still subject to other model misspecification and would not be representative of the broad array of newly issued securities.

[28] Additionally, the current coupon is typically based on front-month contract prices, while a more accurate measure would use back-month contracts, because loans that rate-lock today are typically packaged into TBAs at least two months forward.

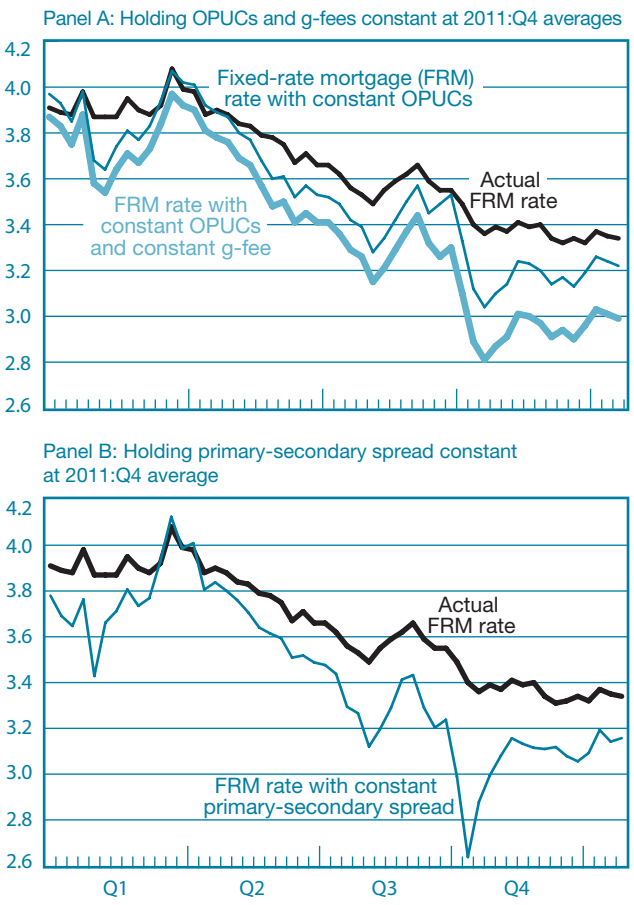
[29] It is important to keep in mind that changes in the secondary yield, even if correctly measured, do not necessarily translate one-to-one into changes in originator margins, which are determined by the TBA prices of different coupons (which in turn determine optimal execution), and also by points paid by the borrower. The primary-secondary spread, even net of g-fees, is thus at best an imprecise measure of originator margins and profitability.

[30] The effective g-fee in our calculation for 2011:Q4 is 28.8 basis points, which then increases to 38.9 basis points for the period January-March 2012 (as the announced increase effective April 1, 2012, is assumed to already be relevant for loans originated at that point), 40.3 basis points for the period April-June 2012, 41.8 basis points for July and August, and then increases by another 10 basis points, to 51.8 basis points, for the rest of 2012 as the December 1 g-fee increase becomes relevant to pricing.

Chart 6

Counterfactual Paths of Mortgage Rates over 2012

Percent



Sources: Bloomberg L.P.; Freddie Mac; authors' calculations.

Note: OPUCs are originator profits and unmeasured costs.

Additionally, it is apparent that in times when rates are stable or increasing, the counterfactual rate with constant OPUCs tends to be close to the actual rate, and most of the gap between the black and the light blue lines comes from the higher g-fees (this is the case, for instance, toward the end of the year). It is during times when rates fall (secondary-market prices increase) that actual rates do not fall as much as they would with constant OPUCs. As we discuss later, this is consistent with originators having limited capacity, which means they can keep rates relatively high and make extra profits. That said, one should not necessarily interpret the counterfactual rate series as indicating "where rates should have been," as this would require a judgment regarding the "right" level of OPUCs. Here, we took the average over 2011:Q4 as our baseline, but if instead we took a lower value, such as the average OPUCs over all of 2011, the dark blue and light blue lines would be significantly lower.

In panel B of Chart 6, we conduct a similar counterfactual rate analysis, but using the primary-secondary spread as the measure of the cost of mortgage intermediation. Holding this spread (measured as the Freddie Mac survey rate minus the Bloomberg current coupon yield) constant, we again get a hypothetical mortgage rate under full pass-through. As shown in panel B, while the overall pattern is similar to the counterfactual rate with constant OPUCs and g-fees in panel A, the series in panel B is more volatile, with the gap between the counterfactual and actual rate spiking at 75 basis points in late September 2012. This volatility of the counterfactual rate and the presence of such large spikes illustrate the imperfect nature of the primary-secondary spread as a pass-through measure.

Over most of 2012, much of the gap between the actual and counterfactual rate derives from the rise in OPUCs.

4. Potential Explanations for the Rise in Costs or Profits

The rest of the article explores in more detail factors that may have driven the observed increase in OPUCs over the period 2008-12. On the cost side, we focus on changes in pipeline hedging costs, putback risk, and possible declines in the valuation of mortgage servicing rights. We also briefly discuss changes in loan production expenses. On the profit side, we focus on potential increases in originators' pricing power due to capacity constraints, industry concentration, or switching costs for refinancers.

4.1 Costs

Loan Putbacks

Originators pay g-fees to the GSEs as an insurance premium; in exchange, the GSEs pay the principal and interest of the loan in full to investors when the borrower is delinquent.

However, mortgage originators or servicers are obligated to repurchase nonperforming or defaulted loans under certain conditions, for example, when the GSEs establish that the loan did not meet their original underwriting or eligibility requirements, that is, if the loan representations and warranties are flawed.[31] The repurchase requests have increased rapidly since the 2008 financial crisis and have been the source of disputes between originators and GSEs. The increased risk to originators that the loan may ultimately be put back to them has been cited as a source of higher costs and thus OPUCs.

How can we assess the magnitude of the contribution of putback costs to OPUCs? To do so, one needs to imagine a stress scenario—not a modal one—with a corresponding default rate, and then assume fractions of putback attempts by the GSEs, putback success, and loss-given-defaults for servicers/lenders forced to repurchase the delinquent loan.

The increased risk to originators that the loan may ultimately be put back to them has been cited as a source of higher costs and thus OPUCs.

To construct a ballpark estimate of the possible putback cost on new loans, we start from the experience of agency loans originated during the period 2005-08. Based on a random 20 percent sample of conventional first-lien fixed-rate loans originated during that period in the servicing data set of LPS Applied Analytics, we find that about 16.5 percent of GSE-securitized mortgages (value-weighted) have become sixty-or-more days delinquent at least once, and 11.5 percent of them have ended in foreclosure.[32] Importantly, these vintages include a substantial population of borrowers with relatively low FICO scores, undocumented income or assets, or a combination of these factors. For instance, the median FICO score was around 735, while the 25th percentile was at 690. In 2012, however, the corresponding values on non-HARP loans were around 770 and 735, respectively.[33]

To account for the tighter underwriting standards on new loans, we focus on the performance of GSE-securitized loans from the 2005-08 vintages with origination FICO of at least 720 and full documentation. Among those, "only" about 8.8 percent have become sixty-or-more days delinquent, and 5.5 percent have ended in foreclosure. Thus, because of today's more stringent underwriting guidelines for agency loans, our expectation in a stress scenario would be for delinquencies, and hence potential putbacks, to be roughly half as large, relative to those experienced by the 2005-08 vintages. Furthermore, we would expect the frequency of putback attempts to be roughly half as large for loans with full documentation as for the overall population of delinquent loans.

We obtain an estimate of the fraction of loans that the GSEs could attempt to force the lender to repurchase from Fannie Mae's 2012:Q3 Form 10-Q, which states (on page 72) that as of 2012:Q3, about 3 percent of loans from the 2005-08 vintages have been subject to repurchase requests (compared with only 0.25 percent of loans originated after 2008). Thus, given that repurchase requests are issued primarily conditional on a delinquency, we would anticipate repurchase requests in a stress scenario to be about one-quarter (0.5 delinquency rate × 0.5 putback rate) as high as those recorded on the 2005-08 vintage, or about 0.75 percent.[34]

Based on repurchase disclosure data collected from the GSEs,[35] it appears that about 50 percent of requests ultimately lead to buybacks of the loan. Furthermore, if we assume a 50 percent loss-given-default (which seems on the high side), this would generate an expected loss to the lender/servicer of:

$$0.75 \text{ percent} \times 0.5 \times 0.5 = 19 \text{ basis points}$$

This estimate, which we think of as being conservative (given the unlikely repetition at this point of large house price declines experienced by the 2005-08 vintages), would imply a putback cost of 19 cents per $100 loan. This cost is modest relative to the widening in OPUCs experienced over the period 2008-12.[36] That said, perhaps the "true" cost of putback risk comes from originators trying to avoid putbacks in the first place by spending significantly more resources on underwriting new loans or on defending against putback

[31] It is also possible that originators need to repurchase incorrectly underwritten loans prior to a loan becoming delinquent. However, the repurchase of nondelinquent loans is likely less costly to originators. The rest of this section therefore focuses on repurchases of delinquent loans.

[32] These statistics are as of November 2012.

[33] Origination LTVs have not changed as dramatically: in 2012, approximately 16 percent of non-HARP loans had an LTV at origination above 80; this is only slightly lower than during the period 2005-08. However, the fraction of loans with second liens was likely higher during the boom period. Also, in 2012 there are no non-HARP Freddie Mac loans with incomplete documentation (this is not disclosed in the Fannie Mae data, but is likely similar).

[34] Without the assumption that full-documentation loans are less likely to be put back, the expected putback rate would be 1.5 percent, resulting in an expected loss of 37.5 basis points.

[35] Source: *Inside Mortgage Finance*.

[36] Furthermore, the FHFA introduced a new representation and warrant framework for loans delivered to the GSEs after January 2013 that relieves lenders of repurchase exposure under certain conditions (for example, if the loan was current for three years). This policy change should further reduce the expected putback cost going forward.

claims. Furthermore, the remaining risk on older vintages is larger than on new loans, and many active lenders are also still subject to lawsuits on nonagency loans made during the boom. It is unclear, however, why these claims on vintage loans should affect the cost of new originations.

Mortgage Servicing Rights Values

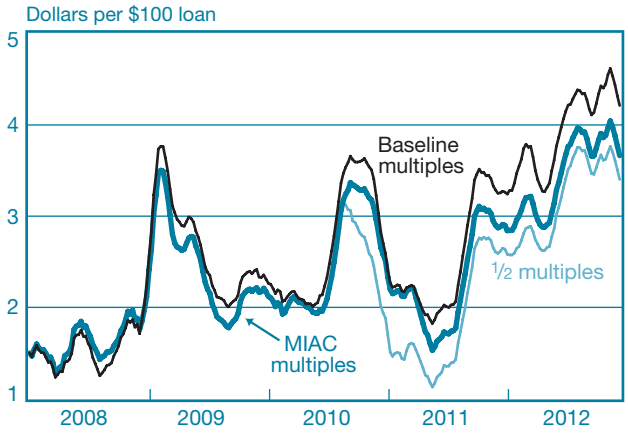
The baseline OPUC calculation assumes constant servicing multiples throughout the sample of 5*x* for base servicing and 4*x* for excess servicing flows. While these are commonly assumed levels, according to market reports, mortgage servicing right (MSR) valuations have declined over the past few years. In this section, we study the sensitivity of OPUCs to alternative multiple assumptions.

We obtain a time series of normal (or base) servicing multiples for production agency MBS coupons from the company Mortgage Industry Advisory Corporation (MIAC).[37] These multiples declined from about 5*x* in early 2008 to about 3.25*x* in November 2012.[38] To evaluate the impact on OPUCs, we repeat our earlier calculation using the MIAC base multiples.[39] The results are shown in Chart 7. Comparing the black (baseline) and dark blue (MIAC) lines, we see that the lower multiple values reduce OPUCs by about sixty cents at the end of 2012, a somewhat significant impact.

Some commentators have attributed the decline in multiples to a new regulatory treatment of MSRs under the 2010 Basel III accord. While the three U.S. federal banking regulatory agencies released notices of proposed rulemaking to implement the accord on June 12, 2012, the introduction of the new rules, originally set for January 2013, has been postponed. Under the June 2012 proposal, concentrated MSR investment will be penalized and will generally receive a higher risk weighting.[40] The long phase-in period for these rules makes it unclear how much the expected tighter regulatory treatment is already affecting MSR multiples. Nonetheless, in order to assess an upper-bound impact on OPUCs, we consider here a more stressed scenario than implied by the MIAC multiples. In this scenario, our baseline multiples are halved starting (for simplicity) with the disclosure by the Basel Committee of the capital rules in July 2010.[41] The resulting eight-week-rolling OPUC series is also depicted in Chart 7. As shown in the chart, following a halving of the MSR multiples, the implied OPUC declines are significant, but still not sufficient to explain the historically high OPUC levels in 2012.

We conclude that lower multiples, while having a sizable impact on OPUCs, can only partially offset their increase over the past few years.

Chart 7
Sensitivity of OPUCs to Alternative Assumptions about Mortgage Servicing Right Multiples



Sources: JPMorgan Chase; Freddie Mac; Fannie Mae; MIAC; authors' calculations.

Notes: The data reflect an eight-week rolling window. MIAC is the Mortgage Industry Advisory Corporation.

[37] These multiples come from MIAC's "Generic Servicing Assets" portfolio and are based on transaction values of brokered bulk MSR deals, surveys of market participants, and a pricing model.

[38] Key drivers of servicing right valuations are expected mortgage prepayments—lower interest rates mean a higher likelihood that the servicing flow will stop due to an early principal payment—and, in the case of base servicing, varying operating costs in servicing the loan, for example, when loans become delinquent. Another important component is the magnitude of the float interest income earned, for instance, on escrow accounts.

[39] We assume a 20 percent discount for excess servicing and keep the g-fee buy-down multiple unchanged at 7*x*. Also, as our MIAC series ends in November 2012, we assume that the multiple in December is identical to that in November.

[40] MSRs will be computed toward Tier 1 equity only up to 10 percent of their value, and risk-weighted at 250 percent, with the rest being deducted from Tier 1 equity. This treatment is significantly more stringent than the status quo that risk-weights the MSRs at 100 percent and limits MSRs to 50 percent of Tier 1 capital of banks (100 percent for savings and loans).

[41] In this alternative scenario, base servicing is now valued at 2.5*x*, while excess servicing is valued at 2*x*. (The GSE buy-down multiple is assumed to stay at 7*x*.) The optimal execution in this exercise again takes into account the lower levels of the multiples.

Pipeline Hedging Costs

For loans that are securitized in MBS, the "mortgage pipeline" is the channel through which an originator's loan commitment, or rate-lock, is ultimately delivered into a security or terminated with a denial or withdrawal of the application. The originators' commitment starts with a rate-lock that typically ranges between thirty and ninety days. This time window appears to have increased significantly in recent years. For example, the time from application to funding for refinancing applications increased from about thirty days in late 2008 to more than fifty days in late 2012 (as shown graphically in section 4.2 below).

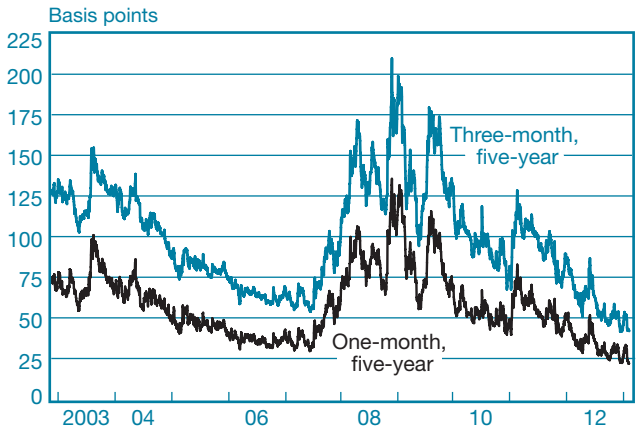
Originators face two sources of risk while the loan is in the pipeline: changes in the prospective value of the loan due to interest rate fluctuations and movements in the fraction of rate-locks that do not ultimately lead to loan originations, referred to as "fallouts."

The first risk—potential changes in the value of the loan due to interest rate movements—can be hedged by selling TBA contracts: at the time of the loan commitment, originators who are long a mortgage loan at the time of the rate-lock can offset the position by selling the yet-to-be-originated loan forward in the TBA market. The calculation in section 3 already takes into account these hedging costs: when computing the OPUC measure, we use the back-month TBA contract price that settles on average about forty-five days following the transaction. To the extent that originators may have been able to sell into the front-month TBA market when the length of the pipeline was shorter, our calculations may understate OPUCs for earlier years by the price difference, or "drop," between the two contract prices. Yet, this drop is typically only about 20 basis points in price space. We conclude that the lengthening of the pipeline does not appear to have had a significant economic impact on the cost of price hedging, and thus the rise in OPUCs experienced over the period 2008-12.

Originators face two sources of risk while the loan is in the pipeline: changes in the prospective value of the loan due to interest rate fluctuations and movements in the fraction of rate-locks that do not ultimately lead to loan originations, referred to as "fallouts."

The second risk is due to movements in the fallout rate. As discussed in section 2, borrowers' terminations may occur involuntarily (if they do not ultimately qualify for the loan or rate offer) or voluntarily. Except for changes in lending standards and house prices, fluctuations in involuntary terminations are largely driven by idiosyncratic factors that are diversified for originators with large-enough portfolios. Movements in voluntary terminations, on the other hand, are mostly due to primary rate dynamics: following the initial rate-lock, mortgage rates may fall, prompting borrowers to pursue a lower rate loan with either the same or a different lender. Common ways to hedge this risk are to dynamically delta-hedge the position using TBAs, using mortgage options or swap options, or a combination of these (or other) strategies.[42] To illustrate, we now consider a hedging example using at-the-money swaptions to gauge the magnitude and time-series pattern of the interest rate hedging cost.

Chart 8
Swaption Price Premia



Source: JPMorgan Chase.

Based on market reports and data from the Mortgage Bankers Association (MBA), normal fallout rates average about 30 percent, and we assume that an originator hedges as much using swaptions. Chart 8 shows the price premium in basis points for swaptions on a five-year swap rate with expirations of one and three months. Conditional on a 30 percent hedging strategy, the cost of protection, when using a three-month expiration, would be about 0.3 x 40 basis points = 12 basis points, or a 12 cent impact on OPUCs. The extension in the length of the pipeline, which may have led originators to go from one-month to three-month expiration, also had a rather small impact on OPUCs.

[42] Correspondent lenders, or small lenders that sell whole loans to the GSEs, can manage the fallout risk by entering into "best-effort" locks with the buyer of the loan. Under this arrangement, the originator does not need to pay a fine for not delivering a mortgage that does not close, unlike under "mandatory delivery." To compensate, the price offered by the buyer of the loan is lower. Thus, in a sense, "best-effort" commitments allow (small) originators to "outsource" the hedging of fallout risk.

More generally and beyond our specific example, implied volatility and option price premia have declined significantly since the fall of 2008, reflecting the lower rate volatility environment. While we do not explicitly consider other, more complex hedging strategies, the lower volatility environment has likely also lowered the cost of these strategies. This is in contrast with the rise in OPUCs over this period. In sum, changing hedging costs does not appear to account for a significant portion of the rise in OPUCs, and at least the cost of hedging fallout risk may in fact have declined during the period 2009-12.

Other Loan Production Expenses

A final possible cost-side explanation for the increase in OPUCs is that other loan production expenses, including costs related to the underwriting of loans and to finding borrowers (sales commissions, advertising, and so on) have increased substantially over the past few years. While it is difficult to obtain a variable loan cost series that can be easily mapped into the OPUC measure, the MBA collects in its Quarterly Mortgage Bankers Performance Report survey information on total loan production expenses that include both fixed and variable costs, such as commissions, compensation, occupancy and equipment, and other production expenses and corporate allocations. With the caveat that the sample of respondents is composed of small- and medium-sized independent mortgage companies, the data indicate a modest increase in loan production expenses over the past few years and a fairly stable pattern of these expenses. For example, total loan production expenses averaged $4,717 per loan in 2008, and $5,163 per loan in 2012:Q3.[43] This modest increase appears unlikely to explain the more than doubling in OPUCs over the period 2008-12.

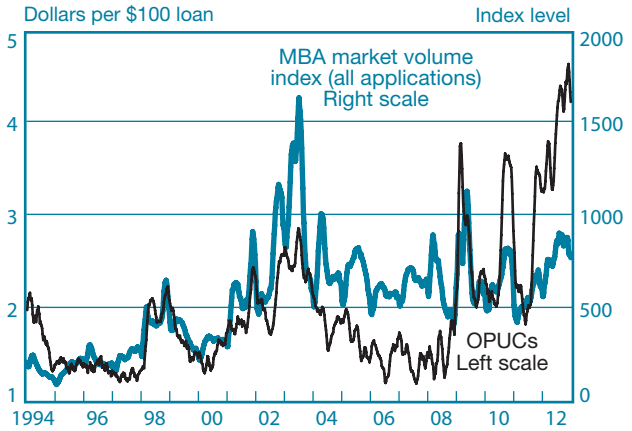
4.2 Industry Dynamics and Originators' Profits

The discussion in the previous subsection appears to indicate that the higher OPUCs on regular agency-securitized loans over the period 2008-12 were not likely driven exclusively, or even mostly, by increases in costs. As a result, the rise in OPUCs during this time could reflect an increase in profits. If so, what are the potential driving forces behind such an increase?

[43] Source: Mortgage Bankers Association, *Press Release Performance Report*, various issues. The numbers cited are gross expenses, not including any revenue such as loan origination fees or other underwriting, processing, or administrative fees.

Chart 9

Originator Profits and Unmeasured Costs (OPUCs) and MBA Application Index



Sources: JPMorgan Chase; Freddie Mac; Fannie Mae; Mortgage Bankers Association (MBA); authors' calculations.

Note: The lines reflect eight-week rolling window averages.

Capacity Constraints

An often-made argument is that capacity constraints in the mortgage origination business have been particularly tight in recent years, and that these constraints become binding when the application volume increases significantly, usually due to a refinancing wave. As a result, originators do not lower rates as much as they would without these constraints, in order to curb the excess flow of applications.

Chart 9 provides some long-horizon evidence on the potential importance of capacity constraints for profits, by plotting our OPUC measure against the MBA application index (including both purchase and refinancing applications). The chart shows that the two series correlate quite strongly: Whenever the MBA application index increases, OPUCs tend to increase, and vice-versa.[44]

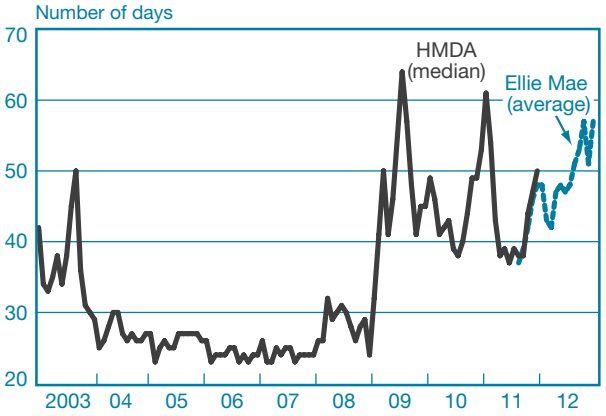
This correlation suggests that capacity constraints play an important role in generating the higher OPUCs. That said, mortgage applications (and other measures of demand and origination activity, such as MBS issuance) were at higher levels in the past, without OPUCs being as high as they were in 2012.

Chart 10 shows some more direct evidence on the potential importance of capacity constraints, by depicting the number of days it takes from the initiation of a refinancing application to the funding of the loan. The chart is based on data from the

[44] Over the period 2004-08, the relationship between the two series appears weaker than elsewhere—OPUCs appear to be on a downward trend over much of that time, even when applications increase.

Chart 10

Time from Refinancing Application to Funding (by Month in Which a Loan Is Funded)



Sources: HMDA (January 2003 to December 2011); Ellie Mae (August 2011 to December 2012).

Notes: HMDA is the Home Mortgage Disclosure Act. HMDA data are restricted to first-lien mortgages for owner-occupants of one-to-four-unit houses or condos.

Home Mortgage Disclosure Act (HMDA), which was available only through 2011 at the time of this writing, and from the Ellie Mae *Origination Insight Report*, which is only available since August 2011.[45] It shows that the median (HMDA) or average (Ellie Mae) number of days it takes for an application to be processed and funded has been substantially higher since 2009 than it was in prior years.[46] The processing time moves in response to the MBA application volume shown earlier; for instance, it reached its maximum after the refinancing wave of early 2009 and increased from less than forty days in mid-2011 to more than fifty-five days by October 2012, as refinancing accelerated over this period. However, to the extent that the HMDA and Ellie Mae data are comparable, it does not appear that it took substantially longer to originate a refinancing loan in 2012 than it did in early 2009, making it difficult to explain the full rise in OPUCs through capacity constraints.[47]

A final interesting question is how rigid capacity constraints may be. Current originators can add staff, but it takes time to train new hires. New originators can enter the market, but entry requires federal and/or state licensing and approval from Fannie Mae, Freddie Mac, and Ginnie Mae to fully participate in the origination process. To the extent that training may take longer than in the past, or that approval delays for new entrants are longer (as anecdotally reported), the speed of capacity expansion may have declined compared with earlier episodes.[48] Another potentially important factor is that the share of third-party originations (by brokers or correspondent lenders) has decreased significantly in recent years (as discussed in footnote 5). Third-party originators may, in the past, have acted as a rapid way to adjust capacity, especially during refinancing waves. In sum, while capacity constraints likely contributed to the rise in OPUCs in recent years, it is unlikely that they were the only source of this rise.

Market Concentration

A second popular explanation for the higher profits in the mortgage origination business is that the market is highly concentrated. It is well known that the mortgage market in the United States is dominated by a relatively small number of large banks that originate the majority of loans. However, as shown in Chart 11, a simple measure of market concentration given by the share of loans made by the largest five or ten originators actually decreased over the period 2011-12, as a number of the large players reduced their market share. Thus, overall market concentration alone seems unlikely to explain high profits in the mortgage business. This would make sense from a theoretical point of view: There is no particular reason why a concentrated market (but with a large number of fringe players, and price competition) should incur large profits.

Overall market concentration alone seems unlikely to explain high profits in the mortgage business.

Recent work by Scharfstein and Sunderam (2013) comes to a different conclusion. The authors argue that looking at national market concentration may mask differential trends in local market concentration, which matters if borrowers shop locally for their mortgages. Using data from 1994 to 2011, the authors find that higher concentration at the county level is

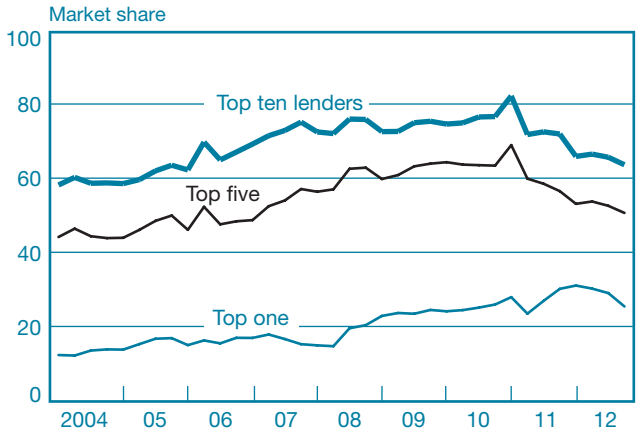
[45] See www.elliemae.com/origination-insight-reports/EMOriginationInsightReportDecember2012.pdf.

[46] The average for HMDA would be higher than the median, but would show similar patterns.

[47] It is interesting to note that the time from refinancing application to funding was significantly lower in 2003, even though application volume was much higher than it was over 2008-12. This is likely driven by tighter underwriting in the recent period compared with during the 2003 refinancing boom.

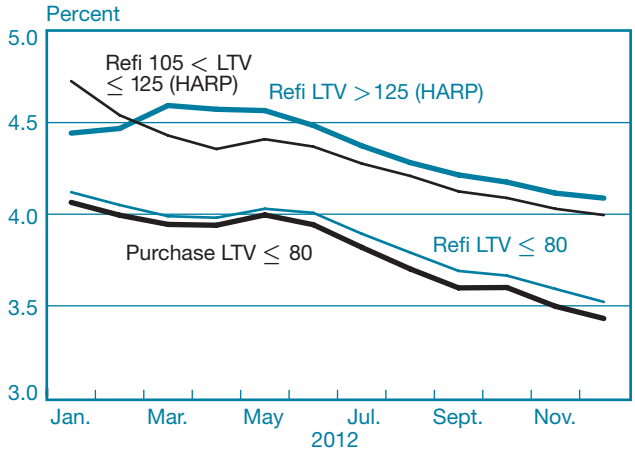
[48] Additionally, existing capacity may have been diverted to defending against putbacks instead of new loan origination.

Chart 11
Origination Market Concentration



Source: *Inside Mortgage Finance*.

Chart 12
Weighted Average Coupons of Different Loan Types



Sources: Fannie Mae; Freddie Mac; eMBS.

Note: The data include thirty-year fixed-rate mortgages with loan amounts less than or equal to $417,000, made to borrowers with a FICO score of at least 720, on owner-occupied one-unit properites.

correlated with a lower sensitivity of refinancing and mortgage rates to MBS yields. It would be interesting to extend their analysis to 2012 to see whether their findings can help explain the increase in OPUCs in that year.

We next turn to an alternative explanation for why originators could make larger profits than in the past, namely that they may enjoy more pricing power on some of their borrowers for reasons unrelated to concentration.

HARP Refinance Loans

A market segment where such pricing power may have been particularly important is the high-LTV segment, which over the past years has been dominated by refinancings through HARP, originally introduced in March 2009. The introduction of revised HARP rules in late 2011, often referred to as "HARP 2.0," led to a significant increase in HARP activity during 2012; the FHFA estimates that in the second and third quarters of 2012, HARP refinancings accounted for about 26 percent of total refinance volume.[49] HARP 2.0 provides significant incentives for same-servicer refinancing (namely, relief from representations and warranties) that are not present to the same extent for different-servicer refinancings. Furthermore, even under identical representation and warranty conditions, a new servicer may be less willing to add high-LTV borrowers to its servicing book, because such borrowers have a higher likelihood of delinquency, which makes servicing high-LTV loans more expensive. For these two reasons, many servicers do not offer HARP refinancing for loans that they are not currently servicing, or only at much worse terms. The result is that the current servicer has significant pricing power over its own high-LTV borrowers looking to refinance.

Is there evidence that lenders can exploit this higher pricing power? The observed note rates for HARP-refinanced loans are at least consistent with this idea. As shown in Chart 12, during 2012 the weighted average coupons (WACs; that is, the loan note rates) on HARP loans with LTVs above 105 tended to be 40-50 basis points higher than those of regular refinancing or purchase loans.[50] Banks earn higher revenues on these HARP loans than on regular loans for two reasons: given the higher note rate, they will typically sell these loans into a pool with a 50-basis-point higher coupon, which usually commands a price premium of around 1.5-2.0 points. Furthermore, thanks to the prepayment protection offered by these pools (as a borrower can only refinance through HARP once), investors are willing to pay a higher price (in the spec-pool market) than for TBA pools; this can add another 1-3 points (depending on the coupon) to the originator's revenue.

[49] See http://www.fhfa.gov/webfiles/24967/Nov2012RefiReport.pdf.

[50] We can also compare WACs on refinancings with LTV between 80 and 95 that are likely to be HARP loans (based on mortgage insurance information) with other loans in the same LTV range that are likely non-HARP loans. On average, the WAC on HARP loans was about 15-20 basis points higher in that range.

Are these higher revenues compensation for higher origination costs for HARP loans? This seems unlikely, as the documentation requirements for HARP loans are in fact significantly lighter than for regular loans. Thus, it is likely that origination costs are *lower*, not higher, for HARP loans relative to regular refinancings.[51]

Another possibility is that high-LTV borrowers are more cash constrained than regular refinancers and thus require higher rebates (negative points) at origination to help cover their closing costs. While this is a possibility, it is unlikely that the difference can offset a significant portion of the additional revenues, especially since closing costs are likely lower than they are for regular loans (thanks, for example, to appraisal waivers).[52]

Finally, for reasons discussed above, the value of base servicing on HARP loans may be significantly lower than that for non-HARP loans with lower LTVs. Even if we assume that the multiple on base servicing drops from 5*x* to 0*x*, however, this would only account for 1.25 points, while, as noted above, revenues are 2.5-5.0 points higher. Furthermore, because HARP borrowers are expected to prepay slowly, the cash flow stream from servicing is in fact more valuable than for regular loans, offsetting part of the higher servicing cost. Also, the expected servicing cost for current servicers declines when loans are refinanced under HARP, as borrowers are less likely to default after the note rate declines (see Tracy and Wright [2012] and Zhu [2012]).

Thus, the evidence strongly suggests that originators have been making larger profits on HARP loans than on regular loans, by being able to exploit their pricing power.

Non-HARP Mortgages

The next question is whether similar pricing power could have contributed to the rise in our OPUCs on regular (non-HARP) loans that seems not fully explained by capacity constraints, as discussed above. While lenders may have pricing power over their HARP borrowers, it is much less clear whether such pricing power may also exist for "regular" loans. Pricing power could arise, for instance, from customers' impediments (actual or perceived) to shop around, an unwillingness of other firms to compete, barriers to entry for new competitors, or a combination of these. Directly measuring originators' pricing power is not a trivial task, and we do not attempt a full analysis here. However, looking at some cross-sectional patterns may suggest some insights.

Chart 12 shows that over 2012, the WAC on non-HARP refinancing loans tended to be slightly larger than it was on purchase loans. This is somewhat surprising if one thinks that the costs of originating a refinance loan are likely lower than those of a purchase loan. In addition, comparing WACs over a longer time period (not shown), it is the case that the positive gap in WACs between purchase and refinancing loans only started emerging in 2010 (and has remained there since); over the period 2005-09, average monthly WACs on refinancing loans were mostly either equal to or below those on purchase loans.[53] However, the WAC divergence could potentially be explained by purchase borrowers paying more points than refinancers; this could be, for instance, because they expect to stay in the mortgage longer or because of tax incentives.[54]

The evidence strongly suggests that originators have been making larger profits on HARP loans than on regular loans, by being able to exploit their pricing power.

One would expect this explanation, if true, to hold across all lenders. However, looking at lender-specific differences in WACs reveals a large variation across lenders. The two panels of Chart 13 show the monthly average WAC for the sixteen largest lenders over 2012 (in terms of number of loans sold to the GSEs), for purchase and refinancing loans separately. We also plot separately the average for all other (smaller) sellers (the thicker lines). We include only thirty-year fixed-rate loans with FICO scores of 720 and higher, and LTVs of 80 or lower, made to single-unit owner-occupiers in order to reduce potential disparities due to differential LLPAs.[55]

[51] Also, the loans with FICO scores of 720 or above that we include in the chart are not subject to loan-level price adjustments under HARP.

[52] Related to this point, it is not the case that HARP note rates are higher because principal amounts are lower than for regular refinancings (as the same fixed closing cost being rolled into the rate will require a larger rate increase for lower principal amounts); controlling for loan amount in a regression basically leaves the estimated differences across loan categories unchanged.

[53] This statement is based on loan-level data from Freddie Mac only, as the Fannie Mae data only became available in 2012.

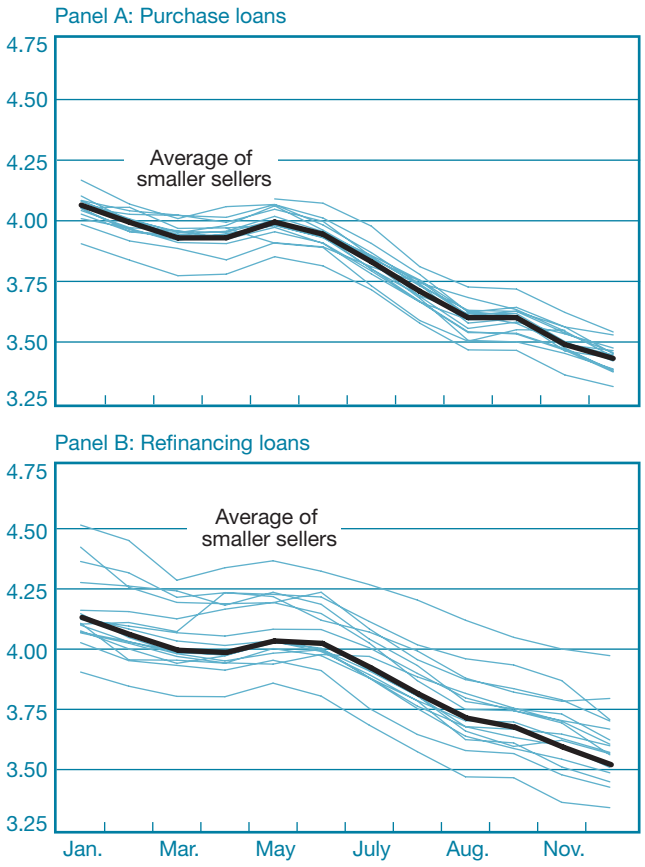
[54] Points paid in cash are fully tax deductible for purchase mortgages in the year the loan is closed. For refinancing mortgages, the deduction is instead spread evenly over the term of the mortgage (for example, thirty years), except if the loan is paid off early, in which case all unused deductions can be taken in the year the loan is paid off. See, for example, www.irs.gov/publications/p936/ar02.html#en_US_2011_publink1000229936.

[55] These calculations are based on the complete set of loan-level disclosures for pools issued in 2012 by Fannie Mae and Freddie Mac.

Chart 13

Dispersion in Weighted Average Coupons across Sellers to Fannie Mae and Freddie Mac, 2012

Percent



Sources: Fannie Mae; Freddie Mac; eMBS.

Note: The data include loans with a FICO score of 720 or higher, an LTV of 80 or lower, an amount less than or equal to $417,000, on owner-occupied single-unit properties, and only for months in which a seller made at least 100 sales.

Panel A of the chart shows that purchase WACs across sellers were quite homogeneous—with the exception of a couple of outliers, most lender WACs lie within a range of approximately 10 basis points. This is consistent with the idea that the purchase mortgage market is quite competitive, as presumably many borrowers shop around (perhaps with the help of their realtor).

Panel B reveals a much larger dispersion for refinancing loans. In particular, while a number of sellers remain concentrated around the thicker line representing the average of smaller players, eight of these large lenders sold loans with WACs that are 15 basis points or more above the thick line in at least one month, and, for six of them, that is the case for at least six out of twelve months.[56] In principle, this observed price dispersion is certainly not inconsistent with the market being competitive; however, under this null hypothesis, it is surprising that the dispersion is so much larger for refinancing loans than for purchases.

As discussed above, during 2012 the HARP program gained significant momentum for high-LTV refinances. A perhaps lesser-known fact is that there exist GSE streamline refinancing programs also for non-HARP loans (with LTV less than 80), with the same cutoff date for eligible mortgages (which must have been delivered to one of the GSEs prior to May 31, 2009). Streamlined refinancing, when done through the institution that currently services the loan, relieves the lender from representation and warranties relating to the borrower's creditworthiness and home value, while a different-servicer refinancing requires more extensive underwriting of the new loan. As a consequence, for borrowers eligible for a streamlined refinancing, there is an advantage to staying with the same servicer/lender, as doing so will reduce the documentation the borrower is required to submit. This, in turn, again creates some pricing power for the current servicer (although likely less so than for high-LTV loans). The population of loans in fixed-rate GSE pools originated prior to June 2009 is large: As of December 2012, about $1.1 trillion of loans were in such pools, relative to an overall Fannie Mae/Freddie Mac fixed-rate universe of about $3.8 trillion. During 2012, about 52 percent of all prepayments came from pools issued prior to June 2009.[57] Therefore, if lenders have pricing power over the refinancings of these loans, this could be a nontrivial contributor to OPUCs.

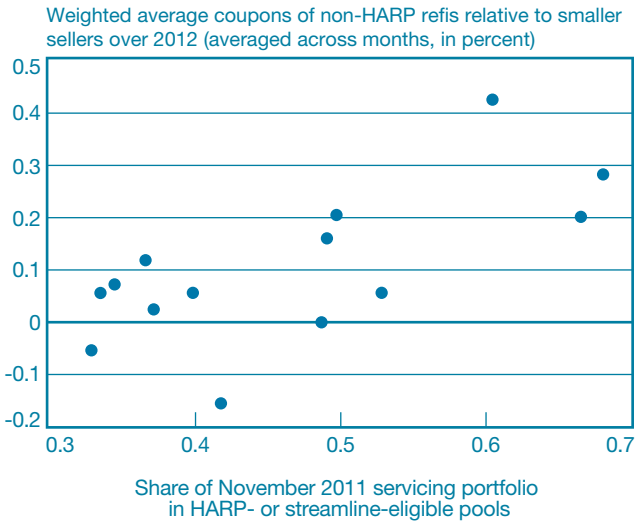
Is there evidence that such pricing power could explain the dispersion in refinancing WACs? Unfortunately, unlike for HARP loans, there is no way for us to observe in the data whether a refinancing was streamlined or not. However, we can look at variation across lenders in the fraction of their servicing portfolio that could potentially be refinanced in a streamlined manner (that is, loans in pools issued prior to June 2009) and correlate this figure with the average WAC of the lenders' non-HARP refinance loans over 2012. Chart 14 shows that there is indeed a positive correlation between the two: The lenders that had a large fraction of potentially streamline-eligible loans in their servicing

[56] With the exception of one of these six lenders, the monthly number of sales of refinancing loans always exceeds 500 loans, meaning that these averages are unlikely to be driven by small-sample noise. Additionally, as above, the result of large WAC dispersion across lenders for refinance loans remains basically unchanged if loan characteristics such as loan amount are added as explanatory variables in a regression framework.

[57] These prepayments include refinancings as well as the loan simply getting paid off (for instance, due to the borrower moving).

Chart 14

Weighted Average Coupons on Regular (Low LTV) Refinance Loans Against Fraction of Servicers' Portfolio Eligible for Streamline Refinancing



Sources: Fannie Mae; Freddie Mac; eMBS.

Notes: HARP- or streamline-eligible pools are pools issued prior to June 2009. The data include only sellers/services with servicing portfolios with more than $1 billion of HARP- or streamline-eligible pools in November 2011. Non-HARP weighted-average coupons are calculated on loans with a FICO score of 720 or higher, an LTV of 80 or lower, an amount less than or equal to $417,000, on owner-occupied single-unit properties.

portfolio at the end of 2011 tend to be those that originated refinance loans with the highest WACs on average over 2012 (that is, those that are above the thick line in panel B of Chart 13). This result is consistent with (though certainly not proof of) originators taking advantage of their pricing power over streamline-eligible borrowers.

5. Conclusions

The widening gap between primary and secondary mortgage rates over the period 2008 to 2012 was due to a rise in originators' profits and unmeasured costs, or OPUCs, as well as increases in g-fees. The magnitude of the OPUCs is influenced by MBS prices, the valuation of servicing rights, points paid by borrowers, and costs such as those from loan putbacks and pipeline hedging.

The rise in OPUCs was mainly driven by higher MBS prices, which were not offset by corresponding increases in measurable costs. Conversely, a decline in the value of mortgage servicing rights may have reduced OPUCs to some extent, and thus contributed to the widening primary-secondary spread. Among harder-to-measure costs, we find that expected putback costs and pipeline hedging likely did not cause a significant portion of the rise in OPUCs. Absent increases in other costs that we cannot measure well, such as operating costs, the rise in OPUCs reflected an increase in originator profits. While market concentration alone does not seem to explain the rise in these profits, capacity constraints do appear to have played a significant role. We also provide evidence suggesting that originators have enjoyed pricing power on some of their borrowers looking to refinance, due to borrowers' switching costs.

Going forward, it will be interesting to study the extent to which interest rate dynamics, capacity expansions, new entry, changes in regulations, and (in the longer term) housing finance reform will affect the pass-through from secondary to primary markets. As illustrated in this article, a number of factors determine this pass-through, and it will therefore be important for policymakers and market participants alike to further improve the measurement and understanding of these factors.

References

Bhattacharya, A. K., W. S. Berliner, and F. J. Fabozzi. 2008. "The Interaction of MBS Markets and Primary Mortgage Rates." Journal of Structured Finance 14, no. 3 (fall): 16-36.

Scharfstein, D. S., and A. Sunderam. 2013. "Concentration in Mortgage Lending, Refinancing Activity, and Mortgage Rates." NBER Working Paper no. 19156, June.

Tracy, J., and J. Wright. 2012. "Payment Changes and Default Risk: The Impact of Refinancing on Expected Credit Losses." Federal Reserve Bank of New York Staff Reports, no. 562, June.

Vickery, J., and J. Wright. 2013. "TBA Trading and Liquidity in the Agency MBS Market." Federal Reserve Bank of New York Economic Policy Review 19, no. 1 (May): 1-19.

Zhu, J. 2012. "Refinance and Mortgage Default: An Empirical Analysis of the HARP's Impact on Default Rates." Unpublished paper, Federal Home Loan Mortgage Corporation. Available at www.ssrn.com/abstract=2184514.