The impact of asymmetric information about collateral values in mortgage lending

Johannes Stroebel*
University of Chicago

Abstract

I empirically analyze the sources and magnitude of asymmetric information between competing lenders in residential mortgage lending. Large property developers often cooperate with vertically integrated mortgage lenders to provide financing offers to buyers of their newly constructed homes. I show that these integrated lenders have superior information about the construction quality of individual homes and exploit this information to lend against higher quality collateral. To compensate for the resulting adverse selection non-integrated lenders charge higher interest rates when competing against an integrated lender.

Keywords: Asymmetric Information, Collateral, Captive Finance, Mortgage

The impact of asymmetric information in financial markets has long been of interest to economists and the recent financial crisis has intensified research on the effects of asymmetric information in mortgage lending and securitization (e.g. Elul, 2011; Keys et al., 2010). Most of this research has focused on asymmetric information about characteristics of the borrower such as their income prospects. The expected return from making a mortgage, however, depends both on the value of the housing collateral as well as on the borrower's ability to make interest payments. Due to the illiquid and heterogeneous nature of housing as an asset it is likely that there is also asymmetric information about collateral values, in particular given the significant resources that mortgage lenders spend on appraisals and inspections to improve their valuation of the house before making a lending decision.

In this paper I empirically analyze the sources of asymmetric information between lenders that compete to originate mortgages used to purchase newly developed properties. In this market, property developers regularly provide home buyers with financing offers through vertically integrated mortgage lenders. These integrated lenders are

*University of Chicago Booth School of Business, johannes.stroebel@chicagobooth.edu; I am indebted to Caroline Hoxby, Monika Piazzesi, Martin Schneider and John Taylor for their encouragement and guidance. Seminar participants at Kellogg, Wharton, Princeton, HBS, UCLA Anderson, Chicago Booth, MIT, MIT Sloan, LSE, LBS, Michigan, Berkeley Haas, NYU Stern, Stanford, Stanford GSB and the Chicago Fed provided insightful comments. I thank Trulia and Buildfax for providing data. Financial support through SIEPR and the Hoover Institution is gratefully acknowledged.

likely to have better information than non-integrated lenders about the value of the house that is used to collateralize the mortgage. For example, an integrated lender might have access to the developer's information about aspects of construction quality that are difficult for non-integrated lenders to observe. By guiding a buyer through the home purchase process an integrated lender might also acquire relevant information about characteristics of the buyer such as their propensity to maintain the property. Perhaps surprisingly, I find that asymmetric information about collateral quality is a significant source of adverse selection in this market. In addition to testing for the presence of asymmetric information and uncovering its sources, I also quantify the impact of this asymmetric information on the cost of mortgages, which I find to be significant.

I first present a theoretical model to analyze the competition between integrated and non-integrated mortgage lenders, and use this model to generate empirical predictions that I subsequently test. In the model, an integrated lender obtains an informative signal about the quality of the housing collateral, while competing lenders only know average collateral quality. The integrated lender conditions its financing offer on its superior information and thereby subjects non-integrated lenders to adverse selection. As true house quality is revealed over time, those homes financed by an integrated lender should thus outperform ex-ante similar homes financed by non-integrated lenders. This effect is bigger when the integrated lender's signal about collateral quality is more precise. Non-integrated lenders need to charge higher interests rate to break even than if they were competing only against equally informed lenders. Interest rates rise by more for borrowers whose repayment is more sensitive to changes in collateral values, for example because they make a smaller downpayment.

The key contribution of this paper is to show empirically that such asymmetric information between competing lenders is in fact an important feature in the financing of newly developed homes, and that it generates the adverse selection predicted by the theoretical model. I construct a dataset of all housing transactions and associated mortgages in Arizona between 2000 and 2011 to track the return of properties following their initial sale. About 85% of new homes are in developments with an active integrated lender, and, when present, the average market share of these integrated lenders is about 73%. I find that in developments with an integrated lender, those houses financed by the integrated lender outperform ex-ante similar houses in the same development financed by non-integrated lenders by about 40 basis points annually. They are also over 40% less likely to enter into foreclosure.

An important finding is that the annual outperformance of the integrated lender's collateral portfolio is larger (about 100 basis points) amongst houses built on "expansive soil," a high clay content soil that makes housing returns more sensitive to unobservable aspects of construction quality such as the care with which the foundation was poured. This result suggests that the construction quality of the housing collateral is a significant source of asymmetric information. The outperformance of houses financed by the integrated lender is also bigger when the borrower makes a larger downpayment, which makes mortgage repayment less sensitive to changes in house prices. As a result, non-integrated lenders find it less necessary to adjust their interest rate offers to avoid the winner's curse and in equilibrium end up lending against lower quality collateral.

I also compare the return and foreclosure probability for the ownership duration of the *second* owner of the house. The relative outperformance of those houses *initially* financed by the integrated lender remains. This result confirms that the outperformance is to a large extent explained by asymmetric information about the housing collateral, not the borrower, since the identity of a possible second owner of the house was not known to *any* lender at the time the mortgage was granted to the initial owner. This specification also rules out that my results are driven by an initial price bundling of the mortgage and the house. Such bundling could be a concern, since any discounts on the house given to customers of the integrated lender would be observationally equivalent to a true collateral outperformance when the house gets subsequently sold. However, any such discounts would be capitalized in the transaction price between the first and second owners and should thus not contaminate the observed collateral return during the ownership of the second owner. In addition, to further test the theory, I analyze the textual description of houses in property listings when these houses are resold by their initial owners. I find that houses initially financed by integrated lenders are significantly less likely to contain descriptions of damage to the property, suggesting that that the integrated lender's outperformance can be best explained by differential depreciation rates of houses, not differential initial pricing.

I also analyze the cost to borrowers in terms of higher interest rates that result from this asymmetric information. I find that non-integrated lenders charge an average interest rate premium of 10 basis points annually for otherwise similar mortgages when competing against an integrated lender. This higher interest rate compensates non-integrated lenders for the adverse selection in the presence of an integrated lender. The interest rate increase is larger, at 23 basis points, for mortgages to purchase houses

built on expansive soil. The return of those houses is particularly sensitive to aspects of construction quality about which the integrated lender could have superior information. As predicted by the model, the interest rate increase is also larger for mortgages with a low downpayment, rising to almost 50 basis points annually for mortgages with a downpayment of less than 3%. For those mortgages the repayment probability is more sensitive to changes in collateral values. Non-integrated lenders thus need to charge higher interest rates to break even when facing adverse selection on collateral quality.

Understanding the sources of asymmetric information is important because such asymmetric information has the potential to disrupt lending markets, in particular during periods of falling house prices (see the models in Fishman and Parker, 2010, and Gorton and Ordonez, 2011). I show empirically that asymmetric information about collateral quality is a key feature of mortgage lending, and that thus such theoretical concerns might be relevant in this market. From a policy perspective, the identification of collateral values as a key source of asymmetric information in mortgage lending helps to develop proposals to improve the functioning of this market. For example, it suggests that better credit scoring technology and the more extensive sharing of borrower information will not address all forms of asymmetric information and that policies to address asymmetric information about collateral quality are also important.

This paper also provides insights into the lending behavior of financial institutions in the pre-crisis period 2000 - 2007. It has sometimes been argued that due to a lack of "skin in the game" generated by securitization, many banks no longer had incentives to distinguish between borrowers and collateral of differential quality, explaining the lower quality of mortgages originated. In contrast, the evidence presented in this paper is highly consistent with lenders actually attempting to price cross-sectional differences in collateral quality in a highly sophisticated manner.

Finally, mortgage lending in new developments provides a rich environment that helps to understand lending competition under asymmetric information in other settings with a similar information structure. One example is the practice of sell-side advisors to offer "stapled financing" packages to buyers in many M&A transactions. Since the bank providing the stapled financing offer advises on the sale of the asset, it presumably has superior information about the quality of the loan collateral. Consequently, one might expect the fear of adverse selection by other lenders to constrain the aggressiveness of their financing offers. This could lead to higher financing costs for buyers of assets that contain a stapled financing offer.

1 Related Literature

This paper relates to an empirical literature that analyzes the impact on interest rates in corporate lending when one bank has superior information about a creditor firm. One set of papers, including Petersen and Rajan (1994) and Berger and Udell (1995), tests whether loan rates increase with relationship duration between borrower and lender as the information advantage of the incumbent bank increases. A second set of papers, including Petersen and Rajan (2002) and Agarwal and Hauswald (2010), considers physical proximity between banks and firms as a source of superior information for nearby lenders. These papers find a negative relationship between distance to the bank and loan rates, consistent with a model of information asymmetries that vary in the distance between lender and borrower. In my paper the information advantage of the integrated lender arises through its relationship with the developer. In addition to analyzing the price impact of the asymmetric information, my data allows me to directly measure differences in collateral quality. Other papers consider the role of competition under asymmetric information in non-financial market settings. Hendricks and Porter (1988) analyze auctions for oil field leases. They find evidence that firms which have previously won neighboring tracts have superior information about the currently auctioned tract and that this is represented in their bidding behavior.

Different aspects of the behavior of integrated lenders have been analyzed in a number of recent papers. Pierce (2011) shows that car leasing firms that are affiliated with a manufacturer have superior information about the timing of new model introductions, which allows them to profitably adjust the lease pricing of existing models. Gartenberg (2011) analyzes whether integrated mortgage lenders lowered their lending standards during the housing boom in order to sell more homes. Consistent with my results, she finds that mortgages granted by integrated lenders were actually *less* likely to default. She concludes that this might be explained by integrated lenders' organizational choices that reduced a loan officer's incentives to approve marginal applications.[1] Agarwal et al. (2011) also document that mortgages made by integrated lenders have lower delinquency rates, and conclude that further research is needed to explain this phenomenon. I argue that integrated mortgage lenders possess and exploit superior

[1]Such incentive effects are a complementary explanation for lower default rates amongst integrated lender mortgages, but do not predict differential capital gains on housing collateral across lender types, in particular by soil quality and loan-to-value ratio. The latter are central predictions of a model with information asymmetries. In addition, I find that non-integrated lenders raise interest rates in response to the adverse selection on collateral quality, something one would not expect if the incentives of the non-integrated lenders' loan officers were to originate mortgages irrespective of collateral quality.

information about mortgage collateral quality, and consider the effect of this asymmetric information on loan pricing. Asymmetric information about property values has been considered in previous empirical work, such as Levitt and Syverson (2008) and Garmaise and Moskowitz (2004). However, most of this research focuses on the impact of asymmetric information on the sales transaction rather than the financing process, which is the focus of the present paper. Finally, my paper relates to the wide literature on testing for asymmetric information in a variety of markets such as insurance and annuities (Chiappori and Salanié, 2000; Finkelstein and Poterba, 2004).

2 Theoretical Model

In this section I present a theoretical model of the competition between differentially informed lenders to provide mortgage financing. The model adapts similar models by von Thadden (2004) and Hauswald and Marquez (2006), as well as Engelbrecht-Wiggans et al.'s (1983) analysis of first-price sealed-bid common value auctions with differentially informed bidders. I first characterize the equilibrium interest rate offers of the integrated and non-integrated lenders. I then simulate the model to generate empirical predictions about each lender's equilibrium collateral quality, and the observed interest rates charged by non-integrated lenders. These predictions are tested in the empirical analysis, which constitutes the key contribution of this paper.

Houses: Houses cost \$1 and can be either of high quality ($\theta = h$) or low quality ($\theta = l$). High quality houses will be worth $H > 1$ with certainty next period. Low quality houses will be worth $L = 0$. Final house value is observable, but house type θ is unknown ex-ante. The fraction of houses that is high quality, q, is common knowledge.

Households: Households are risk-neutral and either live in a purchased house or in rented housing, the cost of which is normalized to zero. Households have no resources and require a mortgage to purchase a house. They are indexed by γ, the probability that they will repay the mortgage when the value of their house falls (i.e. $\theta = l$). A household's γ is common knowledge. The household's expected return from borrowing at rate R is equal to $q(H - R) - (1 - q)\gamma R$, which has to be bigger than the cost of renting. $R(\gamma)_m = \frac{qH}{q+(1-q)\gamma}$ is the maximum interest rate that a household would accept.

Lenders: There are two types of risk-neutral lenders with access to funds at rate $R_f < qH$: an integrated lender that has some private information about the house and N non-integrated lenders that only know q. The private information of the integrated lender consists of a non-conveyable signal $\eta \in \{h, l\}$. The precision of the signal is defined as $\phi = P(\eta = h|\theta = h) = P(\eta = l|\theta = l) > \frac{1}{2}$.

Timing: Households apply to the integrated lender and N non-integrated lenders for a mortgage. All lenders observe γ and q. The integrated lender also observes η. Lenders compete by simultaneously offering loans at interest rate R. Lenders can also choose not to make an offer. Households accept the lowest offer as long as it is below $R(\gamma)_m$.[2]

2.1 Equilibrium

I look for a Bayesian Nash equilibrium. Since the sensitivity of repayment with respect to collateral value, γ, is perfectly observable by all agents, I can solve the equilibrium separately for each value of γ and then compare equilibrium outcomes across γ-types.[3]

Theorem 1 *There are no pure strategy equilibria.*

Proof Proof in Appendix A.

Note that if a pure strategy equilibrium existed, both lenders would have to offer the mortgage at the same interest rate $\tilde{R}$. If one lender offered credit at a rate lower than the other lender, it could increase its payoff by raising its rate by a small ε. However, both lenders offering the same $\tilde{R}$ cannot be an equilibrium. If, conditional on observing η, it is profitable to lend at $\tilde{R}$, then the integrated lender would offer $\tilde{R}-\varepsilon$ and capture the entire market. If, conditional on η, it is unprofitable to lend at $\tilde{R}$, the integrated lender would increase its interest rate offer and subject the less informed lender to a winner's curse, leaving it with an expected loss (von Thadden, 2004).

Theorem 2 *Let $W(R;\eta,\phi,\gamma)$ be the integrated lender's expected revenue from lending at rate R to a type-γ borrower to buy a house with signal η. The interest rate offer game for a type-γ borrower when signal precision is ϕ has a unique mixed strategy equilibrium, such that:*

[2]This timing assumption makes the game resemble a first-price sealed-bid auction in which non-integrated lenders are unable to observe the integrated lender's offer and use this to infer its signal. While this behavior may not represent the optimal search strategy for the consumer, who might benefit from shopping around with the integrated lender's offer, it is a reasonable representation of actual mortgage shopping behavior. Woodward and Hall (2010) find that most borrowers consider no more than two offers. The benefits from more search are so large that they conclude that it must be "confusion about how this market works that caused borrowers to shop too little." Another assumption is that borrowers themselves do not extract information from the integrated lender's offer about the quality of the house they purchase. Since interest rates vary with a large number of characteristics such signal extraction would be extremely complex and beyond the skills of most borrowers.

[3]A standard feature of these models is that the equilibrium bidding strategies of individual non-integrated lenders are indeterminate. What is determinate is the minimum of all non-integrated lenders' bids. Hence solving an equilibrium with many uninformed lenders is equivalent to solving the equilibrium of competition between the integrated lender and one representative non-integrated lender (Engelbrecht-Wiggans et al., 1983).

1. *The non-integrated lender breaks even, the integrated lender earns positive expected profits.*

2. *$\exists\bar{\gamma}$ such that for borrowers with $\gamma < \bar{\gamma}$ the integrated lender rejects all mortgage applications to buy houses when $\eta = l$. When $\eta = h$, the integrated lender randomizes interest rate offers over $[R(\gamma)_a^b, R(\gamma)_m)$ using the following cumulative distribution function:*

$$F_i(R; h, \phi, \gamma) = 1 + \frac{P_i(l)[W(R; l, \phi, \gamma) - R_f]}{P_i(h)[W(R; h, \phi, \gamma) - R_f]}.$$

$R(\gamma)_a^b = \frac{R_f}{q+\gamma(1-q)}$ is the break-even interest rate for lending to a type-γ agent to buy an average quality house. $P_i(\eta)$ is the probability of the integrated lender observing signal η. The integrated lender also makes interest rate offers with a point mass of $1 - F_i(R(\gamma)_m; h, \phi, \gamma)$ at $R(\gamma)_m$. The non-integrated lender randomizes interest rate offers over $[R(\gamma)_a^b, R(\gamma)_m)$ using the following cumulative distribution function:

$$F_n(R; \phi, \gamma) \quad = \quad 1 - \frac{W(R(\gamma)_a^b; h, \phi, \gamma) - R_f}{W(R; h, \phi, \gamma) - R_f}.$$

With probability $1 - F_n(R(\gamma)_m; \phi, \gamma)$ the non-integrated lender does not make an offer.

3. *For borrowers with $\gamma > \bar{\gamma}$ both integrated and non-integrated lenders always offer a mortgage. When $\eta = l$ the integrated lender offers the break-even interest rate $R(\gamma, \phi)_l^b$, defined implicitly by $R_f = W(R(\gamma, \phi)_l^b; l, \phi, \gamma)$. When $\eta = h$ the integrated lender randomizes its interest rate offers over $[R(\gamma)_a^b, R(\gamma)_m]$ using $F_i(R; h, \phi, \gamma)$. The non-integrated lender always randomizes over $[R(\gamma)_a^b, R(\gamma, \phi)_l^b)$ using $F_n(R; \phi, \gamma)$, with a point mass at $R(\gamma, \phi)_l^b$.*

Proof Proof in Appendix A.

2.2 Empirical Predictions from Equilibrium Bank Behavior

To analyze equilibrium outcomes when lenders use the mixed strategies of Theorem 2, I simulate the game for a range of parameter values. This generates predictions about the expected quality of the equilibrium collateral portfolio of each lender, about the equilibrium interest rates, and about how these outcomes vary with different values of γ, the probability of repayment when collateral values fall, and ϕ, the signal precision.[4]

[4]To do this, I consider $100,000$ hypothetical mortgage applicants that apply for financing from the integrated lender and the non-integrated lender. A fraction q of agents apply to buy a house of high quality. When the agent applies to the integrated lender, the lender draws an informative signal η which has known precision ϕ. Both lenders draw an interest rate offer from their equilibrium distribution as defined in Theorem 2. The borrower accepts the lowest offer. The parameters of the economic environment are chosen such that $\bar{\gamma} < 0$, which means that $\gamma > \bar{\gamma}$ and all borrowers will receive an offer. The comparative statics are the same for $0 \leq \bar{\gamma} \leq 1$ and are available on request.

The top row of Figure 1 plots the expected period-2 value of the equilibrium portfolios of houses financed by the two lenders as a function of ϕ and γ. The dashed line represents the integrated lender's portfolio, the solid line the non-integrated lender's portfolio. The dotted line shows the unconditional expected house value, qH. For all values of γ and ϕ the houses financed by the integrated lender are more likely to increase in value than those financed by the non-integrated lender. This is a direct result of the integrated lender conditioning its interest rate offers on the informative signal and the subsequent adverse selection. This implication is formalized in Prediction 1.

Prediction 1: *The average ex-post return of houses financed by integrated lenders is higher than the return of ex-ante similar (conditional on a non-integrated lender's information set) homes financed by non-integrated lenders.*

The bottom row of Figure 1 plots the average interest rate spread over R_f for the non-integrated lender's mortgages (dashed line). It also shows the spread of $R(\gamma)_a^b$, the break-even interest rate for lending against average quality collateral (solid line). When lenders are equally informed about collateral quality, Bertrand competition drives interest rates to $R(\gamma)_a^b$. When competing against a better informed integrated lender, a non-integrated lender lends against below average quality collateral and must charge a higher interest rate to continue to break even. This is formalized in Prediction 2.

Prediction 2: *Non-integrated lenders charge higher interest rates when competing against an integrated lender relative to when competing only against equally informed lenders.*

The left column shows how equilibrium outcomes vary with ϕ, the precision of the integrated lender's signal. The top left panel shows that the expected period-2 value of houses financed by the integrated lender is increasing in ϕ: as the signal becomes more precise the integrated lender is better at identifying high quality collateral. The non-integrated lender correspondingly lends against lower quality collateral. To continue to break even it needs to charge a higher interest rate on the mortgages it makes, as shown in the bottom left panel. These insights are formalized in the following predictions:

Prediction 1(a): *When the integrated lender's information about future returns is more precise (high ϕ), the average ex-post outperformance of the homes financed by the integrated lender is larger.*

Prediction 2(a): *When the integrated lender's information about future returns is more precise (high ϕ), the increase in the interest rate charged by non-integrated lenders when competing against an integrated lender is larger.*

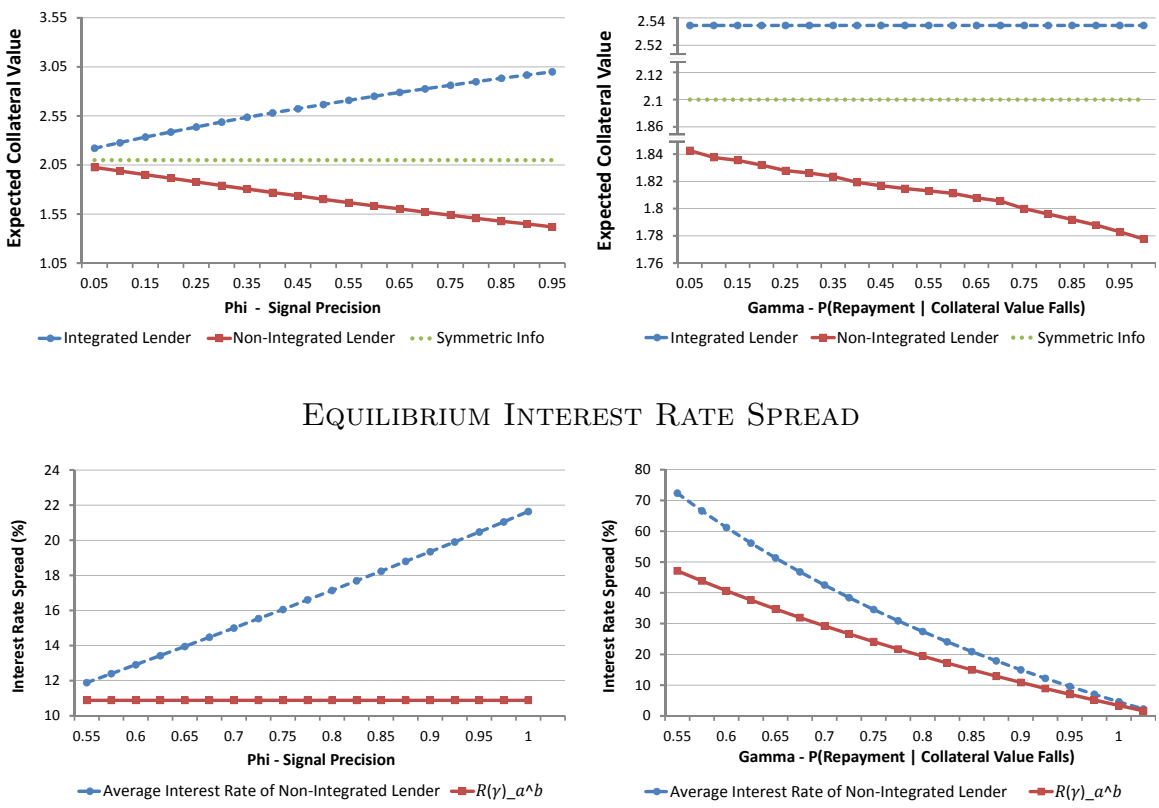


Figure 1: Equilibrium Model Outcomes

Note: The top row plots the expected period-2 price of a house in the integrated lender's equilibrium collateral portfolio (dashed line), the expected period-2 price of a house of average quality (dotted line) and the expected period-2 price of a house in the non-integrated lender's equilibrium collateral portfolio (solid line). The bottom row plots spreads of the average interest rate charged by the non-integrated lender over R_f (dashed line) and the break even rate when lending against average quality collateral, $R(\gamma)_a^b$ over R_f (solid line). In the left column ϕ varies along the horizontal axis. In the right column γ varies along the horizontal axis. If both lenders offer the same interest rate, I resolve the indifference in favor of the non-integrated lender. The model parameters are: $H = 3; q = 0.7; R_f = 1.1$. I set $\gamma = 0.7$ in the left panel and $\phi = 0.7$ in the right panel.

The right column of Figure 1 shows how equilibrium outcomes vary with γ, the sensitivity of the mortgage default probability with respect to changes in collateral values. The top right panel shows that the return of the integrated lender's collateral is unaffected by γ, since the integrated lender only lends when $\eta = h$. The return of the non-integrated lender's collateral declines as repayment becomes less sensitive to collateral values. To follow the intuition for this result it is important to realize that mortgage lenders only care about collateral values to the extent that it influences the repayment probability of the mortgage. When γ is low and the repayment probability is

highly dependent on the value of the collateral, the non-integrated lender is particularly concerned about adverse selection on collateral quality. As a result it offers mortgages at higher interest rates to avoid the winner's curse ("bid shading"), as shown in the bottom right panel. As default probabilities become less sensitive to collateral values, the break-even spread charged by the non-integrated lender declines. Since the integrated lender continues to exploit its superior information to the fullest degree, for larger values of γ the non-integrated lender's equilibrium collateral is of lower quality. Put differently, the less the non-integrated lender shades its bid, the lower the quality of its equilibrium collateral portfolio. These insights are formalized in the following empirical predictions:

Prediction 1(b): *When the mortgage default probability is more sensitive to changes in collateral values (low γ), the ex-post outperformance of houses financed by the integrated lender is smaller.*

Prediction 2(b): *When the mortgage default probability is more sensitive to changes in collateral values (low γ), the increase in the interest rate charged by non-integrated lenders when competing against an integrated lender is particularly large.*

In the following sections I empirically test these predictions of the theoretical model. I show that there is strong evidence for significant adverse selection on collateral quality in the financing of newly developed properties.

3 Possible Sources of Asymmetric Information

Before commencing the empirical analysis I consider the possible sources of superior information of the integrated lender. This information could, in principle, relate to either the quality of the housing collateral or to characteristics of the borrowers.

One component of the integrated lender's superior information about collateral quality concerns aspects of the construction quality of the house which are not observable to buyers and non-integrated lenders at the time of purchase. The Arizona Republic (2001) describes a number of shortcuts in the construction process regularly taken by builders in Arizona which can generate such differences in construction quality:

1. The foundation is often poured without allowing the ground to settle, which saves time but can lead to subsequent shifting and cracking of the foundation.
2. Stucco is often applied too thinly, which can lead to subsequent cracking.
3. Builders sometimes add excess water to the cement mix used for the foundation. This makes it easier and faster to spread, but more subject to cracking later.

The Arizona Republic (2001) discusses the lack of skilled workers to perform delicate construction tasks as a key factor in explaining initially unobservable differences in construction quality. The developer is likely to have superior information about the skill of the work crews working simultaneously on different houses in a development.[5]

A significant proportion of construction-related complaints in Arizona involve insufficient care taken when building on expansive soil. Expansive soils have a high content of clays that absorb large amounts of water into their surfaces. As the expansive soil absorbs water, it swells and exerts high pressure. Differential swelling and subsequent shrinkage of clay occurring under a property that is not properly constructed can result in excessive foundation movements and the cracking of slabs and walls. The Phoenix New Times (2006), in its analysis of construction defects in Phoenix, concludes that: *"As bad as the results [from expansive soil] can be, experts agree that they're entirely avoidable. With proper engineering and careful attention, most soils in Maricopa County could be built on without too much trouble. The problem is that some builders aren't taking the trouble."* Since the integrated lender can know whether the respective subcontractor was sufficiently skilled and experienced to conduct the more delicate procedures, one would expect that adverse selection is particularly prevalent and important amongst houses built on expansive soil. In section 5.6 I exploit differences in the return of houses built on expansive and non-expansive soil to provide evidence for Prediction 1(a), which suggested that amongst houses built on expansive soil, those financed by the integrated lender should outperform particularly.[6]

In addition to superior information about collateral quality, it is possible that in the process of guiding the borrowers through the house purchase process, the integrated lender also obtains superior information about borrower characteristics, some of which might affect the return of the housing collateral. For example, the developer might

[5]It is hard to empirically determine the precise channel through which such information is obtained by the integrated lender. However, given the significant resources spent by lenders on property appraisers and inspectors to acquire information about the quality of a house prior to a lending decision, it seems natural to expect them to acquire additional relevant information from within their own organization. This is particularly likely for developers that co-locate regional sales teams and the integrated lender's loan officers, who often work on-site (and in adjacent offices) at each subdivision (Gartenberg, 2011). Loan officers usually have some discretion in adjusting mortgages rates from rate sheets, by charging overages or underages, which would allow them to adjust pricing based on proprietary information about collateral quality. Black et al. (2003) describe this process in detail.

[6]Problems related to construction quality are not limited to Arizona. A survey by Criterium Engineers (2003) found that of all new homes in the U.S., 21% had problems with roof installations, 15% had problems with the installation of sidings, such as stucco, 23% had problems with the installation of windows and doors and 14% had problems with the construction of the foundation.

learn about the buyer's propensity to maintain the property. However, in the empirical analysis I show that the outperformance of the collateral portfolio of the integrated lender can be best explained by the integrated lender's superior information about initial collateral quality, not borrower characteristics.

4 Data Description

To conduct the empirical analysis, I combine three main datasets. The first dataset contains the universe of ownership-changing deeds in Arizona between 2000 and 2011. The property to which the deeds relate is uniquely identified via the Assessor Parcel Number (APN). The variables in this dataset include property address, contract date, transaction price, type of deed (e.g. Intra-Family Transfer Deed, Warranty Deed, Foreclosure Deed), the type of property (e.g. Apartment, Single-Family Residence) and the name and a classification of buyer and seller (e.g. Husband and Wife, Company). It also reports the amount and the duration of the mortgage and the identity of the mortgage lender. For mortgages with a variable interest rate I also observe the initial rate. The second dataset contains the universe of tax-assessment records for the year 2010. Properties are again identified via their APN. This dataset includes information on property characteristics such as construction year, owner-occupancy status, lot size, building size, and the number of bedrooms and bathrooms. The tax assessment records also include an estimate of the market value of the property for January 2009. The third dataset contains information from the Home Mortgage Disclosure Act's (HMDA) Loan Application Registry, which provides details on every mortgage application in major Metropolitan Statistical Areas. It includes information on the census tract of the house, lender identity, loan amount, property type, and the applicant's income, sex and race. It also records whether the mortgage was sold or securitized within the same calendar year. I merge this dataset to the deeds data as described in Appendix B.2.

I focus on the state of Arizona, which was at the center of the recent boom-bust cycle, and which is an interesting focus of study due to data quality and availability.[7] Since most of the information is originally recorded at the county-level and field

[7]There are a number of reasons to think that my results for Arizona are relevant for understanding mortgage lending in the rest of the U.S. First, most of the large property developers and integrated lenders operate nationally, so I would expect to observe similar behavior by the same actors in other states. Second, as discussed in section 3, problems with construction quality are relevant in developments throughout the U.S. On the other hand, Arizona experienced one of the larger boom-bust cycles in construction during the sample period. This means that my findings might be particularly

population varies widely, not all specifications could be tested on a larger geographic area.[8] Appendix B describes the process of cleaning and merging the data, as well as the identification of integrated lenders. The resulting dataset contains information on 102,818 single-family residences that were sold by developers in 2000 - 2007 and which I can match to assessment records and HMDA data. Summary statistics for the key variables are provided in Appendix B.4.

5 Outperformance of Integrated Lender Collateral

In this section I test for the presence of superior information about collateral quality by the integrated lender. The empirical approach compares the ex-post return of homes financed by an integrated lender to the return of ex-ante similar homes in the same development financed by non-integrated lenders.[9] I measure this return over four different time horizons described below. Each of these time horizons allows me to address a different possible contaminating factor and jointly they provide strong evidence for the presence of asymmetric information about collateral quality by integrated lenders.

5.1 Measuring Collateral Return Using Repeat Sales

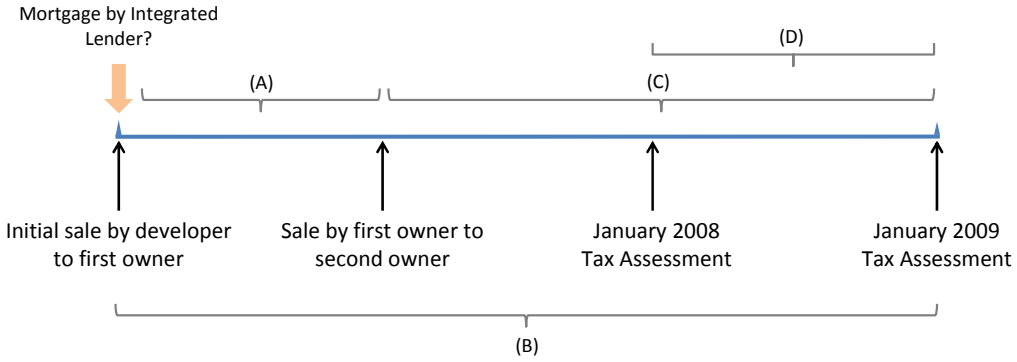
For the first time period I focus on the subset of homes in developments with an integrated lender for which I observe a second armlength transaction subsequent to the initial sale by the developer to the first owner. For each such property I calculate the annualized return between the two sales. This corresponds to period (A) in Figure 2. I then regress this return on observable control variables and a dummy variable IL_i that captures whether the mortgage was granted by the integrated lender. This regression is shown in equation (1). The unit of observation is a house i, first sold in quarter q_1 and resold in quarter q_2. A set of fixed effects for each pair of sales quarters (e.g. first sale in Q1 2000, second sale in Q3 2008) is included as δ_{q_1,q_2}. This controls for general market movements in house prices over time. The vector X_i includes observable

relevant for other states with significant construction booms, such as California and Florida.

[8] For example, a significant number of non-disclosure states do not report transaction prices. Other states, such as Georgia, do not allow me to identify sales by developers. The data from other states such as Maryland does not provide the identity of the mortgage lender. The changes following Proposition 13 in California mean that assessed property values cannot be interpreted to reflect true market values.

[9] Homes in the same development are often very similar to one another, due to developers' common practice of offering a choice from a number of model homes, the interior of which (e.g. the kitchen) is subsequently customized. In its 2004 10-K statement, the homebuilder KB Home describes a development to "typically include two to four different model home design."

Figure 2: Measures of Ex-Post Price Performance



characteristics of the house, the owner and the mortgage. In Table 1 these are added sequentially to the regression. Standard errors are clustered at the developer level.[10]

$$Return_i = \alpha + \kappa IL_i + X_i\beta + \delta_{q_1,q_2} + \epsilon_i \tag{1}$$

Column (1) of Table 1 estimates equation (1) with only county fixed effects included as additional control variables in X_i. The magnitude of κ, the coefficient on IL_i, suggests that houses financed by the integrated lender outperform houses financed by non-integrated lenders by about 40 basis points (0.4 percentage points) annually. This is significant relative to the average annual return during this period of 7.4%.

One concern is that the outperformance detected in column (1) could be the result of a spurious correlation on characteristics of the house or the owner that make it more likely that the mortgage was granted by the integrated lender and that the house increased in value. For example, it could be that certain owners take better care of the house and are more likely to borrow from the integrated lender. To address such concerns, between columns (1) and (5) I add an increasing number of control variables to the vector X_i.[11] Column (2) includes property characteristics such as initial sales price and building size. This captures that houses in different market segments had a different return over the sample period. Column (3) controls for owner characteristics such as income, and mortgage financing characteristics such as the loan-to-value ratio,

[10] Clustering standard errors at the developer level (109 clusters) addresses possible concerns about the correlation of regression residuals across houses built by the same developer by allowing for an arbitrary correlation of residuals of houses built by the same developer.

[11] The precise functional form of the controls is described in Appendix B.4. I do not discuss the coefficients on the control variables, since they are not the focus of my analysis. In a previous version of the paper I showed that, where applicable, these coefficients are consistent with the existing literature.

Table 1: Annualized Collateral Return (%) - Between Repeat Sales - Period (A)

						Forced Moves	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Integrated Lender	0.419***	0.434***	0.506***	0.441***	0.403***	0.418**	0.376*
	(0.155)	(0.150)	(0.135)	(0.119)	(0.113)	(0.170)	(0.198)
Controls (see note)		H	H, B, F	H, B, F T, D1	H, B, F, T, D2		H, B, F, T, D1
R-squared	0.869	0.876	0.878	0.887	0.896	0.885	0.903
$\bar{y}$	7.438	7.438	7.438	7.438	7.438	5.437	5.437
N	30,343	30,343	30,343	30,343	30,343	3,287	3,287

Note: This table shows results from regression (1). The dependent variable is the annualized return of houses between two armslength transactions. I include single-family residences first sold by a developer in 2000 - 2007 in developments with an integrated lender. All specifications include include sales quarter-pair fixed effects and county fixed effects. House characteristics (H) include real initial sales price, building size, lot size, price per square foot, number of garage spaces, average size of bedrooms and bathrooms, whether the house has a pool and whether it is a rental unit. Buyer characteristics (B) include real income, whether the property was purchased by an individual or a couple and whether the owners are Asian or Latino. Financing characteristics (F) include mortgage type, loan-to-value ratio, loan-to-income ratio and mortgage duration. Census tract demographics (T) include median household income and the percentage of adults over 25 with a high school diploma. D1 includes developer fixed effects, D2 includes development fixed effects. Standard errors are clustered at the developer level. Significance Levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

which affect the borrower's ability and incentives to maintain the property. Column (4) includes census tract demographics, such as the median income, as well as developer fixed effects. This prevents the results from being driven by a positive correlation between developer quality and the associated integrated lender's aggressiveness. The coefficient of κ is remarkably stable with respect to the addition of these controls and fixed effects. Similar to Altonji et al. (2005) and others I argue that this reduces the likelihood that the results are driven by selection on unobservable characteristics.[12]

Column (5) adds development fixed effects. Homes in the same development are very similar in terms of school quality, crime and local amenities. Including development fixed effects thus removes further possible biases due to unobservables that might affect housing returns and the propensity of borrowers to select an integrated lender. κ is essentially unchanged by this addition.[13] This result suggests that the

[12]In addition, to the degree that one might expect selection on unobservable buyer characteristics, the more plausible stories suggest that my empirical approach underestimates the true effect of asymmetric information. For example, it might be that less sophisticated agents are more likely to engage in suboptimal mortgage shopping and just accept the integrated lender's offer. If these agents were also less likely to maintain the house, κ would not capture the full extent of asymmetric information.

[13]One might be concerned that the market development of house prices differed significantly by

majority of the outperformance of the integrated lender can be attributed to superior information about characteristics that vary at the property level, such as construction quality. Superior information about characteristics that vary at the development level, such as developers' plans for future nearby developments, does not appear to contribute significantly to the outperformance of the integrated lender's collateral portfolio.

In another test I exclude the dummy IL_i from regression (1) and analyze the residuals ϵ_i by lender type, as plotted in the left panel of Figure 3. This graph shows that the 40 basis points mean return difference is driven by a thicker left return tail for houses financed by non-integrated lenders. This is consistent with a story of asymmetric information about collateral values, where houses financed by non-integrated lenders experience significant structural problems at an above-average frequency.

I also consider the time horizon over which the asymmetric information is revealed. Regression specification (1) implicitly assumes that there is a constant probability of revelation of the integrated lender's initially private information at the house level, which would translate into a constant annualized outperformance at the collateral portfolio level. However, it is also possible that a larger part of the asymmetric information is revealed in the first few years after the property is built. To test this, I re-run regression (1), but instead of including a simple dummy variable for IL_i, I interact this dummy with the number of years between the two sales, as given by regression (2).

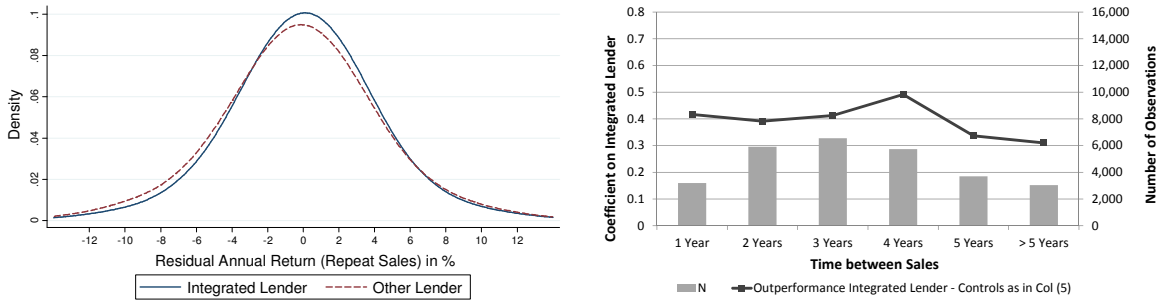
$$Return_i = \alpha + \sum_{j=1}^{6} \kappa_j \times IL_i \times TimeBetweenSales_{i,j} + X_i\beta + \delta_{q_1,q_2} + \epsilon_i \qquad (2)$$

Figure 3 presents the coefficients for a specification that includes the same controls as column (5) of Table 1. There is no clear pattern in the development of the annualized outperformance with respect to the time difference between the two sales. The F-statistics for a Wald test of the equality of all coefficients test is 0.21 (p-value of 0.96). Hence I cannot reject the null hypothesis that the asymmetric information that generates the outperformance of the integrated lender is revealed at a constant rate.[14]

geography in a way that could bias κ. To show that this is not the case, rather than controlling for simple δ_{q_1,q_2} fixed effects, I interact these fixed effects with an ever tighter set of geographic identifiers. When controlling for $\delta_{q_1,q_2} \times county$ fixed effects and including the same covariates as column (5) of Table 1, I estimate κ to be 0.408 (clustered SE = 0.111). When controlling for $\delta_{q_1,q_2} \times city$ fixed effects, κ is equal to 0.393 (clustered SE = 0.137). When controlling for $\delta_{q_1,q_2} \times ZIP$ fixed effects, κ is 0.301 (clustered SE = 0.172).

[14]The longest time between sales in this sample is 11 years. The rate of revelation of the asymmetric information might decline over time for longer time horizons. However, for the current sample the data suggest that regression (1) is correctly specified with respect to the timing of the outperformance.

Figure 3: Density and Timing of Information Release



Note: The left panel plots the density of ϵ_i for regression (1) without the IL_i dummy by lender type. The right panel plots κ_j for regression (2) with the same controls as column (5) of Table 1. The line graph shows the coefficients on the left axis, the bar chart the number of observations in each group.

5.2 Selection into observing repeat sales

One might be worried that the subsample of houses for which I observe a resale is not representative of all newly developed homes, and that such a selection might be correlated with the ϵ in regression (1) in a way that might bias the estimate of κ, the outperformance of the integrated lender's collateral portfolio. To address this concern, columns (6) and (7) of Table 1 restrict the sample to sales pairs where the second sale is precipitated by a plausibly exogenous event, such as the death or divorce of the owners (Appendix B describes how I identify such events). "Forced moves" are identified when the resale is preceeded by a death or divorce of the initial owners. This includes about 10% of all sales pairs, most of them because of a divorce of the initial owners. When measured in the subsample of forced moves, the integrated lender's collateral portfolio outperforms by about the same amount as it does in the full sample.

In addition, in a second analysis of the return of the housing collateral I consider the implied annualized return between the initial sale of a house and its estimated market value in January 2009 (instead of the transaction price at a subsequent sale) as recorded in the tax assessment records. Such an estimated market value is available for *all* houses in the dataset.[15] I run regression (3) for all houses i that were initially sold in month m (e.g. May 2004) in a development with an integrated lender.

$$Return_i = \alpha + \kappa IL_i + X_i\beta + \delta_m + \epsilon_i \tag{3}$$

[15] Appendix B.5 describes the assessment process in Arizona and analyzes assessment accuracy, which I find to be high. One key mechanism through which assessors learn about differential property conditions is through an elaborate complaints process that allows homeowners to appeal their tax assessments if they feel their house is worth less than the assessed value.

Table 2: Annualized Collateral Return (%) - Initial Sale to Assessment - Period (B)

	(1)	(2)	(3)	(4)	(5)	(6)
Integrated Lender	0.392***	0.438***	0.472***	0.393***	0.311***	0.185***
	(0.109)	(0.103)	(0.0951)	(0.101)	(0.086)	(0.048)
Controls (see note)		H	H, B, F	H, B, F, T	H, B, F, T, D1	H, B, F, T, D2
R-squared	0.812	0.829	0.832	0.881	0.892	0.936
$\bar{y}$	-6.349	-6.349	-6.349	-6.349	-6.349	-6.349
N	87,482	87,481	87,481	87,481	87,481	87,481

Note: This table shows results from regression (3). The dependent variable is the annualized return of houses between initial sale and January 2009 assessed market value. I include single-family residences first sold by a developer in 2000 - 2007 in developments with an integrated lender. All specifications include month of sale fixed effects and county fixed effects. Controls for house characteristics (H), buyer characteristics (B), financing characteristics (F) and census tract demographics (T) are defined as in Table 1. D1 includes developer fixed effects, D2 includes development fixed effects. Standard errors are clustered at the developer level. Significance Levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

Table 2 presents the results from regression (3). As before, the outperformance of the integrated lender's collateral portfolio as measured by the coefficient κ on the integrated lender dummy IL_i is not significantly affected by the addition of control variables between columns (1) and (5). Reassuringly, the magnitude of κ using this measure of return is also very similar to the magnitude obtained using the return between repeat sales: houses financed by integrated lenders outperform ex-ante similar houses financed by non-integrated lenders by about 40 basis points annually. Unlike before, however, the inclusion of development fixed effects reduces the measured outperformance of the integrated lender's collateral portfolio. This is explained by the "comparables" methodology used by assessors to calculate the assessed values. This method uses transaction prices of recently-sold similar homes, often from within the same development, to predict the market value of homes that have not recently transacted. This means that the sale of a high quality home financed by the integrated lender will also impact the assessed values of lower quality homes in the same development financed by non-integrated lenders. Without development fixed effects, additional identification comes from the fact that developments with a higher percentage of mortgages financed by the integrated lender should outperform other developments. The fact that κ remains positive after controlling for development fixed effects suggests that the assessor has some success at detecting differential house quality within a development, most likely driven by the use of the appeals process described in appendix B.5.

5.3 Source of Asymmetric Information

In section 3 I argued that the integrated lender might have superior information about characteristics of both the housing collateral and the borrower. Both of these sources of superior information could, in principle, explain why houses financed by an integrated lender outperform those financed by non-integrated lenders. For example, the integrated lender could know more about the construction quality of the property as well as about the propensity of buyers to maintain the home. In this section I provide evidence that asymmetric information about initial collateral quality is likely to be the key driver of the outperformance of the integratd lender's collateral portfolio. To do this I consider the relative return of houses *initially* financed by an integrated lender over the ownership period of the *second* owner of the house. Since the identity of this second owner was unknown to all lenders at the time of granting the initial mortgage, any outperformance of the integrated lender's collateral portfolio over this period can be attributed to superior information about collateral quality.

I first focus on the subset of houses for which I observe at least two armslength sales and calculate the annualized return of these properties between the second sale and the assessed market value in January 2009, as given by period (C) in Figure 2. Columns (1) to (5) in Table 3 show the results of a regression of this return on the identity of the initial mortgage lender. As before, I control for characteristics of the house, the owner and the mortgage, as well as the month of the resale.[16] In columns (4) and (5) I restrict the sample to those houses for which the second sale was prompted by a divorce or death of the initial owners. As in section 5.2, this addresses possible concerns about a non-random selection into observing repeat sales. The magnitude of κ over this time horizon is similar to the magnitudes reported in Tables 1 and 2, with the familiar decline upon the addition of development fixed effects.

In a second specification I focus on those homes for which I observe at least three armslength transactions, and consider the return of the collateral between the second and the third sale. While this specification has the advantage that it only relies on market transactions rather than assessed values, the decline in sales activity in Arizona

[16]The set of characteristics of the second owner and characteristics of the second mortgage is smaller than that of the first owner and mortgage in sections 5.1 and 5.2. This is because a significant number of second purchasers did not use a mortgage. This makes it impossible to retrieve income information from the HDMA data. However, since the integrated lender only interacted with the initial owner, a spurious correlation along characteristics of the second owner seems implausible. I also continue to control for all characteristics of the first owner.

since the crash of 2007/8 means that the number of houses for which I observe three sales is limited. The results are presented in columns (5) - (8) of Table 3. The annual outperformance of houses initially financed by an integrated lender remains at around 40 basis points during the ownership of the second owner. These estimates suggest that the observed outperformance of the integrated lender's collateral portfolio is driven by asymmetric information about the value of the collateral and not the borrower.

Table 3: Annualized Collateral Return (%) - Second Ownerhip Period

	1ST Resale to Assessment					1ST Resale to 2ND Resale		
	All Moves			Forced Moves				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Integrated Lender	0.374***	0.308***	0.170*	0.597**	0.464*	0.598**	0.522*	0.336
	(0.115)	(0.092)	(0.089)	(0.238)	(0.251)	(0.277)	(0.296)	(0.233)
Controls (See Note)	✓	✓	✓	✓	✓	✓	✓	✓
Other Fixed Effects	·	D1	D2	·	D1	·	D1	D2
R-squared	0.893	0.901	0.947	0.876	0.889	0.885	0.886	0.891
$\bar{y}$	-10.85	-10.85	-10.85	-12.53	-12.53	3.32	3.32	3.32
N	18,285	18,285	18,285	1,653	1,653	5,379	5,379	5,379

Note: This table shows results from a regression of the annualized return of houses during the ownership period of the second owner, for single-family residences initially sold by a developer in 2000 - 2007 in developments with an integrated lender. "Integrated Lender" is equal to 1 if the mortgage to the first owner was made by an integrated lender. Columns (1) - (5) measure return from the second sale to the assessed value in January 2009, and restricts to houses resold prior to 2008. Columns (4) and (5) restrict the sample to forced moves as in Table 1. Column (6) - (8) measure return between the second and the third sale. Controls variables include fixed effects for month of resale for columns (1) - (5) and sales-quarter pair for columns (6) - (8). All specifications include fixed effects for construction quarter and county as well as characteristics of the house (H) and census tract (T) as in Table 1 and characteristics of the buyer (B2) and financing (F2). Buyer characteristics include whether the second buyer was an individual or a couple and whether the second buyers are Asian or Latino, in addition to characteristics of the first buyer (B) as in Table 1. Financing characteristics include the loan-to-value ratio of the second mortgage in addition to details of the first buyer's mortgage (F) as in Table 1. D1 includes developer fixed effects, D2 includes development fixed effects. Standard errors are clustered at the developer level. Significance Levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

5.4 Bundling of Home and Mortgage

One concern when measuring collateral return starting from the initial sale by the developer is that price bundling of the house and the mortgage by the developer could contaminate these measures of return. If such bundling involved discounts on the house price given to customers of the integrated lender, it would be observationally equivalent to true collateral outperformance when the house is subsequently sold for its actual value. The results presented in the right panel of Figure 3 partially address

this concern: if the outperformance was indeed driven by an initial price discount that gets subsequently capitalized, the *annualized* outperformance should decline with the time between sales. In addition, the results in section 5.3 show that the outperformance persists over the ownership period of the second owner. Since any initial price discounts for customers of the integrated lender would be capitalized in the sales price between the first and the second owner, it should not contaminate these returns.

A third approach to rule out that the observed outperformance of the integrated lender's collateral is driven by an initial price bundling rather than by the integrated lender's superior information about collateral quality is to look directly for evidence of differential depreciation, rather than to rely on prices to capture this depreciation. To do this I scanned the textual descriptions of all property listings on the online real-estate listings platform Trulia.com between October 2005 and August 2010 for evidence of damage to the property. I identified three categories of evidence for property damage. The first category includes all listings that propose an "as is" sale in which the buyer accepts the house "with all faults," whether or not immediately apparent. The second category includes homes with a description that includes at least one of the phrases "repair," "damage," "broken," "leak," "peeling," "crack," "needs work," "fix-up" and "TLC." The third category includes listings that suggest the home is particularly suited for a special buyer such as a "handyman," "right buyer," and an "investor."

Table 4: Evidence for Property Damage

	"As Is"		Damage Indicator		Special Buyer	
	(1)	(2)	(3)	(4)	(5)	(6)
Integrated Lender	-0.028***	-0.026***	-0.010**	-0.011*	-0.012***	-0.016***
	(0.007)	(0.008)	(0.005)	(0.006)	(0.005)	(0.006)
Controls (See note)	D1	D2	D1	D2	D1	D2
$\bar{y}$	0.138	0.143	0.055	0.063	0.046	0.055
N	11,287	10,732	10,896	9,370	10,746	8,799

Note: This table shows the average marginal effects from a probit regression explaining damage indicators in property listings. I analyze three indicators of property damage: whether the the house is sold "as is" (columns 1 and 2), whether there were details of damage in the description (columns 3 and 4), and whether the property was said to be attractive for "special buyers" (columns 5 and 6). I include single-family residences first sold by a developer in 2000 - 2007 in developments with an integrated lender and which were listed for resale on Trulia.com between October 2005 and August 2010 (but at least one year after the first sale), and which include a textual description. All specifications include month of sale fixed effects, county fixed effects and control for the time between the initial sale and the listing. They also include house characteristics (H), buyer characteristics (B), financing characteristics (F) and census tract demographics (T) as in Table 1. D1 includes developer fixed effects, D2 includes development fixed effects. Standard errors are clustered at the developer level. Significance Levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

The results in Table 4 suggest that amongst all houses listed on Trulia.com, those initially financed by the integrated lender are 2.8 percentage points less likely to propose an "as is" sale, relative to a baseline probability of around 14%. They are also about 1 percentage point less likely to include words that indicate damage to the property (baseline of 6%), and about 1 percentage point less likely to suggest the property is particularly attractive for a special buyer (baseline of 5%). This provides direct evidence that the higher returns of houses financed by the integrated lender are indeed driven by superior collateral quality and not by a bundeling of the home and mortgage.

5.5 Differential Investment in the House

Another possible explanation for the superior return of houses financed by the integrated lender is that the differences in performance are driven by differential investment into the house by owners borrowing from different lenders. To show that this explanation is unlikely to explain the observed outperformance, I now directly test for the presence of differential investment by the homeowners. For homes from Maricopa, Pima and Yuma counties I observe assessment records for 2008 in addition to 2009. For each house in these counties I calculate the implied return between the assessed market values in 2008 and 2009, given by period (D) in Figure 2. I then regress this return on an integrated lender dummy and a set of control variables, as shown in equation (4).

$$Return\_2008_i = \alpha + \kappa IL_i + X_i\beta + \delta_q + \epsilon_i \tag{4}$$

Columns (1) and (2) of Table 5 show the results from this regression. As before, houses financed by an integrated lender outperformed ex-ante similar homes financed by non-integrated lenders by between 40 and 70 basis points (average market price movements over this period were very significant). As in regression (3), which also relies on assessed values, the inclusion of development fixed effects reduces the measured outperformance.

I next test whether this differential return can be explained by differential investment in the house over this period. My first measure of investment is an indicator of whether or not the reported building area of the house changed between the 2008 and 2009 assessor reports, which would provide evidence for construction activity. About 0.4% of properties experienced a change in building area. Second, using data obtained from Buildfax on the universe of all building permits filed in Phoenix, I check for each property whether a permit was filed during 2008, which is the case for about 1% of properties. Both of these variables provide evidence for investment activity at the

Table 5: Annualized Collateral Return (%) - 2008 to 2009 Assessment - Period (D)

	Housing Return 2008			Δ Building Area	Has Permit
	(1)	(2)	(3)	(4)	(5)
Integrated Lender	0.717***	0.373***	0.364***	-0.0004	-0.0020
	(0.215)	(0.131)	(0.126)	(0.0010)	(0.0026)
Δ Building Area			0.960		
			(0.815)		
Controls (See Note)	D1	D2	D2	D2	D2
R-squared	0.606	0.795	0.800	0.0482	0.110
$\bar{y}$	-27.45	-27.45	-27.47	0.0039	0.0097
N	69,834	69,834	69,761	69,811	7,616

Note: The dependent variable in columns (1) - (3) is the return between the assessed market values of January 2008 and 2009 (regression 4), in column (4) a dummy variable whether building size changed in 2008, and in column (5) a dummy variable whether a building permit was filed in 2008. I include single-family residences first sold by a developer in 2000 - 2007 in developments with an integrated lender. All specifications control for the assessed house value in January 2008 in addition to other characteristics of the house (H), the buyer (B), the financing (F) and the census tract (T) as in Table 1. D1 includes developer fixed effects, D2 development fixed effects. Standard errors are clustered at the developer level. Significance Levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

property level. The results in column (3) suggest that, in fact, houses that experienced a change in the building area did have larger returns, though the results are not statistically significant. Columns (4) and (5) test whether houses that were initially financed by an integrated lender were more likely to experience an investment event in 2008. I regress a dummy variable capturing the change in building area or the filing of a permit on whether the house was financed by an integrated lender as well as control variables. There was no statistically significant difference in the probability of investment for houses financed by integrated and non-integrated lenders, and the point estimates even suggest a lower investment activity for houses financed by the integrated lender. This suggests the outperformance of the integrated lender's collateral portfolio was not driven by differential investment by the owners.

5.6 Importance of Asymmetric Information: Soil Quality

Prediction 1(a) stated that the outperformance of the integrated lender's collateral portfolio should be higher when its information about collateral quality is more important in determining housing returns. I test the prediction by exploiting exogenous differences in the type of soil on which houses are built. Section 3 explained that the return of houses built on expansive soil is particularly sensitive to unobservable aspects of construction quality. I use detailed data on the geographic distribution of soil from

the USDA's Soil Survey database to determine which houses are built on expansive soil.[17] Soil expansiveness has significant geographic variation, often within developments. This is seen in Figure 7 in Appendix B.6, which shows the soil distribution in a representative Phoenix development. In regression (5) I include a dummy for "expansive soil" (ES_i) as well as its interaction with the integrated lender dummy.

$$Return_i = \alpha + \kappa_1 IL_i + \kappa_2 ES_i + \kappa_3 IL_i \cdot ES_i + X_i\beta + \delta_{q_1,q_2} + \epsilon_i \quad (5)$$

The results from regression (5), measuring return over periods (A) - (D), are shown in Table 6. The coefficient on ES_i is negative and usually statistically significant, indicating a lower average return for houses built on expansive soil. More importantly, the positive and significant coefficient on the interaction between IL_i and ES_i shows that for houses built on expansive soil, the integrated lender's collateral portfolio outperforms an ex-ante similar portfolio of houses financed by a non-integrated lender by almost one percentage point annually ($\kappa_1 + \kappa_3$). This is further evidence that a significant part of the asymmetric information is likely to relate to the initial construction quality of the housing collateral. κ_1 also remains positive. This result suggests that some of the outperformance of the integrated lender relates to information about characteristics that also affect the return of houses not built on expansive soil (for example, information about the quality of the electric wiring).

For each period I also include a specification with development fixed effects that exploits within-development variation in soil type. The positive coefficient on the interaction remains. The lower statistical significance of the interaction term might result from possible measurement and classification errors that arise from the soil type being inherently continuous, while my measure of soil expansiveness is discrete.[18] The gradual change in soil type implies that soil in different hydrologic groups near classification boundaries will be rather similar. When including development fixed effects, the identification relies more strongly on differences in return for houses closer to soil boundaries and so the actual differences in soil expansiveness are smaller.

[17] The data identify four hydrologic soil groups, which are characterized by their intake of water under conditions of maximum yearly wetness and the maximum swelling of expansive clays. I assign the 10% of houses built on soil in hydrologic group D (more than 40% clay, high shrink-swell potential) to the "expansive soil" category. Expanding the "expansive soil" category to include hydrologic group C (20% - 40% clay), adds another 10% of the observations and does not change the empirical results.

[18] The data documentation states that "the locational accuracy of soil delineations on the ground varies [...]. For example, on long gently sloping landscapes the transition occurs gradually over many feet. Where landscapes change abruptly, the transition will be narrow."

Table 6: Annualized Collateral Return (%) by Soil Type

	Period (A)		Period (B)		Period (C)		Period (D)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Integrated Lender	0.385***	0.371***	0.251***	0.157***	0.204**	0.143	0.508***	0.248**
	(0.122)	(0.108)	(0.086)	(0.048)	(0.092)	(0.088)	(0.174)	(0.108)
Expansive Soil	-0.235	-0.762**	-1.032*	-0.885***	-1.881***	-0.839***	-3.178***	-1.551**
	(0.517)	(0.338)	(0.602)	(0.188)	(0.437)	(0.299)	(1.194)	(0.724)
Integrated Lender × Expansive Soil	0.562**	0.322	0.574***	0.294**	0.785***	0.062	1.159**	0.760***
	(0.267)	(0.226)	(0.204)	(0.125)	(0.244)	(0.148)	(0.445)	(0.264)
Controls	Table 1 Col (4)	Table 1 Col (5)	Table 2 Col (5)	Table 2 Col (6)	Table 3 Col (2)	Table 3 Col (3)	Table 5 Col (1)	Table 5 Col (2)
R-squared	0.887	0.896	0.892	0.936	0.901	0.947	0.623	0.802
$\bar{y}$	7.438	7.438	-6.349	-6.349	-10.85	-10.85	-27.45	-27.45
N	30,343	30,343	87,481	87,481	18,285	18,285	69,834	69,834

Note: This table shows results from regression (5). ES_i is equal to one for houses built on hydrologic soil group D. Columns (1) - (2) correspond to Table 1, columns (3) - (4) correspond to Table 2, columns (5) and (6) correspond to Table 3 and columns (7) and (8) correspond to Table 5. Control variables included as indicated. Standard errors are clustered at the developer level. Significance Levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

5.7 Outperformance and Relevance of Collateral Quality

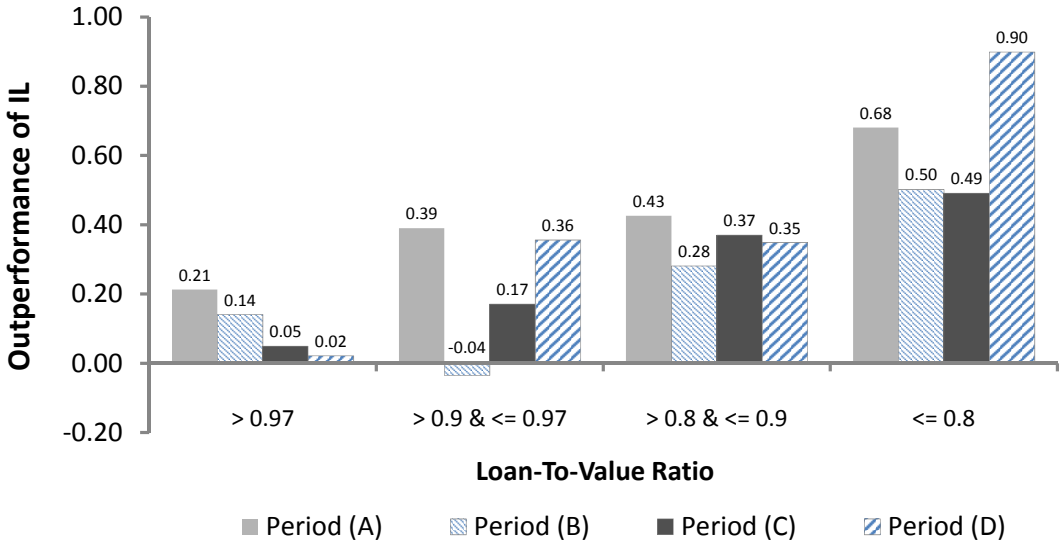
The previous sections showed that the houses financed by the integrated lender outperform otherwise similar houses financed by non-integrated lenders. Prediction 1(b) was that this outperformance should be larger for houses backing mortgages for which repayment is *less* dependent on collateral quality (high-γ mortgages). To test this prediction I use the initial downpayment on the mortgage to proxy for γ. When the downpayment is high the repayment probability is less sensitive to collateral quality, since house prices have to fall by more to create incentives for default. Hence, non-integrated lenders are less concerned about the collateral quality for mortgages with a low loan-to-value (LTV) ratio and offer more aggressive financing (less bid shading). This allows the integrated lender to assemble a particularly attractive collateral portfolio for low LTV ratio mortgages. On the other hand, for high LTV ratio mortgages a small decline in house prices can already generate incentives for default. Non-integrated lenders thus offer less aggressive financing to avoid the winner's curse (more bid shading), which, in equilibrium, improves their average collateral quality. I divide borrowers into four LTV ratio groups: Less than or equal to 80%, between 80% and 90%, between 90% and 97% and above 97%, the upper limit for FHA-insured mortgages.

Regresssion (6) estimates the outperformance of the integrated lender's collateral portfolio for each LTV ratio group.

$$Return_i = \alpha + \sum_{j=1}^{4} \kappa_j IL_i \cdot LTV_{i,j} + \sum_{j=2}^{4} \omega_j LTV_{i,j} + X_i\beta + \delta_{q_1,q_2} + \epsilon_i \tag{6}$$

Figure 4 shows the κ_j coefficients (which estimate the outperformance of the integrated lender's collateral portfolio over each loan-to-value ratio group) when measuring the return over different time horizons. The gray bars show the coefficients when return is measured over period (A) in Figure 2 and I include the same control variables as in column (4) of Table 1. The houses financed by the integrated lender outperform for all LTV ratios. The outperformance is particularly large for houses backing those low LTV ratio mortgages for which a decline in the collateral value does not lead to a significant increase in the default probability. Similar results obtain when measuring the return over periods (B), (C) and (D). A Wald test rejects the null-hypothesis of equality of κ_1 and κ_4. The F-statistics for the test when the return is measured over periods (A) - (D) are 4.20 (p-value of 0.043), 10.92 (p-value of 0.001), 7.74 (p-value of 0.006) and 16.24 (p-value of 0.000) respectively.

Figure 4: Relative Collateral Return of Integrated Lender by LTV Ratio



Note: This graph plot the κ_j coefficients for regression (6). I measure return over periods (A), (B), (C) and (D) in Figure 2 and include the same covariates as column (4) of Table 1, column (5) of Table 2, column (2) of Table 3 and column (1) of Table 5 respectively.

5.8 Collateral Quality - Foreclosure Event

A further test for adverse selection in mortgage lending is to analyze the performance of the mortgage directly. One would expect mortgages backed by low quality collateral to default more frequently than mortgages backed by high quality collateral. Unfortunately the deeds data does not track when a mortgage becomes delinquent. However, I do observe when there is a foreclosure, since foreclosures involve a transfer of ownership to the mortgage lender. For every mortgage made in 2000 - 2007 in a development with an integrated lender I determine whether I observe a foreclosure within 3 years of the initial sale. I then run a probit regression of $Foreclosure3Years_i$ on a dummy variable of whether the loan was made by the integrated lender, IL_i, month of sale fixed effects, δ_m, and control variables X_i as given in regression (7).

$$Foreclosure3Years_i = \alpha + \kappa IL_i + X_i\beta + \delta_m + \epsilon_i \tag{7}$$

Table 7 shows average marginal probit coefficients from regression (7). κ is consistently negative and highly significant. Conditional on observables, a mortgage made by an integrated lender is about one percentage point or 40% less likely to default than a mortgage made by a non-integrated lender. As before, the addition of control variables and fixed effects between specifications (1) and (3) does not affect the magnitude of κ.

Table 7: Relative Foreclosure Probability of Integrated Lender Mortgages

	First Owner				Second Owner	
	(1)	(2)	(3)	(4)	(5)	(6)
Integrated Lender	-0.011***	-0.009***	-0.010***	-0.018***	-0.014**	-0.020***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.006)	(0.001)
Controls (see note)		H,B,F, T,D1	H,B,F, T,D2	H,B,F, T,D2		H,B2,F2, T,D2
$\bar{y}$	0.017	0.018	0.021	0.028	0.056	0.070
N	71,655	68,315	59,303	8,637	10,511	10,723

Note: This table shows average marginal effects from probit regression (7). The dependent variable is whether a foreclosure was observed within 3 years of purchase. I include single-family residences sold by a developer in 2000 - 2007 in developments with an integrated lender. Columns (1) - (4) analyze the foreclosure probability during the first owner's tenure, columns (5) - (6) during the second owner's tenure. Each specifications controls for month of sale fixed effects, county fixed effects and quarter of construction fixed effects. House characteristics (H), buyer characteristics (B), financing characteristics (F) and census tract demographics (T) as in Table 1. Characteristics of the second buyer (B2) and the second mortgage (F2) as in Table 3. D1 includes developer fixed effects, D2 includes development fixed effects. Column (4) restricts the sample to mortgages that were securitized in the year they were originated. Standard errors are clustered at the developer level. Significance Levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

One concern with this analysis is that a foreclosure requires a strategic decision by the lender about whether to foreclose on a delinquent mortgage. Since integrated lenders usually hold many mortgages in the same development, they might be reluctant to initiate a foreclosure if this depresses prices for neighboring homes. One might thus observe fewer foreclosures for integrated lenders without their mortgages performing any better. To address this concern, in column (4) I restrict the sample to mortgages that were securitized within the same calendar year as they were originated, as reported in the HMDA data. For these mortgages, the decision of whether to foreclose is usually outside the discretion of the originator. The magnitude of the effect in this subsample is even larger than in the full sample, suggesting that results are not driven by the integrated lender's concern about the effect of foreclosures on neighborhood prices.

I also explore to what degree the lower default probability of mortgages granted by the integrated lender is explained by superior information about collateral quality or borrower characteristics. To do this, I analyze the probability of the *second* owner of the house entering into foreclosure within 3 years of purchasing the house. I only include those observations with a mortgage-financed second sale, since only those might end up in default. Columns (5) and (6) of Table 7 show the results from the probit regression. This specification compares the default probability of two similar mortgages, neither of which was granted by the integrated lender. The mortgages differ in whether or not they are backed by housing collateral that was *initially* financed by the integrated lender. Mortgages that are backed by such collateral are almost two percentage points less likely to enter into foreclosure than mortgages that are backed by collateral that was initially financed by a non-integrated lender. This difference in default probabilities must be driven by the integrated lenders's superior information about collateral quality not borrower characteristics, since no lender could have had any information about the identity of a possible second owner.

5.9 Robustness Check - Control for Interest Rates

In the previous sections I showed that houses financed by the integrated lender outperformed observationally similar houses in the same development financed by non-integrated lenders. However, I have not so far conditioned on the pricing of the mortgage. This might be a concern if there were fundamental differences in the risk preferences or strategies of integrated and non-integrated lenders. Such differences might explain some of the observed outperformance even if information about collateral qual-

ity was completely symmetric. For example, it could be that integrated lenders are more risk-averse and choose not to lend against low-quality collateral. If non-integrated lenders were more willing to lend against low-quality collateral, but at higher interest rates, we might observe an outperformance of the integrated lender's collateral portfolio even if collateral quality was observed by all lenders. Ideally, I would thus like to control for the pricing of mortgages when analyzing the difference in the return of the two collateral portfolios. Unfortunately, interest rates in Arizona are only recorded when the mortgage is an adjustable rate mortgage or a hybrid-ARM.[19,20]

In this section I present robustness checks for the subset of mortgages for which I observe the initial interest rate. I construct a "mortgage spread" variable that equals the spread of the initial mortgage rate over the average relevant (i.e. adjustable or hybrid-adjustable) interest rate in that month. I then include this spread as an additional covariate in the regressions that produce the key results in Tables 1, 2, 3, 5 and 7. The results are shown in Table 8.

Table 8: Robustness Check - Control For Interest Rate

	Period (A)	Period (B)	Period (C)	Period (D)	Foreclosure
	(1)	(2)	(3)	(4)	(5)
Integrated Lender	0.466*	0.363***	0.481***	0.415**	-0.007**
	(0.237)	(0.108)	(0.166)	(0.175)	(0.003)
Mortgage Spread	-0.454***	-0.172***	-0.097	-0.139*	0.011***
	(0.099)	(0.028)	(0.063)	(0.081)	(0.001)
Controls	Col (4) Table 1	Col (5) Table 2	Col (2) Table 3	Col(2) Table 5	Col (2) Table 7
R-squared	0.888	0.880	0.865	0.586	
$\bar{y}$	5.327	-8.353	-14.38	-27.61	0.038
N	7,968	23,805	3,407	18,534	16,008

Note: This table shows robustness checks, controlling for the initial interest rate for the subset of mortgages for which this information is available. The respective tables and the appropriate control variables are indicated. The dependent variables in columns (1) to (4) are the annualized collateral return over periods (A) - (D) respectively. The dependent variable in column (5) is the probability of observing a foreclosure within three years of the initial sale. Standard errors are clustered at the developer level. Significance Levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

Including interest rates as a control variable reinforces the conclusions about the relative outperformance of the integrated lender's collateral portfolio. Houses financed

[19]Hybrid-ARM mortgages (e.g. a 5/1 ARM is a mortgage with a fixed interest rates for the first five years and then an annually adjusted rate after that) were popular during the recent housing boom.

[20]The interest rate captures the most salient aspect of mortgage pricing. I do not observe other aspects of mortgage pricing such as the closing costs, "lock-in periods" or prepayment penalties.

with a higher interest rate mortgage have a lower return. This result suggests that there are observable aspects of collateral quality that lenders take into account when pricing mortgages. Columns (1) - (4) show that after the inclusion of the interest rate spread as an additional control variable, those houses financed by the integrated lender continue to outperform by about 40 or 50 basis points annually. Column (5) shows that the probability of foreclosure is also larger for higher interest rate mortgages. The mortgages in the integrated lender's portfolio continue to have a significantly lower foreclosure probability than those originated by non-integrated lenders.[21]

6 Interest Rate Response to Integrated Lender

Sections 5 analyzed the effects of adverse selection on the collateral quality of integrated and non-integrated lenders. In this section I test the model's predictions for the equilibrium interest rates charged by non-integrated lenders. Prediction 2 stated that this interest rate should be higher when the non-integrated lender competes against an integrated lender and thus lends against below average quality collateral.[22] To test this prediction I analyze the interest rates charged by non-integrated lenders by running regression (8), which compares developments with and without integrated lenders.

$$MortRate_i = \alpha + \kappa HasIL_i + X_i\beta + \delta_{m,f} + \tau_l + \epsilon_i \tag{8}$$

The dependent variable is the mortgage interest rate (section 5.9 discusses availability). The key explanatory variable, $HasIL_i$, captures whether the non-integrated lender competes against an integrated lender. It is set to one when an integrated lender makes loans in the same development and year. I include month by rate-type (adjustable or hybrid-adjustable) fixed effects, $\delta_{m,f}$, to capture the interest rate environment at the time of making the mortgage. I also include lender fixed effects τ_l. These are important if lenders with different funding sources and strategies are more or less aggressive in

[21] I also run these regressions by including (i) the actual interest rate charged and (ii) the spread over a different base rate (the Federal Funds rate for variable rate mortgages and the average national fixed-rate mortgage rate provided by Freddie Mac's PMMS for hybrid-ARMs) to control for the mortgage interest rate. The conclusions are very similar to the ones presented in Table 8.

[22] In the previous section I showed that in developments with an integrated lender, the integrated lender's collateral outperforms. In a previous version of the paper I also showed that amongst ex-ante similar mortgages made by a non-integrated lender, the collateral backing those mortgages that were made in developments where the non-integrated lender competed against an integrated lender had a significantly lower return. These results are available from the author on request.

their interest rate offers. The regression thus compares the lending behavior of the same lender making similar mortgages to purchase properties in two developments: one in which the developer cooperates with an integrated lender and one in which it does not. Standard errors are clustered at the lender level. This allows for an arbitrary correlation between the residuals of mortgages granted by the same lender.[23]

Table 9: Impact of Integrated Lender on Interest Rates of Non-Integrated Lender

	(1)	(2)	(3)	(4)	(5)	(6)
Has Integrated Lender	0.117**	0.114**	0.098**	0.089**	0.092**	0.077*
	(0.055)	(0.054)	(0.046)	(0.044)	(0.043)	(0.042)
Has Integrated Lender × Expansive Soil						0.150***
						(0.054)
Expansive Soil						-0.089*
						(0.048)
Controls (see note)	·	F	F, H,B	F,H,B,T	F,H,B,T D1	F,H,B,T
R-squared	0.555	0.583	0.590	0.591	0.596	0.591
$\bar{y}$	6.640	6.640	6.640	6.640	6.640	6.640
N	15,587	15,587	15,584	15,584	15,584	15,584

Note: This table shows results from regression (8). The dependent variable is the mortgage interest rate. I include single-family residences sold by a developer in Arizona in 2000 - 2007 that were financed by non-integrated lenders. Each specification includes month × rate-type (adjustable or hybrid-adjustable), county and lender fixed effects. Finance characteristics (F), house characteristics (H), buyer characteristics (B) and census tract demographics (T) as in Table 1. D1 includes developer fixed effects. Standard errors are clustered at the lender level. Significance Levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

The results, which are presented in Table 9, suggest that non-integrated lenders charge about 10 basis points higher interest rates when competing against a better-informed integrated lender. Between columns (1) and (6) I sequentially add control variables. The magnitude of the estimated interest rate increase changes little, which suggests that the results are not driven by a different composition of houses or borrowers in developments with or without an integrated lender. In column (5), when I include developer fixed effects, κ is identified by considering houses built by developers that

[23]There might also be a concern that mortgages granted around the same time should not be considered as independent observations since they might be affected by correlated unobserved shocks that are not picked up by $\delta_{f,m}$. To address this concern, I also computed a second set of standard errors clustered at both the month level and the lender level using a method described by Petersen (2009). This two-level clustering allows for an arbitrary correlation between mortgages made in the same month as well as between mortgages by the same lender. The standard errors clustered at two levels are nearly identical to those reported in Table 9, and are available from the author.

sometimes cooperate with an integrated lender, but do not always do so.[24] Again, the evidence is highly consistent with a story of asymmetric information about collateral quality, even though the lack of exogenous variation in a developer's choice to acquire an integrated lender makes it difficult to make causal statements. In column (7) I add the interaction between $HasIL_i$ and ES_i, the dummy variable capturing whether the house was built on expansive soil. This tests Prediction 2(a) from section 2 which stated that the interest rate premium for competing with an integrated lender should be particularly large for those houses where the integrated lender's information about construction quality has the most impact on future house prices. Consistent with this prediction, the interest rate increase in response to the presence of the integrated lender is more than twice as large for houses built on expansive soil.

The interest rate increase for competing against an integrated lender should also be larger when the adverse selection on collateral quality is more important (Prediction 2(b)). In particular, non-integrated lenders should raise interested rates more for high LTV ratio mortgages, for which a small decline in collateral value precipitates a larger increase in default risk. To test whether this is the case, I run regression (9).

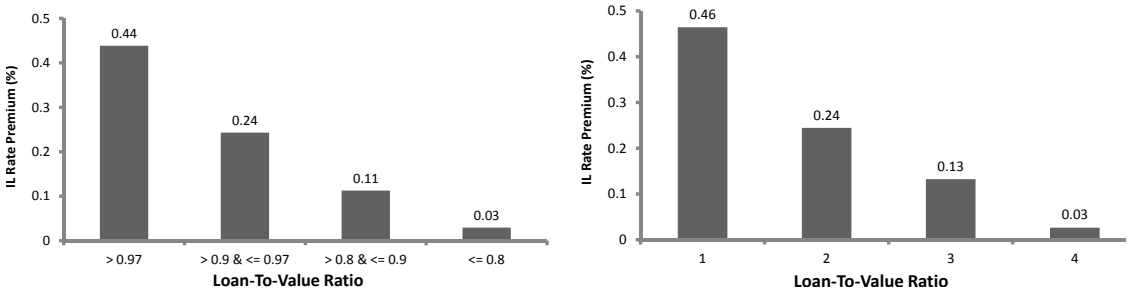
$$MortRate_i = \alpha + \sum_{j=1}^{4} \kappa_j HasIL_i \cdot LTV_{i,j} + \sum_{j=2}^{4} \omega_j LTV_{i,j} + X_i\beta + \delta_{m,f} + \tau_l + \epsilon_i \quad (9)$$

Figure 5 plots the κ_j coefficients from this regression. The regressions for the left and right panels include the same control variables as columns (4) and (5) of Table 9 respectively. The results support the model prediction. For high LTV ratio mortgages the interest rate premium charged by the non-integrated lender when competing against an integrated lender is the largest, at almost half a percentage point annually. The F-statistics for a Wald test of the equality of κ_1 and κ_4 are equal to 16.2 (p-value of 0.00) for the left panel and 17.4 (p-value of 0.00) for the right panel.

While I find that non-integrated lenders adjust interest rates when competing with an integrated lender, these results do not necessarily imply that non-integrated lenders are explicitly aware of the underlying adverse selection. The pricing of mortgages usually relies on statistical models that use past data to compute expected future default probabilities and adjust interest rates accordingly. When mortgages on houses built by a certain developer default more often, the non-integrated lender would charge

[24] I do not include development fixed effects, since these are nearly collinear with $HasIL_i$, which varies at the development-year level. In other words, most developments either do or do not have an integrated lender in all years of operation.

Figure 5: Interest Rate Increase by Non-Integrated Lenders by LTV Ratio



Note: These graphs plot the point estimates of the κ_j coefficients for regression (9), measuring the interest rate increase of a non-integrated lender when competing against an integrated lender for different values of the LTV ratio. The left panel includes the same covariates as column (4) of Table 9. The right panel includes the same covariates as column (5) of Table 9.

higher interest rates for mortgages to purchase those homes. From the perspective of the lender, this could be because the quality of homes built by the developer is generally low or because of adverse selection on collateral quality. To price these mortgages correctly (i.e. so that default realizations do not contradict default expectations), non-integrated lenders would not need to differentiate between these two explanations.

6.1 Magnitude of Effects

Sections 5 discusses the effect of adverse selection on the relative quality of houses financed by integrated and non-integrated lenders. I now consider whether the effect of this difference in collateral quality on the value of the mortgage is consistent in magnitude with the observed changes in the non-integrated lenders' pricing behavior. Relative to lending in a development without an integrated lender, the presence of an integrated lender causes a non-integrated lender to lend against collateral that underperforms by at leat 30 basis points annually. After the average life time of a mortgage (which is about 8-10 years as borrowers move or prepay), a non-integrated lender's collateral in developments with an integrated lender is thus worth about 3% less than if the non-integrated lender did not face the adverse selection. To see whether the increase in the interest rate is adequate compensation for this effect, note that banks regularly conduct similar pricing calculations when deciding how to adjust interest rates for borrowers with different downpayments (a lower initial downpayment has a similar effect on the value of the mortgage as the adverse selection, since both push the borrower's option to default closer to being in the money). I would thus expect the pricing adjustment to the presence of an integrated lender to be roughly similar to the

interest rate increase when raising the initial loan-to-value ratio by 3 percentage points. This indeed appears to be the case. A rate sheet by *US Bank* from June 2008 shows pricing adjustments to changes in the LTV ratio for 5/1 ARMs. For a borrower with a credit score between 700 and 719, increasing the LTV ratio from 70% to 75% and from 75% to 80% involves an increase in the annual interest rate of 10 basis points each. An increase from 80% to 85% involves rate increase of 20 basis points, and an increase in the LTV ratio from 85% to 90% involves a rate increase of 25 basis points. These magnitudes are highly consistent with the interest rate increases detected in section 6.

6.2 Effect of Securitization

During the period under consideration a significant proportion of mortgages were securitized and sold as mortgage-backed securities (MBS). It is often argued that this ability to securitize mortgages led to moral hazard in mortgage origination (Keys et al., 2010), since originators would no longer face the costs of default. If this were true, the same mechanism would also reduce the incentives of non-integrated lenders to avoid the winner's curse. However, while securitization can *reduce* the exposure to an eventual default of the mortgage, there are a number of reasons why it does not *eliminate* it, and why we would expect lenders to continue to price borrower and collateral risk characteristics. First, in private label securitization, which made up 56% of all MBS issued in 2006, the issuer oftentimes retains tranches of varying seniority, generating direct exposure to mortgages. In addition, securitizers usually retain exposure through credit enhancements such as overcollateralization and excess spreads, which are used to cover default losses. Furthermore, Gorton and Souleles (2007) provide evidence that a sponsor's support of securitized products is provided by implicit recourse, since secondary market buyers will not buy from sponsors whose securities frequently default. In addition, the sale of mortgages to investors often includes recourse clauses that require the lender to take back loans if specific events such as borrower default occur.

7 Conclusion

In this paper I analyze the sources and magnitude of asymmetric information between competing lenders in residential mortgage lending. I exploit that property developers often cooperate with vertically integrated mortgage lenders that might have superior information about the quality of the housing collateral and about borrower characteristics. By conditioning their interest rate offers on this superior information, integrated

lenders can subject competing non-integrated lenders to adverse selection. I show that adverse selection on collateral quality is a key feature of the market to finance newly developed homes. In developments with an active integrated lender the portfolio of houses financed by the integrated lender is of above average quality along a number of key dimensions.[25] In particular, its annual return is about 40 basis points higher than that of ex-ante similar houses in the same development financed by non-integrated lenders. By also considering the relative return over the ownership period of the second owner of the house, I show that this result cannot be explained by differential information about borrower characteristics or a bundling of the home sale and the mortgage. This outperformance is particularly large for houses built on expansive soil, which makes housing return more sensitive to construction quality. The outperformance is also larger for mortgages with a low loan-to-value ratio for which repayment is less sensitive to changes in collateral values. I provide further evidence for asymmetric information about collateral quality by considering the textual description of properties in for-sale listings and showing that houses financed by the integrated lender are less likely to experience depreciation events. The adverse selection also translates into higher foreclosure rates for mortgages granted by non-integrated lenders. To compensate for lending against below-average quality collateral, non-integrated lenders charge about 10 basis points higher interest rates when competing against an integrated lender. This interest rate increase is larger for houses built on expansive soil and for houses financed with high loan-to-value ratio mortgages, the repayment of which is more sensitive to changes in collateral values.

My results highlight the pervasive nature of asymmetric information in mortgage markets. While the debate surrounding such asymmetric information usually focuses on information about the borrower, I show that asymmetric information about the mortgage collateral is common and of significant magnitude. This suggests that successful policy proposals aimed at reducing the amount of asymmetric information in mortgage markets should also focus on providing better information about collateral quality, for example by making detailed property inspection records available publicly.

[25] While the activity of integrated lenders in new developments provides a clearly identifiable measure of relative information and thus facilitates an empirical assessment of the sources and magnitude of asymmetric information in mortgage lending, it is likely that there is similar asymmetric information about housing collateral values in lending to purchase existing properties. For example, lenders often acquire superior information about local demand factors that will impact future house prices in a specific geographic or price segment. Such a mechanism is consistent with the results in Loutskina and Strahan (2011), who present evidence that mortgage lenders that are more geographically concentrated produce more private information and make more profitable mortgages on average.

References

Agarwal, S. and R. Hauswald, "Distance and private information in lending," *Review of Financial Studies*, 2010, *23* (7), 2757–2788.

_ , G. Amromin, C. Gartenberg, A. Paulson, and S. Villupuram, "Home-builders, Affiliated Financing Arms and the Mortgage Crisis," *Working Paper*, 2011.

Altonji, J.G., T.E. Elder, and C.R. Taber, "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

Bayer, P., R. McMillan, A. Murphy, and C. Timmins, "A dynamic model of demand for houses and neighborhoods," *NBER Working Paper*, 2011, *17250*.

Berger, A.N. and G.F. Udell, "Relationship lending and lines of credit in small firm finance," *Journal of Business*, 1995, *68* (3), 351–381.

Black, H.A., T.P. Boehm, and R.P. DeGennaro, "Is there discrimination in mortgage pricing? The case of overages," *Journal of Banking & Finance*, 2003, *27* (6), 1139–1165.

Chiappori, P.A. and B. Salanié, "Testing for asymmetric information in insurance markets," *Journal of Political Economy*, 2000, pp. 56–78.

Criterium Engineers, "If it's new, is it good?," http://criteriumhomeinspection.com/new-construction-is-it-good, 2003.

Elul, R., "Securitization and mortgage default," *Federal Reserve Bank of Philadelphia Working Paper*, 2011, pp. 09–21.

Engelbrecht-Wiggans, R., P.R. Milgrom, and R.J. Weber, "Competitive bidding and proprietary information," *Journal of Mathematical Economics*, 1983, *11* (2), 161–169.

Finkelstein, A. and J.M. Poterba, "Adverse selection and the choice of risk factors in insurance pricing: Evidence from the uk annuity market," *Journal of Political Economy*, 2004, *112* (1), 183–208.

Fishman, M.J. and J.A. Parker, "Valuation, adverse selection, and market collapses," *Working Paper*, 2010.

Garmaise, M.J. and T.J. Moskowitz, "Confronting information asymmetries: Evidence from real estate markets," *Review of Financial Studies*, 2004, *17* (2), 405–437.

Gartenberg, C., "Tempted by scope? Homebuilder mortgage affiliates, lending quality and the housing crisis," *Working Paper*, 2011.

Gorton, G.B. and G. Ordonez, "Collateral crises," *Working Paper*, 2011.

_ **and N.S. Souleles**, "Special purpose vehicles and securitization," in M. Carey and R.M. Stulz, eds., *The Risks of Financial Institutions*, University of Chicago Press, 2007.

Hauswald, R. and R. Marquez, "Competition and strategic information acquisition in credit markets," *Review of Financial Studies*, 2006, *19* (3), 967–1000.

Hendricks, K. and R.H. Porter, "An empirical study of an auction with asymmetric information," *American Economic Review*, 1988, *78* (5), 865–883.

Keys, B.J., T. Mukherjee, A. Seru, and V. Vig, "Did securitization lead to lax screening? Evidence from subprime loans," *Quarterly Journal of Economics*, 2010, *125* (1), 307–362.

Levitt, S.D. and C. Syverson, "Market distortions when agents are better informed: The value of information in real estate transactions," *Review of Economics and Statistics*, 2008, *90* (4), 599–611.

Loutskina, Elena and Philip E. Strahan, "Informed and Uninformed Investment in Housing: The Downside of Diversification," *Review of Financial Studies*, 2011, *24* (5), 1447–1480.

Milgrom, P. and R.J. Weber, "The value of information in a sealed-bid auction," *Journal of Mathematical Economics*, 1982, *10* (1), 105–114.

Petersen, M.A., "Estimating standard errors in finance panel data sets: Comparing approaches," *Review of Financial Studies*, 2009, *22* (1), 435.

_ **and R.G. Rajan**, "The benefits of lending relationships: Evidence from small business data," *Journal of Finance*, 1994, *49* (1), 3–37.

_ **and** _ , "Does distance still matter? The information revolution in small business lending," *Journal of Finance*, 2002, *57* (6), 2533–2570.

Phoenix New Times, "Cracked Houses - Homes all over Arizona are falling apart," *Published on March 16, 2006*, 2006.

Pierce, L., "Organizational structure and the limits of knowledge exploitation: Evidence from consumer automobile leasing," *Management Science*, 2011, *forthcoming.*

The Arizona Republic, "Construction Quality Varies Widely," *Published on November 20, 2001*, 2001.

_ , "Appeals of property valuations soar in Maricopa County," *Published on November 10, 2009*, 2009.

von Thadden, E.L., "Asymmetric information, bank lending and implicit contracts: The winner's curse," *Finance Research Letters*, 2004, *1* (1), 11–23.

Woodward, S.E. and R.E. Hall, "Diagnosing consumer confusion and sub-optimal shopping effort: Theory and mortgage-market evidence," *NBER Working Paper*, 2010, *16007.*

A Theory Appendix - NOT FOR PUBLICATION

Define the probability of observing $\eta = h$ as $P_i(h) = q\phi + (1-q)(1-\phi)$ and the probability of observing $\eta = l$ as $P_i(l) = (1-q)\phi + q(1-\phi)$. The probability that a house is high quality conditional on observing $\eta = h$ is $p(h,\phi) = Pr(\theta = H|\eta = h) = \frac{q\phi}{q\phi+(1-q)(1-\phi)}$. The probability that the house is high quality conditional on observing $\eta = l$ is $p(l,\phi) = Pr(\theta = H|\eta = l) = \frac{q(1-\phi)}{(1-q)\phi+q(1-\phi)}$. Define the expected revenue from lending at interest rate R to a type-γ agent wanting to buy a house with signal η as:

$$\begin{aligned} W(R;\eta,\phi,\gamma) &= p(\eta,\phi)R + [1-p(\eta,\phi)]\gamma R \\ &= [p(\eta,\phi)(1-\gamma)+\gamma]R = z(\eta,\phi,\gamma)R. \end{aligned}$$

$z(\eta,\phi,\gamma)$ is the repayment probability of the loan conditional on observing η with signal precision ϕ. $R(\gamma)_a^b = \frac{R_f}{q+\gamma(1-q)}$ is the break-even interest rate when lending to a type-γ agent to purchase an average house. $R(\gamma,\phi)_l^b = \frac{R_f}{z(l,\phi,\gamma)}$ is the break-even interest rate for the integrated lender when lending to a type-γ agent who wants to purchase a house when $\eta = l$.

Theorem 1 *There are no pure strategy equilibria.*

Proof The proof follows by contradiction. Let pure strategies be $R_i(\eta)$ for the integrated lender and R_n for the non-integrated lender. The only possible pure strategy equilibrium is $R_a = R_n = R_i(h) = R_i(l)$. Assume otherwise. If $R_n < R_i(h), R_i(l)$, the non-integrated lender can increase its expected return by offering $R_n' = R_n + \varepsilon$. If $R_i(\eta) < R_n$, the integrated lender can increase its profit by offering $R_i(\eta)' = R_i(\eta) + \varepsilon$. However, each lender offering R_a is also not an equilibrium. If $R_a < R(\gamma)_a^b$, each lender would be better off not offering a mortgage at all. If $R_a > R(\gamma,\phi)_l^b$, the integrated lender would be better off by offering interest rates $R_i(l)' = R_a - \varepsilon$ and $R_i(h)' = R_a - \varepsilon$. If $R(\gamma)_a^b < R_a < R(\gamma,\phi)_l^b$ the integrated lender would be better off offering $R_i(l)' = R_a + \varepsilon$ and $R_i(h)' = R_a - \varepsilon$, subjecting the non-integrated lender to a winner's curse. The non-integrated lender would make a loss in expectation.

Theorem 2 *Let $W(R;\eta,\phi,\gamma)$ be the integrated lender's expected revenue from lending at rate R to a type-γ borrower to buy a house with signal η. The interest rate offer game for a type-γ borrower when signal precision is ϕ has a unique mixed strategy equilibrium, such that:*

1. *The non-integrated lender breaks even, the integrated lender earns positive expected profits.*
2. *$\exists\bar{\gamma}$ such that for borrowers with $\gamma < \bar{\gamma}$ the integrated lender rejects all mortgage applications to buy houses when $\eta = l$. When $\eta = h$, the integrated lender randomizes interest rate offers over $[R(\gamma)_a^b, R(\gamma)_m)$ using the following cumulative distribution function:*

$$F_i(R;h,\phi,\gamma) = 1 + \frac{P_i(l)[W(R;l,\phi,\gamma) - R_f]}{P_i(h)[W(R;h,\phi,\gamma) - R_f]}.$$

$R(\gamma)_a^b = \frac{R_f}{q+\gamma(1-q)}$ is the break-even interest rate for lending to a type-γ agent to buy an average quality house. $P_i(\eta)$ is the probability of the integrated lender observing signal η. The integrated lender also makes interest rate offers with a point mass of $1 - F_i(R(\gamma)_m; h, \phi, \gamma)$ at $R(\gamma)_m$. The non-integrated lender randomizes interest rate offers over $[R(\gamma)_a^b, R(\gamma)_m)$ using the following cumulative distribution function:

$$F_n(R; \phi, \gamma) \quad = \quad 1 - \frac{W(R(\gamma)_a^b; h, \phi, \gamma) - R_f}{W(R; h, \phi, \gamma) - R_f}.$$

With probability $1 - F_n(R(\gamma)_m; \phi, \gamma)$ the non-integrated lender does not make an offer.

3. *For borrowers with $\gamma > \bar{\gamma}$ both integrated and non-integrated lenders always offer a mortgage. When $\eta = l$ the integrated lender offers the break-even interest rate $R(\gamma, \phi)_l^b$, defined implicitly by $R_f = W(R(\gamma, \phi)_l^b; l, \phi, \gamma)$. When $\eta = h$ the integrated lender randomizes its interest rate offers over $[R(\gamma)_a^b, R(\gamma)_m]$ using $F_i(R; h, \phi, \gamma)$. The non-integrated lender always randomizes over $[R(\gamma)_a^b, R(\gamma, \phi)_l^b)$ using $F_n(R; \phi, \gamma)$, with a point mass at $R(\gamma, \phi)_l^b$.*

To find the unique mixed strategy equilibrium I follow a number of steps in similar proofs in Hauswald and Marquez (2006), von Thadden (2004) and others. Let $F_i(R; \eta, \phi, \gamma)$ represent the cumulative distribution function (cdf) of the integrated lender's distribution of interest rate offers R for a type-γ borrower wanting to buy a house with signal η when the integrated lender's signal precision is ϕ. Let $F_n(R; \phi, \gamma)$ be the cdf of the non-integrated lender's distribution over interest rate offers R for a type-γ borrower wanting to buy a house when the integrated lender's signal precision is ϕ. Both $F_i(R; \eta, x, \gamma)$ and $F_n(R; \phi, \gamma)$ are continuous, strictly increasing and atomless on a common support $[\underline{R}, \bar{R}]$ (see von Thadden, 2004). For each signal precision ϕ there is a marginal household with $\gamma = \bar{\gamma}$ to whom it is no longer valuable to lend at the highest possible rate if the collateral signal is negative. This cutoff is defined as the solution to $R(\bar{\gamma}, \phi)_l^b = R(\bar{\gamma})_m$. $\bar{\gamma}$ is increasing in ϕ: for higher signal precision, the probability that house is truly $\theta = l$ when the integrated lender observes $\eta = l$ is higher. $\bar{\gamma}$ is also decreasing in q and H.

Since a less informed bidder cannot profit from a sealed-bid auction against a better-informed competitor,[26] the non-integrated lender must break even in equilibrium. This allows us to calculate the lower bound of the support. When offering $\underline{R}$ the non-integrated lender wins almost surely and since it needs to make a profit of 0, we have

[26] The set-up analyzed here is similar to the first-price sealed bid common value auction analyzed by Milgrom and Weber (1982). There the authors show that when the information set of the less informed competitor is less finely partitioned, the less informed lenders will make zero profit in equilibrium.

$\underline{R} = R(\gamma)_a^b$. The upper bound of the distribution, $\bar{R}$, depends on γ. When $\gamma \geq \bar{\gamma}$ a repeated undercutting argument similar to Bertrand competition shows that for $\eta = l$ the integrated lender offers $R(\bar{\gamma}, \phi)_l^b$ and makes zero profit. When $\eta = h$, the integrated lender mixes offers on the support of $[R(\gamma)_a^b, R(\gamma, \phi)_l^b)$. For $\gamma < \bar{\gamma}$ the integrated lender never makes an offer if $\eta = l$ and mixes over $[R(\gamma)_a^b, R(\gamma)_m)$ when $\eta = h$. For any ϕ, the common support is thus given by $[R(\gamma)_a^b, \min\{R(\gamma, \phi)_l^b, R(\gamma)_m\})$. The expected profit for the integrated lender from offering an interest rate R when $\eta = h$ (recalling that the integrated lender will make zero profits if $\eta = l$) is:

$$\begin{aligned} \pi_i(R; h, \phi, \gamma) &= \text{Probability of winning} \times \text{Expected Profit when Winning} \quad (10) \\ &= [1 - F_n(R; \phi, \gamma)] \times [W(R; h, \phi, \gamma) - R_f] \end{aligned}$$

The expected profit for the non-integrated lender from offering interest rate R is:

$$\begin{aligned} \pi_u(R; \phi, \gamma) &= [(\text{Prob. } i \text{ has } \eta = l) \times (\text{Expected Profit when } i \text{ has } \eta = l)] + \quad (11) \\ &\quad [(\text{Prob. } i \text{ has } \eta = h) \times (\text{Prob. of winning}) \times \\ &\quad (\text{Expected Profit when } i \text{ has } \eta = h)] \\ &= P_i(l)[W(R; l, \phi, \gamma) - R_f] + P_i(h)[1 - F_i(R; h, \phi, \gamma)][W(R; h, \phi, \gamma) - R_f] \end{aligned}$$

Since the non-integrated lender must break even, we have that $\forall(R, \gamma) : \pi_u(R; \phi, \gamma) = 0$. In addition, since the mixing distributions are strictly increasing, equilibrium profit for each lender must be the same for every interest rate offered on the support: $\pi_i(R; h, \phi, \gamma) = \bar{\pi}(\phi, \gamma)$. If we now evaluate $\pi_i(R; h, \phi, \gamma)$ at the lower bound of the support, since $F_n(R(\gamma)_a^b; \phi, \gamma) = 0$, we have that $\bar{\pi}(\phi, \gamma) = W(R(\gamma)_a^b; h, \phi, \gamma) - R_f$. Plugging this into equation (10) and solving for $F_n(R; \phi, \gamma)$ gives:

$$F_n(R; \phi, \gamma) = 1 - \frac{W(R(\gamma)_a^b; h, \phi, \gamma) - R_f}{W(R; h, \phi, \gamma) - R_f} \quad (12)$$

Similarly, solving equation (11), by setting $\pi_u(R; \phi, \gamma) = 0$ gives:

$$F_i(R; h, \phi, \gamma) = 1 + \frac{P_i(l)[W(R; l, \phi, \gamma) - R_f]}{P_i(h)[W(R; h, \phi, \gamma) - R_f]} \quad (13)$$

Since both lenders randomize over the full support of the distribution functions, they cannot profitably deviate from their mixed strategies. Hence, the preceding distributions represent the unique equilibrium for a borrower of type γ.

Probability of making an offer: $\gamma < \bar{\gamma}$

For $\gamma < \bar{\gamma}$, $F_i(\bar{R}; h, \phi, \gamma) = F_i(R(\gamma)_m; h, \phi, \gamma) < 1$. Hence the integrated lender randomizes over $[R(\gamma)^b_a, R(\gamma)_m)$ for $\eta = h$ houses, without any atoms, but with point mass at $R(\gamma)_m$, where the mass is equal to $1 - F_i(R(\gamma)_m; h, \phi, \gamma)$.[27] The integrated lender never bids for $\eta = l$ agents and bids with probability 1 for $\eta = h$ agents. The non-integrated lender bids with probability $F_n(R(\gamma)_m; \phi, \gamma) < F_i(R(\gamma)_m; h, \phi, \gamma) < 1$ for all agents. With probability $1 - F_n(R(\gamma)_m; \phi, \gamma)$ the non-integrated lender does not make an interest rate offer and the household gets rationed.

Probability of making an offer: $\gamma \geq \bar{\gamma}$

For $\gamma \geq \bar{\gamma}$, both lenders always make an offer to the borrower. I argued above that for $\eta = l$ the integrated lender always offers credit at $R(\gamma(\phi), \phi)^b_l$, making zero profit. For $\eta = h$ we have $F_i(R(\gamma, \phi)^b_l; h, \phi, \gamma) = 1$, since $R_f = W(R(\gamma, \phi)^b_l; l, \phi, \gamma)$ and $\bar{R} = R(\gamma, \phi)^b_l$ for $\gamma \geq \bar{\gamma}$. Hence the informed lender will make an offer by randomizing over the full support without atoms. Similarly, $F_n(R(\gamma, \phi)^b_l; \phi, \gamma) < 1$, so the uninformed lender will also randomize over the full support, with a mass point of $1 - F_n(R(\gamma, \phi)^b_l; \phi, \gamma)$ at $R(\gamma, \phi)^b_l$.

B Data Appendix - NOT FOR PUBLICATION

I begin with a dataset that contains 3.34 million ownership-changing deeds recorded in Arizona between 2000 and 2011. The data include both armslength market transactions, as well as transfers in divorce, estate settlements and foreclosures. For each deed with sufficient information to uniquely identify the property, the address is geocoded to determine the property's precise location. For 91.7% of deeds the address information is sufficiently detailed to determine the exact latitude and longitude. For another 2.1% of deeds the street-number is missing and a latitude and longitude is assigned that locates the property at the geographic midpoint of the street. The 6.2% of deeds with insufficient address information to assign a location are dropped (many of them refer to the sale of vacant land). I then merge each deed via its asssessor parcel number (APN) and county to the underlying property's tax assessment record for the year 2010.

[27] In order for a lender to not make an interest rate offer in some instances, it must be indifferent between bidding and not bidding. Since the integrated lender makes a profit in expectation when making an offer on the $\eta = h$, it is never indifferent between bidding and not bidding, which generates expected profits of zero. Thus, unlike the non-integrated lender, it will never not bid.

B.1 Data Cleaning + Identifying Transaction Types

Armslength Transactions: I identify all deeds that contain information about armslength transactions in which both buyer and seller act in their best economic interest. This ensures that transaction prices reflect the market value of the property. I include all deeds that are one of the following: "Grant Deed," "Condominium Deed," "Individual Deed," "Warranty Deed," "Joint Tenancy Deed," "Special Warranty Deed," "Limited Warranty Deed" and "Corporation Deed." This excludes intra-family transfers and foreclosures. I drop all observations that are not a Main Deed or only transfer partial interest in a property. This leaves 1.73 million armslength transactions.

Newly Developed Single-Family Residences: Amongst the armslength transactions I identify mortgage-financed purchases of newly developed properties. This includes all deeds in which the seller is identified as a company or a partnership, but that are not REO resales (i.e. sales by a bank following a foreclosure). I exclude sales in which the construction date of the house (as reported in the assessor data) precedes the sales date by more than two years. These transactions usually involve a developer that renovates and resells existing properties. I also exclude transactions where the buyer is identified as a company. In addition, I only consider single-family residences, which make up about 85% of newly developed properties in Arizona. This leaves me with 240,803 observations. For each newly developed property I collect subsequent armslength sales to track the their future return.[28]

Divorce and Death: I identify those repeat sales pairs for which I observe a divorce or death of the owners up to six months before the second sale. I identify divorces through the presence of an "Intra-Family Transfer & Dissolution" deed that transfers property rights from initially joint ownership to one of the initial owners. The death of an owner is identified if either (i) the seller on a deed is classified as an "estate", "executor", "deceased" or "surviving joint owner" or (ii) if I observe one of the following: "Affidavit of Death of Joint Tenant" or "Executor's Deed."

Foreclosures: I mark those properties that experience a foreclosure within three years of the initial sale by the developer. A foreclosure event is identified (i) if the deed is either a "REO Repossession", "REO Resale", "Foreclosure Deed", "Deed in Lieu of Foreclosure", "Trustee's Deed" or (ii) when the buyer is identified as a "beneficiary."

[28] I exclude repeat sales pairs for which the time difference between the two sales is less than 270 days. Such sales often precede or follow the redevelopment of a property. For similar reasons, the Case-Shiller house price index excludes transaction pairs with less than six months time difference.

Data Cleaning: I identify houses in the same development by combinations of seller identity and census tract. I only consider houses that were first sold before 2008 and are located in developments with more than 30 units. I drop a few observations that are likely to have misreported loan or sales price details (i.e. when the sales price is less than $\$25,000$ or more than \$10 million and when the LTV ratio is more than 1.3 or less than 0.3). In addition, I only keep observations with a full set of control variables in the assessor data.[29] This leaves me with $158,785$ observations.

B.2 Deeds Data to HMDA Merge

I next merge the deeds to data from the Home Mortgage Disclosure Act (HMDA). This allows me to obtain additional characteristics of the home owners, as well as information on the subsequent securitization of mortgages. The HMDA is a mortgage-level dataset and identifies a mortgage by year, census tract, mortgage amount and mortgage lender.[30] Bayer et al. (2011) use these characteristics to merge a dataset similar to my deeds data to the HMDA. This allows them to uniquely match about 70% of all sales. I use additional characteristics to improve match rates and quality. First, both the deeds and HMDA data report whether mortgages are FHA-insured or VA-guaranteed. Second, HMDA data identifies whether a house is purchased as a rental property, while the assessor data has information about whether it was owner-occupied in 2009. Third, HMDA data contains information about whether the mortgage was applied for by a male, female, or two applicants. The deeds data also identifies purchasers as male, female or a married couple. Fourth, the HMDA data has information about the race and ethnicity of applicants. In the deeds data I do not have this information, but I do observe the names of buyers. I match the surnames of buyers to the 1000 most common Asian and Latino surnames from the 2000 U.S. Census. Using these four additional characteristics allows me to confirm $64,947$ unique matches. Despite the use of additional match variables, my unique match rate is lower than the one reported by Bayer et al. (2011). There are a number of reasons for this: First, since integrated lenders make a significant number of mortgages in new developments, the power of using lender identity to merge deeds to HMDA data declines. Second, lenders in new developments might be more likely to fall below the asset reporting threshold. For my

[29] This primarily drops observations from Pima county (city of Tuscon), which does not usually provide lot size and building size in the assessment records. See discussion in Appendix B.4.

[30] The Federal Reserve's Regulation C, which governs the HMDA, applies to most depository institutions with a branch office in a metropolitan area. Banks below \$39 million in assets are exempt from reporting requirements, as are nondepository institutions with assets below \$10 million.

main data set, for those mortgages where more than one match is possible, I match each deed randomly to one of the possible records in the HMDA data. I can merge a total of 102,818 deeds to HMDA data. In a previous version of the paper I showed that the key empirical results are robust to considering (i) only the sample of houses with a unique HMDA merge, and (ii) the full sample of houses in my data, without requiring an HMDA merge and without conditioning on owner characteristics.[31]

B.3 Identifying Integrated Lenders

To identify integrated lenders, I follow a number of steps: First, developers usually own their integrated lenders (e.g. the developer "Shea Homes" owns "Shea Mortgage"). For each developer, I determine whether there is joint ownership with its largest lender, using OneSource North American Business Browser and SEC filings. If I can confirm joint ownership, I assign the lender to be the integrated lender of this developer. This allows me to identify 45,266 mortgages granted by integrated lenders. I also analyze instances in which the market share of a single lender in a development exceeds 50%, but in which the developer does not own this lender. In these cases, I also assign the lender to be integrated, which assigns another 18,550 transactions to have mortgages granted by an integrated lender. Using this process of identifying integrated lenders, 85.1% of newly built houses are in a development with an integrated lender. For houses in developements with an integrated lender, the integrated lender has a market share of 72.9%. I believe that this process of identifying integrated lenders is appropriate: when analyzing the distribution of the market share of the largest lender for lending to purchase *existing* homes, I find that there are essentially no census tracts in which the largest lender has a market share in excess of 35%. Consequently, in any development in which a lender attains more than 50% of all mortgages, it is very likely that this lender is only able to obtain such a market share through an integrated lender arrangement.[32] In a previous version of the paper I reported robustness checks that show that the results do not change when only considering integrated lenders identified through joint ownership with the developer. These results are available from the author.

[31] Not conditioning on owner characteristics should not contaminate my results since section 5.1 shows that there is no selection into the integrated lender portfolio along observable owner characteristics that also affect the housing return.

[32] Additional lenders identified through this channel are usually independent companies that specialize in providing financing for developers in a integrated lender role, such as *IMortgage*, which states on its website: "We partner with homebuilders across the country to establish and manage their mortgage operations. We originate, underwrite, process and close mortgages on newly constructed homes."

B.4 Summary Statistics

Table 10 shows how the observations in my data are distributed over time and across counties. It includes all observations with HMDA-merge and which contain the full set of covariates. Summary statistics are split up for developments with and without an integrated lender. For developments with an integrated lender, the results are given separately for the integrated lender and for other non-integrated lenders. The top panel shows that the majority of observations are from Maricopa and Pinal county, which constitute the Phoenix MSA. Pima county (including Tuscon) only contributes a few observations. This is because for Pima I only observe building and lot size for a small number of observations in the assessment data. These variables are important controls in my main specifications. In order to estimate all models with a common sample, observations with missing data on home characteristics were dropped.[33] The bottom panel shows the distribution of observations by year of sale. The number of newly developed properties sold increased up to 2005, the peak of Arizona's housing boom, and then declined markedly during the financial crisis.

Table 11 shows summary statistics for the control variables used in the regressions. Most of these controls are not included linearly in the regression, but by splitting them into groups of values represented by dummy variables. This allows a more flexible functional form. The results are not sensitive to the exact definition of groups. All dollar amounts are in year-2000 dollars.

House Characteristics: Controls for initial sales price are included by adding dummy variables for $\$10,000$ buckets. Lot size and building size are controlled for by adding dummy variables for 20 equally sized groups. To control for garage spaces, I add a dummy variable for each possible value.

Borrower and Financing Characteristics: Income is controlled for by adding dummy variables for 50 equally sized groups. The loan-to-income (LTI) ratio is included by adding dummy variables for mortgages with LTI ratio ≤ 1.5, between 1.5 and 2, between 2 and 2.5, between 2.5 and 3, between 3 and 3.5 and > 3.5. The loan-to-value (LTV) ratio is included by dummy variables for mortgages with an LTV $\leq 80\%$, between 80% and 90%, between 90% and 97% and $> 97\%$.

Census Tract Demographics: I control for the median income as well as the proportion of adults over 25 with at least a high school diploma. These are from the

[33] A robustness check shows that the results are unaffected when including observations from Pima county and dropping the control variables with incomplete field population from the empirical model.

Table 10: Number of Observations by County and Year

	No Integrated Lender		Has Integrated Lender				*Total*
			Integrated Lender		*Other Lender*		
	No.	%	No.	%	No.	%	No.
County							
Cochise	94	63.9	35	23.8	18	12.2	147
Coconino	154	100.0	0	0.0	0	0.0	154
Maricopa	12,367	15.1	50,263	61.3	19,377	23.6	82,007
Mohave	27	26.0	52	50.0	25	24.0	104
Pima	9	20.0	20	44.4	16	35.6	45
Pinal	1,113	6.1	13,184	71.7	4,081	22.2	18,378
Yavapai	609	70.8	171	19.9	80	9.3	860
Yuma	963	85.8	91	8.1	69	6.1	1,123
Total	15,336	14.9	63,816	62.1	23,666	23.0	102,818
Year Sold							
2000	1,896	20.4	5,327	57.3	2,075	22.3	9,298
2001	2,088	17.7	7,596	64.4	2,111	17.9	11,795
2002	1,759	16.0	7,231	65.7	2,009	18.3	10,999
2003	2,060	16.2	8,083	63.5	2,582	20.3	12,725
2004	2,700	16.4	9,238	56.0	4,567	27.7	16,505
2005	2,729	17.0	9,155	57.2	4,134	25.8	16,018
2006	1,320	9.4	8,792	62.8	3,877	27.7	13,989
2007	784	6.8	8,394	73.1	2,311	20.1	11,489
Total	15,336	14.9	63,816	62.1	23,666	23.0	102,818

Note: This table shows the number of observations in the primary dataset used in this paper. It includes observations with a successful HDMA merge and a full set of covariates.

2005 - 2009 estimates of the American Community Survey. I control for census tract demographics by including dummy variables for the following (roughly equally sized) median income groups: $\leq$ \$35k, \$35k - \$50k, \$50k - \$65k, \$65k - \$75k, \$75k - \$100k and $\geq$ \$100k. Dummy variables for high-school graduation rates are: $\leq 75\%, 75\% - 80\%, 80\% - 90\%, 90\% - 95\%, \geq 95\%$.

B.5 Tax Assessment Process in Arizona

Arizona Revised Statues (A.R.S) 42-11054 (C) tasks tax assessors to annually compute the so-called "full cash value" of each residential property. A.R.S 42-11001(6) specifies the full cash value to be "synonymous with market value, which means the estimate of value that is derived annually by using standard appraisal methods and techniques." The full cash value provided in the tax assessment records is set at 82% of the assessed market value for residential properties. The procedure for arriving at these valuations is described by the assessor of Mohave County, Arizona, as follows: "Between January

Table 11: Summary Statistics of Control Variables

	No Integrated Lender			Has Integrated Lender			
	Mean	SD	Med.	Mean	SD	Med.	ΔIL
Housing Characteristics							
Sales Price (k year-2000$)	232.3	133.4	192.4	204.9	100.0	179.4	-3.14
Lot Size (Sqft)	8,594	5,191	7,694	7,306	3,498	6,576	-44.5
Building Area (Sqft)	2,280	855	2,111	2,164	735	2,023	-53.1
Price / Sqft (k year-2000$)	32.8	19.3	27.8	33.1	15.5	29.3	-0.42
Garage Spaces	1.97	1.10	2	1.82	1.06	2	-0.01
Total Rooms	7.59	1.91	7	7.18	1.70	7	-0.14
Has Pool	0.31			0.21			0.01
Owner Occupied	0.78			0.79			0.06
Financing Characteristics							
LTV Ratio	0.83	0.14	0.80	0.85	0.14	0.84	-0.002
LTI Ratio	2.52	1.06	2.44	2.69	1.09	2.61	0.19
Mortgage Duration (Years)	29.4	3.48	30	29.6	2.83	30	-0.22
FHA Insured	0.09			0.16			0.07
VA Insured	0.04			0.04			0.009
Jumbo Mortgages	0.008			0.002			-0.01
Borrower Characteristics							
Income (k year-2000$)	90.2	76.4	70.6	75.8	60.1	61.5	-11.4
Single Person	0.34			0.38			-0.06
Latino	0.13			0.13			-0.02
Asian	0.04			0.04			-0.01
Census Tract Demographics							
Med. Census Tract Inc. (k $)	76.5	22.1	76.9	75.8	17.7	72.2	
Highschool Grad. Rate	0.90	0.09	0.93	0.89	0.08	0.91	

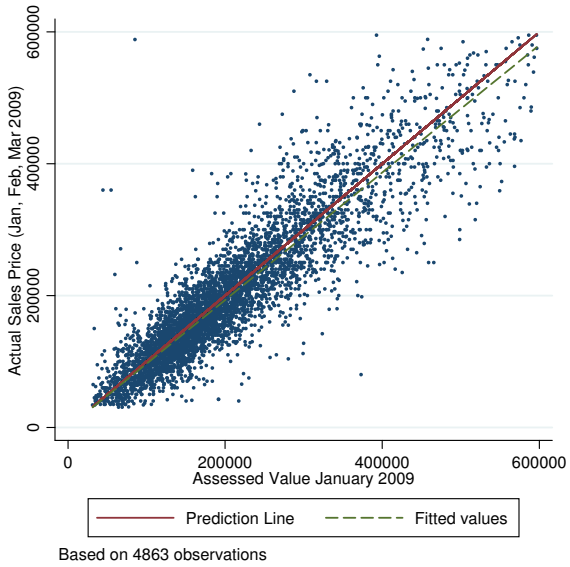
Note: This table shows mean, median and standard deviation for control variables. The last column shows the coefficient on IL in a regression $Characteristic_i = \alpha + \beta IL_i + \psi_{Year \times Development}$ for developments with an integrated lender.

and March of each year, the Assessors Office is required (by Arizona State Statute) to notify property owners of their assessed values for the following tax year. For residential and land parcels, this is accomplished by first collecting sales data in the area in which a property is located. Elements of comparability such as location, view, size, quality and condition are taken into consideration, and a mass appraisal mathematical model is used to arrive at each parcel's value. The market is driven by actual sales that have

occurred in a time window established by Department of Revenue guidelines. Increases or decreases in sale prices impact the final assessed valuation."

There exist a number of procedures through which a homeowner can challenge a tax assessment if she feels that the house was valued too highly. The appeals process provides a mechanism through which the assessor obtains information about differential depreciation of housing units. According to The Arizona Republic (2009), in 2009 there were 19,801 assessment appeals in Maricopa County, up from 17,213 in 2008 and 13,251 in 2007. This means that about 1.3% of valuations get appealed annually. In 2008, Maricopa County assessors reduced property valuations by a total of $1.9 billion, while the second stage of the appeals process, the Arizona State Board of Equalization, reduced valuations by an additional $2 billion. In the following I test how well these assessed values in Arizona capture true market values. To do this I consider those properties that were sold in an armslength transaction between January and March 2009 (in this section I use all sales in Arizona, not just those pertaining to newly developed properties). I compare the transaction price with the assessed value in January 2009. In Figure 6, each dot represents such a transaction.

Figure 6: Quality of Assessment Values



Note: This figure test for the accuracy of the estimated market value in the assessment data. Each dot represents an observation of a house that was sold in the first three months of 2009 and for which I observe an assessed value in January 2009. On the horizontal axis is the assessed value and on the vertical axis the corresponding transaction price. The solid line represents the 45° line. The dashed line represents the linear prediction of a regression of sales price on transaction price.

The solid line represents the 45° line - if assessments were 100% correct, all observations would lie on this line. It is not suprising that there is a significant spread around the 45° line. Unlike homogenous goods such as stocks and bonds, houses are heterogeneous assets that are sold in a search market. By adjusting the time that a seller is prepared to wait, she can influence the final transaction price. The dashed line represents the prediction from an ordinary least squares regression. The fact that it is very close to the 45° line suggests that on average assessed values capture current market values reasonably well.

B.6 Soil Data

Figure 7 shows a map of a representative housing development in Arizona. Each blue circle (●) and red cross (✚) represents a sale by one of two developers building in this development that appears in my dataset. The right panel also presents the soil type for each house. Houses built on the light gray, striped land are built on expansive soil while houses built on the dark green land are not built on expansive soil.

Figure 7: Map of Representative Development and Soil Type