The Failure of Models That Predict Failure: Distance, Incentives and Defaults*

Uday Rajan† Amit Seru‡ Vikrant Vig§

August 2, 2010

*For helpful comments and discussions, we thank Viral Acharya, Anat Admati, Sumit Agarwal, Patrick Bolton, Charles Calomiris, John Cochrane, Douglas Diamond, Martin Eichenbaum, Dirk Jenter, Anil Kashyap, Tobias Moskowitz, Loriana Pelizzon, Amiyatosh Purnanandam, Raghuram Rajan, Andrei Shleifer, Jeremy Stein, Ilya Strebulaev, Richard Thaler, Robert Vishny, Luigi Zingales, and participants at seminars at Bank of London, Berkeley, Board of Governors, BYU, Chicago FRB, Columbia, Harvard, Houston, LSE, Michigan, Michigan State, MIT Sloan, Naples, Philadelphia FRB, Stanford, UCLA, Utah and at the AEA, ALEA, Caesarea Center, EFA, ISB, LBS/LSE Credit Risk, NBER Behavioral, NBER Summer Institute, Southwind Finance and WFA conferences. Seru thanks the Initiative on Global Markets at the University of Chicago for financial support. Part of this work was undertaken when Seru was a visiting scholar at Sorin Capital Management. We are also indebted to Tanmoy Mukherjee for extensive discussions. All errors are our responsibility.

†Ross School of Business, University of Michigan, E-mail: urajan@umich.edu.

‡Booth School of Business, University of Chicago, E-mail: amit.seru@chicagobooth.edu.

§London Business School, E-mail: vvig@london.edu.

The Failure of Models That Predict Failure: Distance, Incentives and Defaults

Abstract

Statistical default models, widely used to assess default risk, are subject to a Lucas critique. We demonstrate this phenomenon using data on securitized subprime mortgages issued in the period 1997–2006. As the level of securitization increases, lenders have an incentive to originate loans that rate high based on characteristics that are reported to investors, even if other unreported variables imply a lower borrower quality. Consistent with this behavior, we find that over time lenders set interest rates only on the basis of variables that are reported to investors, ignoring other credit-relevant information. The change in lender behavior alters the data generating process by transforming the mapping from observables to loan defaults. To illustrate this effect, we show that a statistical default model estimated in a low securitization period breaks down in a high securitization period in a systematic manner: it underpredicts defaults among borrowers for whom soft information is more valuable. Regulations that rely on such models to assess default risk may therefore be undermined by the actions of market participants.

I Introduction

The true quality of a loan is not directly observed by third parties such as regulators or rating agencies. It is common in such settings to analyze historical data and uncover a statistical relationship between observed characteristics of a loan and its long-term quality. Regulators setting capital requirements use statistical models to forecast default rates on loans made by banks, as do rating agencies that assess the risk of underlying collateral in instruments such as CDOs.[1] The central thesis of this paper is that these predictive models are subject to a Lucas critique: they fail to account for a change in the relationship between observable characteristics of a good and its long-term quality that is caused by a fundamental change in the behavior of economic agents that produce the good.[2]

We demonstrate this phenomenon in the context of subprime mortgage loans issued in the US over the period 1997–2006. A notable feature of this period is a change in the nature of lending due to securitization, from “originate and hold” to “originate and distribute.” By increasing the distance between a home-owner and the ultimate investor, securitization changes the incentives of lenders in the following manner. The contract between investors of securitized loans and a lender is based on only a subset of observable characteristics of the loans. Some information that is potentially verifiable (perhaps at a cost) is excluded from the contract, and is not reported to investors.[3] Consequently, there is moral hazard: the lender originates loans that rate high based on the characteristics that affect its compensation, even if the unreported information implies a lower quality.[4] Keeping fixed the characteristics of loans observable to investors, the quality of the loan pool worsens. Although rational investors will anticipate this effect and price loans accordingly, statistical models estimated on past data ignore the change

[1]The Basel II guidelines mention that regulators may either allow banks to use their own probability of default (PD) models or use those offered by an independent third party such as a rating agency. Each rating agency typically has its own statistical model. For example, the Standard and Poor’s web site mentions an S&P LEVELS® 6.1 Model that estimates defaults on subprime mortgage loans.

[2]In the context of monetary policy, Gali and Gertler (2007) discuss the impact the Lucas critique has had on models used by central bankers across the world.

[3]Bolton and Faure-Grimaud (2009) and Tirole (2009) argue that contracts will be endogenously incomplete when there are costs involved in verifying or processing information. Along similar lines, Stein (2002) draws a distinction between hard (verifiable) and soft (unverifiable) information. One can think of the latter as being verifiable only at an infinite cost; it cannot be communicated to a third party, and so cannot be contracted on.

[4]A similar tension exists in the multi-tasking framework of Holmström and Milgrom (1990): an agent compensated for specific tasks ignores other tasks that also affect the payoff of the principal.

in the mapping between observable characteristics and likelihood of default. As a result, regulators or other market participants naïvely using such models will make systematically erroneous inferences.

We show that all these phenomena are exhibited in our data set, which covers the majority of securitized subprime loans across all lenders in the US. There is a progressive increase over time in the securitization rate of subprime mortgage loans during the sample period. Although the data set contains all variables transmitted from the lender to the investor, we find that the interest rate on new loans relies increasingly on a small set of variables. Specifically, the R^2 of a regression of interest rates on borrower FICO scores and loan-to-value (LTV) ratios increases from 9% for loans issued in the period 1997–2000 to 46% for 2006 loans. Thus, over the years, the variables not reported to the investor become less important in determining the interest rate on a loan. Furthermore, conditioning on the FICO score, the standard deviation of interest rates on loans shrinks over time. The latter effect occurs especially for borrowers with low FICO scores, on whom unreported information (such as the likelihood of future income shocks) is more likely to be important.

The increased (decreased) reliance of interest rates on variables that are reported (unreported) to investors is confirmed more directly by examining data from a single large subprime lender, New Century Financial Corporation (NCFC).[5] The advantage of this data set over our main sample is that for each loan it includes an internal rating, which is a summary measure of the quality of a loan as perceived by NCFC. We expect the internal rating to be important in determining the interest rate on a loan.[6] The internal rating is available to us, but is not reported to an investor who buys the loan. Similar to the trend in the overall market, the proportion of newly-issued loans that are securitized by NCFC shows an increase over the years. Strikingly, we find that the interest rates offered by NCFC also rely increasingly on the FICO score and LTV ratio over time, at the expense of the internal rating measure. That is, in pricing a loan to a consumer, NCFC steadily reduces its dependence on variables that are not reported to the investor.

As it is costly to acquire information, we expect lenders to stop collecting information

[5]In 2006, NCFC had the second-highest market share in the US subprime mortgage market. See, for example, "New Century, Biggest Subprime Casualty, Goes Bankrupt," bloomberg.com, April 2, 2007.

[6]In the context of subprime auto loans, Einav, Jenkins and Levin (2008) show that the profitability of dealerships at a lender increases when they improve their use of information and employ an internal rating to screen borrowers and price loans.

they are no longer using. The result is a worsening in the quality of the loan pool, since borrowers are no longer being screened along the dimensions of information not being reported to the investor. Consequently, the mapping between interest rates and default behavior should contemporaneously change with the securitization level. To examine this prediction we assess whether the interest rate becomes a noisier predictor of loan defaults over time. Indeed, we find that it does.

More broadly, we expect the change in lender behavior will alter the data generating process by transforming the mapping from all observables to loan defaults. To illustrate this effect, we estimate a baseline statistical model of default for loans issued in a period with a low degree of securitization (1997–2000), using information reported by the lender to the investor. We show that the model underpredicts defaults on loans issued in a regime with high securitization (2001 onward), thus exemplifying the Lucas critique. The degree of underprediction is progressively more severe as securitization increases, indicating that for the same observables, the set of borrowers receiving loans worsens over time.

We expect the prediction errors to be particularly high when information not reported to the investor is valuable in assessing the quality of the borrower; that is, for borrowers with low FICO scores and high LTV ratios. Indeed, we find a systematic variation in the prediction errors, which increase as the borrower's FICO score falls and the LTV ratio increases. As a placebo test, we estimate a default model for low-documentation loans over a subset of the low securitization era, and examine its out-of-sample predictions on loans issued in 1999 and 2000 (also a low securitization period). The statistical model performs significantly better than in our main test, and in particular yields prediction errors that are approximately zero on average.

Our findings on the performance of a statistical default model may perhaps be influenced by other macro factors that have changed over time with securitization. We perform two cross-sectional tests to rule this out. First, we separately consider loans with full documentation and loans with low documentation. Full-documentation loans include information on a borrower's income and assets, so that there is less unreported information on such borrowers. As a result we expect the change in lender behavior to have less of an impact for these loans. Indeed, we find that the prediction errors from the default model in the high securitization era are lower for full-documentation loans. Second, following Keys, et al. (2010a, b), we exploit the fact that the ease and likelihood of securitization is greater for low-documentation loans with FICO scores just above 620 compared to those with FICO scores just below 620. In this narrow FICO range *within* low-documentation loans, there are no observable differences across the two sets of loans except the FICO scores. Yet, the prediction errors are greater for loans above 620 than loans below 620. That is, there is a greater change in the mapping between observables and loan defaults among loans that are easier to securitize.

A fall in house prices, or more broadly an economic decline, will contribute to an increase in default rates. It is important to note that we consider defaults only within two years of a loan being issued. Thus, we find that the default model underpredicts errors even in a period in which house prices were increasing (i.e., for loans issued in 2001–2004). Nevertheless, to explicitly account for the effect of house prices, we consider a stringent specification that both estimates the baseline model over a rolling window, and also explicitly accounts for the effects of changing house prices. We determine the statewide change in house prices for two years *after* the loan has been issued and include it as an explanatory variable in the default model (i.e., we assume perfect foresight on the part of regulators estimating the default model). Approximately 50% of the prediction error survives the new specification, and the qualitative results remain: a default model estimated in a low securitization regime continues to systematically underpredict defaults in a high securitization regime.

Our work directly implies that regulations based on statistical models will be undermined by the actions of market participants. For instance, the Basel II guidelines assign risk to asset classes relying in part on probability of default models.[7] We highlight the role of incentives in determining the riskiness of loans, and in turn affecting the performance of models used to determine capital requirements. Our findings suggest that a blind reliance on statistical default models will result in a failure to assess and regulate risks taken by financial institutions. Indeed, the regulation itself must be flexible enough for regulators to be able to adapt it to changing market circumstances (see Brunnermeier, et al. (2009) for another argument for flexible regulation).

More broadly, we identify a dimension of model risk (i.e., the risk of having an incorrect model) that cannot be corrected by mere application of statistical technique. The term "model risk" is often understood to refer to an incomplete set of data, conceptual errors in a model, or both. The focus in the literature has thus been on testing the consistency and robustness

[7]See, for example, Basel Committee on Banking Supervision (2006). Kashyap, Rajan and Stein (2008) provide a detailed perspective on the role of capital requirements in the subprime crisis.

of inputs that go into statistical models. Collecting more historical data, possibly on extreme (and rare) events, is a key corrections that is frequently suggested. However, when incentive effects lead to a change in the underlying regime, the coefficients from a statistical model estimated on past data have no validity going forward, regardless of how sophisticated the model is or how well it fits the prior data. Indeed, aggregating data from different regimes may exacerbate the problem.

Although a naïve regulator may not understand that the lending regime has changed, we expect that rational investors will price loans accurately in either regime. Our hypotheses do not depend in any way on investors being boundedly rational.[8] However, if investors too are naïve, prices of loans or CDO tranches will fail to suitably reflect the default risk in a given loan pool. If anything, this will exacerbate the tendency of lenders to stop screening borrowers on unreported information, leading to even greater underprediction of defaults.

The rest of this paper is organized as follows. We explain our hypotheses in Section II. The primary data set is described in Section III, and Section IV details the findings on interest rates increasingly relying on information reported to investors. Section IV.C describes our results on the data from New Century Financial Corporation. In Section V, we evaluate statistical default models, and Section VI discusses a few issues related to our findings. Section VII elaborates on the connections of our work with the existing literature and discusses policy implications of our findings.

II Hypothesis Development

The first step in the securitization of loans in the subprime mortgage market typically consists of an outright sale by the original lender to a third party. A lender who retains a loan has an incentive to acquire hard and soft information about the borrower and the property to determine the riskiness of the loan. However, a lender who sells a loan will focus only on the variables that are included in the contract and reported to investors. Variables that are excluded consist of both soft information, which cannot be verified by a third party, and information that can be communicated to investors but is costly for them to verify.

More formally, for loan application i made at time t, let X_{it} be a vector of hard information

[8] While we are agnostic on whether investors mis-predicted the riskiness of loans in the build-up to the subprime crisis, there is emerging evidence that CDO tranches may have been mispriced (see, for example, Faltin-Traeger, Johnson and Mayer, 2010).

variables reported to the investor if the loan is securitized, and Z_{it} a vector of variables the lender uses when it retains most of its loans but does not report to the investor if loans are securitized. The variables in Z_{it} will include hard information variables that are quantified and contained in the lender's own files, and soft information variables that are observed neither by investors nor the econometrician. On each loan application, the lender has two decisions to make: whether to approve the application, and, if it does extend a loan, what interest rate to charge. Let A_{it} be a binary variable set to 1 if the application is approved and 0 otherwise, and let r_{it} denote the interest rate on the loan.

A lender's incentives to acquire and use information not reported to investors will then depend on the ease with which it securitizes loans on average.[9] As Keys, et al. (2010b) document, the ease of securitization can have multiple dimensions, including the probability or likelihood that a loan issued by a lender will be securitized and the average time taken to sell a loan. In this paper, for brevity we use the terms "high level of securitization" or "high securitization regime" to more generally mean a greater ease of securitization along all dimensions.

Intuitively, in a low securitization regime, both the approval decision and the interest rate will depend on the variables X_{it} and Z_{it}. That is, we can write

$$\begin{aligned} A_{it} &= f(X_{it}, Z_{it}) \\ r_{it} &= g(X_{it}, Z_{it}). \end{aligned}$$

As the level of securitization increases, a lender transits from a regime in which it retains most of the loans it issues to one in which it sells most of its loans. As it is costly to acquire information and the lender's own compensation on sold loans does not depend on the unreported variables Z_{it}, in a high securitization regime the lender stops collecting these variables. Its decisions therefore depend only on X_{it}, the variables that are reported to the investor. That is,

$$\begin{aligned} A_{it} &= \tilde{f}(X_{it}) \\ r_{it} &= \tilde{g}(X_{it}), \end{aligned}$$

[9] We assume that, at the time a loan is issued, the lender does not know whether it will be securitized. In the subprime market, investors are typically offered a basket of loans and choose a subset of the basket. In addition, there is some quality checking through a comparison of loans sold by a lender and loans retained by it. It is difficult for lenders to cherry-pick loans to retain. See Keys, et al. (2010a) for a discussion on this point.

where we use the notation $\tilde{f}$ and $\tilde{g}$ to indicate that the mapping from the reported variables X_{it} to both the approval decision and the interest rate has changed after securitization.

Our first hypothesis is that, with increasing securitization, a focus on the variables X_{it} reported to the investor will lead to the offered interest rate relying to a greater extent on these variables. In a low securitization regime, if the interest rate is regressed only on the reported variables, the estimated equation is $r_{it} = \hat{g}(X_{it})$. Since the interest rate also depends on the omitted variables Z_{it}, such a regression should provide a poor fit. In a high securitization regime, such a regression should yield a better fit, since the lender uses only X_{it} in setting the interest rate.

We test this hypothesis in two ways in Section IV. First, we regress the interest rate on a loan on the FICO score and LTV ratio. We predict that the explanatory power of the right-hand side variables (i.e., the R^2 of the regression) will increase over time, since the level of securitization increases dramatically through time. Our second test considers the converse: if interest rates depend more on reported information as securitization increases, they must depend less on unreported information. Thus, keeping fixed the level of the reported variables such as the FICO score and the LTV ratio, interest rates should exhibit less dispersion at higher levels of securitization.

Our second hypothesis is that, once the lender starts to ignore the unreported information Z_{it} in its own decision on whether to offer a loan, the quality of the loan pool will worsen at any given level of the reported variables X_{it}. This effect occurs because there is a pooling across borrower types along the dimension of the unreported information Z_{it}.[10] In other words, the change in lender behavior alters the data generating process, with the mapping between default behavior and observables (including interest rates) changing as securitization increases.

This hypothesis is tested in two ways. First, we examine if interest rates become a worse predictor for defaults as securitization increases by considering a regression of loan defaults on interest rates year-by-year across our sample. The prediction is that the R^2 of this regression will fall over time as securitization increases, and that the coefficients will change in a manner consistent with a worsening in the quality of the loan pool.

Second, we directly examine the predictions of a statistical default model. We begin by constructing a statistical default model and fitting it to data from a low securitization period

[10] Rajan, Seru and Vig (2010) provide a formal theoretical model that develops the intuition, building on the literature on loan sales (such as Gorton and Pennacchi, 1995).

(1997–2000). Keeping the coefficients fixed, we then determine the predicted defaults for loans issued from 2001 onward. We expect the prediction errors (i.e., actual minus predicted defaults) to be positive on average, increase with securitization and to be larger for borrowers on whom the unreported information is more informative about quality (in particular, borrowers with low FICO credit scores and high loan-to-value ratios). It is important to note that the exact nature of the statistical model used to assess our prediction is not important. The changed mapping between observables and defaults should show up in any statistical model that does not account for changed lender behavior.

In Appendix A, we explain how the change in the data generating process can be understood using the selection model framework of Heckman (1980). The essence of the argument is that a regulator and rating agencies only see approved loans, which are a selected sample. As noted earlier, the approval process changes with lender incentives and behavior. Consequently, as securitization increases one expects the change in lender behavior to affect the loans that are selected into the approved pool, thereby altering the mapping from observables to defaults.

III Data

We use two sets of data in our analysis. Here, we describe the primary data set, which is used in the bulk of the paper. A second data set consisting of loans from a single lender, New Century Financial Corporation, is described more fully in Section IV.C.

Our primary data set contains loan-level information on securitized non-agency mortgage loans. The data include information on issuers, broker dealers, deal underwriters, servicers, master servicers, bond and trust administrators, trustees, and other third parties. As of December 2006, more than 8,000 home equity and nonprime loan pools (over 7,000 active) that include 16.5 million loans (more than 7 million active) with over \$1.6 trillion in outstanding balances are included. Estimates from the data vendor suggest that as of 2006, the data cover over 90% of the subprime loans that have been securitized. As Mayer and Pence (2008) point out, there is no universally accepted definition of "subprime."[11] Broadly, a borrower is classified as subprime if she has had a recent negative credit event. Occasionally, a lender signals a borrower with a good credit score is subprime, by charging higher than usual fees on a loan. In our data, the vendor identifies loans as subprime or Alt-A (thought to be less risky

[11]Chomsisengphet and Pennington-Cross (2006) provide a history of the subprime market.

than subprime, but riskier than agency loans).

The data set contains all variables obtained from the issuer by the investor, including the loan amount, maturity, loan-to-value (LTV) ratio, borrower credit score, interest rate, and other terms of the loan contract. The FICO credit score is a summary measure of the borrower's credit quality. This score is calculated using information about the borrower's credit history (such as the amounts of various types of debt outstanding), but not about her income or assets (see, for example, Fishelson-Holstein, 2004). The software used to generate the score from individual credit reports is licensed by the Fair Isaac Corporation to the three major credit repositories, TransUnion, Experian, and Equifax. FICO scores provide a ranking of potential borrowers by the probability of having any negative credit event in the next two years. Probabilities are rescaled as whole numbers in a range of 400–900 (though nearly all scores in our data are between 500 and 800), with a higher score implying a lower probability of a negative event.

The loan-to-value ratio (LTV) of the loan, which measures the amount of the loan expressed as a percentage of the value of the home, also serves as a signal of borrower quality. For borrowers who do not obtain a second lien on the home, the LTV ratio provides a proxy for wealth. Those who choose low LTV loans are likely to have greater wealth and hence are less likely to default.

Borrower quality can also be gauged by the extent of documentation collected by the lender when approving the loan. The various levels are categorized as full, limited or no documentation. Borrowers with full documentation provide verification of income as well as assets. Borrowers with limited documentation provide no information about income and some information about their assets. No-documentation borrowers provide no information about income or assets. In our analysis, we combine limited- and no-documentation borrowers and call them "low-documentation" borrowers. Our results are unchanged if we remove the small proportion of loans which have no documentation.

Other variables include the type of the mortgage loan (fixed rate, adjustable rate, balloon or hybrid), and whether the loan is provided for the purchase of a principal residence, to refinance an existing loan, or to buy an additional property. We present results exclusively on loans for first-time home purchases. We ignore loans on investment properties, which are more speculative in nature, and likely to come from wealthier borrowers. The zip code of the property associated with each loan is included in the data set. Finally, there is also information about the property being financed by the borrower, and the purpose of the loan. As most loans in the data set are for owner-occupied single-family residences, townhouses, or condominiums, we restrict the loans in our sample to these groups. We also exclude non-conventional properties, such as those that are FHA or VA insured, pledged properties, and buy down mortgages.

We report year-by-year summary statistics on FICO scores and LTV ratios in Table I. The number of securitized subprime loans increases more than fourfold from 2001 to 2006. This pattern is similar to that described by Demyanyk and Van Hemert (2009) and Gramlich (2007). The market has also witnessed an increase in the proportion of loans low (i.e., limited or no) documentation, from about 25% in 1997 to about 45% in 2006.

Table I: Summary Statistics, Primary Data Set

Year	Number of Loans	Proportion with Low Documentation (%)	Mean Loan-To-Value Ratio (%)	Mean FICO Score
1997	24,067	24.9	80.5	611
1998	60,094	23.0	81.5	605
1999	104,847	19.2	82.2	610
2000	116,778	23.5	82.3	603
2001	136,483	26.0	84.6	611
2002	162,501	32.8	85.6	624
2003	318,866	38.9	87.0	637
2004	610,753	40.8	86.6	639
2005	793,725	43.4	86.3	639
2006	614,820	44.0	87.0	636

LTV ratios have gone up over time, as borrowers have put in less equity into their homes at the initial purchase. The average FICO score of individuals who access the subprime market has been increasing over time, from 611 in 1997 to 636 in 2006. This increase in the average FICO score is consistent with a rule-of-thumb leading to a larger expansion of the market above the 620 threshold as documented in Keys et al. (2010a,b). Though not reported in the table, average LTV ratios are lower and FICO scores higher for low-documentation loans, as compared to the full-documentation sample. This possibly reflects the additional uncertainty lenders have about the quality of low-documentation borrowers. The trends for loan-to-value ratios and FICO scores in the two documentation groups are similar.

IV Increased Reliance on Reported Information

In Table II, we report the proportion of newly-issued subprime mortgage loans that are securitized in each period. The second row shows the overall securitization rate in the market, and the third row the securitization rate for a single lender, New Century Financial Corporation (NCFC). As shown in the table, both the overall market and NCFC experience a steady increase in the securitization rate over time. For the overall market, the securitization rate climbs from 37% in the period 1997–2000 to 76% in 2004, and even higher in 2006. A common explanation for this trend (see, for example, Greenspan, 2008) is a surge in investor demand for securitized loans over this period. Due to an unprecedented budget surplus, the US Treasury engaged in a buyback program for 30-year bonds in 2000–01, and ceased to issue new 30-year bonds between August 2001 and February 2006.[12] Coincidentally, there was a rapid increase in CDO volume over this period, with a significant proportion containing subprime assets.[13]

Table II: Securitization Rate Over Time (%)

Year	1997–2000	2001	2002	2003	2004	2005	2006
Overall Market	37	58	62	66	76	79	85
NCFC Loans	41	50	77	88	92	85	96

Note: The yearly securitization proportion for the overall market is obtained from *Inside B&C Lending*, a publication that has extensive coverage of the subprime mortgage market. Data on NCFC securitization rates comes from the origination and servicing loan files that encompass all lending activities of NCFC from 1997 to 2007.

The bulk of our tests compare outcomes across time, and examine whether incremental effects of increased securitization can be observed in the aggregate data. We consider the period 1997–2000 to be a low securitization regime, and the period 2001 and later to involve high securitization.[14] In what follows, we use the term "year-by-year" regression to refer to separate regressions for the combined period 1997–2000 and for each year from 2001 to 2006.

It is important to remember that lenders in this market are heterogeneous, and include

[12] "30-Year Treasury Bond Returns and Demand Is Strong," the *New York Times*, Feb 9, 2006.

[13] The volume of CDOs issued in 2006 reached $386 billion, with home equity loans (largely from the subprime sector) providing for 26% of the underlying assets (from "Factbox - CDOs: ABS and other sundry collateral," reuters.com, June 28, 2007).

[14] In the overall market, the securitization rate over the period 1997 to 2000 remains between 33 and 41%. Since the volume of loans in each year in this period is also lower than in the later years, we combine these years in the rest of our analysis.

commercial banks, thrifts, independent mortgage companies, and bank subsidiaries (see, for example, Gramlich, 2007). We expect that different lenders would cross over from a low to a high degree of securitization at different points of time. In addition, there may be new lenders entering the market over time. In both cases, we expect a lender securitizing a large proportion of loans to rely primarily on the variables reported to investors when setting the interest rate on a loan. In the time series for the aggregate loan market, such behavior will imply that in the whole sample, the interest rate on new loans relies increasingly on the reported variables over time. We proceed to test this prediction.

IV.A Regression of Interest Rate on Reported Variables: All Subprime Securitized Loans

A direct way to capture the importance of the reported variables on the lender's behavior is to consider the R^2 of a year-by-year regression of interest rates on new loans on key variables. An increase in the R^2 of the regression over time indicates an increased reliance on reported variables, whereas a decrease suggests an increased reliance on variables not reported to the investor.

We estimate the following regression year-by-year as our base model:

$$r_i = \beta_0 + \beta_{FICO} \times FICO_i + \beta_{LTV} \times LTV_i + \epsilon_i. \tag{1}$$

Here, r_i is the interest rate on loan i, $FICO_i$ the FICO score of the borrower, LTV_i the LTV ratio on loan i, and ϵ_i an error term.

We report β_{FICO}, β_{LTV} and the R^2 of the regression in Table III. Consistent with our first prediction, column 5 of the table shows that there is a dramatic increase in the R^2 of this regression over the years. Starting from about 9% in 1997–2000, the R^2 increases to 46.7% by the end of the sample. As expected, β_{FICO} is consistently negative (higher FICO scores obtain lower interest rates), and β_{LTV} is consistently positive (higher LTV ratios result in higher interest rates).[15]

We next add dummy variables for three important features of the loan contract as explanatory variables to the base model: whether the loan is an Adjustable Rate Mortgage

[15] Note that the variance of FICO and LTV observed in the sample varies across years. As a result the coefficients across years are not readily comparable. We re-ran the base model after standardizing the interest rate, FICO score, and LTV ratio. The trend in R^2 is similar for the standardized regression, and, as may be expected from the increase in the R^2, the coefficients increase in magnitude over time.

Table III: Reliance of Interest Rates on FICO Scores and LTV Ratios, All Securitized Loans

Year	Base Model Coefficients		No. Obs.	Adjusted R² (%) of Various Models		
	β_{FICO}	β_{LTV}		Base Model	With Additional Contract Variables	Including Top 102 Lenders Only
1997–2000	-0.009***	0.033***	305,786	8.98	11.38	8.40
	(.0001)	(.0003)				
2001	-0.012***	0.038***	136,483	19.49	22.74	20.13
	(.0001)	(.0004)				
2002	-0.011***	0.071***	162,501	17.42	26.43	15.66
	(.0001)	(.0001)				
2003	-0.012***	0.079***	318,866	29.72	41.26	33.29
	(.0001)	(.0001)				
2004	-0.010***	0.097***	610,753	36.85	45.39	41.00
	(.0001)	(.0001)				
2005	-0.009***	0.110***	793,725	43.91	50.14	52.82
	(.0001)	(.0001)				
2006	-0.011***	0.115***	614,820	46.67	50.83	46.72
	(.0001)	(.0001)				

Note: Standard errors are in parentheses. *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level.

(ARMs generally have low initial "teaser" rates), whether the loan has low documentation (full-documentation loans have lower interest rates), and whether there is a prepayment penalty. The R^2 of the enhanced model is reported in the column 6 of Table III. The added dummy variables somewhat improve the R^2 of the regression, but clearly preserve the trend, with the R^2 increasing from 11.4% in 1997–2000 to 50.8% in 2006. Although not reported in the table, the coefficients on the FICO score and LTV ratio for the regressions in the last two columns of the table are similar to those of the base model.

One concern may be that the results in the base model are driven by a change in lender composition over time rather than a change in lender behavior. To alleviate this concern, we estimate the base model using a fixed set of lenders across the sample period. There are several thousand lenders in the sample, each identified by name.[16] Most lenders are small; the largest 102 lenders account for 78% of the data, and the largest 700 lenders for approximately 90% of

[16]The process of matching lenders to loans is somewhat cumbersome, since the same lender is sometimes referred to by slightly different names. For example, New Century Financial Corporation is sometimes referred to as New Century, NCF, and NCFC.

the data. We re-run the regression including only the top 102 lenders, and report the results in the last column of Table III. As seen from the table, the R^2 displays the same trend as in the base model, suggesting that underlying our results is a change in lender behavior.

Finally, we also estimate equation (1) separately for loans with low documentation and those with full documentation, to ensure that our results are not being driven simply by a change in the composition of loans over time. The trend in the R^2 is similar across both sets of loans. For brevity, the results are not reported in the table.[17]

Overall, in the low securitization regime (1997–2000), the variables reported to the investor explain very little variation in interest rates. The clear suggestion is that the unreported variables are particularly important in these years. As the securitization regime shifts, the same reported variables explain a large amount of variation in interest rates. Our results are thus consistent with the notion that the importance of variables not reported to the investor in determining interest rates on new loans declines with securitization.[18]

IV.B Shrinkage of the Distribution of Interest Rates

Another way to test the relationship between included information and interest rates is to consider the dispersion of interest rates at different values of a reported variable. We calculate the standard deviation of interest rates at each FICO score and track it over time. Let $\sigma_{it} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(r_{ijt} - \bar{r}_{it})^2}$, where r_{ijt} is the interest rate on the j^{th} loan with FICO score i in year t, and $\bar{r}_{ijt} = \frac{1}{N}\sum_{j=1}^{N} r_{ijt}$ is the mean interest rate. We pool observations into FICO score buckets of 10 points starting from a score of 500 and ending at 800 (i.e., the buckets are FICO scores 500-509, 510-519,...). We then estimate the following regression separately for each bucket b:

$$\sigma_{bt} = \alpha_b + \beta_b \times t + \epsilon_{bt}, \tag{2}$$

[17] Another factor to consider is that, during the sample period, there were some bank mergers. As banks become become large, interest rates will depend more on hard information, due to the effects identified by Stein (2002). To rule out this explanation, we re-estimate equation (1) only for banks that did not engage in mergers over the sample period, and obtain similar results.

[18] In a different context, Cole, Goldberg and White (1998) and Liberti and Mian (2009) find that loan offers to firms by large banks and at higher levels within a bank are more sensitive to financial statement variables, consistent with the notion of Stein (2002) that soft information cannot be communicated up the hierarchy within a firm. Berger, et al. (2005) conduct a more indirect test on small business lending and find consistent results.

where t indexes year and ϵ_{bt} is an error term. The coefficient β_b captures how the dispersion of interest rates within each FICO score bucket changes over time. We expect β_b to be large and negative for low FICO scores, i.e., we expect a shrinkage of dispersion in interest rates at low FICO scores. Information not reported to investors is likely to be more important in assessing the quality of such borrowers, compared to those with high FICO scores.

For loans at low FICO scores (500–599), we find β_b to be about -0.15 (which translates to about a 6.8% reduction per year in the dispersion of interest rates). For higher FICO scores (600 and above), β_b is about -0.05 (a 2.5% reduction per year in the dispersion of interest rates).[19] The magnitude of shrinkage can also be interpreted relative to the mean interest rate. Across sample years, the mean interest rate is 9.2% at FICO scores 500–599 and 8.1% at FICO scores 600 and higher. Thus, scaling the degree of shrinkage by the mean interest rate yields the same results.

We conduct an additional test to rule out the hypothesis that the shrinkage in the dispersion of interest rates may occur due to standardization of mortgage contract terms over time. We extend equation (2) to condition for shrinkage in the dispersion of not just the loan-to-value ratio, but also other contractual terms (including whether the loan is an ARMs and the presence of a prepayment penalty) at each FICO score in each year. The results of this estimation (unreported for brevity) are similar to those reported in Table VII.

IV.C Evidence from a Single Lender: New Century Financial Corporation

In our primary data set, we do not observe variables that are *not* reported to investors, so we cannot directly demonstrate that the reliance on these variables reduces over time. We now examine data from a single lender, New Century Financial Corporation (NCFC), which both confirm and enhance our findings. NCFC was a large subprime mortgage lender that filed for bankruptcy in April, 2007.

The NCFC data have two distinctive features that allow us to test our first hypothesis more extensively. First, the data contain both accepted and rejected loan applications, and both securitized loans and loans retained by NCFC. This allows us to directly consider the accept/reject decision, and also to compute the proportion of securitized loans in each year. Second, and more importantly, the dataset includes several variables that are not passed to investors but are observed by NCFC. Most important of these is an internal rating measure,

[19] Table VII in Appendix B reports the β_b coefficient for each FICO bucket.

which we call “Rating”. The Rating on each loan is a summary measure based on all information about the borrower and property observed by NCFC. The latter information includes variables that were passed on to investors (such as the FICO score and the LTV ratio). The Rating measure ranges between 1 (best quality loan) and 20 (worst quality loan). Importantly, the measure is correlated with numerous variables contained in the NCFC data set (and therefore observed by NCFC) that are not reported to investors, including whether the borrower is self-employed, is married, has been referred by an existing customer, and has other debt in addition to the mortgage. We expect the rating measure to also capture soft information observed by NCFC but unobservable to both investors and the econometrician (such as a loan officer's assessment of default likelihood based on a personal interview with the borrower).

In second row of Table II, we report the proportion of loans issued by NCFC each year that are securitized. The results are consistent with the trend in the overall market: the proportion of securitized loans increases from 41% in the period 1997–2000 to 92% in 2004 and 96% in 2006. The overall summary statistics for securitized loans issued by NCFC are also similar to those reported for the aggregate market in Table I. For example, the mean FICO score is 611 in the period 1997–2000, and 636 in 2006. Similarly, the mean LTV ratio is 79% in 1997–2000 and 85% in 2006.

To examine whether NCFC increasingly relies on the variables reported to the investor (specifically, the FICO score and the LTV ratio) in setting the interest rate on new loans, we estimate our base model in equation (1) on first-lien loans in the NCFC data, applying the same filters as in the main sample. The results are shown in Panel A of Table IV. The increase in the R^2 of the regression, from 10.8% in 1997–2000 to 28.1% in 2004, has a similar pattern to that shown for the aggregate market in Table III, though the magnitude of the increase is somewhat smaller.

We now conduct two tests which directly provide evidence that the internal Rating measure, which encapsulates several of the variables not reported to investors, increasingly becomes less important in the decisions made by NCFC. In the last column of Panel A of the Table IV, we show the R^2 of the regression when Rating is added as an explanatory variable. The improvement in R^2 over the base model is about 50% for the period 1997–2000, and falls to 5% or less in the years 2004 through 2006. The results are therefore strongly consistent with NCFC abandoning its internal rating measure in setting interest rates, and relying instead on

Table IV: Results from New Century Financial Corporation Data

Panel A: OLS regression of interest rate on FICO and LTV

Year	Base Model Coefficients		Observations	Adjusted R^2 (%)	
	β_{FICO}	β_{LTV}		Base Model	Model Including Internal Rating
1997–2000	-0.0053***	0.014***	21,553	10.8	16.3
	(0.0001)	(0.0008)			
2001	-0.0072***	0.013***	7,302	12.9	18.9
	(0.0002)	(0.0016)			
2002	-0.0084***	0.009***	15,092	19.5	24.5
	(0.0001)	(0.0010)			
2003	-0.0085***	0.020***	33,690	25.1	28.6
	(0.0001)	(0.0006)			
2004	-0.0075***	0.050***	63,174	28.1	29.3
	(0.0001)	(0.0005)			
2005	-0.0062***	0.060***	84,002	23.9	24.4
	(0.0001)	(0.0005)			
2006	-.0064***	0.066***	82,163	27.4	28.0
	(0.0001)	(0.0005)			

Panel B: Logit regression of accept/reject decision on internal rating measure

Year	β_{Rating}	Observations	Pseudo-R^2 (%)
1997–2000	-0.053***	60,049	1.00
	(0.002)		
2001	-0.059***	14,905	1.12
	(0.004)		
2002	-0.070***	29,656	1.08
	(0.003)		
2003	-0.097***	71,188	0.76
	(0.004)		
2004	-0.075***	154,893	0.21
	(0.004)		
2005	-0.080***	199,369	0.16
	(0.004)		
2006	-0.056***	210,856	0.09
	(0.004)		

Note: Standard errors are in parentheses. *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level. The β_{Rating} in Panel B is a logit coefficient.

the FICO score and the LTV ratio.[20]

Next, we estimate a logit regression of the accept or reject decision on the internal rating

[20] Although not reported in the table, the coefficients on FICO score and LTV ratio are similar to those in the base model.

measure. The regression equation here is

$$Accept_{it} = \Phi(\beta_0 + \beta_{\text{Rating}}\, Rating_{it}), \tag{3}$$

where $Accept_{it}$ is a binary variable equal to 1 if loan application i at time t was accepted, and 0 otherwise, $Rating_{it}$ is the internal rating of application i at time t, and $\Phi(\cdot)$ is the logistic distribution function. The results are reported in Panel B of Table IV. While the coefficient on Rating remains statistically significant in each year of the sample, the pseudo-R^2 of the regression falls from 1% or higher in the period 1997 through 2002 to 0.2% in 2004 and 0.09% in 2006. Therefore, over time, the internal rating measure becomes less important in the selection process for new loans.[21]

One may conjecture that the patterns observed both in the main and the NCFC data merely reflect that the FICO score is becoming a better predictor of defaults over time. If that were correct, lenders would need to collect and use less additional information in later years. However, we should then find that the FICO score becomes a better predictor of contemporaneous defaults over time. We estimate a logit regression of loan default within 24 months of origination on the FICO score, and find the exact converse. The pseudo-R^2 of the regression progressively falls from about 5% (3.9%) in 1997–2000 to 0.01% (1.1%) in 2006 in the main (NCFC) data. Thus, we find that over time the FICO score becomes a poorer rather than a better predictor of loan defaults.

V Empirical Default Model

We now consider the effect of securitization on mortgage defaults. Following the arguments in Section II, we have two predictions on the default rates of loans. First, the ability of the interest rate to predict defaults should fall over time as information not being reported to the investor is no longer collected by the lender. Thus, in a year-by-year regression of default rates on interest rates, the R^2 should decrease over time. Second, the quality of the loan pool should worsen, keeping fixed the observable characteristics of a loan. To test this second prediction, we estimate a baseline statistical model using observables from a low securitization regime. We expect this baseline model to underpredict defaults under high securitization for borrowers on

[21]Consistent with our other results, the accept/reject decision increasingly relies on the FICO score and LTV ratio over time. In a similar vein, when we regress loan defaults on the internal rating measure, we find that the measure progressively becomes a noisier predictor of defaults.

whom information not reported to investors is likely to be important in assessing quality; i.e., borrowers with low FICO scores and high LTV ratios.

V.A Ability of Interest Rates to Predict Defaults

We examine the default experience of loans by issue year, assigning a variable $\textit{Actual Default}_{it} = 1$ if loan i issued in year t defaults within 24 months of issue, and zero otherwise. Here, default is defined to be the event that the loan is delinquent for at least 90 days. FICO scores are designed to predict negative credit events over the next two years.[22] Further, 24 months is before the first reset date of the most common types of ARMs in this market. We therefore restrict attention to defaults that occur within 24 months of loan origination.

The actual default experience on a loan in the two years beyond issue will depend on many factors, including local and macro-economic conditions and idiosyncratic shocks to the borrower's financial status. At the time the loan is issued, the interest rate on the loan reflects the lender's estimate of the overall likelihood the loan will default at some later point. It captures both what the lender knows about the riskiness of the borrower and the lender's forecast about future economic conditions that may influence default. Thus, we expect that the interest rate on a loan will be the most important predictor of whether the loan defaults.

Our hypothesis is that the interest rate loses its ability to predict defaults over time. We expect the loss of predictive ability to be more pronounced when the information not reported to the investor is more economically relevant, that is, for low-documentation loans and loans to borrowers at the lower part of the credit distribution. We therefore consider low- and full-documentation loans separately in our test, and focus on the change in sensitivity of defaults to interest rates for borrowers at the 25^{th} percentile of the FICO score distribution.

We estimate the following year-by-year logit model:

$$\text{Prob}(\text{Actual Default}_{it} = 1) \; = \; \Phi(\beta_0 + \beta_r r_{it}), \tag{4}$$

where r_{it} is the interest rate on loan i issued at time t.

Table V shows the estimated coefficients and the pseudo-R^2 values. First, consider Panel A, which reports on low-documentation loans. Observe that the pseudo-R^2 consistently falls

[22] Holloway, MacDonald and Straka (1993) show that the ability of FICO scores observed at loan origination to predict mortgage defaults falls by about 25% once one moves to a three-to-five year performance window.

Table V: Contemporaneous Default Regressions

Panel A: Low-documentation Loans

Year	β_r	Constant (β_0)	Pseudo-R^2 (%)	Observations
1997–2000	0.282***	-4.996***	2.43	65,895
	(0.00920)	(0.0965)		
2001	0.333***	-5.159***	3.42	35,110
	(0.0112)	(0.113)		
2002	0.224***	-4.079***	2.54	52,967
	(0.00709)	(0.0689)		
2003	0.224***	-4.023***	2.21	123,766
	(0.00514)	(0.0442)		
2004	0.159***	-3.215***	1.12	248,839
	(0.00341)	(0.0282)		
2005	0.127***	-2.331***	0.73	343,581
	(0.00247)	(0.0208)		
2006	0.111***	-1.444***	0.65	270,284
	(0.00231)	(0.0215)		

Panel B: Full-documentation Loans

Year	β_r	Constant (β_0)	Pseudo-R^2 (%)	Observations
1997–2000	0.211***	-4.065***	1.94	231,103
	(0.00376)	(0.0409)		
2001	0.243***	-4.051***	2.61	98,751
	(0.00506)	(0.0534)		
2002	0.177***	-3.344***	1.88	107,648
	(0.00437)	(0.0422)		
2003	0.240***	-3.856***	2.93	194,010
	(0.00355)	(0.0307)		
2004	0.199***	-3.268***	1.83	360,646
	(0.00261)	(0.0212)		
2005	0.140***	-2.451***	0.92	448,422
	(0.00215)	(0.0177)		
2006	0.0858***	-1.689***	0.38	343,393
	(0.00216)	(0.0199)		

Note: Both panels show logit coefficients. Standard errors are in parentheses. *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level.

from 3.42% for 2001 vintage loans to 1.12% for 2004 vintage loans and 0.65 for 2006 vintage loans. Further, at the 25^{th} percentile of the FICO score distribution, a 1 standard deviation change in interest rate implies a change in default rate of about 4.2% in 2001, 2.0% in 2004 and 1.7% in 2006. That is, there is a decline in the sensitivity of defaults to interest rates in the later years of the sample, suggesting that interest rates are not responding as much to changes in the riskiness of a borrower. Of course, defaults on loans issued in 2005 and 2006 are high from July 2007 onward because of the financial crisis. Although these two years are arguably special, it is important to note that the trends in both R^2 and the marginal effects of the coefficients are observable even over the period 2001–2004.

The results on full-documentation loans are shown in Panel B of Table V. Among loans of vintage 2001 through 2004, there is no monotone pattern in the R^2 of the regression. Loans issued in 2005 and 2006 display the same trend as exhibited by low-documentation loans. Importantly, the marginal effect of the coefficients evaluated at the lower part of the credit distribution again suggests a progressive reduction in the sensitivity of interest rates to default risk. At the 25^{th} percentile of the FICO score, the marginal effect of a 1 standard deviation change in the interest rate on the default rate is about 3.8% in 2001, 2.7% in 2004 and 1.9% in 2006.

V.B Failure to Predict Failure: A Statistical Default Model

We now test whether the mapping between observables reported to the investors and loan defaults has changed, by evaluating how a statistical default model estimated on historical data from a low securitization regime performs as securitization increases. In particular, we examine if the statistical model produces positive errors on average, and whether these errors exhibit the systematic variation with observables predicted by our hypothesis.

V.B.1 Main Test

For our first direct test of default predictions, we consider the period 1997–2000 to be a low securitization era, and the period 2001–2006 to be a high securitization one. We estimate the following logit model on all securitized loans in our primary data set issued in the period 1997 to 2000:

$$\text{Prob}(\text{Actual Default}_i = 1) \;=\; \Phi(\beta \cdot X_i + \beta^{Low} \cdot I_i^{Low}\, X_i). \tag{5}$$

Here, X_i is a vector that includes the interest rate on the loan, the FICO credit score of the borrower, the LTV ratio, an ARM dummy, and a prepayment penalty dummy. I_i^{Low} is a dummy set to 1 if loan i has low documentation and 0 otherwise. We also include state fixed effects in the regression. This model resembles the LEVELS® 6.1 Model used by S & P. As mentioned before, what is important here is not the exact specification of the model, but its use of historical information without regard to the changing incentives of agents who produce the data. The latter feature is common to most models used by rating agencies or regulators.

Panel A of Table VI shows the estimated coefficients on the interest rate, FICO score and LTV ratio from the baseline model. A low interest rate and high credit score are both associated with lowering the probability that the borrower will default in the subsequent two years, for both full-documentation and low-documentation loans.

Next, we use the coefficients of the baseline model to predict the probability of default for loans issued from 2001 to 2006, where default again is an event that occurs up to two years after a loan is issued. Concretely, let $\hat{\beta}_{1,t}$ and $\hat{\beta}_{1,t}^{Low}$ be the coefficients estimated from equation (5) for the baseline model over the period 1 to t (where year 1 is 1997 and year t is 2000). Then, for $k = 1, 2, \cdots, 6$, we estimate the predicted probability that a loan i issued at $t + k$ will default in the next 24 months (keeping the baseline coefficients fixed) as $Predicted\ Default_{i,t+k} \equiv \mathrm{Prob}(\widehat{\mathrm{Default}}_{i,t+k} = 1)$, where:

$$\mathrm{Prob}(\widehat{\mathrm{Default}}_{i,t+k} = 1) = \Phi(\hat{\beta}_{1,t} \cdot X_{i,t+k} + \hat{\beta}_{1,t}^{Low} \cdot I_{i,t+k}^{Low} X_{i,t+k}).$$

We then examine the actual default experience of loans issued in each of years 2001 to 2006. The prediction error is computed as $Prediction\ Error_{i,t+k} = Actual\ Default_{i,t+k} - Predicted\ Default_{i,t+k}$.

If there is systematic underprediction at low FICO scores and high LTV ratios, the prediction error should decline in magnitude as the FICO score increases and LTV ratio falls. To check this, we estimate yearly the regression for borrower i in year $t + k$ (where $t = 2000$ and $k = 1, 2, \cdots, 6$) as follows:

$$Prediction\ Error_{i,t+k} = \alpha + \beta_{FICO} \times FICO_{i,t+k} + \beta_{LTV} \times LTV_{i,t+k}.$$

Panel B of Table VI reports the coefficients on the FICO scores and LTV ratio for loans issued in each of the years 2001 to 2006. As can be observed from columns 2 and 3, the coefficient β_{FICO} is negative while β_{LTV} is positive and significant across 2001 to 2006. The magnitudes seem large. For instance, an increase in one standard deviation in the FICO score (about 70 points) leads to a reduction in the prediction error of about 33.5% for 2006 loans. Similarly, a one standard deviation increase in LTV ratio (about 10%) leads to a reduction in prediction error of about 9.4% for 2006 loans.

Column 6 of Panel B in Table VI confirms that the prediction errors are positive. Further, the average prediction error increases over time as securitization increases, implying that the fit of the baseline model worsens over time. Moreover, the magnitudes of the prediction errors are large relative to actual defaults (reported in the last column). For instance, among loans

Table VI: Default Model: Failing to Predict Failure

Panel A: Coefficients of Baseline Model in Low Securitization Regime, 1997–2000

$FICO$	r	LTV	$I^{Low}\times$ $FICO$	$I^{Low}\times$ r	$I^{Low}\times$ LTV	Pseudo R^2 (%)	No. Obs.
-0.009***	0.231***	0.003***	0.001***	-0.043***	-0.008***	7.05	267,511
(0.0001)	(0.006)	(0.001)	(0.0001)	(0.016)	(0.001)		

Panel B: Prediction Errors during High Securitization Regime.

	β_{FICO} ($\times 10^{-3}$)	β_{LTV} ($\times 10^{-2}$)	No. Obs.	Pseudo R^2 (%)	Mean Prediction Error (%)	Actual Defaults (%)
2001	-0.123***	0.052***	128,772	0.05	3.96***	16.0
	(0.018)	(.010)				
2002	-0.197***	0.082***	152,057	0.15	4.70***	14.1
	(0.015)	(.010)				
2003	-0.428***	0.077***	308,340	0.61	5.01***	11.9
	(0.010)	(0.010)				
2004	-0.621***	0.061***	596,485	0.97	7.79***	13.9
	(0.008)	(0.004)				
2005	-1.341***	0.143***	788,299	3.90	14.67***	21.1
	(0.030)	(0.007)				
2006	-1.120***	0.190***	608,559	1.60	25.49***	33.2
	(0.012)	(0.005)				

Note: Standard errors are in parentheses. *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level.

of 2004 vintage, the mean prediction error of 7.8% reflects an underprediction of about 55% on actual defaults of 13.9%.

As a confirmation that prediction errors are positive, we plot the Epanechnikov kernel density of mean prediction errors over time.[23] If the relationship between defaults and observables has not changed since the baseline period, one would expect the average of the mean prediction error across the entire sample to be approximately zero. However, as is clear from Figure 1, the distributions show that on average the mean prediction error has been positive in each year.

[23]Plotting each of the error data points results in a dense figure with a large file size. To ensure manageable file sizes, all the kernel density figures in the paper are constructed as follows. For each year, across all loans at each FICO score, we determine the mean prediction error. We then plot the kernel density using the mean errors at each FICO score. We also plotted the densities weighing the errors by the actual number of loans at each FICO score. The plots look similar.

Moreover, the distribution of the mean prediction error progressively shifts to the right over time, as securitization becomes more prevalent in the subprime market. Of course, we expect macro-economic effects to shift the distribution of errors to the left or the right. However, as seen from the figure, there are remarkably few observations with negative mean prediction errors, even in years in which the economy was doing well and house prices were increasing. Instead, the vast majority of prediction errors are positive in each year.

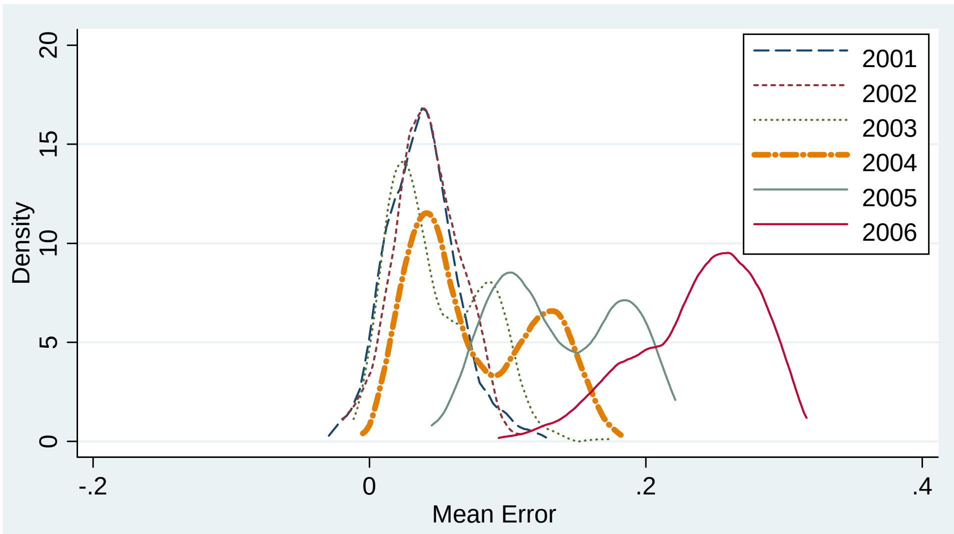


Figure 1: Kernel Density of Mean Prediction Errors Over Time, All Loans

Our test above estimates the coefficients of the model in the window 1997 to 2000, and considers the prediction errors in the period 2001 to 2006. As seen from Table II, there is a steady increase in securitization over the latter period. Hence, an alternative way to conduct this test is to use as much historical data as available for each year to tease out the incremental effect of additional securitization on the prediction errors of a default model. Using a rolling window, we predict defaults for loans issued in years 2005 and 2006, which allows the baseline model to include a few years of data from the high securitization regime. Thus, we expect the prediction errors to be smaller. For 2005 loans, the baseline model is estimated over the period 1997 to 2004, and for 2006 loans the base period is 1997 to 2005.[24] The results are qualitatively

[24] This is a stringent specification. We track default on loans issued in 2004 until the end of 2006 and on loans issued in 2005 until end of 2007. As a result, the rolling window estimation incorporates adverse forward information in the baseline model. Consequently, the errors we obtain from such a model will be smaller than those obtained by a regulator using only data available in real time.

similar, though the magnitudes of the errors are reduced. The average prediction error in this specification is 8.3% for 2005 loans (compared to 14.7% in the baseline specification) and 15.1% for 2006 loans (compared to 25.5% in the baseline specification).

Our results are also robust to the introduction of lender fixed effects in the baseline regression model in equation (5). We re-estimate the model adding lender fixed effects for the largest 700 or so lenders, which comprise 90% of securitized loans over the entire sample period. The results on prediction errors are essentially similar to those reported in Table VI and shown in Figure 1. For brevity, these results are not reported in the paper. The important conclusion is that our results on defaults are also not driven by a change in lender composition over the sample period, but rather hold within each lender.[25]

V.B.2 Cross-Sectional Tests

Since our results so far rely on changes in the data over time, our findings could be influenced by macro factors other than securitization levels that may have also changed over time. We next consider two cross-sectional tests that help rule out such concerns.

Full- and low-documentation loans

Fixing a FICO score and LTV ratio, unreported information should be more important for low-documentation loans. Thus, all else equal, a default model fitted during a low securitization era should perform better (in terms of default predictions in the high securitization period) on full-documentation loans compared to low-documentation loans. Importantly, the distribution of full- and low-documentation loans across zip codes is similar. To check this, we sort the volume of each kind of loan by zip code over 2001–2006, and consider the top 25% of zip codes in each case (which contribute over 60% of the volume of each kind of loan). A large proportion of zip codes (about 82%) are common across the two lists. In Figure 6 in Appendix B, we plot the top 25% of zip codes for each kind of loan. As can be seen, there is substantial overlap across the two kinds of loans. Thus, under the assumption that low- and full-documentation borrowers are equally sensitive to changes in the economy, any differential effects across the two kinds of loans are insulated from macroeconomic and zip-code level shocks to employment

[25] As separate confirmation, we perform the same exercise on loans issued by NCFC, and obtain qualitatively similar results. For instance, the mean prediction errors for low-documentation loans computed using model (5) is about 3.2% in 2001 and progressively increases to about 17% in 2006. For brevity, we do not report the details.

and house prices.

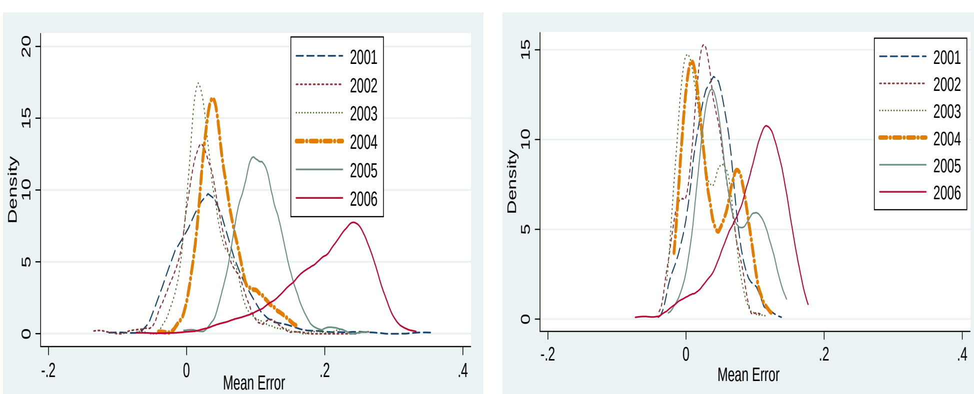
To evaluate how prediction errors vary across the two kinds of loans, we use a rolling window specification and fit separate baseline models for full- and low-documentation loans. That is, for predicting default probabilities on loans issued in year $t+1$, the baseline model is estimated over years 1 through t, where year 1 is 1997. For each kind of loan $s = Low, Full$, the baseline specification is a logit model of the form

$$\text{Prob}(\text{Default}_i^s = 1) \;=\; \Phi(\beta_{1,t}^s \cdot X_i^s),$$

where the vector X_i is the same as described earlier in this section. Let $\hat{\beta}_{1,t}^s$ be the estimated coefficients from this regression. The predicted default probability for loans issued in year $t+1$ is then estimated as

$$\text{Prob}(\widehat{\text{Default}}_{i,t+1}^s = 1) = \Phi(\hat{\beta}_{1,t}^s \cdot X_{i,t+1}^s),$$

Figures 2 (a) and (b) plot the Epanechnikov kernel density of mean prediction errors at each FICO score over time separately for full and low-documentation loans. The plots suggest that, as predicted, the prediction errors are larger for low-documentation loans than for full-documentation loans. For completeness, we report the mean prediction errors for full- and low-documentation loans in Table VIII in Appendix B. The mean errors are substantially higher among low-documentation loans for loans issued in 2003 and later.



(a) **Low-documentation Loans** (b) **Full-documentation Loans**

Figure 2: Kernel Density of Mean Prediction Errors for Low- and Full-Documentation Loans with a Rolling Estimation Window

Loans on either side of the 620 threshold

Although the test on low- versus full-documentation loans is cross-sectional, one can still argue that borrowers who choose low-documentation loans are more sensitive to house price and macro movements than those who choose full-documentation loans. To disentangle this concern from our hypothesis that lender behavior has changed, we need a cross-section of borrowers who are similar in terms of their contractual characteristics but exogenously differ in the likelihood that their loans will be securitized.

Following guidelines set by FNMA and FHLMC in the mid-1990s, a FICO score of 620 has become a threshold below which it is difficult to securitize low documentation loans in the subprime market. Keys, et al. (2010a,b) document that the ease and likelihood of securitization is greater for low documentation loans with FICO scores just above 620 (call these 620^+ loans) compared to those with FICO scores just below 620 (620^- loans). Importantly, other observable borrower and loan characteristics are the same across the two sets of loans (see Keys, et al., 2010a). This allows us to construct a cross-sectional test for borrowers *within* the low-documentation market.

In particular, our test compares the prediction errors on 620^+ low-documentation loans to those on 620^- low-documentation loans, where 620^+ includes FICO scores from 621 to 630 and 620^- includes FICO scores from 610 to 619. We exploit the feature that, in this narrow FICO range, there are no observable differences across the two sets of loans except the FICO scores. Any differences in prediction errors across the two groups therefore cannot be explained by trends in house prices or other macro-economic effects. Our hypothesis is that, since 620^- loans are harder to securitize, the lender will originate these loans paying more attention to information not reported to investors, relative to 620^+ loans. As a result, the mapping between observables and defaults should change more above a FICO score of 620, implying that the prediction errors should be lower for 620^- loans relative to 620^+ loans.

For brevity, we conduct this test averaging the prediction errors (at each FICO score) for all low-documentation loans issued in the period 2001-06.[26] The baseline model used is the model in equation (5), estimated on only 620^+ and 620^- loans. The kernel densities of the mean prediction errors are shown in Figure 3. The prediction errors are indeed lower for 620^- loans (16.6%) than 620^+ loans (18.2%). The difference in mean errors of 1.6% is statistically significant at the 1% level. Even though 620^- loans are harder to securitize relative to 620^+

[26]The results are similar (though smaller in magnitude) when the test is repeated for 2001-04 loans.

loans, over time the ease of securitization for the entire market (including 620^- loans) increases. Further, among low-documentation loans, 620^- loans are at the low end of the FICO spectrum. As a result, the prediction errors over the entire period 2001–2006 are high for 620^- loans as well.

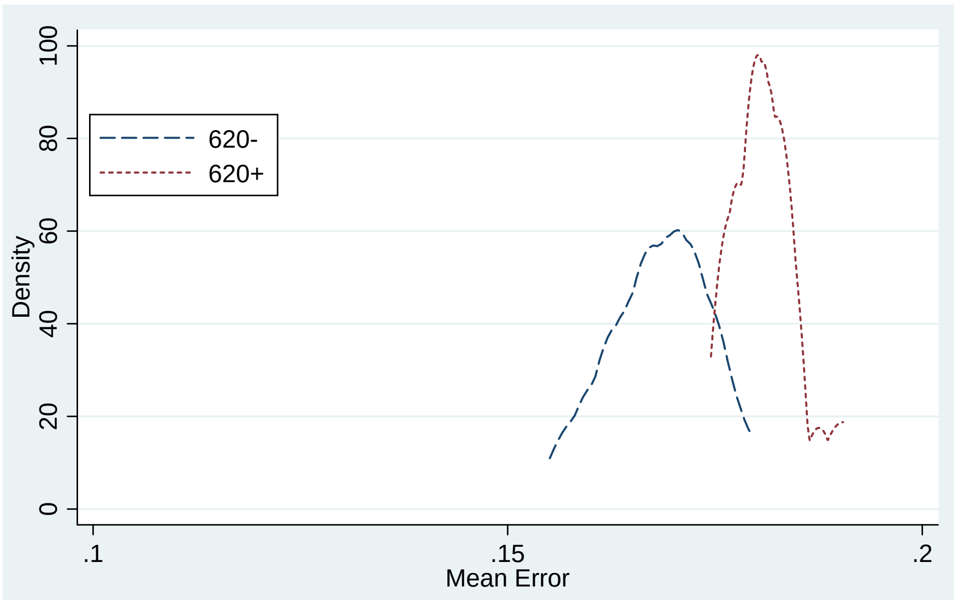


Figure 3: Mean Prediction Errors for 620^+ and 620^- Low-documentation Loans

V.B.3 Placebo Test: Predictability of Defaults in Low Securitization Regime

Across different years in the low securitization regime, there should be no substantive change in a lender's incentives to collect information about a borrower or property. Thus, the mapping between observables and defaults should be approximately similar from year to year. This argument forms the basis of a placebo test in which we assess whether a default model estimated during a low securitization regime generates small prediction errors in another period with relatively *low* securitization.

To conduct the test, we predict defaults on low-documentation loans issued in 1999 and 2000, using a baseline model estimated from 1997 and 1998 for 1999 loans, and 1997 through 1999 for 2000 loans (i.e., employing a rolling window). The results are reported in Table IX. The mean prediction error is not significantly different from zero, and is also substantially smaller in magnitude than the mean errors reported in Table VIII for years 2001 and beyond.

The same result is confirmed in Figure 4, where we plot the kernel distribution of the mean prediction error at each FICO score. In contrast to Figure 1, the mean errors are centered around 0. We also regress the prediction errors on FICO score and LTV ratio for each year 1999 and 2000, and report the coefficients in Table IX in Appendix B. In contrast to the results in Table VI, the β_{FICO} and β_{LTV} coefficients are insignificant, suggesting that there is no systematic underprediction by the baseline model. Thus, the control test is consistent with our hypothesis.

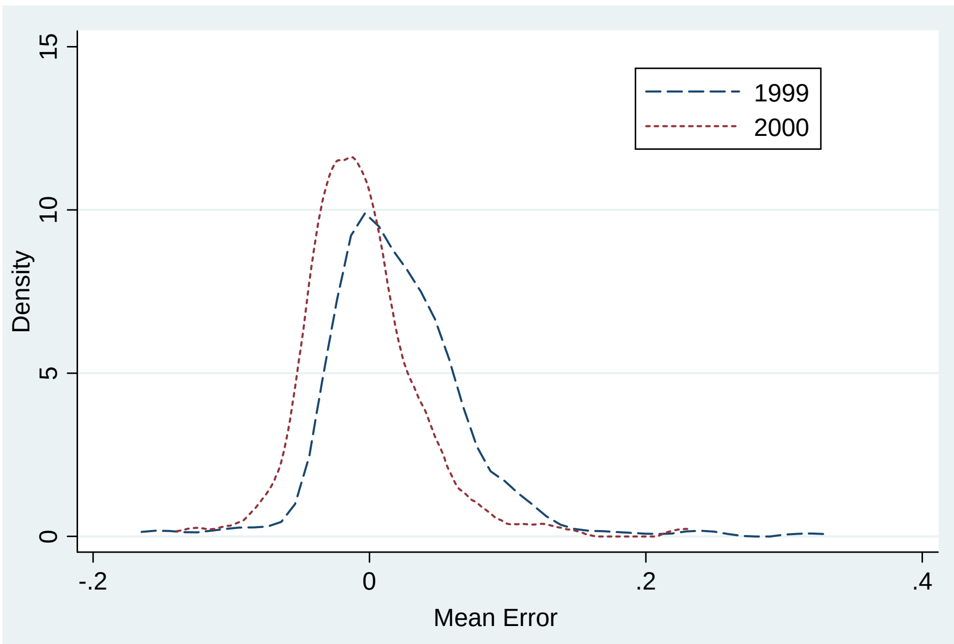


Figure 4: Placebo Test—Kernel Density of Mean Prediction Errors in Low Securitization Period (Low-Documentation Loans Only, Rolling Estimation Window)

VI Further Tests on Statistical Default Models

VI.A Effect of Changes in House Prices

Our results on loan defaults in Section V.B.1 ignore the effect that changing house prices may have had on defaults. There is no doubt that a fall in house prices is partly responsible for the surge in defaults for loans issued in 2005 and 2006 (see, for example, Mayer, Pence, and Sherlund, 2009, and Mian and Sufi, 2009). However, only in August 2007 did the composite (i.e., national level) Case-Shiller index indicate a fall from its value 24 months earlier. As a result, loans issued in 2004 and before did not suffer from a fall in house prices over the next 24 months, yet as shown in Table VI and Figure 1, the prediction errors from a default model remain high.[27] Further, in our comparison between 620^+ and 620^- loans, both sets are subject to exactly the same effects of changing house prices. The same is true of the our comparison between full- and low-documentation loans, since the distribution of both kinds of loans across zip codes is similar. In this section, we explicitly include the future change in house prices at the state level as an explanatory variable.

For each loan, we construct a house price appreciation (HPA) variable as follows. We begin with the state-level quarterly house price index constructed by the Office of Federal Housing Enterprise Oversight. For each state s, a house price index for each year t, $h_{s,t}$, is constructed as a simple average of the indices over four quarters. Consider loan i issued in state s in year t. The house price appreciation variable for loan i is set to the growth rate of house prices over the next two years, $HPA_i = \frac{h_{s,t+2} - h_{s,t}}{h_{s,t}}$. We include HPA_i in the vector of loan characteristics X_i in both the baseline and predictive regressions. Our specification is stringent: It clearly includes more information than available to an econometrician at the time the forecast is made and will soak up more variation in defaults than a prediction made in real time (in other words, the specification assumes the regulator or rating agency has perfect foresight.

We re-estimate the baseline model (5) after including the HPA variable (both by itself and interacted with I^{Low}, the low-documentation dummy) on the right-hand side. We then predict default probabilities for loans issued in each of the years 2001 through 2006. A rolling window is used for this estimation, so default probabilities for loans issued in year $t+1$ are predicted based on coefficients estimated over years 1 through t, where year 1 is 1997. In Figure 5, we plot the Epanechnikov kernel density of mean prediction errors (computed at each FICO score) in each year 2001 through 2006. For ease of comparison, the figure has six panels, each panel showing the kernel density of mean out-of-sample prediction errors in

[27]There are two possible explanations for borrowers defaulting when house prices increase. First, over 70% of the loans in our sample have a prepayment penalty, increasing the transaction cost to a borrower of selling the house. Second, some borrowers who experience an increase in home prices may be taking out additional home equity loans, effectively maintaining a higher LTV ratio than reported in the sample. The latter effect is consistent with our story, since information on whether a borrower may be credit-constrained in the future and take out additional home loans is soft information potentially observable by a lender but not reported to the investor.

a given year with and without including house price appreciation as an explanatory variable, using a rolling estimation window in each case.

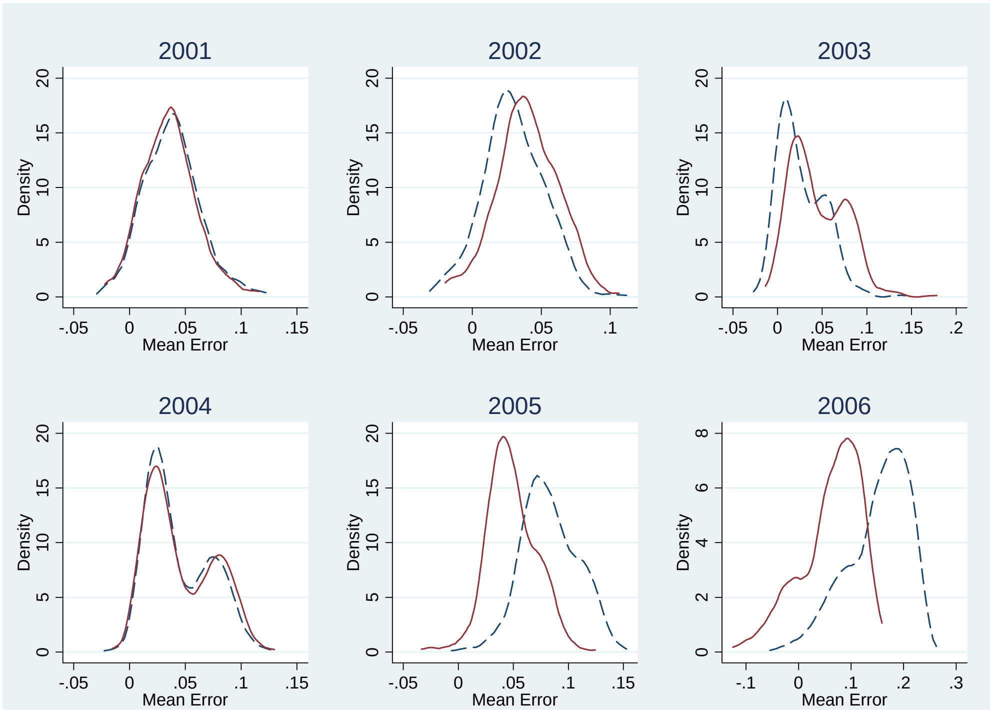


Figure 5: Kernel Density of Mean Prediction Errors With (Solid Lines) and Without (Dashed Lines) House Price Effects, Rolling Estimation Window

Two observations emerge from the figure. First, for 2001–2004 loans, there is not much difference in the two kernel densities. In fact, for 2002–2003 loans, including the house price effect slightly magnifies the prediction errors. Second, the prediction errors for loans issued in 2005 and 2006 are indeed reduced in magnitude when the effect of house prices is included. In particular, using a rolling window for estimating the baseline model, the mean prediction error for 2005 loans falls from 8.3% to 4.9% when HPA is included as an explanatory variable, and for 2006 loans falls from 15.1% to 6.1%. Thus, for these two years, approximately 50% of the mean prediction error survives over and above the effect of falling house prices. Therefore, even after accounting fully for the effect of falling house prices on defaults, the prediction errors exhibit patterns consistent with our predictions. It continues to be striking how few of the mean errors are less than zero across the entire period 2001–2006.[28]

[28]In unreported tests, we repeat the analysis low- and full-documentation loans after including the house

VI.B Were Investors Fooled?

Our analysis is largely agnostic on whether investors priced loans fairly in the build-up to the subprime crisis. Importantly, our predictions obtain even when both lenders and investors are fully rational, with the latter incorporating the worsening of the loan pool into prices paid to lenders.[29] Nevertheless, suppose investors are boundedly rational and price loans using default predictions from a naïve method. Loan prices will then be too high, especially for borrowers on whom the unreported information is an important predictor of quality. Lenders now have an even stronger incentive to ignore the unreported information in approving loans and setting interest rates. As a result, the tendency of a statistical model to underpredict defaults for these borrowers will worsen.

It is important to consider whether investors rationally anticipated the increase in defaults implied by our results: with rational investors, asset prices can be used to fine tune regulation.[30] A direct test of investor rationality is difficult to conduct. We do not have data on the pricing of CDO tranches backed by subprime mortgage loans. As an indirect test, we consider the subordination levels of AAA tranches for new non-agency pools consisting of loans originated in 2005 and 2006. We have already shown (Figures 1 and 5) that a statistical default model most severely underestimates actual defaults in 2005 and 2006. The subordination level measures the magnitude of losses an equity tranche can absorb, before the principal of the AAA tranches is at risk. Thus, if rating agencies were correctly forecasting future defaults, the subordination levels in the pools must have a positive correlation with the prediction errors of the default model (otherwise the tranches should not have been rated AAA). At best, we find a weak relationship, suggesting that rating agencies were unaware of or chose to overlook the underlying regime change in the quality of loans issued as securitization increased. Figure 7 in Appendix B shows the subordination level plotted against the mean prediction error of the pool.

These results are consistent with the work of Ashcraft, Goldsmith-Pinkham and Vickery (2010), who find that during this period subordination levels do not adjust enough to reflect

price effect. For loans issued in 2001–2004, the results are similar to those reported in the cross-sectional test described earlier. For loans in 2005 and 2006, the magnitudes of the prediction errors are reduced for both groups of loans, but the errors continue to be larger for low-documentation loans.

[29] See, for example, Rajan, Seru and Vig (2010) for a model with rational investors that delivers the predictions we test.

[30] See, for example, Hart and Zingales, "To Regulate Finance, Try the Market," *Foreign Policy*, March 30, 2009.

the increased riskiness of originated loans. Similarly, Benmelech and Dlugosz (2009) and Griffin and Tang (2008) argue that ratings of CDO tranches were aggressive relative to realistic forward-looking scenarios. More directly, Faltin-Traeger, Johnson and Mayer (2010) consider the pricing of CDO tranches, and find that the ability of spreads to predict future downgrades is weak across all tranches. There is therefore suggestive evidence that some classes of subprime-backed securities were mispriced by investors.[31]

VII Conclusion

Establishing a liquid market for a complicated security requires standardization of not just the terms of the security, but also of the fundamental valuation model for the security, both of which help investors to better understand the security. Inevitably, the process of constructing and validating a model will include testing it against previous data. We argue in this paper that the growth of the secondary market for a security can have an important incentive effect that affects the quality of the collateral behind the security itself. The associated regime change will imply that even a model that fits historical data well will necessarily fail to predict cash flows, and hence values, going forward.

While we focus on a particular statistical default model, similar models are widely used by market participants for diverse purposes such as making loans to consumers (for example, using the FICO score), assessing capital requirements on lenders and determining the ratings of CDO tranches. Our critique applies to all such models, since they all use historical data in some manner to predict future defaults without accounting for the impact of changed incentives of participants that generate the data. Importantly, the effects we document are systematic and stronger for borrowers with low FICO scores and low-documentation. Since the loans we analyze represent the underlying collateral for CDOs and subsequent securitization, the errors cannot be diversified away. The phenomenon we examine is therefore different from the much-discussed argument that correlations (but not levels) of loan defaults had been mis-estimated.

The inescapable conclusion of a Lucas critique is that actions of market participants will undermine any rigid regulation. What can market participants do to better predict the future? Agents such as regulators setting capital requirements or rating agencies will take some time to

[31] As another example, once loan defaults had increased in the 3rd quarter of 2007, in November 2007 Standard and Poor's adjusted their default model to reduce the reliance on the FICO score as a predictor of default (Standard & Poor's, 2007).

learn about the exact magnitudes of relevant variables following a regime change. Nevertheless, we certainly expect them to be aware that incentive effects may lead to such a regime change, which can systematically bias default predictions downward. An adaptive learning approach that places more weight on recent data may help in such a setting.[32] Once sufficient data has accumulated in the new regime, a statistical model can be reliably estimated (until the regime changes yet again). During the learning phase, however, participants need to be particularly aware that predictions from the default model are probabilistic and the set of possible future scenarios has expanded in an adverse way. Thus, the assessment of default risk must be extra conservative during this period.

We expect that the agents in the market will eventually learn that the regime has changed. The challenge for regulators in particular is to recognize such shifts in real time and take appropriate actions. If investors are rational, market prices should reflect the risk of assets and could be used by regulators to assess default risk. Another alternative is to use a structural approach. As Gali and Gertler (2007) highlight, monetary policy evaluation has moved beyond reduced form statistical models to using structural models in a rigorous way. In the regulatory context, perhaps a regulator can require greater disclosure of data collected by a lender, even if not reported to an investor. Such data can then be used in a structural framework to properly determine the default risk of loans by accounting for changes in the behavior of agents in response to a change in incentives (for example, by augmenting the statistical default model with a selection equation, as highlighted in Appendix A).

[32] See Malmendier and Nagel (2009) for a perspective on adaptive learning about inflation.

References

[1] Ashcraft, Adam, Paul Goldsmith-Pinkham and James Vickery (2010), "MBS Ratings and the Mortgage Credit Boom," FRB New York Staff Reports, No. 449.

[2] Basel Committee on Banking Supervision (2006), "International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Comprehensive Version," http://www.bis.org/publ/bcbs128.pdf.

[3] Benmelech, Efraim and Jennifer Dlugosz (2009), "The Alchemy of CDO Credit Ratings," *Journal of Monetary Economics* 56(5): 617–634.

[4] Berger, Allen N., Nathan H. Miller, Mitchell A. Petersen, Raghuram G. Rajan, and Jeremy C. Stein (2005), "Does Functional Form Follow Organizational Form? Evidence from the Lending Practices of Large and Small Banks," *Journal of Financial Economics* 76(2): 237–269.

[5] Bolton, Patrick and Antoine Faure-Grimaud (2009), "Satisficing Contracts," forthcoming, *Review of Economic Studies.*

[6] Brunnermeier, Marcus, Andrew Crockett, Charles Goodhart, Avinash Persaud and Hyun Shin, "The Fundamental Principles of Financial Regulation," *11th Geneva Report on the World Economy.*

[7] Chomsisengphet, Souphala and Anthony Pennington-Cross (2006), "The Evolution of the Subprime Mortgage Market," *FRB of St. Louis Review*, 88:1, 31-56.

[8] Cole, Rebel, Lawrence Goldberg and Lawrence White (1998), "Cookie-Cutter Versus Character: The Micro Structure Of Small Business Lending By Large And Small Banks," *FRB Chicago Working Paper.*

[9] Demyanyk, Yuliya and Otto Van Hemert (2009), "Understanding the Subprime Mortgage Crisis," Forthcoming, *Review of Financial Studies.*

[10] Einav, Liran, Mark Jenkins and Jonathan Levin (2008), "The Impact of Information Technology on Consumer Lending," Working paper.

[11] Faltin-Traeger, Oliver, Kathleen Johnson and Christopher Mayer (2010), "Issuer Credit Quality and the Price of Asset-Backed Securities," *American Economic Review*, 100(2): 501–505.

[12] Gali, Jordi and Mark Gertler (2007), "Macroeconomic Modeling for Monetary Policy Evaluation," *Journal of Economic Perspectives* 21(4): 25–45.

[13] Fishelson-Holstein, Hollis (2005), "Credit Scoring Role in Increasing Homeownership for Underserved Populations," in Retsinas and Belsky, eds., *Building Assets, Building Credit: Creating Wealth in Low-Income Communities*, Washington, D.C.: Brookings Institution Press.

[14] Gorton, Gary B. and George G. Pennacchi (1995), "Banks and Loan Sales: Marketing Nonmarketable Assets," *Journal of Monetary Economics*, 35, 389-411.

[15] Gramlich, Edward (2007), "Subprime Mortgages: America's Latest Boom and Bust," Washington, D.C.: *The Urban Institute Press.*

[16] Greenspan, Alan (2008), Testimony before House Committee of Government Oversight and Reform, Oct 23, 2008.

[17] Griffin, John M. and Dragon Y. Tang (2009), "Did Subjectivity Play a Role in CDO Credit Ratings?", Working Paper, SSRN.

[18] Heckman, James (1980), "Varieties of Selection Bias," *American Economic Review* 80, 313-318.

[19] Holloway, Thomas, Gregor MacDonald and John Straka (1993), "Credit Scores, Early-Payment Mortgage Defaults, and Mortgage Loan Performance," Working Paper, FHLMC.

[20] Holmström, Bengt and Paul R. Milgrom, "Multi-Task Principal-Agent Problems: Incentive Contracts, Asset Ownership and Job Design," *Journal of Law, Economics and Organization* 7 (Special Issue): 24–52.

[21] Kashyap, Anil, Raghuram Rajan and Jeremy Stein (2008), "Rethinking Capital Regulation," Paper prepared for the FRB Kansas City Symposium, Jackson Hole.

[22] Keys, Benjamin J., Tanmoy K. Mukherjee, Amit Seru and Vikrant Vig (2010a), "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans," *Quarterly Journal of Economics* 125(1): 307–362.

[23] Keys, Benjamin J., Tanmoy K. Mukherjee, Amit Seru and Vikrant Vig (2010b), "620 FICO, Take II: Securitization and Screening in the Subprime Mortgage Market," Working Paper, University of Chicago.

[24] Liberti, Jose and Atif Mian (2009), "Estimating the Effect of Hierarchies on Information Use", *Review of Financial Studies* 22(10): 4057–4090.

[25] Malmendier, Ulrike and Stefan Nagel (2009), "Learning from Inflation Experiences," Working paper.

[26] Mayer, Christopher and Karen Pence (2008), "Subprime Mortgages: What, Where, and to Whom?" NBER Working Paper No. 14083.

[27] Mayer, Christopher, Karen Pence and Shane Sherlund (2009), "The Rise in Mortgage Defaults," *Journal of Economic Perspectives* 23(1): 27–50.

[28] Mian, Atif and Amir Sufi (2009), "The Consequences of Mortgage Credit Expansion: Evidence from the U.S. Mortgage Default Crisis," *Quarterly Journal of Economics* 124(4): 1449–1496.

[29] Olsen, Randall (1980), "A Least Squares Correction for Selectivity Bias," *Econometrica* 48(7): 1815–1820.

[30] Rajan, Uday, Amit Seru and Vikrant Vig (2010), "Statistical Default Models and Incentives," *American Economic Review Papers and Proceedings* 100(2): 506–510.

[31] Tirole, Jean (2009), "Cognition and Incomplete Contracts," *American Economic Review* 99(1): 265–294.

[32] Sheather, S.J. and M.C. Jones (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society, Series B* 53: 683–690.

[33] Standard & Poor's (2007), "Standard & Poor's Enhances LEVELS® 6.1 Model," News release, November 9, 2007, available at www2.standardandpoors.com.

[34] Stein, Jeremy (2002), "Information Production and Capital Allocation: Decentralized versus Hierarchical Firms," *Journal of Finance*, 57(5): 1891–1921.

Appendix

A Selection Model

In this appendix, we use the selection model framework of Heckman (1980) to discuss our hypothesis that the mapping between observables and loan defaults will change with securitization. Recall that X_{it} consists of variables reported by the lender to the investor and Z_{it} of variables observed by the lender but not reported to the investor. For convenience, assume that X_{it} and Z_{it} are both non-negative scalars, denoted respectively by x_{it} and z_{it}. For example, x_{it} could be the FICO score of the borrower and z_{it} could be a summary statistic based on other hard and soft information available to the lender.

A regulator or rating agency has the same information as the investor, and is interested in evaluating the quality of the loan based on x_{it}. Let d_{it} represent a default event on loan i issued at time t. A contemporaneous default regression may be estimated as:[33]

$$d_{it} \quad = \alpha + \beta x_{it} + \epsilon_{it}, \tag{6}$$

where ϵ_{it} is a mean zero error term with variance σ_ϵ^2.

In a low-securitization regime, the lender approves a loan application if either x_{it} is high or x_{it} is low but z_{it} is high. That is,

$$A_{it} = 1 \quad \text{if and only if} \quad \gamma z_{it} + \delta x_{it} + \eta_{it} > 0,$$

where η_{it} is a mean-zero error term with variance σ_η^2. The regulator, rating agencies and the investors only observe approved loans (i.e., $A_{it} = 1$).

Assume that the conditional expectation of ϵ_{it} given η_{it} is linear in η_{it}, and the correlation between ϵ_{it} and η_{it} is ρ. Then, we can write $\epsilon_{it} = \rho(\eta_i - \bar{\eta})\frac{\sigma_\epsilon}{\sigma_\eta} + \omega_{it}$, where ω_{it} is uncorrelated with η_{it}. Therefore, $E(d_{it} \mid x_{it}, A_{it} = 1) = \beta x_{it} + \frac{\rho\sigma_\epsilon}{\sigma_\eta} E(\eta_{it} \mid \eta_i > -\gamma z_{it} - \delta x_{it})$.

In the spirit of Olsen (1980), assume that η_{it} is uniformly distributed over $[-1, 1]$. Then, $E(\eta_{it} \mid \eta_i > -\gamma z_{it} - \delta x_{it}) = \frac{1-\gamma z_{it}-\delta x_{it}}{2}$. It follows that

$$E(d_{it} \mid x_{it}, A_{it} = 1) \quad = \quad \beta x_{it} + \frac{\rho\sigma_\epsilon}{2\sigma_\eta}[-\delta x_{it} - \gamma z_{it} + 1].$$

[33] Although default is a binary event, in this section we use a linear regression specification for expositional simplicity. The analysis is similar with a logit or probit specification. Our actual regressions in Section V use the logit model.

Therefore, when equation (6) is estimated, the relationship between the observed coefficient β^* and the true coefficient β may be written as $\beta^* = \beta + \frac{\rho\sigma_\epsilon}{2\sigma_\eta}[-\delta Var(x_{it} \mid A_{it} = 1) - \gamma Cov(x_{it}, z_{it} \mid A_{it} = 1)]$. Here, $Var(x_{it} \mid A_{it} = 1) > 0$. Further, the selection equation implies on average that, for high values of x_{it}, $A_{it} = 1$ even when z_{it} is low. However, for low values of x_{it}, on average $A_{it} = 1$ only when z_{it} is high. Thus, $Cov(x_{it}, z_{it} \mid A_{it} = 1) < 0$. Let $B_\ell = \beta - \beta^* = \frac{\rho\sigma_\epsilon}{2\sigma_\eta}[\delta Var(x_{it} \mid A_{it} = 1) + \gamma Cov(x_{it}, z_{it} \mid A_{it} = 1)]$ denote the bias in the low-securitization regime.

Next, consider a high securitization regime. Here, the lender bases its decisions on hard information variables that are reported to the investor, downplaying information it may have used in a low securitization regime. In the extreme case, if z_{it} is completely ignored, the selection equation changes to:

$$A_{it} = 1 \quad \text{if and only if} \quad \delta_h x_{it} + \eta_{it} > 0,$$

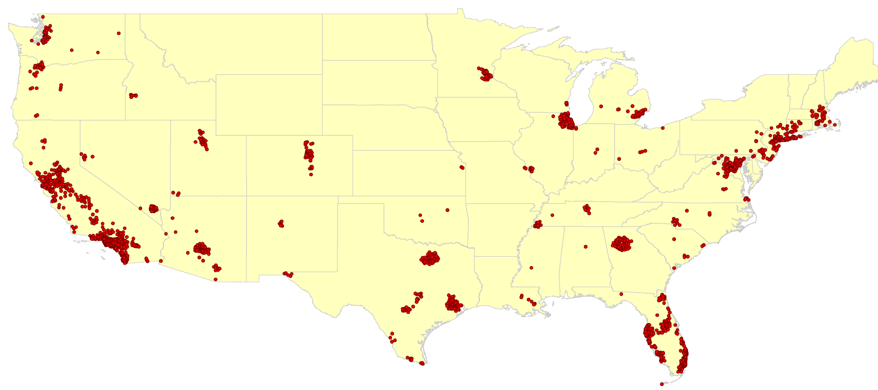
where δ_h is sufficiently greater than δ to ensure that the minimum value of x_{it} at which a loan is granted is the same in both regimes. That is, even when x_{it} is small, on average the loan is granted regardless of the value of z_{it}. Here, $Cov(x_{it}, z_{it} \mid A_{it} = 1) = 0$. Therefore, the bias in the high-securitization regime may be represented as $B_h = \frac{\rho\sigma_\epsilon}{2\sigma_\eta}\delta_h Var(x_{it} \mid A_{it} = 1)$, where we assume that $Var(x_{it} \mid A_{it} = 1)$ is similar in both regimes.

Since the true coefficient β is negative (that is, when the FICO score x_{it} is high, a default is less likely), the estimated coefficient in the low-securitization regime (say β^*_ℓ) is closer to zero due to additional covariance term than the coefficient in the high-securitization regime (β^*_h). Therefore, if β^*_ℓ is used to forecast defaults for low values of x_{it}, it will underestimate defaults.[34] Since defaults themselves are more likely at low values of x_{it}, the overall effect is to underpredict defaults in the high-securitization era.

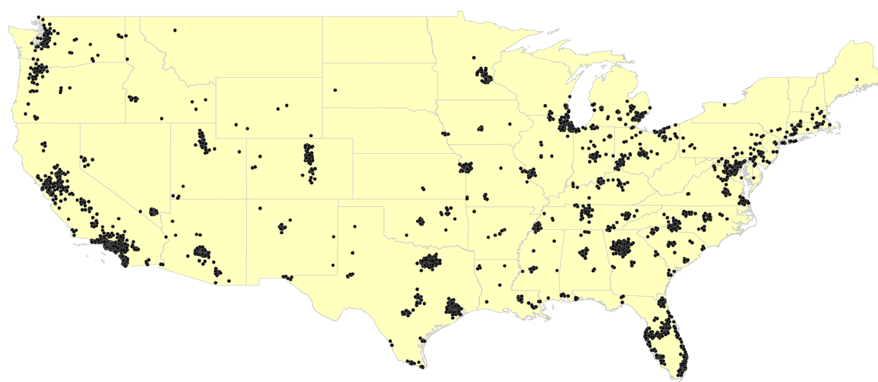
Overall, then, our argument is that regulators, rating agencies and investors only see approved loans, which by definition have survived a selection process. The selection process for loans changes when the incentives of the lender change. Consequently, as securitization increases, one expects that the behavior of the lender will change. This changes the selection process, thereby altering the mapping from observables to loan defaults.

[34] In other words, the bias with respect to the true coefficient changes across the two regimes. In particular, since $Cov(x_{it}, z_{it} \mid A_{it} = 1) < 0$ in the low-securitization regime and $\delta_h > \delta$, it follows that $B_h > B_\ell$.

B Additional Figures and Tables



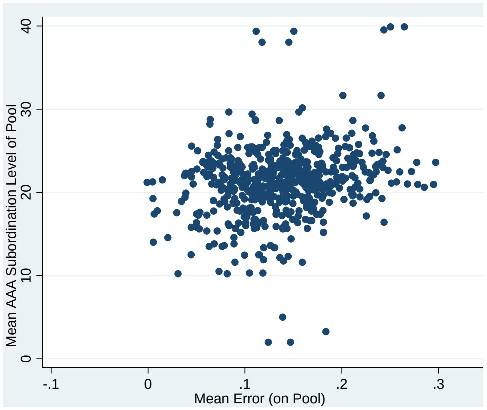
(a) **Low-documentation Loans**



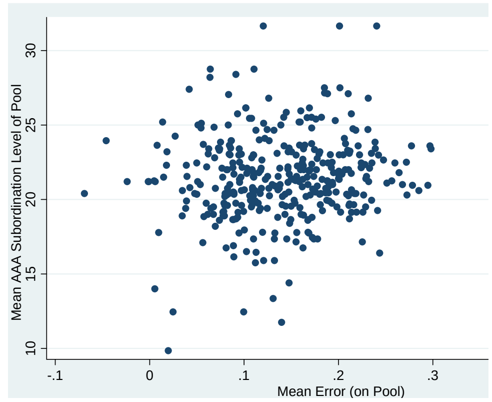
(b) **Full-documentation Loans**

Figure 6: Top 25% of Zip Codes for Subprime Loans, 2001–2006

These figures display the top 25% of zip codes (by number of loans) in which low-documentation (top; figure (a)) and full-documentation (bottom; figure(b)) subprime mortgage loans issued made over the period 1997–2006. These zip codes contribute over 60% of the volume of subprime loans in the respective category. The figure shows that there was substantial overlap of zip codes across the two kinds of loans, with concentrations in places such as California, Florida and the North-East.



(a) **2005 Loans**



(b) **2006 Loans**

Figure 7: Pool Subordination Level and Mean Prediction Error

These figures present the scatter plot of mean subordination level of AAA tranches in a pool against the mean prediction error of defaults in that pool for loans issued in 2005 (figure (a)) and 2006 (figure (b)). To highlight whether there is a relationship between subordination levels and prediction errors on default, we consider only pools for which prediction errors (i.e., actual defaults − predicted defaults given the baseline model) are likely to be high: we restrict attention to pools with at least 30% low-documentation loans. Subordination level information is obtained from Bloomberg and cross-checked with information provided in the Intex database. Prediction errors are computed using the baseline model in equation (5). The figure suggests that there is no relationship between the prediction errors from the default model and subordination levels of AAA tranches.

Table VII: Shrinkage in the Distribution of Interest Rates

FICO	β_b	Std. Err.	R^2 (%)
500	-0.212***	(0.019)	53
510	-0.191***	(0.013)	67
520	-0.214***	(0.013)	71
530	-0.179***	(0.011)	71
540	-0.17***	(0.009)	74
550	-0.151***	(0.010)	69
560	-0.146***	(0.008)	75
570	-0.126***	(0.009)	65
580	-0.062***	(0.009)	31
590	-0.052***	(0.008)	25
600	-0.035***	(0.008)	14
610	-0.037***	(0.008)	17
620	-0.035***	(0.007)	17
630	-0.023***	(0.006)	10
640	-0.023***	(0.005)	13
650	-0.043***	(0.007)	23
660	-0.049***	(0.009)	22
670	-0.06***	(0.009)	27
680	-0.047***	(0.008)	22
690	-0.058***	(0.010)	25
700	-0.05***	(0.011)	16
710	-0.059***	(0.012)	19
720	-0.055***	(0.010)	21
730	-0.101***	(0.013)	35
740	-0.085***	(0.012)	33
750	-0.071***	(0.016)	14
760	-0.066***	(0.015)	15
770	-0.045***	(0.013)	9
780	-0.059***	(0.015)	11
790	-0.064***	(0.019)	9
800	-0.065***	(0.032)	3

We report estimates from regression of yearly standard deviation of interest rates at each FICO score on time. The regressions are estimated separately in buckets of ten FICO points, in the range 500 to 800. The sample period is from 1997–2006.

Table VIII: Default Model—Mean Prediction Errors for Low- and Full-Documentation Loans with a Rolling Estimation Window

	Low-Documentation (%)	Full-Documentation (%)	Difference (%) (Low-Doc − Full-Doc)
2001	3.40	3.80	-0.40
2002	2.78	2.79	-0.01
2003	3.20	2.21	0.99***
2004	5.17	3.51	1.66***
2005	10.58	5.85	4.73***
2006	20.11	9.84	10.27***

We report the mean prediction errors for low and full-documentation loans issued from 2001 through 2006. The estimation uses a rolling window approach with separate baseline models for low-documentation and full-documentation loans. That is, the predictions for year $t+1$ are based on a model estimated over the years 1 through t, where year 1 is 1997. ***, ** and * represent that differences are significant at the 1%, 5% and 10% levels respectively.

Table IX: Default Model Placebo Test—Low Securitization Years, Low-Documentation Loans Only with a Rolling Estimation Window

Panel A: Coefficients of Baseline Model in Low Securitization Regime

	$FICO$	r	LTV	Pseudo-R^2 (%)	No. Obs.
1997-1998	-0.009***	0.249***	-0.008***	8.11	16,002
	(0.0005)	(0.034)	(0.003)		
1997-1999	-0.007***	0.259***	-0.003*	7.94	33,868
	(0.003)	(0.022)	(0.001)		

Panel B: Prediction Errors during Low Securitization Regime.

	β_{FICO} ($\times 10^{-3}$)	β_{LTV} ($\times 10^{-2}$)	No. Obs.	Pseudo R^2 (%) (%)	Mean Prediction Error (%)	Actual Defaults (%)
1999	0.039	0.026	17,866	0.01	0.91	11.0
	(0.038)	(.023)				
2000	0.039	-0.026	24,591	0.01	0.97	11.9
	(0.034)	(.020)				

We report estimates from a baseline default model estimated for low-documentation loans issued in 1997 and 1998 in Panel A. Panel B reports the β coefficients from a regression of prediction error on FICO score and LTV ratio for loans issued in 1999 and 2000, and also reports the mean prediction errors for each vintage. *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level.