



U.S. Department of Housing and Urban Development
Office Of Policy Development & Research



**A Study of
Market Sector Overlap
and Mortgage Lending**

A Study of Market Sector Overlap and Mortgage Lending

Prepared for:

U.S. Department of Housing and Urban Development
Office of Policy Development & Research

Prepared by:

David T. Rodda
Jody Schmidt
Satyendra Patrabansh

Abt Associates Inc.
Cambridge, MA

Contract C-OPC-21895
Task Order CHI-TO003

May 2005

Acknowledgments

This report was prepared by Abt Associates Inc. under a contract with the U.S. Department of Housing and Urban Development. The Department acknowledges with thanks the work of the study's authors, David Rodda, Jody Schmidt, and Satyendra Patrabansh. We would also like to acknowledge the contributions of Abt Associates staff members Emily Finnin Ma who provided data preparation work, Ken Lam who provided careful programming, and David Hoaglin who provided statistical advice. In addition, the Department acknowledges with thanks the contributions of Robert F. Cotterman and Holly Krier at Unicon Research Corporation who developed and conducted the default modeling. The report benefited from the expert advice from HUD staff, including Harold Bunce, John Gardner, Theresa DiVenti, and Bill Reeder. The analysis in this report builds on a previous HUD contract that included the development of a database by Abt Associates staff working with Jay Schultz, HUD, and Ismail Mohamed and Ron Hanson, Titan Systems, Inc.

The contents of this report are the views of the contractor and do not necessarily reflect the views or policies of the U.S. Department of Housing and Urban Development or the U.S. Government.

PREFACE

In 1992 legislation, Congress mandated that HUD provide information on the affordable lending activities of Fannie Mae and Freddie Mac, the two major government-sponsored enterprises (GSEs) in the mortgage market. At that time, Congress said there was a “vacuum of information” surrounding the GSEs’ mortgage purchases, and called for HUD to collect data and conduct research on the GSEs’ mortgage purchase activity. One priority for HUD has been to initiate an active research program with respect to the GSEs, including both in-house and contract research. Much of this research was used in the development of the three Affordable Housing Goal Regulations issued in 1995, 2000, and 2004.

This report continues this line of GSE research by comparing the characteristics of GSE-purchased loans with mortgages originated or insured by other sectors of the mortgage market, such as the Federal Housing Administration (FHA). This so-called overlap analysis clarifies the role of the GSEs in providing credit support for low-income and minority families. Intuitively, overlap refers to the set of loans that could have gone to either market sector, for example, either insured by FHA or sold to the GSEs.

The overlap question is one of great interest in mortgage policy discussions but has not been satisfactorily dealt with because of the lack of a database covering FHA-insured, GSE-purchased, and other conventional loans that includes borrower credit history and other mortgage loan data, such as the loan-to-value ratio. Thus, this study extends previous research by including data on borrower credit history, loan-to-value ratios, and other underwriting variables for conventional loans, privately-insured loans, GSE-purchased loans, and FHA-insured loans. It examines the extent of overlap between the mortgage market sectors, particularly the FHA and the two GSEs, Fannie Mae and Freddie Mac. The study finds that different market sectors serve distinct segments of the population. For example, one of the main findings of this study is that only about ten percent of FHA-insured loans have risk characteristics similar to GSE-purchased loans. Compared with GSE-purchased loans, FHA-insured loans are characterized by lower borrower credit scores and higher loan-to-value ratios (i.e., lower downpayments), and are more targeted to lower-income and minority borrowers.

Table of Contents

Executive Summary	ix
A Study of Market Sector Overlap and Mortgage Lending	ix
Section 1: Background and Literature Review.....	1
Literature Review.....	3
Research Questions.....	6
Section 2: Data Preparation and Description	9
HMDA Data.....	9
Experian Data.....	9
Data Cleaning and Record Selection	14
Explanatory Variables	20
Subsetting the Data.....	33
Section 3: Origination Model and Market Sector Overlap	37
Confidence Interval Measure of Overlap.....	37
Tolerance Limit Methods of Overlap.....	40
Parametric Tolerance Limits	41
Non-Parametric Tolerance Limits	42
Origination Models and the Application of Overlap Methods.....	43
FHA vs. PMI within GSE.....	63
FHA vs. All PMI	65
FHA vs. Subprime	73
GSE vs. Depository Lenders	78
Section 4: Default Model and Market Sector Overlap.....	85
Default Model on FHA Performance Data	85
References	107
Appendix: Notes on Variables and Calculations	

List of Exhibits

Exhibit 1: Analysis of Matched and Unmatched Experian and HMDA Loans from 11 MSAs	12
Exhibit 2: Methodology for Limiting HUD Data sets	14
Exhibit 3: PMI Loans by LTV Ratio	15
Exhibit 4: GSE Loans With & Without PMI by LTV Ratio	16
Exhibit 5: FHA Loans by Experian Dwelling Unit Size	17
Exhibit 6: Number of FHA Loans Above FHA Loan Limits	18
Exhibit 7: Observations with Missing Data.....	18
Exhibit 8: Subprime Loans by Mortgage Market Sector.....	19
Exhibit 9: Distribution of Loans by Purchaser Type	20
Exhibit 10: Comparison of Characteristics by Dataset.....	21
Exhibit 11: Analysis of Loans in Mortgage Market Sectors	23
Exhibit 12: Comparison of Fixed-Rate vs. Adjustable-Rate Mortgage.....	25
Exhibit 13: Analysis of FHA-Eligible Loans in Mortgage Market Sectors	27
Exhibit 14: Analysis of 80-100% LTV Loans in FHA and GSE Market Sectors	29
Exhibit 15: Weighted Distribution of FICO Scores	35
Exhibit 16: Weighted Distribution of LTV Ratios	36
Exhibit 17: FHA/GSE Overlap Based on the Confidence Interval Method.....	39
Exhibit 18: Replication of FHA vs. PMI Model From HUD 1995 Report	44
Exhibit 19: Origination Model Results on FHA vs. GSE Loans	47
Exhibit 20: Higher LTV Has Positive Effect on Probability of FHA.....	48
Exhibit 21: Distribution of Predictions of the GSE and FHA Loans	50
Exhibit 22: Percent of Loans in Overlap by Confidence Interval Method	51
Exhibit 23: Characteristics of Loans In and Out of Overlap by Confidence Interval Method	54
Exhibit 24: Percent of Loans in Overlap by Parametric Tolerance Interval Method	56
Exhibit 25: Characteristics of Loans In and Out of Overlap by Parametric Tolerance Method.....	57
Exhibit 26: Percent of Loans in Overlap by Non-Parametric Tolerance Interval Method	59
Exhibit 27: Characteristics of Loans In and Out of Overlap by Non-Parametric Tolerance Method ..	60
Exhibit 28: Comparison of Origination Model Results for Loans in FHA/GSE Sectors and Subset of Loans in Non-Parametric Overlap	62
Exhibit 29: Compensating Risk Factors for Overlap Loan by Non-Parametric Tolerance Method.....	63
Exhibit 30: Origination Model Results on FHA vs. PMI within GSE	64
Exhibit 31: Distribution of Predictions of the GSE with PMI and FHA Loans	66
Exhibit 32: Characteristics of FHA and PMI Loans In and Out of Overlap by Non-Parametric Tolerance Interval Method.....	67
Exhibit 33: Origination Model Results on FHA vs. PMI	69
Exhibit 34: Distribution of Predictions of All PMI and FHA Loans.....	70
Exhibit 35: Characteristics of FHA and PMI Loans In and Out of Overlap by Non-Parametric Tolerance Interval Method.....	71
Exhibit 36: Origination Model Results on FHA vs. Subprime.....	74
Exhibit 37: Distribution of Predictions of the Subprime and FHA Loans	75
Exhibit 38: Characteristics of FHA and Subprime Loans In and Out of Overlap by Non-Parametric Tolerance Interval Method.....	76
Exhibit 39: Origination Model Results on GSE vs. Depository Lender.....	79

Exhibit 40: Distribution of Predictions of the GSE and Depository Loans.....	81
Exhibit 41: Characteristics of GSE and Depositories In and Out of Overlap by Non-Parametric Tolerance Interval Method.....	82
Exhibit 42: Default Model Estimated on FHA Data Originated in 1992, 1994 and 1996.....	86
Exhibit 43: Distribution of Predictions of the GSE and FHA Loans	88
Exhibit 44: Percent of Loans in Default Model Market Sectors by Parametric Tolerance Intervals ...	89
Exhibit 45: Analysis of Loans in Default Model Market Sectors (Parametric Tolerance Intervals)....	90
Exhibit 46: Percent of Loans in Default Model Market Sectors by Non-Parametric Tolerance	92
Exhibit 47: Analysis of Loans in Default Model Market Sectors (Non-Parametric Tolerance)	93
Exhibit 48: Origination Model with Predicted Risk of Default Replacing FICO	96
Exhibit 49: Distribution of Predictions of the GSE and FHA Loans	97
Exhibit 50: Analysis of Loans in Market Sectors of Origination Model with Predicted Risk of Default Replacing FICO (Non-Parametric Tolerance Intervals)	98
Exhibit 51: Tract Counts Corresponding to Loan Counts Among FHA and GSE Loans	102
Exhibit 52: Tract Level Descriptive Statistics of FHA Eligible FHA and GSE Loans	103
Exhibit 53: Regression of FHA Share as a Percent Among FHA Eligible FHA and GSE Loans	105

Appendix Exhibits

Exhibit A.1: Notes on Calculations	A-1
Exhibit A.2: Analysis of Matched and Unmatched Experian and HMDA Loans in Baltimore.....	A-6
Exhibit A.3: Analysis of Matched and Unmatched Experian and HMDA Loans in Chicago	A-8
Exhibit A.4: Analysis of Matched and Unmatched Experian and HMDA Loans in Cleveland.....	A-10
Exhibit A.5: Analysis of Matched and Unmatched Experian and HMDA Loans in Denver	A-12
Exhibit A.6: Analysis of Matched and Unmatched Experian and HMDA Loans in Los Angeles... A-14	
Exhibit A.7: Analysis of Matched and Unmatched Experian and HMDA Loans in Oakland	A-16
Exhibit A.8: Analysis of Matched and Unmatched Experian and HMDA Loans in Philadelphia... A-18	
Exhibit A.9: Analysis of Matched and Unmatched Experian and HMDA Loans in Portland	A-20
Exhibit A.10: Analysis of Matched and Unmatched Experian and HMDA Loans in St Louis	A-22
Exhibit A.11: Analysis of Matched and Unmatched Experian and HMDA Loans in Tampa.....	A-24
Exhibit A.12: Analysis of Matched & Unmatched Experian & HMDA Loans in Washington DC	A-26
Exhibit A.13: Analysis of Loans by MSA	A-28
Exhibit A.14: Analysis of Loans by MSA	A-30
Exhibit A.15: Analysis of Loans by MSA	A-32
Exhibit A.16: Origination Model Results by MSA, FHA vs. GSE	A-35

Executive Summary

A Study of Market Sector Overlap and Mortgage Lending

This study examined the extent of overlap between the mortgage market sectors, particularly the Federal Housing Administration (FHA) and the two government-sponsored enterprises (GSEs), Fannie Mae and Freddie Mac. Intuitively, overlap refers to the set of loans that could have gone to either market, for example, either insured by FHA or sold to the GSEs. Home Mortgage Disclosure Act (HMDA) data were supplemented with Experian data to provide FICO credit scores and house values (used to calculate loan-to-value (LTV) ratios). The years covered were 1998 to 2000. There were 11 MSAs for which the match rate between Experian and HMDA loans was at least 56 percent. The matched, loan-level data were divided by mortgage market sector and tabulated to compare borrower, loan, property and neighborhood characteristics. For example, GSE loans have the highest average FICO score (726) followed by private mortgage insurance (PMI) loans (712), depositories (699), FHA (643) and subprime (637). The share of loans in low-income areas (tract income below 90 percent of median income) is nearly the reverse order: subprime (35 percent), FHA (33 percent), depositories (25 percent), PMI (21 percent) and GSE (16 percent).

In order to determine which loans were selected for each mortgage sector, a series of origination models were estimated. For the choice between FHA and GSE, the most important factors were LTV, FICO, payment-to-income ratio and borrower race/ethnicity. The model produces a predicted probability that a particular loan will be FHA insured; this predicted probability conveniently compresses all the credit and non-credit factors into a single dimension. The distribution of predictions for FHA loans was overlaid on the distribution of predictions for GSE loans and the overlap was measured to be 11 percent of the combined set of loans. In other words, 11 percent of the loans in the combined FHA/GSE market have very similar characteristics, such that the model could not distinguish whether they were FHA or GSE loans. In terms of the FHA portfolio and depending on the overlap methodology, between 10 and 14 percent of FHA loans fall in the overlap region. In other words, 10-14 percent of FHA loans have characteristics that are similar to GSE loans. On examination, there are minor differences in income, FICO and LTV among the overlap loans, but overall the FHA loans in the overlap region look remarkably similar to the GSE loans. It appears that the FHA overlap loans were as qualified as many GSE-purchased loans.

A default model was also estimated using separate FHA performance data (that included FICO scores) and then applied to the Experian/HMDA matched data. The coefficients from the default model were used to assign a risk score to the matched loans. This risk score was highly correlated with the FICO credit score, but the risk score alone could not explain the choice of FHA vs. GSE loans. The origination model with its full complement of variables (including the FICO score) does a better job of assigning loans between the market sectors. Still, with respect to overlap, virtually the same results are obtained when the risk score replaces the FICO score in the origination model.

Two methods of overlap measurement were tried. The confidence interval method determines the boundaries of the overlap region based on the 95 percent confidence interval around each loan's predicted probability. If that confidence interval does not include either 0 or 1 (for example, prediction not clearly GSE or FHA), the loan falls in the overlap region. The confidence interval approach has the advantage of being intuitive, in that, when the model cannot determine with 95

percent confidence that the loan is either FHA or GSE, it falls in the overlap region. As explained in the text, there are two issues with the confidence interval method. First, imprecise models with wide confidence intervals create very narrow overlap regions. Second, and more problematic, the overlap region may be distorted by the extreme values of the marginal distribution that do not stand out relative to the combined market distribution. In other words, some high risk GSE loans may look very similar to FHA loans, but may not be representative of the typical underwriting for GSE loans. The tolerance interval method solves the problem of outliers by trimming off the top and bottom five percent for each market participant. A parametric version of the tolerance interval is well-suited when the individual distributions are normal (Gaussian). The preferred method for measuring the overlap of non-normal distributions is the non-parametric tolerance interval method based on order statistics. This method, which is used below, trims the outliers and does not assume normality.

The following table summarizes the measurements of overlap regions for pairs of mortgage market sectors. For each pair, the first row provides the combined overlap (share of similar loans relative to the market or combined sample). The second row gives one participant's share (typically FHA) of the overlap as a percentage of the combined sample. The third row presents FHA's overlap loans as a percentage of the total number of FHA loans. For example, in the FHA vs. GSE market using FHA-eligible loans with LTV between 80 and 100 percent, the combined overlap by the confidence interval method is 20 percent — made up of 9 percent FHA loans and 11 percent GSE loans. The FHA overlap loans are 14 percent of the FHA portfolio. When measured by the non-parametric tolerance interval method, FHA overlap loans are 10 percent of the FHA portfolio. The summary table also provides overlap measurements between FHA and other market sectors. For example, 15 percent of FHA-insured loans are similar in characteristics to privately-insured (PMI) loans.

The implication of this research is that about 10 percent of FHA borrowers have risk characteristics similar to GSE borrowers. It appears that these FHA borrowers could qualify for conventional loans. The measures of overlap presented in this report can serve as a baseline for comparison over time. The GSEs have increased their purchases of LTV loans above 95 percent since HUD conducted the first GSE-FHA overlap study in 1995. In addition, recent GSE commitments to buying more subprime loans indicate there will likely be increased overlap between the FHA and GSE markets in the future.

Summary Table of Overlap Regions

Overlap in Originations		
<u>FHA vs. GSE</u>		
Combined Overlap	20% ^a	11%
FHA Share of Overlap	9% ^b	7%
FHA Overlap loans rel. to FHA Distribution	14% ^c	10%
<u>FHA vs. PMI within GSE</u>		
Combined Overlap		10%
FHA Share of Overlap		7%
FHA Overlap loans rel. to FHA Distribution		9%
<u>FHA vs. all PMI</u>		
Combined Overlap		18%
FHA Share of Overlap		10%
FHA Overlap loans rel. to FHA Distribution		15%
<u>FHA vs. Subprime</u>		
Combined Overlap		13%
FHA Share of Overlap		9%
FHA Overlap loans rel. to FHA Distribution		13% ^d
<u>GSE vs. Depositories</u>		
Combined Overlap		78%
Dep. Share of Overlap		37%
Dep. Overlap loans rel. to Depository Portfolio		73%

^a Interpreted as follows: 20 percent of the FHA-eligible loans from the combined FHA and GSE distributions fall in the overlap region.

^b Interpreted as follows: 9 out of the 20 percentage points in the combined overlap are FHA loans and the remaining 11 percentage points are GSE loans.

^c Interpreted as follows: 14 percent of the FHA distribution of loans (not the combined set, but just the FHA loans) are in the overlap region.

^d The overlap between FHA and subprime is smaller than expected because the LTVs for FHA loans (average 97 percent) are much higher than for subprime purchases (average 81 percent).

Section 1: Background and Literature Review

This document is the Final Report for HUD Contract C-OPC-21895, Task Order CHI-T0003. The purpose of this research is to investigate the extent of overlap between the mortgage market sectors, especially between the Federal Housing Administration (FHA) and the Government Sponsored Enterprises¹ (GSE) loans. Loan level data matched under a previous contract were used to estimate the degree of overlap among home purchase mortgage originations in 11 metro areas. Overlap is defined as home purchase loans that could have gone to either the FHA or the GSE sector. A second component of the research is to estimate a default risk score, based on FHA loan performance, and apply that model to a large set of Home Mortgage Disclosure Act (HMDA) loans reported between 1998 and 2000.

This study finds that about 11 percent² of the combined FHA-eligible GSE and FHA loans with loan-to-value (LTV) ratios between 80 and 100 percent are in the overlap region. In terms of the FHA portfolio, 10 percent of the FHA loans fall in the overlap region. The main implication from this work is that, although it remains rather modest, the overlap between the FHA and GSE mortgage sectors has been increasing as the GSEs strive to meet their housing goals. With more flexible underwriting, made possible by automated underwriting models, the GSEs are more likely to compete with FHA for high loan-to-value (LTV) and high payment-to-income (PTI) loans. As the GSE housing goals are increased, requiring the GSEs to purchase a higher share of loans from low-income and minority borrowers would be expected to increase the overlap between the sectors. With increased competition from the GSEs, FHA lenders may seek to preserve market share by competing with the subprime market for qualified loans. Overall, the increased competition should benefit consumers who will have more choice, if they are willing to shop for the best loan terms.

The Final Report is organized as follows. Section 1 contains the background and a brief literature review that motivates the research questions for the remaining sections. Section 2 describes the data sources and preparation. Home Mortgage Disclosure Act (HMDA) data include information on the loan amount, the income, race and ethnicity of the borrower, and the census tract location of the newly-mortgaged property. HMDA data do not contain credit scores, which are needed to assess the risk of the loan. HMDA also omits house values, which means an important measure of equity, the loan-to-value ratio, cannot be calculated. HUD purchased from Experian data that included credit scores and house values on over 1 million home purchase records from 24 MSAs that were matched to HMDA originations for the origination years 1998 to 2000. The match rates were not uniform across MSAs. The 11 MSAs with the best match rates were used,³ and Section 2 describes how the loans were selected for the subsequent analysis.

¹ In this document, the GSEs refer to Fannie Mae and Freddie Mac. The Federal Home Loan Banks are excluded. It is highly likely that some GSE loans are not conventional in terms of prime risk quality. As shown in Exhibit 7, there are 1,116 GSE loans that are also designated as subprime. The GSE categorization is based on purchaser type 1 or 3 in the HMDA data.

² The 11 percent overlap is based on the preferred overlap method of non-parametric tolerance limits explained in Section 3. The corresponding amount from the confidence interval method is 20 percent, but the tolerance limits method has a stronger theoretical foundation and is less sensitive to model fit.

³ For the 11 included MSAs, the match rate ranged from 0.56 in Baltimore to 0.66 in Portland. The other MSAs are: Chicago, Cleveland, Denver, Los Angeles, Oakland, Philadelphia St. Louis, Tampa and Washington, DC.

The overlap is measured in several different ways and most of the approaches begin with an origination model presented in Section 3.⁴ The origination model is a logistic regression in which the dependent variable is the probability of a loan being insured by FHA relative to the alternative of being sold to the GSEs. Other combinations, such as FHA vs. private mortgage insurance (PMI), FHA vs. subprime, and GSE vs. depository lenders, are also estimated. Taking the initial example of FHA vs. GSE, the overlap region is the subset of home purchase loans for which the 95 percent confidence interval around the predicted probability does not include either 0 or 1. In other words, if the model prediction is not 95 percent certain that the loan is insured by FHA or sold to a GSE, then that loan is defined to be in the overlap region. An alternative definition of overlap is derived from tolerance intervals in which the overlap is the set of loans between the lower limit of the FHA distribution and the upper limit of the GSE distribution. Both parametric and nonparametric methods for setting the tolerance intervals are described in detail.

Overlap can also be viewed in terms of risk of default, as presented in Section 4. In this case, Unicon estimated the probability of claim after 3 years from origination using FHA performance data. From that default model, a default risk score could be predicted for every loan in the matched data. The distribution of FHA loans is shifted to the right along the risk scale relative to the GSE distribution, but there is considerable overlap between the distributions. This suggests that a lot of FHA loans are no riskier than the loans purchased by the GSEs. However, predictions of FHA vs. GSE status based only on the risk score are not very accurate. If the risk score is used in place of the credit score in the origination model, the predictions are virtually the same as with the credit score and the overlap is back to the values in Section 3. Thus, the risk score is a good proxy for the credit score, but apparently non-credit factors are important in determining whether a loan is insured by FHA or sold to the GSEs. Non-credit factors (e.g., neighborhood factors such as center city location and average family income in the census tract) contribute significantly to choice of mortgage sector. These neighborhood variables may reflect the prospect for future property value appreciation.

Although the matched data have only 3 years of originations and 11 MSAs, the loan level data can be organized by census tract. Section 5 contains estimates of the variation in FHA market share across 4,240 tracts representing neighborhood housing markets. Much of the explanatory power of the models comes from the separate indicators for each MSA, but within an MSA the percent minority in the census tract, the percentage of household heads aged 15 to 24 years, the average FHA default rate of the census tract, and the median household income in the census tract are positively related to FHA market share. The median house value of the census tract and the percentage of elderly owners in the census tract tend to reduce FHA market share. These are tentative results that suggest more work needs to be done to fully explain the variation in FHA and GSE market shares.

The Appendix contains more extensive tables with variable definitions and calculation methods along with separate tabulations for each MSA.

⁴ The GSEs do not originate loans, but rather purchase loans originated by the primary market lenders. FHA does not originate loans either, but rather insures loans originated by lenders. The text refers to the origination model to designate that the information about the loan comes from the origination and to distinguish it from the subsequent default model, where the designation is based on FHA claims.

Literature Review

1995 HUD Study. The baseline for this analysis on market overlap comes from a 1995 HUD report titled *An Analysis of FHA's Single-Family Insurance Program*. In that report overlap between FHA and PMI was measured as the loans with LTV between 80 and 95 percent with loan amounts below the FHA loan limits. Loans with LTV below 80 percent did not need private mortgage insurance and loans with LTV above 95 percent could not qualify for private mortgage insurance (at least, at that time). However, most FHA loans had LTV ratios above 95 percent. Moreover, conventional loans typically had to have a payment-to-income ratio below 28 percent and a debt-to-income (DTI) ratio below 36 percent. The corresponding guidelines for FHA were PTI of 29 percent and DTI of 41 percent, but there was considerable flexibility that allowed loans to exceed those guidelines. Given the difference in underwriting guidelines and the strictness of PMI guidelines at that time, most FHA loans did not qualify as conventional loans. However, about 1/3 of FHA loans did have LTV between 80 and 95 percent and thus were considered roughly comparable to conventional loans under the FHA loan limits. Overlap, in the 1995 HUD study, was based on the share of FHA loans that met conventional lending guidelines for LTV and PTI. In that sense, the study selected loans with the potential for overlap. The authors did not predict what subset of those FHA loans had the combination of characteristics that made them indistinguishable from conventional loans, at least for a statistical model's point of view.

In HUD's 1995 study, a linear probability model was estimated on the potential overlap loans to determine which characteristics affected the probability of the loan becoming insured by FHA versus PMI. The factors that increased the probability of a loan being FHA-insured were: payment-to-income ratio, first-time homebuyer, black race of borrower, Hispanic ethnicity of borrower, center city location of property, tract median family income relative to MSA median family income, and the percent minority households in the tract. Factors that decreased the probability of a loan being FHA-insured were: borrower age, loan amount relative to FHA loan limit, LTV, borrower income relative to MSA median family income, and other race of borrower (white being the reference group). Given that high LTV is a distinguishing feature of FHA loans, it is a little surprising that LTV had a negative effect on the probability of FHA. However, the selection of potential overlap loans excluded the high (over 95 percent) LTV loans in FHA, but included the mass of conventional loans at the maximum LTV of 95 percent. A further subdivision of loans into high, medium, and low cost areas based on FHA loan limits, showed that the coefficients on tract income and center city location became negative in the high cost areas. Unfortunately, the data on 1993 originations (used in the 1995 study) did not include credit scores.

The implication of the HUD report was that there existed little overlap between the FHA and conventional mortgage markets. FHA tended to serve low-income, minority, young and first-time homebuyers. Rather than treating FHA and conventional as separate markets, HUD promoted increased purchases of loans from low-income and minority borrowers through the GSE housing goals.

Other Studies. The following papers did not measure overlap, but they did provide valuable information about the specifications for loan choice models (including neighborhood effects) and they developed a theory about the FHA market sector relative to the conventional market.

Berkovec, Canner, Gabriel, and Hannan (1998) wrote a well-known paper on discrimination in which they used 1987-89 FHA performance data to test whether minorities have lower default rates than non-minorities. The theory, based on Becker (1971), was that minorities would have to meet a higher standard than whites for their loans to be accepted by a discriminating lender. The authors found that minorities have a higher probability of default, indicating that lenders did not discriminate. This study has some similarities to Berkovec et al. in that a logistic regression is estimated with FHA data and the impact of market concentration is measured using Herfindahl-Hirschmann indices. However, this study does not focus on discrimination, but rather the overlap between the FHA and conventional originations. In addition, the data are more recent (1998-2000) and include FICO scores, which were not available for the earlier discrimination study.

A subsequent study by Cotterman (2002) replicated the Berkovec et al. (1998) analyses on FHA data from a more recent time period, but also included credit scores in parallel analyses. When credit scores were introduced, the estimated minority coefficients tended to fall, often becoming statistically insignificant and sometimes changing sign. When credit scores were included, the empirical results no longer gave unambiguous support for the notion that lenders do not discriminate. The change in results also suggests that the original work suffered from omitted variable bias, despite considerable efforts to control for omitted variables.

In 1998, Onder investigated the neighborhood factors affecting FHA market share with a two-stage model. In the first stage the probability of a loan being FHA is regressed on loan level characteristics and a set of tract dummies. In the second stage, the coefficients on the tract dummies are regressed on a set of tract characteristics. The idea is to determine which characteristics are associated with the sign and size of the tract fixed effect. The research showed that minority composition was not significant, and there was a negative relation between tract median family income and the likelihood of being FHA-insured. Interestingly, census tract income had a positive effect on FHA for values below \$30,000 and negative effect above \$30,000. Also, although the level of minority composition was not significant, an upward change in minority share greater than 15 percent over the previous decade was a positive factor in a loan being insured by FHA. High rent levels had a negative relation to FHA, but rent increases were positively associated with FHA. Similarly, high vacancy rate had a negative relation to FHA, but vacancy rate increases had a positive association to FHA. On a national pool of 35,464 tracts, the R-square was 0.39, which jumped to 0.72 when the specification included 333 MSA dummies. Apparently, there remained significant differences between MSAs even after controlling for an extensive list of individual and neighborhood effects.

Pennington-Cross and Nichols (2000) filled in a gap left by earlier research by including new data on credit history. A national loan level sample included originations from 1995 and 1996 and represented 306 MSAs. They showed that there was a considerable overlap between FHA and conventional loans in terms of the distribution of credit scores. Their research also showed that the credit score was an important ingredient in loan choice. The probability of a loan being FHA declined with higher credit scores.

Using the same data, Ambrose and Pennington-Cross (2000) estimated an FHA market share equation using logistic regression. Concerned that LTV may be jointly chosen with FHA, the researchers estimated an instrumental variable equation as a first stage and used the predicted LTV in the FHA market share equation. The coefficient results showed the following metropolitan area factors had a positive association with FHA market share: unemployment rate, segregation of blacks, percent underserved, and FHA loan limit relative to the median house price for the MSA. The negative

coefficients were for: 1-year and 10-year house price change, annual volatility in house prices, and higher minority share. The unexpected result on minority share may be because underserved is also included in the specification and minority share is an important component in underserved status. From these results, the authors concluded that FHA market share was higher in cities with greater economic risk characteristics. GSE purchase rates were fairly insensitive to local economic conditions.

Ambrose, Pennington-Cross, and Yezer (2002) provided a more detailed theoretical explanation of the interface between FHA and conventional market sectors. They assumed that loans could be ordered by a single risk factor and conventional underwriting determined the upper limit on acceptable risk. FHA has more lenient underwriting standards, so the higher risk loans rejected by conventional lenders may be acceptable to FHA. They call this the FHA wedge. In this view, the amount of overlap between the markets is quite small. A few loans may go to FHA that could have qualified for conventional lending, but these are basically a mistake due to insufficient shopping by the borrower or steering by the lender. The reason it is considered a mistake is that mortgage insurance for conventional loans is less expensive than FHA mortgage insurance (at least it is for loans with an LTV ratio less than or equal to 95 percent). FHA charges a higher premium that corresponds to the higher risk and claim rate for most loans in the FHA wedge. So, if a loan could qualify as a conventional loan, it would be less expensive for the borrower to have a conventional loan than an FHA loan. Historically, private mortgage insurance was not available for such high-risk loans, though in recent years the insurance has become available at a higher rate than FHA charges. Thus, they conclude that the overlap between FHA and conventional loans comprises a small set of loans. If borrowers with low-risk FHA loans had conducted a more thorough search, they would have realized their mistake and pursued a conventional loan.

Building on the data and analysis in their 2000 paper, Ambrose, Pennington-Cross, and Yezer (2002) estimate an FHA market share model (at the metropolitan area level) with measures for cyclical risk and permanent risk. The cyclical risk factors include local unemployment rate and the percent change in delinquent bank loans. The permanent risk factors include volatility in house price appreciation, average default rate over the past six years, share of low-income households, and the percent of loans with loan amount relative to income greater than three. The findings showed positive coefficients for: change in unemployment rate, change in delinquency rate, average delinquency rate, volatility in house prices, share of incomes below \$20,000, and percent black. The variables with negative coefficients were current and lagged house price change, loan-to-income greater than three, and black segregation (Gini coefficient). It is interesting to note that the house price volatility, black segregation, and percent black had reversed signs in the 2000 paper.

The authors conclude that conventional underwriting does not adjust to local risk factors in order to maintain market share. Rather, non-price credit rationing by conventional lenders leaves FHA with the role of maintaining the mortgage credit supply in declining housing markets. These effects from the 1995-1996 data may have been less apparent during the 1998-2000 period when housing markets were more uniformly strong. Indeed, both data sets relied on cross-sectional variation rather than a full cycle or major regional recession.

Freeman, Galster and Malega (2003) provide an in-depth empirical analysis of the secondary mortgage market impacts on underserved areas of Cleveland during 1993-1999. Based on single family home sales by census tract, the researchers found that secondary mortgage purchases, particularly by non-GSE buyers, had a positive effect on the number of sales transactions with a one

year lag. The increase in purchases did not affect sales prices, though there is some evidence that non-GSE purchases of refinances did boost prices one to two years later. Gyourko and Hu (2002) did a broader study in 20 major metropolitan areas and found a spatial mismatch between GSE purchases in low-income and minority areas and the demand for affordable housing. These studies indicate a modest degree of competition for loans from low-income and minority borrowers, but do not quantify the degree of overlap, i.e., how many loans could have qualified for GSE and non-GSE purchases.

To measure market overlap, this study follows the loan choice literature using logistic regression. However, an alternative approach using discriminant analysis is described by Amemiya (1985, pp. 281-285). The maximum likelihood approach is robust to non-normal covariates, though it may not be as efficient as discriminant analysis in some cases. This comparison is left for future research.

Recently there have been several media announcements from the GSEs that they intend to increase their purchases of loans from low-income and minority households, as well as subprime loans. For example, the National Mortgage News (Oct. 25, 2004) reported on an interview with the new CEO of Freddie Mac, Richard Syron, in which Mr. Syron is quoted as saying (p. 86), "I'd like to be more aggressive in the minority and Hispanic markets, yes. We will push. It's what we are supposed to be doing but it's good business." Regarding subprime, Mr. Syron said the Freddie Mac credit losses are about one percent on subprime loans and he feels the company can afford to take on more credit risk. Reported in Origination News (www.originationnews.com/plus/#4 on 10/21/2004), Eugene McQuade, Chief Operating Officer at Freddie Mac, announced at the America's Community Bankers convention in Washington, DC, that the company was simplifying its A-minus loans and related low-downpayment products. At the same convention, Franklin Raines, Fannie Mae chairman and chief executive, declared that Fannie Mae intended to be more aggressive in serving the subprime market. He said the subprime market is estimated to be \$323 billion and growing. Mr. Raines said, "We estimate that about half of the subprime borrowers have only slightly blemished credit and are just a notch away from qualifying for Fannie Mae's prime conventional financing." The main point is that as the GSEs become more aggressive about purchases of low-income and minority loans, it is highly likely that this will entail more overlap with FHA and subprime lenders that have traditionally served those borrowers.

Research Questions

Given this background from the literature, the goal of this study was to update and broaden the overlap findings from the literature. Previous work was updated by estimating FHA market share models on more recent data that included controls for credit scores. The results are broader because they included mortgage market sectors for GSEs, depository lenders, FHA, and subprime. A limited attempt at explaining FHA market shares based on tract level information was also attempted. The research questions were the following:

- 1) What are the borrower, loan, property and neighborhood characteristics associated with each mortgage market sector?
- 2) What factors determine the market shares captured by each mortgage sector?
- 3) Is there a significant degree of overlap between FHA and GSE sectors such that those loans could have gone to either FHA or GSE?
- 4) How much overlap is there in terms of default risk?

- 5) If the credit score is replaced by a default risk score in the origination model, does that attenuate the importance of non-credit factors?

Section 2: Data Preparation and Description

The data used in this analysis come from three sources: HMDA (1998-2000), Experian (1998-2000) and Census (1990). HMDA data provide nearly a complete set of loans for metropolitan areas.⁵ Although HMDA data provide the loan amount, borrower race, tract location, and much more, it does not include credit scores or house values (needed to calculate LTV). To bridge this gap, loan record data were purchased from Experian for a select set of MSAs and HUD merged the Experian data with the HMDA data. Additional neighborhood information was obtained from the 1990 Census. This section describes the process of selecting, merging and cleaning the data along with tabulations of the data used in the origination models. More detailed information about individual variables or tabulations by MSA can be found in the Appendix.

HMDA Data

Depository and other financial institutions report their mortgage loan activity to the Federal Financial Institutions Examination Council (FFIEC), which makes a subset of the data available to the public for analysis.⁶ The HMDA data comprise the most comprehensive source for mortgage lending information and HMDA data were used as the benchmark for weights. For each loan, there is information on loan type (especially conventional vs. FHA), loan purpose (home purchase, improvement, refinancing or multifamily dwelling), action taken (originated, denied, withdrawn) and type of purchaser (GSE, Ginnie Mae, commercial bank, etc.). For this research, newly-originated home purchase loans were selected. Besides a listing of loans, HMDA data provide the MSA and tract location, which makes it possible to merge in other data, particularly Census data, at the tract level. In addition, HMDA data include income and race/ethnicity of the borrower, which enabled this study to focus on low-income and minority borrowers.

Experian Data

Unfortunately, HMDA data do not include two pieces of information crucial to the assessment of risk and the underwriting process, namely credit score and house value. These data were obtained from a private vendor, Experian Information Solutions, Inc., under a previous HUD contract.^{7,8} Twenty-four metropolitan areas were selected for their diversity of geographic location, housing appreciation rates, housing prices, and broad representation of the nation. Within each MSA, the census tracts were

⁵ According to the HMDA website (<http://www.ffiec.gov/hmda/default.htm>), in 2000 all depository institutions with assets exceeding \$30 million and a metropolitan office that originated a home purchase loan or refinancing secured by a single family home must report their loans. Other for-profit mortgage lending institutions with home purchase loan originations at least 10 percent of total loan originations, a metro office, and assets of \$10 million or originated at least 100 loans, must also report their mortgage loans to HMDA.

⁶ More complete description of HMDA data can be found in, *A Guide to HMDA Reporting: Getting It Right!* Published and updated frequently by the Federal Financial Institutions Examination Council (<http://www.ffiec.gov/hmda/guide.htm>).

⁷ The previous HUD contract was C-OPC-18571, Task Order 9.

⁸ Much more information on the Experian data is recorded in the “Experian Data Report,” HUD, Policy Development and Research, May 30, 2003.

stratified according to underserved status (based on income and percent minority in 1999). Underserved tracts typically have fewer mortgage loans per year, so those tracts were sampled at a higher rate than the served tracts. A prioritized list of tracts was given to Experian and they extracted all the home purchase loans from those tracts that originated in 1998 to 2000 (over 1 million loans). The loan information came from county recorders. The credit score is the FICO score based on the borrower characteristics at approximately the time of the origination.

Researchers at HUD merged the Experian loan data with the HMDA data using all the loan level data available.⁹ First, the match is based on geography. Both data sets have state, county, tract and MSA information geocoded. In addition, both databases have variables for loan amount, race, gender and loan type (conventional, FHA, Veterans Administration and Farmers Home Administration). The matching process goes through six iterations in which the best matches are removed and the remaining records are compared using fewer variables or wider bounds for a match. For example, in the first iteration, race, loan amount, gender, and loan type must all be equivalent for the loans to qualify as a match. In the next iteration, the race variable is dropped from the matching requirement. In the following round, race is brought back and loan amount is dropped. By iteration four, race and loan amount are dropped. Then in iteration five, race, loan amount and gender are dropped. Finally in iteration six, race, loan amount, gender and loan type are dropped. The matches are screened for unacceptable matches (race/ethnicity is different, loan amount differs by more than \$3000, gender is different or both missing, or loan type does not match). There is also a tie-breaking protocol used in case where more than one loan record qualifies for a match. Out of the original 24 MSAs, the 11 MSAs with the best match rates were selected for analysis.

Weights are assigned to the matched loans so that the sum of the matched loans equaled the sum of the HMDA loans in each tract. The weight starts with a base weight according to the probability of selecting the tract multiplied by the probability of selecting the loan within a particular tract. Separate weights are needed for the served and underserved strata. To correct for missing loans and non-matches, adjustment factors are assigned to ensure the weighted total of matched loans equals the HMDA totals. For example, after the matching is completed, if the weighted total of loans for a tract is 90 percent of the HMDA total, then an adjustment factor of 1.11 is applied so that the weighted sample matches the HMDA total. Final weights are associated with each matched loan and used in the tabulations presented below. Separate weights have not been designed for the FHA-eligible subset or the FHA vs. GSE subset of loans. The assumption is that the tract level weights for the full, matched sample is adequate for subsets of loans drawn from the full sample.

A comparison between the matched and non-matched loans is shown in Exhibit 1 for the variables used to select the sample and conduct the regression analysis. The analogous tables by MSA are in the Appendix. To the extent that HMDA is representative of the universe of loans and the weights are designed to match HMDA, then the averages and distributions for the matched data were representative of the home purchase loans in the 11 MSAs during 1998-2000. Differences between the unmatched and matched data do not imply that the matched data were incorrect or unrepresentative because the unmatched data were unweighted. If a higher percentage of Experian loans had been successfully matched, that would have resulted in different unweighted values and

⁹ The data matching was done by Ismail Mohamed and Ron Hanson of Titan Systems, Inc. and Jay Schultz of the Office of Policy Development and Research, HUD.

different weights, but not necessarily different weighted values. Lacking more complete data, the weighted, matched sample provides the most representative estimates for the population values.

For the pooled sample of 11 MSAs, the median income in the matched sample, \$55,000, is similar to the median income in HMDA, even though the weighted mean in the matched sample is much higher. As noted, there are extreme values in the reported incomes. The average FICO score in the matched sample (696) is higher than the unmatched Experian data, but a quarter of the unmatched data were missing FICO. FICO was a required field for the matching process. To make the distributions of unmatched Experian data more comparable to the matched data, the distributions are calculated on the non-missing loans. The matched data have a higher share of whites, but a lower share of race missing. The age distribution is very similar after adjusting for the 29 percent missing in the unmatched data. The average loan amount is higher in the matched data. However, the ratio of the loan amount to the FHA loan limit is essentially the same in the matched and unmatched data.

The distribution of LTV is shifted higher in the matched data, which has an average LTV of 84 (vs. 71 percent in the unmatched).¹⁰ It is possible that the higher average LTV in the matched data accentuate the degree of overlap between FHA and GSE. Another notable difference is the lower percentage of new construction in the matched data compared to the unmatched data, 11 vs. 19 percent respectively.

The neighborhood characteristics in the matched and HMDA data are generally close. This is no surprise because the weights are at the tract level and should eliminate any substantial differences at the neighborhood level. The unmatched sample has a lower share of minority neighborhoods than the Experian data and a higher share of loans in low-cost MSAs.

Overall, the largest difference that would affect the study results is the higher average LTV in the matched data. Unfortunately, there are no LTV data in HMDA for comparison. If the weighted matched data have an upward bias for LTV, it is likely that the estimates for overlap are also biased upwards.

¹⁰ As shown in Exhibit 1, 22 percent of the matched sample had an LTV greater 98 percent, compared with 9 percent in the unmatched sample.

Exhibit 1: Analysis of Matched and Unmatched Experian and HMDA Loans from 11 MSAs (1998-2000)

Characteristics	Experian Unmatched (Unweighted)	Experian/HMDA Matched (Unweighted)	Experian/HMDA Matched (Weighted)	HMDA Unmatched	All HMDA
Borrower Characteristics					
Unweighted Number of Borrowers	239,529	347,732	347,732	1,589,133	1,936,865
Weighted Number of Borrowers			1,980,080		
Average Annual Income	\$71,052	\$60,986	\$63,145	\$69,041	\$67,565
Median Annual Income	\$62,500	\$52,000	\$55,000	\$59,000	\$58,000
Average Annual Income (Trimmed Top 1%)	\$69,527	\$58,440	\$58,458	\$65,827	\$64,485
% Estimated Income Information	0%	2%	2%	2%	2%
Average FICO	690	694	696		
% With FICO < 620	22%	20%	20%		
% With FICO 620 - 679	18%	17%	17%		
% With FICO => 680	61%	62%	63%		
% Missing FICO Information	25%	0%	0%		
% White		71%	67%	63%	65%
% Black		10%	11%	9%	9%
% Hispanic		10%	12%	10%	10%
% Other		6%	7%	8%	7%
% Missing Race Information		3%	3%	10%	8%
% Female	21%	28%	28%	26%	26%
% Male	79%	72%	72%	67%	68%
% Missing Gender Information	7%	0%	0%	6%	5%
% Age 19-34	31%	32%	32%		
% Age 35-49	51%	49%	50%		
% Age 50-64	15%	15%	14%		
% Age >65	4%	4%	3%		
% Missing Age Information	29%	0%	0%		
Loan Characteristics					
Average Loan Amount	\$124,847	\$122,385	\$128,684	\$119,205	\$119,776
Average LTV %	71%	84%	84%		
% With LTV <= 90	72%	56%	55%		
% With LTV 90 - 96	12%	15%	15%		
% With LTV 97 - 98	6%	7%	8%		
% With LTV > 98	9%	22%	21%		
% Missing LTV Information	37%	0%	0%		
Average Ratio of Loan Amount to FHA Loan Limit	74%	72%	73%		
% With LoantoFHA Ratio <=.5	24%	25%	23%		
% With LoantoFHA Ratio of .6 - 1	56%	58%	59%		
% With LoantoFHA Ratio of 1.1 - 1.2	12%	10%	11%		
% With LoantoFHA Ratio > 1.2	8%	7%	6%		

Exhibit 1 (cont.): Analysis of Matched and Unmatched Experian and HMDA Loans from 11 MSAs (1998-2000)

Charateristics	Experian Unmatched (Unweighted)	Experian/HMDA Matched (Unweighted)	Experian/HMDA Matched (Weighted)	HMDA Unmatched	All HMDA
Average PTI	19%	20%	20%	17%	18%
% Originated in 1998	23%	25%	32%	32%	31%
% Originated in 1999	35%	35%	34%	34%	34%
% Originated in 2000	43%	40%	34%	34%	35%
Mortgaged Property Characteristics					
% Old Construction	81%	89%	89%		
% New Construction	19%	11%	11%		
% Missing Construction Information	9%	0%	0%		
% Unit Size 1	96%	96%	95%		
% Unit Size 2	1%	1%	2%		
% Unit Size 3	2%	2%	2%		
% Unit Size 4	1%	1%	1%		
Borrower Neighborhood Characteristics (1990 Census)					
% In Underserved Tracts		34%	32%	31%	32%
% Not in Underserved Tracts		66%	68%	67%	67%
% Missing Underserved Tracts		0%	0%	2%	1%
% in High Cost Cities	10%	10%	12%	13%	12%
% in Average Cost Cities	78%	84%	85%	85%	85%
% in Low Cost Cities	12%	6%	3%	2%	3%
% In Center City	25%	28%	27%	27%	27%
% Not in Center City	75%	72%	73%	71%	71%
% Missing Center City Information	0%	0%	0%	2%	1%
Average 5-year Appreciation Lagged 1 year	117%	117%	113%	112%	113%
% In Area with Depreciation	4%	5%	11%	12%	11%
% In Area with Appreciation up to 20%	63%	63%	71%	73%	71%
% In Area with Appreciation over 20%	33%	33%	18%	15%	18%
% In <90% Relative Income Tracts	26%	26%	23%	24%	24%
% In 90-120% Relative Income Tracts	32%	33%	34%	32%	32%
% In >120% Relative Income Tracts	42%	41%	43%	44%	44%
% In <10% Minority Tracts	51%	50%	42%	40%	41%
% In 10-30% Minority Tracts	29%	30%	32%	34%	33%
% In >30% Minority Tracts	20%	20%	26%	25%	24%
% Missing Minority Tract Information	0%	0%	0%	2%	1%

Note: Loans in Boston MSA and Jumbo loans are excluded from this analysis. Share of missing is so large for the unmatched Experian loans that the distributions are calculated for the nonmissing loans in column 1.

Data Cleaning and Record Selection

In building analysis files from the matched Experian-HMDA data, there are several selections, as shown in Exhibit 2. The total number of matched records for 12 MSAs is 393,643, but Boston did not have loan type and jumbo loans are excluded from all analysis, leaving 347,732 loans in 11 MSAs. After excluding non-fixed-rate mortgages and loan amounts greater than the FHA loan limits, the FHA-eligible file contains 238,158 loans. A final selection is made for LTV between 80 and 100 percent and FHA or GSE loans to reach a working file of 114,780 loans. The LTV is calculated simply as the reported loan amount divided by the reported house value. The FHA designation is based on a match to the complete set of FHA loans. The GSE designation is based on the purchaser type in HMDA being either 1 (Fannie Mae) or 3 (Freddie Mac). Although there are FHA loans in the GSE portfolios, in the regression analysis the FHA loans are taken out of the GSE group.

Exhibit 2: Methodology for Limiting HUD Data sets

	Count of Loans		
	Add	Subtract	Net
Start with 'hmda_ex_selected_pmsa_1998_2000'	315,625		315,625
Append 'hmda_ex_chicago_1998_2000'	51,900		367,525
Append 'hmda_ex_la_1998_2000'	26,118		393,643
Exclude Boston loans (MSA<>1120)		12,009	381,634
Exclude remaining jumbo loans (Conform=1)		33,902	347,732
	393,643	45,911	347,732
Exclude remaining non-fixed rate mortgages (ex_rate_type<>B, U, or V)		61,847	285,885
Exclude remaining loans above FHA limit		47,727	238,158
		109,574	238,158
Exclude remaining loans not between 80-100% LTV		80,146	158,012
Limit to FHA loans or loans in pur_type=1 (FNMA) or 3 (FHLMC)		43,232	114,780
		123,378	114,780

In the large analysis file (347,732 loans), there are 61,922 PMI loans. Traditionally, mortgage insurance was required by the GSEs for loans with LTV greater than 80 percent. Although most PMI loans have LTV greater than 80 percent, 28 percent of the PMI loans have LTV less than or equal to 80 percent, as shown in Exhibit 3. The lower panel shows that most of the below-80 percent loans are in the range of 70 to 80 percent, as expected, but there are some much lower. No attempt was made to impute or exclude those loans in the PMI analysis; that is, they were included.

Exhibit 3**PMI Loans by LTV Ratio****On All Matched, Conforming Loans, Weighted
(n=347,732; weighted sample=1,980,080)**

LTV Ratio	Number of Loans with PMI	Percent of Loans with PMI
<=80	17,615	28%
81-85	3,280	5%
86-90	9,740	16%
91-95	17,403	28%
96-97	10,033	16%
>97	3,851	6%
Total PMI Loans	61,922	100%

Low LTV PMI Loans by LTV Ratio**On All Matched, Conforming Loans, Weighted
(n=347,732; weighted sample=1,980,080)**

LTV Ratio	Number of Loans with PMI
<=10	17
11-20	19
21-30	60
31-40	172
41-50	394
51-60	848
61-70	1,443
71-80	14,662

The intersection of PMI with GSE loans is shown in Exhibit 4. Out of 115,798 total GSE loans, 69 percent or 79,777 are without PMI and 31 percent or 36,021 have PMI. Most of the GSE loans without PMI have LTV less than or equal to 80 percent, and most of the GSE loans with PMI have LTV greater than 80 percent. However, there are many cases (10,206) of GSE loans below 80 percent LTV with PMI and GSE loans above 80 percent LTV without PMI (23,975). There are even a substantial number (1,797) of GSE loans with PMI above 97 percent LTV. Apparently, PMI is not a hard constraint on the GSEs for their high-LTV purchases and the GSEs are using other forms of credit enhancement. It is likely that there are more than 48 percent of GSE loans above 80 percent with PMI, but they were not successfully matched either between Experian and HMDA or PMI and HMDA.

Exhibit 4: GSE Loans¹ With & Without PMI by LTV Ratio
On All Matched, Conforming Loans, Weighted
(n=347,732; weighted sample=1,980,080)

LTV Ratio	Without PMI	Percent Without PMI	With PMI	Percent With PMI
<=80	55,802	85%	10,206	15%
>80	23,975	48%	25,815	52%
Total GSE Loans	79,777	69%	36,021	31%

LTV Ratio	Without PMI	Percent Without PMI	With PMI	Percent With PMI
<=80	55,802	85%	10,206	15%
81-85	6,205	76%	1,933	24%
86-90	4,325	43%	5,793	57%
91-95	6,625	39%	10,458	61%
96-97	3,575	38%	5,834	62%
>97	3,245	64%	1,797	36%
Total GSE Loans	79,777	69%	36,021	31%

1. Percent of GSE Loans with an LTV above 80% that do not have PMI: 48%

There are 93,606 total FHA loans with dwelling unit size as reported by Experian in Exhibit 5. All of these loans are supposed to be associated with single-family properties (1 to 4 units), but a sizable portion of the units is reported to be in buildings with 5+ units. Given that out of the entire set of FHA loans, only 347 exceed the loan limits for a single unit that do not exceed the loan limits for 4 units, the loans in 5+ unit buildings are treated as single-unit loans. Loans in 2- to 4-unit buildings are assumed to be loans for multiple units and the FHA loan limits are adjusted accordingly.

Exhibit 5: FHA Loans by Experian Dwelling Unit Size

Dwelling Unit Size	"EX_UNIT_SIZE" Category	Number of FHA Loans	Percent of FHA Loans
Single	C, Missing	79,533	85%
Duplex	D	1,332	1%
3-unit	E	1,568	2%
4-unit	F	848	1%
Larger	G	4,920	5%
Larger	H	797	1%
Larger	I	148	0%
Larger	J	613	1%
Larger	K	71	0.1%
Larger	L	1,569	2%
Larger	M	922	1%
Larger	N	576	1%
Larger	O	709	1%
TOTAL FHA LOANS		93,606	100%

Percent of FHA Loans for Units in Dwellings Larger than 4 units: 11%

"EX_UNIT_SIZE"		Number of	Percent of
Category	Loan Limit Size	FHA Loans	FHA Loans
C, G-O, Missing	1	89,858	96%
D	2	1332	1%
E	3	1568	2%
F	4	848	1%
TOTAL FHA LOANS		93,606	100%

FHA eligibility is based on FHA loan limits, which can change throughout the year. This analysis uses the loan amount reported in HMDA, which is rounded to the nearest \$1000. The FHA designation comes from a match to FHA records, so those loans are expected to be below the FHA loan limits at the time of origination. However, as shown in Exhibit 6, there are 2,345 FHA loans (2.5 percent of 93,606 total FHA loans) above the FHA loan limits for the date and MSA of origination. The data appear to be inconsistent. The limits could have been increased until nearly all the FHA loans were included, but the concern was that would include many GSE loans that were actually ineligible at their time of origination. Therefore, the more conservative approach for FHA eligibility of applying the FHA loan limit as of the origination date was chosen, even though this excludes 2,345 FHA loans.

Exhibit 7 shows the number of missing observations associated with a few of the variables that are not completely available. By far, the biggest problem of incomplete data is due to race/ethnicity not being reported on HMDA.

Exhibit 6: Number of FHA Loans Above FHA Loan Limits¹

On All Matched, Conforming Loans, Weighted
(n=347,732; weighted sample=1,980,080)

MSA	FHA Loans Above FHA Loan Limits	FHA Loans Above FHA Loan Limits2
	Unrounded	Rounded Up to Nearest \$1000
0720	90	77
1600	213	190
1680	122	117
2080	613	553
4480	285	252
5775	144	125
6160	131	99
6440	320	281
7040	97	80
8280	405	334
8840	272	237
Total	2,692	2,345

FHA Loan Limit Source: FHA_limits_holly.xls.

1. Number of FHA loans above loan limits, excluding loans not in ex_unit_sizes C, D, E, F or missing. Loan limits assigned as follows by ex_unit_size: if C or missing, then Unit Size 1, if D then Unit Size 2, if E then Unit Size 3, if F then Unit Size 4, if G or higher then Unit Size 1 (assumed single unit in larger building). In 1998, FHA loan limits for units in categories D, E and F are estimated based on loan limits for category C. D limit = C limit*1.28. E limit = C limit*1.55. F limit = C limit*1.92.

2. Loan limits based on loan limits as of October 21, 1998; December 1999; and January 2000. Note that 2000 loan limits do not change. Loans are compared to loan limits according to HMDA Action Date.

Exhibit 7: Observations with Missing Data

On All Matched, Conforming Loans, Weighted
(n=347,732; weighted sample=1,980,080)

Missing Data	Number of Loans	Reason
OBS_POST_WGHT	169	No weights. Model ignores these observations.
Race	10,665	BO_Race in categories 0, 7 or 8.
Center City	86	No information provided by Unicon for 2 tracts in our data.
Age	721	No age or age range information.

There are 19,330 subprime loans in the large analysis file, as shown in Exhibit 8. The subprime designation is based on the lender name (HUD determined which agencies were predominantly subprime lenders in each origination year). Most of the subprime loans (53 percent) have a purchaser type of life insurance company or other, although 41 percent are held by depositories. A small share (5 percent) are insured by FHA and 6 percent are purchased by GSEs. It is quite possible there are errors in the designation of subprime loans and that the subprime mortgage sector is considerably larger than represented in this data set. Most subprime loans are refinance loans, which are not considered in this research.

**Exhibit 8: Subprime Loans¹ by Mortgage Market Sector²
 On All Matched, Conforming Loans, Weighted
 (n=347,732; weighted sample=1,980,080)**

Sector	Number of Subprime Loans	Percent of Subprime Loans
FHA	994	5%
PMI	1,507	8%
Other Investor	10,333	53%
Depository	7,833	41%
GSE	1,116	6%
TOTAL Subprime Loans³	19,330	

1. Subprime data downloaded from www.huduser.org/datasets/manu.html. Merged on compressed AGENCY and RESP_ID. SUBPRIME=1 if lender classified as a primarily subprime lender in year of loan origination.

2. Sectors assigned as follows:

FHA EX_LOAN_TYPE_HUD2=F
 PMI PMI_FLAG="Y"
 Other Investor PUR_TYPE=7, 9
 Depository PUR_TYPE=0, 5, 6, 8
 GSE PUR_TYPE=1,3

3. Total does not equal sum of subprime loans in each sector because the sectors overlap, e.g, PMI overlaps with GSE.

The cross tabulation of HMDA purchaser type with FHA, PMI and subprime mortgage sectors is shown in Exhibit 9. The FHA column shows that 1.6 percent of the FHA loans were purchased by Fannie Mae (usually as part of a batch sale). Typically most FHA loans are insured by Ginnie Mae, but according to purchaser type only 46.5 percent are reported under Ginnie Mae. The PMI column shows that most PMI loans are purchased by the GSEs (33.7 percent by Fannie Mae and 24.1 percent by Freddie Mac), but a substantial fraction (42 percent) are held by other institutions.

Exhibit 9: Distribution of Loans by Purchaser Type
On All Matched, Conforming Loans, Weighted
(n=347,732; weighted sample=1,980,080)

	<u>HMDA</u> <u>Purchaser Type</u>	<u>FHA</u>	<u>PMI</u>	<u>Subprime</u>
0	Loan not originated or sold in calendar year	10.6%	20.2%	24.9%
1	Federal National Mortgage Association	1.6%	33.7%	2.3%
2	Government National Mortgage Association	46.5%	0.0%	0.2%
3	Federal Home Loan Mortgage Association	0.1%	24.1%	4.7%
4	Farmers Home Administration	0.2%	0.0%	0.0%
5	Commercial Bank	2.9%	1.1%	1.5%
6	Savings Bank or Association	0.6%	0.8%	0.4%
7	Life Insurance Company	0.1%	0.0%	0.0%
8	Affiliate Institution	2.5%	7.5%	11.0%
9	Other	34.9%	12.6%	54.9%

Explanatory Variables

The goal was to use the combined Experian/HMDA data, along with census tract level data and OFHEO MSA level house price appreciation data, to capture the most important factors in the underwriting decision. The FICO score provided credit history information, and house value made it possible to calculate LTV. In addition, borrower income made it possible to calculate the payment-to-income ratio, and demographics provided measures of age, gender, and race. The following description provides more details about how each explanatory variable was calculated and then how these variables were used to subset the data for analysis of mortgage sectors. A data dictionary is also provided in Appendix Exhibit A.1.

In Exhibits 10 through 14, borrower, loan, property, and neighborhood characteristics are evaluated to determine how the pools of loans change as the data were categorized according to the area of greatest potential overlap between FHA and the GSEs. Note that Exhibits 10 to 13 continue on a second page.

Exhibit 10: Comparison of Characteristics by Dataset
Weighted

	Matched, Conforming, Fixed- Rate, FHA-Eligible FHA & GSE Loans with LTV 80-100%	Matched, Conforming, Fixed-Rate, FHA- Eligible Loans	Matched, Conforming Loans
Borrower Characteristics			
Unweighted Number of Borrowers	114,780	238,158	347,732
Weighted Number of Borrowers	674,238	1,369,923	1,980,080
Average Annual Income	\$52,438	\$55,749	\$63,145
Median Annual Income	\$47,000	\$49,000	\$55,000
Average Annual Income (Trimmed Top 1%)	\$50,697	\$53,596	\$60,552
% Estimated Income Information	1%	2%	2%
Average FICO	668	693	696
% With FICO <620	29%	21%	20%
% With FICO 620-680	21%	17%	17%
% With FICO =>680	50%	62%	63%
% Missing FICO Information	0%	0%	0%
% White	57%	65%	67%
% Black	16%	12%	11%
% Hispanic	19%	14%	12%
% Other	5%	6%	7%
% Missing Race Information	3%	3%	3%
% Female	31%	31%	28%
% Age 19-34	38%	33%	32%
% Age 35-49	49%	49%	50%
% Age 50-64	11%	14%	14%
% Age >65	2%	4%	3%
% Missing Age Information	0%	0%	0%
Loan Characteristics			
Average Loan Amount	\$116,716	\$113,209	\$128,684
Average LTV %	95	85	84
% With LTV<=80	6%	38%	42%
% With LTV 80-90	12%	11%	13%
% With LTV 90-96	17%	15%	15%
% With LTV 96-98	14%	9%	8%
% With LTV>98	51%	27%	21%
% Missing LTV Information	0%	0%	0%
Average Ratio of Loan Amount to FHA Loan Limit	64%	63%	73%
% With LoantoFHA Ratio <=.5	25%	28%	23%
% With LoantoFHA Ratio of .5 - 1	75%	72%	59%
% With LoantoFHA Ratio of 1 - 1.2	0%	0%	11%
% With LoantoFHA Ratio of >1.2	0%	0%	6%

Exhibit 10 (cont.): Comparison of Characteristics by Dataset
Weighted

	Matched, Conforming, Fixed- Rate, FHA-Eligible FHA & GSE Loans with LTV 80-100%	Matched, Conforming, Fixed-Rate, FHA- Eligible Loans	Matched, Conforming Loans
Average PTI	21%	20%	20%
% Originated in 1998	31%	32%	32%
% Originated in 1999	35%	34%	34%
% Originated in 2000	34%	34%	34%
Mortgaged Property Characteristics			
% New Construction	8%	9%	11%
% Unit Size 1	95%	95%	95%
% Unit Size 2	2%	2%	2%
% Unit Size 3	2%	2%	2%
% Unit Size 4	1%	1%	1%
Borrower Neighborhood Characteristics (1990 Census)			
% In Underserved Tracts	43%	36%	32%
% In High Cost MSA	11%	11%	12%
% In Medium Cost MSA	86%	86%	85%
% In Low Cost MSA	3%	3%	3%
% In Center City	29%	28%	27%
Average 5-yr House Price Appreciation, Lagged 1 year	112%	113%	113%
% In Area with Depreciation	12%	10%	11%
% In Area with Appreciation up to 20%	74%	73%	71%
% In Area with Appreciation over 20%	14%	17%	18%
% In Tracts with Income <90% of MSA Income	29%	26%	23%
% In Tracts with Income 90 - 120% of MSA Income	39%	36%	34%
% In Tracts with Income =>120% of MSA Income	31%	38%	43%
% In <10% Minority Tracts	34%	41%	42%
% In 10-30% Minority Tracts	33%	32%	32%
% In =>30% Minority Tracts	32%	27%	26%

Exhibit 11(cont.): Analysis of Loans In Mortgage Market Sectors

On All Matched, Conforming Loans, Weighted

(n=347,732; weighted sample=1,980,080)

	GSE Purchased Loans			Loans Held by Depository Lenders			Other Investors			Not Mutually Exclusive		
	All GSE	GSE With PMI	GSE No PMI	All Depository	Depository With PMI	Depository No PMI	All Other Investors	Other Investors With PMI	Other Investors No PMI	All with PMI	All FHA	All Subprime
	pur_type=1,3			pur_type=0,5,6,8			pur_type=7,9			pmi_flag=1	fha_loan=1	subprime=1
Average Ratio of Loan Amount to FHA Loan Limit	80%	81%	80%	71%	75%	70%	70%	80%	68%	79%	63%	62%
% With LoantoFHA Ratio <= .5	18%	15%	18%	27%	21%	29%	25%	16%	26%	17%	27%	38%
% With LoantoFHA Ratio of .5 - 1	57%	59%	56%	56%	59%	55%	62%	60%	62%	59%	70%	50%
% With LoantoFHA Ratio of 1 - 1.2	16%	16%	16%	11%	13%	11%	9%	17%	8%	15%	3%	8%
% With LoantoFHA Ratio of >1.2	10%	9%	10%	6%	6%	5%	4%	8%	3%	8%	0%	3%
Average PTI	19%	20%	19%	19%	20%	19%	20%	20%	20%	20%	22%	18%
% Originated in 1998	36%	38%	35%	28%	29%	28%	31%	31%	31%	34%	30%	26%
% Originated in 1999	31%	30%	31%	37%	40%	36%	34%	31%	34%	33%	36%	34%
% Originated in 2000	33%	32%	34%	35%	31%	36%	35%	38%	35%	32%	34%	39%
Mortgaged Property Characteristics												
% New Construction	12%	10%	12%	10%	10%	10%	11%	11%	10%	10%	8%	7%
% Unit Size 1	96%	95%	96%	95%	94%	95%	96%	95%	96%	95%	95%	95%
% Unit Size 2	1%	1%	1%	2%	2%	2%	2%	1%	2%	2%	2%	1%
% Unit Size 3	2%	2%	2%	2%	3%	2%	2%	2%	2%	2%	2%	2%
% Unit Size 4	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
Borrower Neighborhood Characteristics (1990 Census)												
% In Underserved Tracts	22%	26%	21%	33%	34%	33%	40%	29%	42%	29%	48%	47%
% In High Cost MSA	10%	8%	11%	14%	12%	14%	14%	11%	15%	10%	11%	17%
% In Medium Cost MSA	87%	88%	87%	83%	85%	83%	84%	86%	83%	87%	87%	79%
% In Low Cost MSA	3%	3%	3%	3%	3%	3%	2%	2%	2%	3%	3%	4%
% In Center City	24%	27%	23%	29%	31%	28%	28%	27%	28%	28%	29%	36%
Average 5-yr House Price Appreciation, Lagged 1 year	114%	114%	114%	113%	114%	113%	112%	114%	112%	114%	111%	113%
% In Area with Depreciation	10%	10%	10%	10%	9%	10%	14%	12%	14%	10%	11%	14%
% In Area with Appreciation up to 20%	72%	71%	72%	70%	70%	70%	69%	68%	69%	70%	75%	64%
% In Area with Appreciation over 20%	19%	19%	18%	21%	21%	21%	17%	20%	16%	20%	14%	22%
% In Tracts with Income <90% of MSA Income	16%	19%	14%	25%	26%	25%	27%	20%	28%	21%	33%	35%
% In Tracts with Income 90 - 120% of MSA Income	32%	34%	30%	33%	33%	33%	36%	34%	36%	34%	40%	33%
% In Tracts with Income =>120% of MSA Income	53%	47%	55%	42%	41%	43%	37%	46%	36%	45%	28%	33%
% In <10% Minority Tracts	49%	48%	50%	44%	43%	44%	35%	44%	34%	46%	31%	31%
% In 10-30% Minority Tracts	33%	33%	33%	30%	31%	30%	32%	34%	32%	32%	32%	31%
% In =>30% Minority Tracts	18%	20%	17%	26%	26%	26%	33%	23%	34%	22%	37%	38%

Exhibit 12: Comparison of Fixed-Rate vs. Adjustable-Rate Mortgage
On All Matched, Conforming Loans, Weighted
(n=347,732; weighted sample=1,980,080)

	Fixed-Rate Loans	Adjustable-Rate Loans
	ex_rate_type <> B,U,V	ex_rate_type = B,U,V
Share of Loans	83%	17%
Borrower Characteristics		
Unweighted Number of Borrowers	285,885	61,847
Weighted Number of Borrowers	1,649,056	331,024
Average Annual Income	\$62,203	\$67,838
Median Annual Income	\$54,000	\$59,000
Average Annual Income (Trimmed Top 1%)	\$59,723	\$64,785
% Estimated Income Information	2%	2%
Average FICO	700	677
% With FICO <620	19%	26%
% With FICO 620-680	16%	20%
% With FICO =>680	65%	55%
% Missing FICO Information	0%	0%
% White	67%	67%
% Black	10%	11%
% Hispanic	12%	10%
% Other	7%	8%
% Missing Race Information	3%	3%
% Female	28%	29%
% Age 19-34	32%	33%
% Age 35-49	50%	50%
% Age 50-64	15%	14%
% Age >65	3%	3%
% Missing Age Information	0%	0%
Loan Characteristics		
Average Loan Amount	\$127,247	\$135,844
Average LTV %	84	83
% With LTV<=80	41%	49%
% With LTV 80-90	12%	19%
% With LTV 90-96	16%	12%
% With LTV 96-98	8%	5%
% With LTV>98	23%	15%
% Missing LTV Information	0%	0%

Exhibit 12 (cont.): Comparison of Fixed-Rate vs. Adjustable-Rate Mortgage
On All Matched, Conforming Loans, Weighted
(n=347,732; weighted sample=1,980,080)

	Fixed-Rate Loans	Adjustable-Rate Loans
	ex_rate_type <> B,U,V	ex_rate_type = B,U,V
Average Ratio of Loan Amount to FHA Loan Limit	72%	74%
% With LoantoFHA Ratio <=.5	24%	23%
% With LoantoFHA Ratio of .5 - 1	59%	59%
% With LoantoFHA Ratio of 1 - 1.2	11%	13%
% With LoantoFHA Ratio of >1.2	6%	5%
Average PTI	20%	20%
% Originated in 1998	34%	19%
% Originated in 1999	33%	39%
% Originated in 2000	32%	42%
Mortgaged Property Characteristics		
% New Construction	11%	11%
% Unit Size 1	95%	95%
% Unit Size 2	2%	2%
% Unit Size 3	2%	3%
% Unit Size 4	1%	1%
Borrower Neighborhood Characteristics (1990 Census)		
% In Underserved Tracts	33%	32%
% In High Cost MSA	11%	17%
% In Medium Cost MSA	86%	80%
% In Low Cost MSA	3%	3%
% In Center City	27%	28%
Average 5-yr House Price Appreciation, Lagged 1 year	113%	115%
% In Area with Depreciation	11%	10%
% In Area with Appreciation up to 20%	72%	65%
% In Area with Appreciation over 20%	17%	25%
% In Tracts with Income <90% of MSA Income	23%	23%
% In Tracts with Income 90 - 120% of MSA Income	34%	33%
% In Tracts with Income =>120% of MSA Income	43%	44%
% In <10% Minority Tracts	43%	40%
% In 10-30% Minority Tracts	32%	33%
% In =>30% Minority Tracts	25%	27%

Exhibit 13: Analysis of FHA-Eligible Loans in Mortgage Market Sectors

On Matched, Conforming, Fixed-Rate FHA-Eligible Loans, Weighted
(n=238,158; weighted sample=1,369,923)

	GSE Purchased Loans			Loans Held by Depository Lenders			Other Investors			Not Mutually Exclusive		
	All GSE	GSE With PMI	GSE No PMI	All Depository	Depository With PMI	Depository No PMI	All Other Investors	Other Investors With PMI	Other Investors No PMI	All with PMI	All FHA	All Subprime
	pur_type=1,3			pur_type=0,5,6,8			pur_type=7,9			pmi_flag=1	fha_loan=1	subprime=1
Share of Loans	33%	10%	23%	30%	5%	25%	22%	2%	20%	17%	34%	4%
Borrower Characteristics												
Unweighted Number of Borrowers	77,895	24,452	53,443	70,646	10,789	59,857	51,735	5,193	46,542	40,435	79,175	10,419
Weighted Number of Borrowers	445,372	136,499	308,873	407,958	61,981	345,977	302,835	28,708	274,127	227,199	465,118	54,874
Average Annual Income	\$59,710	\$55,723	\$61,472	\$56,296	\$52,258	\$57,019	\$54,226	\$56,820	\$53,955	\$54,916	\$48,384	\$57,065
Median Annual Income	\$54,000	\$52,000	\$55,000	\$48,000	\$48,000	\$48,000	\$48,000	\$52,000	\$47,000	\$51,000	\$44,000	\$49,000
Average Annual Income (Trimmed Top 1%)	\$57,731	\$54,730	\$59,171	\$53,756	\$51,146	\$54,299	\$51,968	\$55,605	\$51,613	\$53,866	\$46,718	\$54,443
% Estimated Income Information	1%	1%	1%	3%	4%	3%	3%	7%	3%	3%	1%	13%
Average FICO	726	714	731	698	706	696	673	713	669	712	642	653
% With FICO <620	9%	12%	8%	20%	14%	20%	28%	12%	30%	12%	38%	37%
% With FICO 620-680	13%	16%	12%	16%	17%	16%	20%	16%	20%	16%	24%	18%
% With FICO =>680	78%	73%	80%	64%	69%	63%	52%	73%	50%	72%	37%	45%
% Missing FICO Information	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
% White	76%	73%	77%	66%	66%	66%	57%	73%	56%	71%	49%	51%
% Black	5%	7%	4%	12%	9%	13%	16%	7%	17%	7%	22%	22%
% Hispanic	8%	11%	7%	12%	16%	11%	19%	11%	20%	13%	24%	12%
% Other	8%	6%	9%	7%	5%	7%	5%	5%	5%	6%	3%	6%
% Missing Race Information	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	2%	8%
% Female	28%	27%	29%	32%	30%	32%	31%	29%	32%	28%	33%	34%
% Age 19-34	30%	34%	27%	31%	35%	31%	34%	34%	34%	34%	39%	27%
% Age 35-49	49%	50%	49%	49%	49%	49%	50%	50%	50%	50%	49%	54%
% Age 50-64	17%	13%	18%	15%	13%	16%	13%	13%	13%	13%	10%	15%
% Age >65	5%	3%	5%	4%	3%	5%	3%	3%	3%	3%	2%	4%
% Missing Age Information	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Loan Characteristics												
Average Loan Amount	\$118,706	\$120,645	\$117,849	\$107,541	\$114,740	\$106,251	\$114,949	\$124,029	\$113,998	\$119,462	\$111,413	\$98,472
Average LTV %	79	88	75	83	89	81	89	88	89	88	97	82
% With LTV<=80	55%	23%	69%	44%	24%	48%	29%	22%	29%	23%	5%	51%
% With LTV 80-90	15%	22%	13%	12%	18%	11%	9%	23%	7%	21%	2%	17%
% With LTV 90-96	21%	44%	11%	15%	36%	11%	11%	46%	8%	42%	5%	16%
% With LTV 96-98	5%	10%	3%	10%	15%	9%	10%	7%	11%	11%	15%	4%
% With LTV>98	3%	2%	4%	18%	7%	20%	41%	2%	46%	3%	72%	11%
% Missing LTV Information	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Exhibit 13 (cont.): Analysis of FHA-Eligible Loans In Mortgage Market Sectors

On Matched, Conforming, Fixed-Rate FHA-Eligible Loans, Weighted
(n=238,158; weighted sample=1,369,923)

	GSE Purchased Loans			Loans Held by Depository Lenders			Other Investors			Not Mutually Exclusive		
	All GSE	GSE With PMI	GSE No PMI	All Depository	Depository With PMI	Depository No PMI	All Other Investors	Other Investors With PMI	Other Investors No PMI	All with PMI	All FHA	All Subprime
	pur_type=1,3			pur_type=0,5,6,8			pur_type=7,9			pmi_flag=1	fha_loan=1	subprime=1
Average Ratio of Loan Amount to FHA Loan Limit	66%	68%	65%	60%	64%	59%	63%	68%	62%	67%	62%	55%
% With LoantoFHA Ratio <= 5	24%	20%	25%	34%	27%	35%	28%	21%	29%	22%	29%	44%
% With LoantoFHA Ratio of .5 - 1	76%	80%	75%	66%	73%	65%	72%	79%	71%	78%	71%	56%
% With LoantoFHA Ratio of 1 - 1.2	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
% With LoantoFHA Ratio of >1.2	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Average PTI	19%	20%	19%	19%	20%	19%	21%	20%	21%	20%	21%	17%
% Originated in 1998	34%	35%	34%	30%	31%	30%	31%	29%	31%	33%	31%	28%
% Originated in 1999	32%	32%	32%	36%	40%	35%	33%	31%	34%	34%	35%	35%
% Originated in 2000	34%	33%	34%	34%	29%	34%	35%	39%	35%	33%	34%	37%
Mortgaged Property Characteristics												
% New Construction	9%	8%	10%	9%	7%	9%	9%	8%	9%	8%	8%	7%
% Unit Size 1	94%	94%	94%	95%	93%	95%	95%	95%	95%	94%	95%	94%
% Unit Size 2	2%	2%	2%	2%	2%	2%	2%	1%	2%	2%	2%	2%
% Unit Size 3	3%	3%	2%	2%	3%	2%	2%	2%	2%	3%	2%	3%
% Unit Size 4	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
Borrower Neighborhood Characteristics (1990 Census)												
% In Underserved Tracts	26%	31%	24%	37%	39%	37%	44%	32%	45%	33%	49%	48%
% In High Cost MSA	10%	8%	11%	12%	12%	12%	14%	12%	14%	10%	11%	15%
% In Medium Cost MSA	88%	89%	87%	85%	85%	85%	84%	86%	84%	87%	86%	81%
% In Low Cost MSA	3%	3%	2%	3%	3%	3%	2%	2%	2%	3%	3%	4%
% In Center City	26%	29%	24%	30%	34%	29%	29%	30%	29%	31%	30%	36%
Average 5-yr House Price Appreciation, Lagged 1 year	114%	115%	114%	113%	114%	112%	111%	114%	111%	114%	111%	112%
% In Area with Depreciation	9%	9%	10%	9%	9%	9%	14%	12%	14%	9%	12%	13%
% In Area with Appreciation up to 20%	73%	72%	73%	73%	72%	73%	71%	69%	71%	72%	75%	68%
% In Area with Appreciation over 20%	18%	19%	17%	19%	19%	19%	15%	19%	14%	19%	13%	19%
% In Tracts with Income <90% of MSA Income	19%	22%	17%	28%	31%	28%	30%	23%	30%	25%	34%	37%
% In Tracts with Income 90 - 120% of MSA Income	35%	37%	34%	35%	35%	35%	38%	37%	38%	36%	40%	34%
% In Tracts with Income =>120% of MSA Income	47%	40%	49%	37%	35%	37%	33%	40%	32%	39%	26%	30%
% In <10% Minority Tracts	48%	46%	49%	44%	41%	44%	33%	42%	33%	44%	30%	33%
% In 10-30% Minority Tracts	33%	32%	33%	29%	30%	29%	32%	34%	32%	32%	32%	30%
% In =>30% Minority Tracts	20%	22%	19%	27%	29%	27%	35%	24%	36%	24%	38%	36%

Exhibit 14: Analysis of 80-100% LTV Loans In FHA and GSE Market Sectors
On Matched, Conforming, Fixed-Rate FHA-Eligible FHA & GSE Loans with LTV 80-100%, Weighted
(n=114,780; weighted sample=674,238)

	GSE Purchased Loans			Not Mutually Exclusive	
	All GSE	GSE With PMI	GSE No PMI	All FHA	All Subprime*
	pur_type=1,3			fha_loan=1	subprime=1
Share of Loans	36%	16%	20%	65%	1%
Borrower Characteristics					
Unweighted Number of Borrowers	40,600	18,673	21,927	75,462	1,361
Weighted Number of Borrowers	242,176	108,889	133,287	440,104	8,033
Average Annual Income	\$59,624	\$55,899	\$62,667	\$48,467	\$48,145
Median Annual Income	\$54,000	\$52,000	\$56,000	\$44,000	\$42,500
Average Annual Income (Trimmed Top 1%)	\$57,860	\$54,945	\$60,390	\$46,792	\$46,977
% Estimated Income Information	1%	1%	1%	1%	18%
Average FICO	713	710	715	642	642
% With FICO <620	12%	12%	12%	38%	39%
% With FICO 620-680	16%	17%	15%	24%	21%
% With FICO =>680	73%	71%	73%	38%	40%
% Missing FICO Information	0%	0%	0%	0%	0%
% White	71%	70%	72%	50%	44%
% Black	7%	7%	6%	21%	25%
% Hispanic	11%	13%	10%	24%	21%
% Other	7%	6%	9%	3%	6%
% Missing Race Information	3%	4%	3%	2%	4%
% Female	28%	27%	28%	33%	40%
% Age 19-34	35%	36%	34%	39%	32%
% Age 35-49	49%	50%	49%	49%	55%
% Age 50-64	13%	12%	14%	10%	11%
% Age >65	3%	2%	3%	2%	2%
% Missing Age Information	0%	0%	0%	0%	0%
Loan Characteristics					
Average Loan Amount	\$124,978	\$122,685	\$126,852	\$112,017	\$104,497
Average LTV %	90	92	88	98	95
% With LTV <=80	17%	3%	28%	0%	6%
% With LTV 80-90	28%	27%	29%	2%	12%
% With LTV 90-96	39%	55%	26%	5%	23%
% With LTV 96-98	10%	12%	8%	16%	9%
% With LTV >98	6%	2%	8%	76%	51%
% Missing LTV Information	0%	0%	0%	0%	0%

Exhibit 14 (cont.): Analysis of 80-100% LTV Loans In FHA and GSE Market Sectors
On Matched, Conforming, Fixed-Rate FHA-Eligible FHA & GSE Loans with LTV 80-100%, Weighted
(n=114,780; weighted sample=674,238)

GSE Purchased Loans			Not Mutually Exclusive	
All GSE	GSE With PMI	GSE No PMI	All FHA	All Subprime*
pur_type=1,3			fha_loan=1	subprime=1

Average Ratio of Loan Amount to FHA Loan Limit	69%	68%	69%	62%	57%
% With LoantoFHA Ratio <=.5	19%	20%	18%	28%	41%
% With LoantoFHA Ratio of .5 - 1	81%	80%	82%	72%	59%
% With LoantoFHA Ratio of 1 - 1.2	0%	0%	0%	0%	0%
% With LoantoFHA Ratio of >1.2	0%	0%	0%	0%	0%
Average PTI	20%	20%	20%	21%	20%
% Originated in 1998	33%	34%	33%	30%	26%
% Originated in 1999	33%	32%	33%	36%	34%
% Originated in 2000	34%	34%	34%	34%	40%
Mortgaged Property Characteristics					
% New Construction	8%	7%	9%	8%	4%
% Unit Size 1	94%	94%	94%	95%	94%
% Unit Size 2	2%	2%	2%	2%	2%
% Unit Size 3	3%	3%	3%	2%	3%
% Unit Size 4	2%	2%	2%	1%	2%
Borrower Neighborhood Characteristics (1990 Census)					
% In Underserved Tracts	31%	33%	29%	49%	56%
% In High Cost MSA	11%	9%	12%	11%	15%
% In Medium Cost MSA	86%	87%	85%	86%	83%
% In Low Cost MSA	3%	3%	2%	3%	1%
% In Center City	29%	31%	27%	29%	39%
Average 5-yr House Price Appreciation, Lagged 1 year	113%	114%	113%	111%	108%
% In Area with Depreciation	11%	10%	11%	12%	16%
% In Area with Appreciation up to 20%	74%	73%	75%	75%	79%
% In Area with Appreciation over 20%	15%	17%	14%	13%	5%
% In Tracts with Income <90% of MSA Income	22%	24%	20%	33%	44%
% In Tracts with Income 90 - 120% of MSA Income	37%	38%	37%	40%	38%
% In Tracts with Income =>120% of MSA Income	41%	38%	43%	26%	18%
% In <10% Minority Tracts	41%	42%	41%	30%	32%
% In 10-30% Minority Tracts	35%	34%	35%	33%	31%
% In =>30% Minority Tracts	24%	25%	23%	37%	37%

Notes:

* The subprime classification in our analysis is not mutually exclusive from other categories.

The average annual income¹¹ of the borrowers among all of the matched conforming¹² loans is over \$63,000 (see Exhibit 10). However, the distribution of incomes in this data set suggests that a number of outliers artificially raise this average. Both a median income, as well as an average income derived from a trimmed data set is provided to offer a potentially more accurate picture of borrowers' incomes.¹³ Additionally, the borrowers' average credit score, or FICO, is provided along with a distribution of FICO scores.¹⁴ Breakdowns of the race, age, and sex of the borrowers are also displayed.

In addition, the data were classified based on several key characteristics of the loans. The average loan amount is calculated based on the HMDA loan amount variable. However, to get a better sense of the potential risk of the loans in each data set, several variables are provided, including the average loan to value ratio (LTV), the average ratio of loan to FHA loan limit, and the average payment to income ratio (PTI). A distribution of LTV and loan to FHA loan limit ratios is provided, as well as a breakdown of originations by year.

The loan-to-value ratio was calculated by dividing the Experian loan amount by the Experian sale amount (i.e., purchase value). The ratio of loan amount to a rounded¹⁵ FHA loan limit is calculated by comparing the HMDA loan amount to the corresponding FHA loan limit based on origination date, assigned dwelling size,¹⁶ and MSA of origination. In the first column of Exhibit 10, the 51 percent of loans with LTV greater than 98 percent may seem high, but the sample of FHA-eligible loans is dominated (65 percent) by FHA loans. Only 6 percent of GSE loans in the FHA-eligible sample have such high LTVs and only 2 percent of the full sample of GSE loans has LTVs above 98 percent.

The payment-to-income ratio is a measure of the annual loan payment relative to the annual income of the borrower. In order to calculate the payment-to-income ratio, the annual payment of the mortgage must be estimated. It is assumed that all loans were 30-year mortgages with a fixed interest rate based on the national average contract interest rate for fixed-rate mortgages in the month of sale.¹⁷ This estimated annual payment is then compared to the borrower's annual income. (In fact,

¹¹ Annual income is based on the HMDA Annual Income variable. However, roughly two percent of the loans were missing annual income information. All but seven of these loans did, however, have an income range code provided in the Experian Income variable. The midpoint of each range was used as an estimate of the borrowers' income. Those in the top income range of >\$250,000 were assigned an income of \$250,000.

¹² Conforming loans in this context refers to loans below the GSE conforming loan limit (i.e., non-jumbo loans), which was \$252,700 in 2000.

¹³ The outliers are not excluded from the rest of the analysis.

¹⁴ FICO is not collected in HMDA and is only available because of the merging of HMDA with the Experian data.

¹⁵ The HMDA loan amounts are reported rounded to the nearest \$1,000, so the FHA loan limits are rounded up to the nearest thousand to avoid inadvertently classifying loans as ineligible for FHA.

¹⁶ See above discussion of dwelling size assignments.

¹⁷ Contract Interest Rates-Monthly National Averages for All Homes, Fixed-Rate Mortgages in 1998, 1999, and 2000. Federal Housing Finance Board Monthly Interest Rate Survey: www.fhfb.gov/mirs/mirs_downloads.htm.

17% of the matched conforming loans are not on fixed-rate terms. The above annual payment estimation does not capture the changing annual payment associated with adjustable-rate or balloon loans).

Highlighted mortgaged property characteristics include the share of loans that are for new construction as well as the assumed unit size.

In addition to the above characteristics of the borrowers, loans, and properties, information is provided about the metropolitan areas and neighborhoods of the mortgaged properties, including tract “underserved area” status, relative cost of the MSA based on FHA eligibility limits, center city location, MSA property value growth, and tract minority percentage and tract income relative to the MSA median.

In the matched, conforming data set, 27 percent of the loans were originated in center city tracts¹⁸ and 32 percent were originated in tracts that were “underserved”¹⁹ in the 1990 Census.

The distribution of loans into high, average and low cost MSAs is based on whether loans originated in a MSAs with the highest or lowest FHA loan limits at the time of origination.²⁰ Only three percent of the loans in the data originated in “Low Cost” MSAs, meaning MSAs that had the lowest FHA loan limits at the time of origination.

In order to evaluate whether loans originated in areas of high or low property value growth, the average 5-year house price change (lagged one year) was calculated. MSAs with higher and lower levels of appreciation were based on the OFHEO House Price Index.²¹ Over 70 percent of loans in the largest analysis file were originated in MSAs with property value appreciation up to 20 percent over five years (lagged one year) compared to only 11 percent of loans originated in areas with negative appreciation over five years.

In addition, 26 percent of the loans were originated in tracts with a population at least 30 percent minority in the 1990 Census, and 23 percent were originated in tracts with a median household income that was less than 90 percent of the MSA median household income.

Exhibit 11 examines the pool of loans associated with each mortgage market “sector” in the matched conforming data. As described above, the GSE sector is based on the HMDA secondary purchaser type (Purchaser Type=1 or 3). The HMDA purchaser type is also used to determine the loans that were held by depository lenders (Purchaser Type =5, 6, 8 or 0) and those that were sold to other types

¹⁸ Tracts were assigned as “center city” tracts by Unicon based on the 1990 Census.

¹⁹ Tracts are designated by HUD as underserved if they have a median household income no more than 90 percent of the MSA median household income or if they have a population that is at least 30 percent minority with a median household income that is no more than 120 percent of the MSA median household income.

²⁰ In 1998, FHA loan limits varied within an MSA by county. From 1999 on, HUD indexed the base FHA loan limit at 48 percent of the conforming loan limit and the maximum FHA one-family loan limit for “high cost” areas at 87 percent of the conforming loan limit, depending on the median house price in the county or MSA. FHA loan limits from FHA_limits_holly.xls.

²¹ OFHEO House Price Index, Q3 1993.

of investors (Purchaser Type =7 or 9). Additionally, the loans are segmented into those with private mortgage insurance (PMI) or with FHA insurance, and those originating with lenders designated as primarily subprime lenders in the year of origination.²² It is important to note that some loans in the last three columns of Exhibit 11, labeled “Not Mutually Exclusive,” are also included in the first nine columns. In other words, there are a few FHA loans in GSE portfolios or held by depositories. There are also FHA loans originated by lenders designated as subprime lenders. In the context of Exhibit 11, those loans are allowed to appear in both categories. In the origination models, a mutually exclusive categorization is created by excluding FHA loans from the GSE category or excluding subprime loans from the FHA category.

Using this method of segmentation, the GSE sector has borrowers with the highest FICO scores (average FICO of 726) and the highest incomes (median income of \$62,000), and the loans with the lowest LTV ratios (average LTV of 80 percent). In contrast, FHA-insured borrowers have the lowest annual incomes (median income of \$45,000) and the loans with the highest LTV ratios (average LTV of 97 percent). The pool of FHA-insured loans also contains the highest proportion (48 percent) of loans originated in underserved tracts.

Subsetting the Data

In order to evaluate the loans with the greatest likelihood of being in the FHA and GSE overlap, the analysis first excluded the 61,847 adjustable rate loans from the matched conforming data set (see Exhibit 12). This choice was made, in part, because the PTI calculation for this set of loans would not capture the changing annual payments over time and there was a desire for a more consistent set of loans for this analysis. Furthermore, these loans should be modeled separately because they are not directly comparable to fixed-rate loans. However, the set of adjustable loans is already fairly small and, after cutting it into competing market sectors, would create subsets that were too small for reliable results.

An examination of the loans with adjustable rates shows that these loans tend to go to borrowers with higher incomes and lower FICO scores, tend to be for a larger loan amount, and are more likely to originate in high cost MSAs than fixed-rate loans. For example, California is a high-cost area with a disproportionate share of ARMs. Borrowers prefer ARMs in high-cost areas because, given a certain income, they can qualify for a larger loan amount under an ARM than a FRM. Overall, the remaining subset of fixed-rate loans has very similar characteristics to the full matched conforming set.

The matched conforming loans were further subdivided to include only fixed-rate, FHA-eligible loans, that is, loans under the corresponding FHA loan limit. This remaining set of conforming, fixed-rate, FHA-eligible loans contains 238,158 observations (See Exhibit 10). The loans in this pool tend to go to borrowers with lower incomes and that are more likely to be female than those in the set of all conforming loans. The loans in this set are also smaller, on average. However, segmenting this data into mortgage market sectors reveal that the pool of FHA-insured loans in the set of fixed-rate, FHA-eligible loans is not substantially different from the pool of FHA-insured loans in the set of all conforming loans (see Exhibit 13). But, not surprisingly, the characteristics of the loans in the other sectors changed more dramatically. For example, the median income of GSE borrowers dropped

²² Subprime data were downloaded from www.huduser.org/datasets/manu.html and merged into this study's data on compressed Agency and Resp_ID variables. Subprime=1 if lender was classified as a primarily subprime lender in the year of loan origination.

from \$62,000 to \$54,000, the average GSE loan amount dropped from \$139,103 to \$118,706, and the GSE average ratio of loan amount to FHA loan limit dropped from 80 to 66 percent.

While all of these loans are eligible for FHA, the fixed-rate, FHA-eligible loans were further subdivided to include only FHA and GSE loans with an LTV of 80–100 percent in order to focus on the loans most likely to be considered by both FHA and the GSEs. The 114,780 loans in this data set have an average FICO score of 668, down substantially from the average of 696 in the set of all matched, conforming loans (see Exhibit 10). The average LTV of 95 percent is appreciably higher than the average of 84 percent in the larger set. Furthermore, 43 percent of the pool of FHA and GSE loans in the 80-100 percent LTV range originated in underserved tracts. In contrast, only 32 percent of all matched, conforming loans originated in underserved tracts. The borrowers in the FHA/GSE working data set also have lower incomes, are more likely to be a minority, and are younger than the borrowers in the larger set of all conforming loans.

Overall, the pool of loans in the FHA and GSE data sets appear to have higher potential risk characteristics. The pool of FHA-insured loans in this data set (see Exhibit 14) does not appear to be significantly different from the pool of FHA-insured loans in the larger data set.²³ However, the pool of GSE loans in this data set is substantially different from GSE loans in the larger set along many different dimensions. FICO scores are 13 points lower, loans are over \$14,000 smaller, the average LTV increased from 80 to 90 percent, the average ratio of loan amount to FHA loan limits fell from 80 to 69 percent, and the share of GSE loans originating in underserved tracts increased from 22 to 31 percent.

Given the focus on overlap, it is interesting to see how much overlap there is in credit scores and LTV. Exhibit 15 shows the high degree of overlap in FICO scores (94 percent by the non-parametric overlap method explained below or a KS statistic²⁴ of 0.37). Certainly, GSE loans have higher FICO scores on average, but there are many FHA loans with higher FICO scores than some GSE loans. Exhibit 16 shows much less overlap in the LTV distributions. The GSE loans are bunched around 80, 90 and 95 percent LTV, whereas FHA loans are almost entirely 96 percent and above. Many FHA borrowers who appear to be low-risk based on credit score, may actually be high-risk based on LTV. A regression model is needed to control for all these differences and to estimate the predicted probability of being FHA along a single dimension.

²³ Note that Exhibit 14 includes only loans that are either GSE or FHA loans. The subprime loans listed are also either GSE or FHA loans. The data in Exhibit 14 are the same as those used for the FHA vs. GSE origination models to follow.

²⁴ The KS (Kolmogorov-Smirnov) statistic measures the degree of separation between two distributions. It is sensitive to small differences in distributions, so all the KS statistics are significant. Larger values indicate more separation and correspond to less overlap.

**Exhibit 15: Weighted Distribution of FICO Scores
In FHA & GSE Dataset**

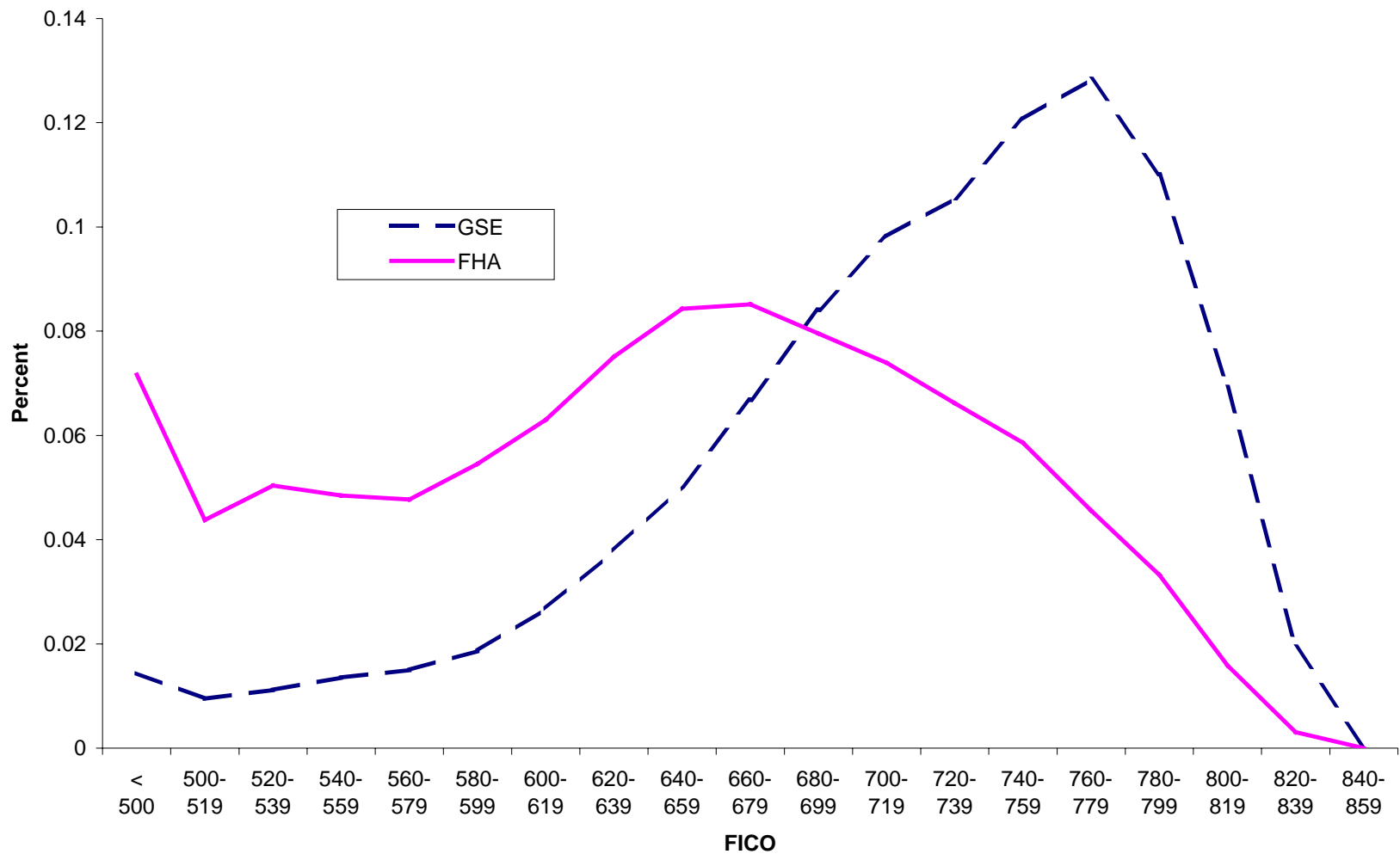
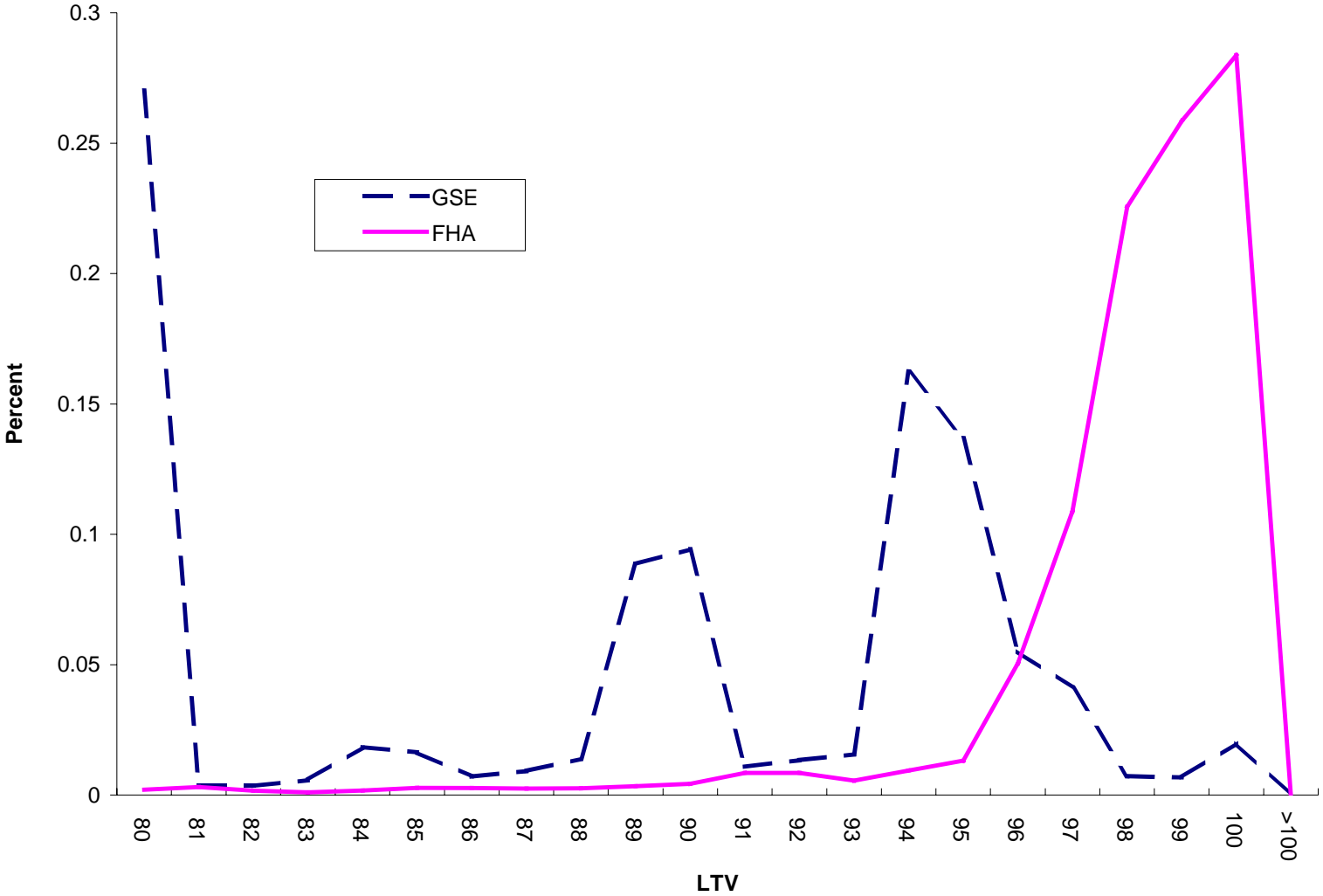


Exhibit 16: Weighted Distribution of LTV Ratios
In FHA & GSE Dataset



Section 3: Origination Model and Market Sector Overlap

To measure an overlap between FHA and GSE, or any two mortgage market sectors, it is useful to summarize all the different characteristics of a loan, borrower, property, and neighborhood into a single dimension. The origination model is a regression of the probability of a loan being insured by FHA relative to being purchased by a GSE, controlling for the observed characteristics that could significantly affect the predicted probability.²⁵ The predicted probability (or the log odds translation of the predicted probability²⁶) is the single dimension summary that represents the FHA wedge described by Ambrose, Pennington-Cross, and Yezer (2002). If a sharp distinction exists between low-risk GSE loans and higher-risk FHA loans, there would be little overlap between the predicted probabilities for GSE and FHA loans. However, if loans with the same characteristics are nearly as likely to go to GSE as FHA, then there is considerable overlap. Perhaps the borrowers did not shop carefully or were steered into FHA when they could have qualified for a conventional loan. The immediate aim is not to explain why there is overlap, but rather to devise methods to measure the degree of overlap. This section describes three methods for measuring overlap and then applies those methods to the interfaces between FHA vs. GSE, FHA vs. PMI, FHA vs. subprime, and GSE vs. depository lenders.

Confidence Interval Measure of Overlap

The first method for measuring overlap is based on the confidence interval around the predicted probability. The predicted probability of FHA can range from zero to one and each prediction has a standard error for each prediction. A 95 percent confidence interval (± 1.96 times the standard error) indicates the reliability of the prediction. If the confidence interval around the predicted probability of a loan being FHA contains zero (or nearly zero), the true probability of a loan being FHA is statistically indistinguishable from zero. To be 95 percent certain that a loan has a non-zero probability of being FHA, the lower bound of the confidence interval should be greater than zero. Conversely, if the confidence interval contains one, the true probability of a loan being FHA is not statistically distinguishable from one. In between those extremes are loans with confidence intervals that contain neither zero nor one. The loans that are statistically the same as definitely FHA or definitely GSE have been excluded. The remaining loans could have gone to either FHA or GSE and, thus, those loans are defined as the overlap region.

Technically, the confidence interval should not fall outside of the zero-one interval, but practically the interval measured as 1.96 times the standard error does fall below zero for some predictions with very low probabilities or above one for predictions with very high probabilities. The limit is arbitrary. For a more stringent definition of overlap, the boundaries for the confidence intervals could be 0.05

²⁵ To simplify the interpretation, only two alternatives are assumed for each model. It is further assumed that the results for those two alternatives would not be affected by the presence of other alternatives, that is, the independence of irrelevant alternatives (IIA).

²⁶ If p is the predicted probability, then $\log(p/(1-p))$ is the log odds. The log odds has a range from negative infinity to positive infinity compared to the range of the predicted probability of 0 to 1. Moreover, the distribution of log odds is more Gaussian or bell-shaped which makes it easier to visualize the overlap between distributions.

and 0.95 instead of 0 and 1. The boundaries of zero and one were chosen to emphasize the idea that, if there is 95 percent confidence that the probability “includes” zero of being FHA, then it is unlikely that FHA is competitive for that loan. The point is that confidence intervals “close” to or encompassing the endpoints (zero or one) indicate the extreme probabilities, either the loan is highly likely to be FHA or GSE. The intervals between the extremes indicate cases more evenly divided and less definite in the outcome. For example, the loans in the middle are nearly equally likely to go to either FHA or GSE. The confidence interval is used to trim off the cases with more definite outcomes leaving a set of overlap cases that have at least a modest chance of either outcome.

The boundaries are somewhat arbitrary, but the goal is to identify loans with probabilities evenly divided between the two possible outcomes. Those loans could have gone to either FHA or GSE which is the core idea of the overlap region. In this formulation, the boundaries chosen are that the lower limits of the confidence interval exceed 0 and the upper limits fall below 1. Other valid boundaries could have been chosen as a way to identify the loans with probabilities in the middle. In fact, the tolerance limit approach (explained below) is an example of a good alternative for defining the boundaries. The size of the overlap region is somewhat different, but the basic finding of a significant overlap region is robust to the method or boundaries chosen.

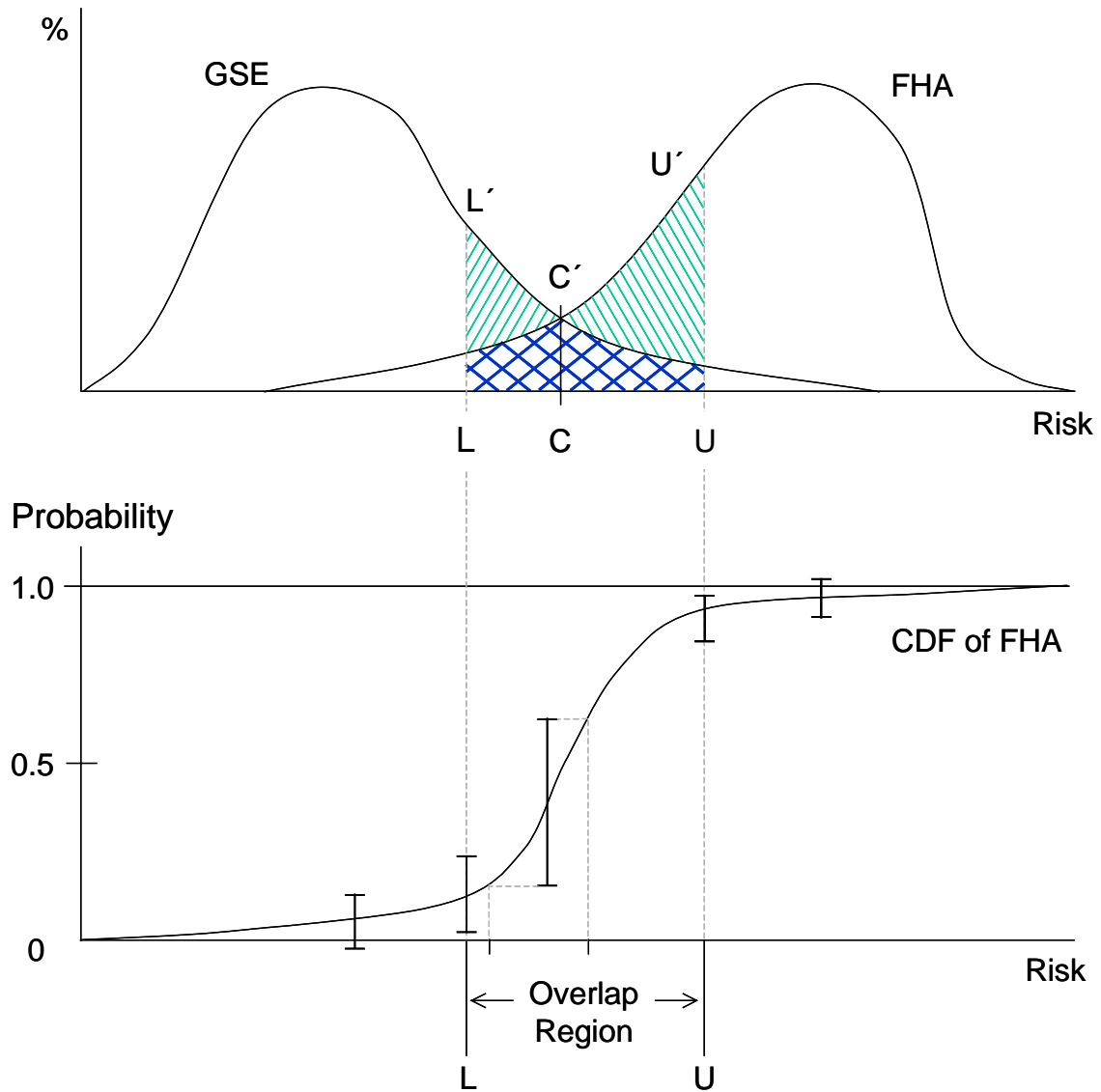
Note, the size of the overlap region is linked to the precision of the predictions and the goodness-of-fit of the origination model. A poor model fit, perhaps due to errors in the data, will lead to wide standard errors around the model predictions. All the predictions with confidence intervals reaching 0 or 1 are excluded from the overlap. Wide standard errors mean wide confidence intervals and more predictions being excluded from the overlap. Thus, a poor model fit is associated with a narrower overlap region. The “true” size of the overlap region may be larger, but limitations in the data or model specification can bias downward the estimate.

The diagram in Exhibit 17 shows overlapping distributions along the risk dimension. The GSE distribution falls mostly in the low-risk region and FHA distribution falls mostly in the higher-risk region. The portion of the distributions that overlaps corresponds to the range, shown in the lower panel, where the lower end of the confidence interval is greater than zero and the upper end of the interval is less than one.²⁷ As shown in the lower panel of Exhibit 17, the probability of a loan being FHA is aligned with higher default risk.

²⁷ The lower panel of Exhibit 17 only shows the probability for FHA, whereas the upper panel shows the marginal distributions for both GSE and FHA. The probability of GSE would be 1 minus the probability of FHA and thus the mirror image of the FHA probability crossing at 0.5. The distinction to note is the upper panel shows the marginal distributions according to actual outcome, whereas the lower panel translates those outcomes into predicted probabilities in the same way that a logistic regression does. If there was perfect correspondence between actual outcomes and model predictions, then the CC’ crossing point would correspond to the 50 percent probability.

Exhibit 17: FHA/GSE Overlap Based on the Confidence Interval Method

Distributions represented in percentage terms so equal in size.



One disadvantage of the confidence interval method is that the width of the confidence interval is based on an arbitrarily chosen alpha value. In this case, an alpha of 0.05 was used, but that is based on the traditional preference of 95 percent confidence intervals used in a different context. Another issue is that a poor model with wide confidence intervals could have a very small overlap region, whereas greater uncertainty is expected to lead to greater overlap.²⁸ On the other hand, the tolerance limit methods (explained below) have the characteristic that poorly fitting models have greater

²⁸ As shown in Exhibit 19, the goodness-of-fit for the origination model for FHA vs. GSE is quite good (percent concordant is 96.1, pseudo-R² is 0.64 and combined K-S statistic is 0.83).

overlap. In this respect, the confidence interval method is more conservative in its measure of overlap.²⁹

A more important distinction between the methods is how the extreme values are handled. Consider again the overlapping distributions shown in Exhibit 17. The confidence interval method excludes the extreme high values of the FHA distribution (on the right) and the extreme low values of the GSE distribution (on the left). However, the high values of the GSE distribution and the low values of the FHA distribution may be close enough to the center of the combined sample that they are included in the overlap region. In bulk purchases, the GSEs do acquire some high-risk loans, and it is likely that some loans appear higher risk than they truly are due to errors in FICO score or LTV reported. For the purpose of determining which FHA loans could qualify for conventional underwriting, the overlap boundaries should not be distorted by the high-risk cases that do not represent typical GSE underwriting. Similarly, some low-risk FHA loans may be unusual cases or reporting errors. These outliers are unusual because they are not extremely high or low risks in the joint distribution, but rather extreme values relative to the marginal distributions, GSE or FHA. Therefore, a methodology is needed for trimming outliers from the marginal distributions before setting the overlap region boundaries. After the trimming, the overlap can be defined as the range between the FHA lower bound and the GSE upper bound.

Tolerance Limit Methods of Overlap

The concept of tolerance limits can help determine where the lower and upper bounds should be placed. Tolerance limits are statistical constructs designed to cover a fixed proportion of the population with a stated level of confidence using the sample data on hand. For example, a company that manufactures a particular engineering product can identify the tolerance limits such that, on average, a fixed proportion of the products may be expected to have a quality falling between the limits. Tolerance limits have also been used in the field of medicine. For instance, to identify the “normal” lower and upper limits for a particular physiological function (such as heart rate), a physician may base his/her estimates on a large sample of healthy subjects. Tolerance limits can be constructed so that a large proportion, say 90 percent, of the population will have a heart rate falling between the bounds with a stated level of confidence (such as 95%).

It is important to distinguish the concept of tolerance limits from *confidence intervals*. Confidence limits/intervals are statistical bounds within which a given population parameter, such as the mean, is expected to lie with a stated level of confidence. Tolerance limits, on the other hand, are statistical bounds within which a fixed proportion of the population is expected to lie with a stated level of confidence. In order to identify a set of typical cases, rather than the true value of a statistic, tolerance limits should be used. For example, the tolerance limits can be set so that 90 percent of the

²⁹ An alternative research approach has been suggested whereby mortgage markets are stratified according to FHA’s competitive position. For example, if FHA garnered less than 5 percent of the loans, the market would be designated as low competitiveness for FHA, whereas if the FHA market share exceeded 25 percent then it could be designated as high competitiveness. Once the markets are defined as low, medium and high, then further analysis could be done to determine just which characteristics of the markets and borrowers are associated with each market segment. Moreover, the degree of potential FHA competitiveness can be determined by the predicted FHA probability. In a local market, if the actual market share deviates from the predicted share, this difference can lead to an investigation of the market in which FHA does particularly well or poorly. This approach is left for future research.

loans meet FHA underwriting requirements. By trimming the top and bottom 5 percent tails of the FHA distribution, the remaining loans represent the range of loans that meet the typical FHA underwriting requirements. The same process can be applied to the GSE distribution of loans to determine a 90 percent set of loans that meet the typical GSE underwriting requirements. Then the overlap is the set of loans that meet both the FHA and GSE underwriting requirements. Given that FHA underwriting allows for riskier loans than GSE underwriting, the overlap set is comprised of loans above the lower limit for FHA and below the upper limit for GSE.

There are two types of tolerance limits – parametric and non-parametric.³⁰ The parametric version assumes that the underlying distribution is normal, whereas the non-parametric makes no assumption about either the normality or symmetry of the distribution. The non-parametric version is preferred in the mortgage origination context because the distributions are neither normal or symmetric. Moreover, the non-parametric version efficiently trims off the extreme cases, which is important because the results should not be distorted by outliers or recording errors.

Parametric Tolerance Limits

The parametric tolerance limits require the strong assumption that the sample data were drawn from a normally distributed, i.e., symmetric, population. Below, the calculation formula for the limits is described. Readers interested in the derivation should consult the texts referenced in the footnote.

Suppose there are a series of measurements Y_1, Y_2, \dots, Y_N . Let \bar{Y} and S be the sample mean and sample standard deviation of the distribution. Then, the upper and lower tolerance limits that cover P percent of the population measurements with α level of confidence are:

$$Y_L = \bar{Y} - kS$$

$$Y_U = \bar{Y} + kS$$

where

$$k = \sqrt{\frac{(N-1)\left(1 + \frac{1}{N}\right)Z_{(1-P)/2}^2}{\chi_{\alpha, N-1}^2}}$$

$Z_{(1-P)/2}$ is the critical value if the standard normal distribution is exceeded with probability $(1-P)/2$ and $\chi_{\alpha, N-1}^2$ is the critical value of the chi-square distribution with $N-1$ degrees of freedom that is exceeded with probability α .

³⁰ This discussion is drawn heavily from the following sources: Howe, W. G. (1969). "Two-sided Tolerance Limits for Normal Populations - Some Improvements", *Journal of the American Statistical Association*, 64, pages 610-620. *Selected Techniques of Statistical Analysis for Scientific and Industrial Research and Production and Management Engineering*. Edited by Churchill Eisenhart et al. NY: McGraw-Hill Book (1947). *Sturdy Statistics: Nonparametric and Order Statistics*. Frederick Mosteller and Robert E.K. Rourke. MA: Addison-Wesley (1973).

Non-Parametric Tolerance Limits

Non-parametric tolerance limits relax the assumption that the sample data must come from a normally distributed population, particularly a symmetric distribution. The calculation is based on order statistics.³¹ A parametric tolerance starts from the middle and goes outward plus or minus the same amount. A non-parametric tolerance starts at the extreme values and proceeds inward. Below, the derivation is skipped and only a sketch of the computation steps is provided.

Suppose there are a series of measurements Y_1, Y_2, \dots, Y_N . Let Y_i and Y_{N-i+1} be the lower and upper limits that cover P percent of the population measurements with α level of confidence. According to the distribution theory of order statistics, the expected proportion of measurements in the interval (Y_i, Y_{N-i+1}) is $P = (N-2i+1)/N$, and the standard deviation of Y is $\sigma_Y = \sqrt{P(1-P)/(N+2)}$.

Let P^* be the desired proportion and Z_α be the α -quantile from a standard normal distribution. To construct the tolerance limits Y_i and Y_{N-i+1} , the goal is to find the smallest P and the corresponding i such that:

$$\frac{P - P^*}{\sqrt{P(1-P)/(N+2)}} \geq |Z_\alpha| \quad (1)$$

and

$$P = (N-2i+1)/N \quad (2)$$

Equations (1) and (2) together form a system that is quadratic in P . Rather than solving this system of equations by formula, the optimal values of P and i are obtained by the method of iteration. Specifically, one starts with a very small value (.0001) for P and increases it by 0.0001 until the right-hand-side of equation (1) is greater than $|Z_\alpha|$. Once the optimal value of P is solved, the value of i can be obtained from equation (2).³²

It is important to emphasize that the order statistics are applied to the predicted probabilities, which is a continuous distribution. The main purpose of the order statistics is to trim off the tails of the marginal distributions for FHA and GSE, or whichever two outcomes are competing. The remaining overlap of the trimmed distributions is based on “normal” underwriting conditions rather than the exceptional cases, which may dominate the tails. The method of non-parametric tolerance limits is robust in that it makes no assumptions about the underlying distributions for FHA or GSE. Just as with the confidence interval approach, the particular boundaries are arbitrary. Instead of the central 90 percent, the central 95 percent or the central 80 percent could have been chosen. Different order

³¹ *Sturdy Statistics: Nonparametric and Order Statistics*. Frederick Mosteller and Robert E.K. Rourke. MA: Addison-Wesley (1973).

³² Stata 8 is used to perform the iteration and search for the P and i values in this analysis. For the pooled sample of GSE and FHA loans that contain 114,780 observations, the optimal P and i values can be found in approximately 4 minutes of CPU time.

statistics would affect the size of the overlap. However, the main point is that a significant (although modest) overlap exists and the results are robust to different methodologies. Small changes in the definition of the boundaries would not eliminate the basic finding of an overlap region.

Origination Models and the Application of Overlap Methods

The first set of regression models form a bridge from the HUD 1995 Report by replicating the linear probability model using the original specification on 1998-2000 origination data. Exhibit 18 shows a progression of models. The first model replicates the specification used in the HUD 1995 Report (p. 6-24) on FHA-eligible loans with LTV between 80 and 95 percent. The R^2 in the replication (0.0427) is even lower than the original model ($R^2=0.0927$). This result is probably because even fewer FHA loans in 1998-2000 than in 1993 have loans with LTV below 95 percent. The coefficients in the replication model are somewhat different than the original model, but perhaps closer to expected. LTV is positive in the replication model, as opposed to the unexpected negative sign in the original model. The underserved indicator is still positive, but the component tract variables for income and percent minority are insignificant. The center city indicator is now negative. First-time homebuyer is not available in the new data. Otherwise, the signs and significance of the replication model are similar to the original model.

The next model in Exhibit 18 expands the range of loans to include all FHA-eligible loans with LTV greater than 80 percent. This change boosted the LTV coefficient more than tenfold and increased the model R^2 from 0.04 to 0.46. Switching to logistic regression brings few changes except the underserved indicator is no longer significant. Adding the credit score in the next model improves the model fit, as lower credit scores are strongly associated with FHA loans. Credit score also makes tract percent minority significant, attenuates the LTV coefficient and makes insignificant the ratio of loan to FHA loan limit. Incorporating the weights revives the tract income and percent minority variables and switches the sign on the high cost MSA indicator. The other changes and improvement in fit are modest.

