The Prevalence and Impact of Misstated Incomes on Mortgage Loan Applications*

McKinley L. Blackburn
Department of Economics
University of South Carolina
Columbia, SC 29208
(803) 777-4931
blackbrn@gmail.com

Todd Vermilyea
Federal Reserve Bank of Philadelphia
Ten Independence Mall
Philadelphia, PA 19106
(215) 574-4125
todd.vermilyea@phil.frb.org

July 2010 (Revised)

JEL Codes: G01, G21

Keywords: mortgage lending; liar loans; housing crisis

Abstract

Misstatement of income on mortgage loan applications (the "liar-loan" problem) is thought to have been a contributor to the boom and bust of mortgage markets. We provide nationwide measurements that reflect the degree to which incomes on mid-2000 home-purchase mortgage loan applications were overstated relative to the actual incomes of mortgage applicants. Our results suggest a substantial degree of income overstatement in 2005 and 2006, one consistent with the average mortgage application overstating income by almost 20 percent. We find the tendency to misstate income was influenced by securitization markets. We find limited evidence that income overstatement played a role in subsequent mortgage defaults.

*Helpful comments were provided by seminar participants at the Federal Reserve Bank of Philadelphia. We thank Andrew Kish for outstanding research assistance. The views expressed in this paper do not necessarily represent those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

Try these handy steps to get SISA findings …

(3) If you do not get [approval], try resubmitting with slightly higher income. Inch it up $500 to see if you can get the findings you want. Do the same for assets.

It's super easy! Give it a try!

Internal Memo Circulated at JPMorgan Chase, as reported in *The Oregonian*

I. Introduction

After a number of years of rapid house-price appreciation, the third quarter of 2006 saw the beginning of a major decline in prices, suggesting that the mid-2000s housing boom may have been fed by a speculative bubble.[1] Accompanying the fall in house prices has been an increase in mortgage-payment delinquencies. A major part of this boom-and-bust episode seems to have been over-lending to individuals unable to make payments and who lacked substantial home equity. With hindsight, both lenders and borrowers entered into contracts that in hindsight seemed excessively risky.

One of the explanations that is commonly offered for the increased rates of lending in the mid 2000s is a lack of diligence in documenting income on mortgage loan applications by lending institutions. The quote above is from a memo from that time, and refers to attempts by loan officers to get Chase's automated underwriting software to approve "stated income/stated assets" (SISA) applications, applications that allowed income and assets to be stated by the applicant without verification. The mid-2000s saw an increased use of "low-doc" or "no-doc" lending in which the traditional verification processes regarding income sources were no longer part of the loan application process. Historically, these type loans were marketed to high-income individuals who were self-employed or had highly variable income. However, over the 2000s

[1] The second quarter of 2006 was the peak for the S&P/Case-Shiller national house price index. This price index had fallen 20 percent from its peak by the third quarter of 2008.

this characterization appears to have changed dramatically, as low-doc and non-doc lending increased substantially. The prevalence of this type of lending is now thought to have given scope for applicants (and their loan officers or brokers) to massage income levels on applications so as to meet standards required in underwriting software. While stories of this type of activity have been noted, no academic study has clearly documented the prevalence or importance of income misstatement in the boom period for conventional mortgage lending.

Our study provides nationwide measurements that reflect the degree to which incomes in mid-2000 home-purchase mortgage loan applications were overstated relative to the true incomes of mortgage applicants. We do so by comparing reports on incomes of mortgage applicants from two different data sources. One data source – the Home Mortgage Disclosure Act data – allows us to measure incomes as reported on actual home mortgage applications. The second data source – the American Housing Survey – provides incomes of new home purchasers measured outside the loan application process. Data from both sources are examined from the period 1995-2007. Our findings suggest that, while reports of income between these two sources do differ in any year, this difference is relatively stable over time. The primary exception is around 2005 and 2006, in which there was an increase in the reported incomes on mortgage loan applications relative to those reported in the housing survey.

We are able to construct measures of the degree of income overstatement across a large sample of MSAs in the U.S. In so doing, we are able to examine potential borrower/lender/MSA characteristics that might have helped contribute to income overstatement. We also develop simple models of delinquency rates during 2008 and 2009, in which income misstatement is allowed to be a potential explanation. While we do find a simple correlation between income overstatement and higher delinquency rates, this correlation does not hold up when other factors are incorporated.

II. Mortgage Loans in the 2000s

A. Subprime and Alt-A Loans

Prior to the housing boom, the mortgage market was dominated by "conforming" home-purchase loans that met certain credit, income, and loan-limit guidelines. One advantage of conforming loans is the ability for resale to one of the government-sponsored enterprises (GSEs) – namely, Fannie Mae and Freddie Mac. An important component of meeting the GSE guidelines was documentation on the applicant's employment, income, and debts. As a result, before 2000 the large majority of home-purchase mortgage loans were "full-doc" loans with a thorough investigation of the applicant's debt and income situation. A small minority of loans were "low-doc" or "no-doc," with the usual explanation for these type loans being that the lender was reasonably assured of the borrower's capacity to repay the loan without this documentation.

During the run-up in house prices, the mortgage market saw an important weakening in the dominance of loans that met conforming guidelines. Non-conforming loans – consisting of jumbo, subprime, and Alt-A loans -- all became more prevalent as the decade continued. Borrowers for jumbo loans typically meet the "prime" standards for being purchased by the GSEs, but loan amount on the mortgage exceeds the limit imposed by the GSEs (this limit was $417,000 in 2006). Although this limit was increased over time, it did not keep pace with house price appreciation in many markets, and so increasingly limited the ability of loans to meet conformability standards in high-price markets. More important were the increases in subprime and Alt-A loans. As shown in Table 1, subprime loans grew from roughly 9 percent of mortgage loan value in 2001 to 24 percent in 2006, while Alt-A loans grew from 3 percent in 2001 to 16 percent in 2006. The rate of increase in the importance of this type lending was largely concentrated in the 2004-2005 period, with Alt-A loans in particular increasing six-fold in their importance over this two-year period.

Both subprime and Alt-A loans fail to meet the traditional conforming standards of the GSEs, though for different reasons. Subprime loans are typically targeted towards borrowers with poor credit histories. However, the 2000s saw an increased use of subprime mortgages to finance borrowers with somewhat better credit scores than in the past, but who were attempting to finance purchases that would leave the mortgage with a high loan-to-value ratio, or a high debt-to-income ratio (see Foote, et al., 2008). The desire to avoid full-documentation requirements was also noted as an increasingly common motivator for subprime mortgages in this decade, although this desire was perhaps a greater motivation for the growth in Alt-A loans. Borrowers on Alt-A loans typically have good credit histories (though they may still be less than perfect), but desire nontraditional loan or underwriting terms. Novel payment structures – such as interest-only or negatively amortizing payments – were common for Alt-A loans, and this characteristic combined with less-than-full-doc requirements likely allowed the purchases of homes by owners that would not have occurred under conforming standards. As reported in Ashcraft and Scheurmann (2008), 65 percent of Alt-A loans were less than full-doc in 2001, with this percentage growing to 84 percent by 2006. By comparison, only 28 percent of subprime loans were less than full-doc in 2001, this percentage increasing to 42 percent by 2006.

The distinction between subprime and Alt-A mortgages is not uniformly defined.[2] By the mid-2000s, the large majority of mortgage loans originated were eventually packaged with other loans for sale to private investors in securitized form. For prime loans, this had been the case for

[2] Indeed, the turn "subprime" has often been used to refer to any loan that would be in either the subprime or Alt-A class as we describe them.

many years, with most of the mortgage-backed securities issued by the GSEs. Securitization was less common for subprime loans and, in particular, for Alt-A loans in the early part of the 2000s. However, by 2005, 74 percent of subprime loan origination value was securitized, and 87 percent of Alt-A loans (by comparison, 82 percent of prime loans were securitized). As argued by Ashcraft and Schuermann (2008), several sources of friction potentially arise between financial market participants in the securitization of nonconforming loans, leading to a substantial disconnect between the motives of the borrower and originator and the desires of the investor who ends up holding the security. The disconnect between borrower and ultimate investor was perhaps enhanced by the growing tendency for less-than-full-doc loans to go through mortgage brokers rather than retail lenders (Green, 2008). As a result, by 2005, investors were holding securities that they may have mistakenly felt were almost risk-free, with the underlying assets consisting of poorly underwritten nonprime loans that were at substantial risk if housing prices were to fall.[3]

Shiller (2008) places much of the blame for the mid-2000s housing bubble on subprime lending, citing a high growth in house prices at the lower end of the house-price distribution in San Francisco in the mid-2000s as supporting evidence. Along this line, Table 2 reports percentage changes in the S&P/Case-Shiller house price index for large MSAs (for which indices are provided) over the 2000-2005 period, where the changes are broken down by whether the house sold was originally in the bottom tier (bottom third), middle tier, or upper tier for that MSA. With only two exceptions (Las Vegas and Phoenix), growth in the bottom tier was more rapid than growth in the middle or top tiers. And in many cases – notably Boston, New York, San Diego, and San Francisco – the growth at the bottom tier was almost twice as fast as at the

[3] Geradi et al. (2008) provide evidence that market analysts in the mid 2000s appreciated the consequences of a nationwide reversal of house price appreciation, but rated the likelihood of this occurrence as very low.

top. Shiller argues that this would be expected if the growth in subprime loans was a major contributing factor to the housing boom. It might also suggest that the growth in less-than-full-doc loans could have been concentrated in individuals at the bottom end of the income and house-price distribution.

Mortgage lending fell precipitously as signs appeared suggesting the end of the housing bubble. The S&P/Case-Shiller national index peaked in the second quarter of 2006, and was rapidly falling by late 2007.[4] By early to mid 2007, it was apparent that the subprime market was in crisis. While initially the crisis was associated with the failure (or takeover) of smaller subprime lenders in late 2006, the larger lenders in the subprime market started to experience major financial difficulties throughout 2007. This was associated with a decline in the number of mortgage loans, in particular among nonprime loans: as noted in Mayer et al. (2009), the rate of of Alt-A lending fell 40 percent from 2006 to early 2007, while the rate of subprime lending fell 70 percent over that same period.

B. Performance of Loans Made During the Housing Bubble

Former Federal Reserve Bank Governor Randall Kroszner has pointed to the prevalence of "stated-income" loans as a "clear culprit" in the rise in mortgage problems.[5] Mayer et al. (2009) note that the growth in no-doc and low-doc loans was indicative of a slackening in underwriting standards, and that default-rate increases have been particularly high for these type loans. Sanders (2008) provides additional evidence that banks were reporting a weakening of

---

[4] An alternative index from the Office of Federal Housing Enterprise Oversight peaked one year after the S&P/Case-Shiller index. However, the former index is much less sensitive to changes in house prices associated with non-conforming loans, which may explain the differential movements between the two.

[5] In remarks to the National Association of Hispanic Real Estate Professionals Legislative Conference in March 2008, Governor Kroszner noted that "When we looked closely at why so many borrowers had mortgages that they struggled to repay so soon after taking out the loan, the prevalence of 'stated-income' lending was a clear culprit. Substantial anecdotal evidence indicates that failing to verify income invited fraud. Moreover, when we looked at the loan-level data we saw a clear correlation between 'low-doc' or 'no-doc' lending and performance problems, particularly early payment defaults."

underwriting standards in the 2004-2006 period. He also notes that delinquency rates for subprime mortgages were falling in the 2000s until 2006 (delinquency rates were steady for prime loans). Green (2008) finds falling delinquency rates from 2001 to 2005 among both low-doc and full-doc subprime ARM loans (for loans with a set of fixed credit characteristics). However, he notes that the delinquency rates started to rise with the 2005 vintage, though only for the low-doc subprime loans. It would seem that this fall in delinquencies before 2005 could have contributed to a growing confidence in the subprime part of the market, and hence an increased desire to make loans to borrowers that in an earlier period might have been considered too risky. Contributing to this growing confidence was the rapid house price appreciation occurring in the mid-2000s, as an expected increase in house prices could make attractive a loan that would be considered unprofitable in a period of stable expected prices. The appearance of exotic mortgage products – for example, Alt-A loans that allowed interest-only (or even less-than-interest) payments – became increasingly common in the mortgage market, again indicative of a belief that house price appreciation could improve the loan-to-value ratio without there being direct contributions to principal through loan payments.[6]

In explaining the rise in mortgage default rates since 2006, Haughwout et al. (2008) point to changes in the economy as the most important factor, with falls in house prices as the dominant explanation.[7] However, a reduction in underwriting standards also appears to have played a role. Foote et al. (2008) note that, although FICO scores were actually rising among subprime borrowers in the early-to-mid 2000s, on net the creditworthiness of these loans was

[6] A belief of many borrowers in the subprime market may have been that an increase in the house price would enable a refinance at a lower rate, as the loan-to-value ratio would be increased at this time. Mortgage payments that might have been unsustainable for the full 30 or more years of a loan would then be refinanced to a lower, sustainable payment.

[7] Haughwort et al. only attempt to explain rising default rates within subprime, or within Alt-A loans. Of course, one factor in the increase in default rates is the shift in lending towards more nonprime loans, which have always had higher default rates (see Sanders, 2008).

still falling due to decreases in requirements for loan-to-value, debt-to-income, and documentation status. They argue that a combination of forces may have been important to the severity of the decline – the reason why the turnaround in house prices had such a large effect on mortgage default was that the 2005 vintage of subprime loans had been underwritten with reduced standards (such as high loan-to-value ratios) that put these loans at substantially higher risk if the house-price boom were to unravel.

In the mid-2000s, it became common knowledge in the mortgage industry that “stated-income” and other less-than-full-doc loans opened up the possibility of a substantial misstatement of income, and the term “liar loan” was adopted by those analysts that were concerned about this possibility (see, for example, Harney, 2005). This overstatement of incomes may have led to loans being made to borrowers that would not have been considered creditworthy if true income had been utilized. If this were the case, the resulting fall in house prices left these applications at particular risk, and so the rate of default should be expected to be particularly high where liar loans were more prevalent. An alternative view is that participants in mortgage loan markets were aware of the liar loan problem and took this into account in their lending decisions, implying that an increase in liar loans need not have led to higher delinquency rates.

An open question is why the mortgage industry may have condoned this increase in less-than-full-doc lending. The evidence seems to suggest lending institutions were weakening their underwriting standards in many dimensions of mortgage lending (for example, credit score, loan to value requirements, debt-to-income requirements) at the same time. It may have been that many lending institutions desired to weaken standards to an even greater degree than their official policies seemed to indicate. Rather than officially set a less-restrictive debt-to-income threshold (or abolish this requirement altogether), lenders might prefer to maintain official standards while increasing their use of stated income loans. In packaging these loans to sell to investors, it may have looked better if mortgages met normal underwriting standards, even if this appearance was based on unsubstantiated characteristics.

No academic study has previously examined system-wide evidence for the U.S. on the existence and magnitude of potential income overstatement in home mortgage loans. There have been a small number of previous compilations of limited evidence on income misstatement, but these generally use small or select samples. An interesting example is cited by Gimein (2008), referring to Steven Krystofiak's testimony to the Federal Reserve that 60 percent of 100 stated-income loans that he examined appeared to overstate income by at least 50 percent. In the following, we combine evidence from two large nationwide datasets to assess the relevance of income overstatement on mortgage loans in the mid-2000s. We also examine whether markets that seem to exhibit a tendency for income overstatement also had higher delinquency rates in the subsequent years. To our knowledge, ours is the first attempt to use U.S. data on incomes across MSAs to measure the importance and impact of potential income overstatement in this period.

III. Data Sources

A. Incomes from AHA and HMDA

The basic data set for our analysis of actual incomes among owners of newly-purchased homes is the American Housing Survey. The AHS is conducted by the U.S. Bureau of the Census, and we use data from every year available starting in 1995. The data are collected at a household level, and the primary purpose is to obtain information on the quality of housing. The AHS provides respondent's reports on various measure of income for the previous year, as well as basic demographic information such as education, age, and MSA location. Further, respondents are asked for a variety of information concerning any mortgage loan that may have been associated with a home purchase, including information on whether the loan was insured by the government, the interest rate on the loan, and the lien status of the loan. To concentrate on recent participants in the mortgage market, we restrict out sample to homeowners who moved into their home and obtained a new mortgage within the year-long period prior to the AHS interview, and focus on purchasers of single-unit dwellings located in an MSA.

The AHS actually consists of two different survey mechanisms. National surveys incorporate a nationwide sampling scheme (of roughly 55,000 households), while metro surveys focus on about 14 MSAs in a given year (with at least 3,200 households sampled within each MSA in a particular year). In recent years, national surveys have typically been conducted in odd-numbered years and metro surveys in even-numbered, though neither survey was conducted in 2000, 2006 and 2008, and both surveys were conducted in 2007. The surveys are conducted in a 3-to-7 month period in the middle of the year. The MSAs included in the metro survey are cycled from a list of 47 large MSAs, with some slight variation in this list over time. As a result, the particular MSAs represented in our sample will vary over time, with the number of MSAs represented considerably higher when using the AHS from an odd-numbered year.

As our desire is to measure the extent to which reported incomes reflected true incomes among loan borrowers, we also require data on incomes reported on mortgage applications. The information reported on a loan application is not collected in the AHS, so we estimate this income using the Home Mortgage Disclosure Act (HMDA) data.[8] All mortgage lending

[8] Studies that have examined recent problems in the mortgage market have primarily relied on secondary-market individual-loan data, such as FirstAmerican CoreLogic's Loan Performance data, which carries extensive information on loan characteristics and outcomes for securitized loans. However, this is a select data set (only securitized loans are included), and in any case does not include income *per se* as one of its loan characteristics.

institutions with offices in metropolitan areas are required to report data on home-mortgage applications, and the coverage of applications is almost universal. While the HMDA requirement has existed since 1975, the level of information required was substantially expanded starting with the 1990 HMDA. Beginning with that year, lending institutions were required to provide application-level data on applicant income, loan request, sex and ethnicity of the applicant(s), property location, and the decision outcome on the application. While the data are not suitable for reliably assessing the creditworthiness of applicants – for instance, no information is collected on the property value or the applicant's debt or credit history – it is ideal for purposes of measuring incomes as stated on originated loan applications.

We primarily use the HMDA data to construct measures of average incomes as stated on loan applications across homeowners with newly-purchased homes within specific MSAs. In so doing, we restrict our sample of HMDA applications to those that were: (1) originated; (2) home-purchase; and (3) owner-occupied. A small number of loans report income as zero, so we treat this as a missing value and exclude these applications in our calculations.[9] In cleaning the data, we also excluded applications with requested loan amounts of zero, or action dates that were either before the application date, or more than a year after that date. We begin our analysis with applications originated in 1995.

HMDA and AHS do no provide an identifier that allows us to simply match loans across the two data sources.[10] Therefore, we use common characteristics to create a set of all possible HMDA matches for each AHS new homeowner observation. In particular, we use geography,

[9] We also top code incomes at an upper limit of $10 million dollars, as a similar top code is imposed on the comparison data in the AHS.

[10] In an earlier version of the paper, we performed matches based only on MSA and date, comparing average incomes in the same AHS/HMDA samples at the MSA level. While only using data through 2005, these results were consistent with the evidence reported in Table 3 concerning the existence and magnitude of income misstatement.

race, government insurer, and the move/loan date to perform our matches. Both datasets provide an indication of whether a mortgage insured by FHA, VA, or Farmer Mac, so we match on any government insurer of the loan. Both sources provide detail on the location of the home, with the MSA being the most specific definition common to both data sets. The precise definition of each MSA does vary over time, so in our matching we attempted to define MSAs within the HMDA to match as closely as possible the MSA definition used in the corresponding AHS.[11] In both datasets, the definition of racial categories also changes over time, and by the end of the sample period both data sets allow the reporting of multiple racial categories for the same individual. For each AHS observation, we take all racial classifications listed of any adult in the household and match to any HMDA application where the applicant or co-applicant shares any racial classification. Both AHS and HMDA provide some information on race, as well. In the HMDA we have data on the applicant, and a co-applicant if present. The timing of the loan was also used to match across HMDA and AHS, where for each AHS loan we matched to any HMDA loan with an origination date that was within a period starting two months before and one month after the reported AHS move date.

Finally, we made use of the reported loan amount on the application. Given that there may be some reporting error concerning the loan amount, we did not require loan-amount matches to be exactly equal from both AHS and HMDA. Rather, for each AHS loan, we matched to any HMDA application (with the same MSA, insurer status, date, and race) that had a loan amount within a certain range of the AHS reported loan amount. In most of our empirical results, we set the bandwidth at $2000, so any otherwise-matching loan from the HMDA with loan amount within $2000 of the AHS loan amount is used in forming the predicted loan-application

---

[11] Although the definition of an MSA will vary slightly over time – for example, the counties included in the Philadelphia MSA have changed – we use a single dummy indicator across the years to indicate that MSA in our regression analysis.

income for that AHS loan. If there is more than one loan matched from HMDA, a kernel-weighted average is taken of the incomes from those HMDA loans using an Epanechnikov kernel function. We also computed averages using higher bandwidths -- $5000, and $10,000. One limitation of the AHS is that loans are top-coded at $999,999, making it difficult to match to HMDA on the basis of loan amount for these observations, and so we exclude those AHS observations. Starting with 11,913 AHS observations with otherwise complete information, we only lose 12 observations due to top-coding. Using our narrowest bandwidth ($2000) reduces the number of observations to 10,572, although using the broader bandwidth of $10,000 only adds back 262 observations to this total.[12] The number of matched HMDA applications was generally large – 79 HMDA applications for each AHS observation, on average.

A final choice necessary for comparing AHS and HMDA incomes is the particular measure of AHS income to use. We would naturally want the measure that reflects all income sources that would be reported on a mortgage application, but this can vary across lending institutions (and over time), due to variations in policies and procedures for underwriting mortgages.[13] The AHS provides several income definitions that could be used to measure gross annual income. Our results are based on total family income, for the family listed as containing the householder. We have also tried other, more narrowly defined measures of income – for example, an income definition that consists of the joint annual wage and salary income for a married couple (if the householder is a partner in a marriage) or the householder's wage and salary income for an unmarried householder – with largely similar results. One aspect of AHS total family income is that it is on average quite close to the HMDA income matches, especially

[12] In practice, we also excluded observations in which the loan amount was less than half AHS income, in which AHS income was zero or less, as well as observations from the Honolulu MSA. This left us with 10,171 matched AHS observations in our basic set of regressions using a $2000 bandwidth.

[13] For example, one institution might include income from the rental of a room or garage in the new home, while another does not. Treatment of investment income might also vary across institutions.

in years before 2005, while the alternative based just on salary income tends to have a large discrepancy on average in those years.[14] We recognize that these measures do not represent a perfect comparison, and so anticipate a bias in using AHS income as a measure in what might be expected for HMDA income in any given MSA/quarter (even if there were no income misstatement in the HMDA). However, we also operate under the assumption that this bias did not change dramatically during the mid-2000 period.

Figure 1 presents a time-series plot of our basic average incomes (in real terms) from the two data sources. As average incomes do vary across MSAs, we present means that remove MSA-related variation in incomes in each survey.[15] In any given year, the average incomes from the two sources are not equal, but at least before the mid-2000s the two paths are reasonably close. AHS income does exceed HMDA income in 2000, but this comparison is based on a small number of observations as there was no AHS survey in 2000. The two lines clearly start to differ in 2005, with HMDA average income considerably above AHS income. This difference continues in 2006. The two averages start to come back together in 2007, though a clear difference is still suggested. Almost all of the 2007 observations are from the first half of that year (due to the nature of the AHS sampling) so these results are from a period in which market participants were gradually becoming aware of problems in the subprime mortgage market. While we do explore the statistical significance and possible explanation of these differences in the next section, we find the pattern interesting: while the HMDA incomes continue to grow in 2005 and 2006, AHS incomes over this same period are falling in real terms – consistent with a

[14] The average difference between HMDA and AHS family income is less than $3000 before 2005, while the average difference between HMDA and our AHS salary measure is over $13,000.

[15] In particular, we estimate separate regressions for HMDA and AHS incomes with annual dummies, along with dummies for MSA quarter of the year. We then use the annual dummy coefficients to adjust the starting incomes in 1995 for each income source.

shift of the distribution of actual incomes toward individuals with low incomes without any accompanying shift reflected in the distribution of incomes stated on mortgage applications.

B. Other Variables Used in the Analysis

In our statistical analyses, we make use of several additional loan factors that might be associated with a tendency to income overstatement. For one, we construct annual measures of house-price changes within MSAs. While data on median home prices is available from the National Association of Realtors, we do not use this measure as it has no control for changes in the quality of housing. We instead use the house-price index that was constructed by the Office of Federal Housing Enterprise Oversight (OFHEO), based on changes in house prices in repeat sales or refinances (OFHEO was subsumed in the newly created Federal Housing Finance Agency, which now maintains the index). We take a four-quarter average (in any year) of the OFHEO index in any given MSA, and calculate the percentage change in that average from one year prior. One limitation of the OFEHO index is that it is only constructed for houses in which mortgages are sold to either Fannie Mae or Freddie Mac (see Calhoun, 1996). The S&P/Case-Shiller index is constructed in a similar manner – in fact, OFHEO based their index on the methods developed by Case and Shiller. The S&P/Case-Shiller index is not restricted to Fannie and Freddie loans in its coverage, but it covers far fewer MSAs.

The HMDA data also contain a variable that indicates whether the loan was sold by the end of the year in which the action was taken. We measure the percentage of originated loans in an MSA that were sold to Fannie Mae, Freddie Mac, or Ginnie Mae, which we refer to as “government sold.” The HMDA data also provide the underlying regulator of the financial institution that originally made the loan, so we are able to assess income reporting in the banking vs. non-banking parts of the market. The AHS provides a number of useful pieces of individual information, including age and education of all family members (education is a coded variable for various degrees or highest grades obtained, which we use to construct a years of education measure), and the mortgage rate paid on the loan.

In our final set of regressions we use as dependent variables serious delinquency rates on all outstanding mortgages in an MSA. These are measured as of the end of the year in each MSA, and represent mortgages that are either 60 days past due on a payment or are in foreclosure. The source for the delinquency rates is the Lender Processing Services Applied Analytics data (formerly known as McDash), which represent mortgages serviced by a number of large loan servicers in the market (covering roughly two-thirds of all outstanding mortgages). These data include loans sold to GSEs, as well as subprime and Alt-A loans that may have been securitized. All loans in the data have an indication of whether or not the loan was generated through a subprime lending channel, so we also calculate delinquency rates for this group of "subprime" loans. We point out that this group basically consists of loans that are non-conforming to GSE standards, so in fact includes Alt-A along with the narrower definition of subprime. In these regressions, we use unemployment rates as a measure of MSA-level economic activity, which are taken from the Local Area Unemployment Statistics provided by the Bureau of Labor Statistics.

IV. Empirical Results on the Extent of Income Misstatement

A. Empirical Model and Estimation Approach

For a varying sample of new mortgage loans over time, we have observations on the income of new home buyers from two different surveys. Our expectation is that data from the HMDA source will tend to show an overstatement of income for new homeowners relative to data from the AHS source, especially in the mid-2000s period in which nonprime loans became more prevalent. In this section, we develop an empirical model that allows us to test if the difference in repeated incomes from the two surveys increased in the mid-2000s, relative to preceding years.

As discussed earlier, there are reasons to anticipate that incomes between the two surveys could differ for a given AHS loan even without the increase in income overstatement generally thought to have occurred in the mid 2000s. As such, we want to develop a statistical model that allows for a difference in income (between the surveys) that can vary both across geographic areas at a point in time, and within an area over time. To do so, we assume that the income of the ith loan (y_{ismt}) follows:

$$y_{ismt} = \tau_{sm} + \eta_{st} + \varepsilon_{ismt} \tag{1}$$

where s is a subscript indicator for the survey, m indicates the MSA, and t indicates the quarter in which the loan was finalized. In this model, the income from the AHS survey is that reported by the respondent, while the income from the HMDA is the matched average income obtained for that AHS loan amount (as detailed in the previous section). The equation allows the MSA (τ_{sm}) and time (η_{st}) effects to vary between the surveys. Given equation (1), we can difference across the surveys (HMDA values minus AHS values) to provide

$$\Delta y_{imt} = \Delta\tau_{m} + \Delta\eta_{t} + \Delta\varepsilon_{imt} \quad . \tag{2}$$

In this equation, the difference in average incomes across surveys is the dependent variable. Our focus is on the values for $\Delta\eta_{t}$, which reflect differences across the surveys in time effects for income. While the equation could be estimated with a complete set of time effects, this does not provide a frame of reference with which to argue that the mid-2000 effects differ from the early part of the sample. So we instead restrict the temporal path of these effects by initially assuming a possible linear trend for this survey difference, but one that allows deviations from trend in certain quarters. In particular, we assume

$$\Delta\eta_{t} = \gamma_0 + \gamma_1 t + \gamma_2' QS_t \qquad (3)$$

where QS is a vector of dummies representing “suspect” quarters that might differ from the early years, and γ_2 measures deviations from the trend in those suspect quarters. Equation (3) can be substituted into the one above to provide our basic estimating equation

$$\Delta y_{imt} = \Delta\tau_m + \gamma_0 + \gamma_1 t + \gamma_2' QS_t + \Delta\varepsilon_{imt} \ .$$

which is estimated by a fixed-effects type estimator (incorporating MSA effects), so as to identify the primary parameter of interest γ_2. The error term in the equation ($\Delta\varepsilon_{imt}$) is likely to be heteroskedastic, given the varying number of matched loans from the HMDA used to calculate HMDA income, and is also possibly correlated over time for observations from the same MSA. These potential problems are addressed by using standard errors that are robust to heteroskedasticity, and allow an arbitrary correlation across errors at the MSA level. There is naturally measurement error in the HMDA income measures, as we are not matching perfectly to each AHS household in constructing our application-income measures. However, this should be a random source of error, and so would be subsumed in the equation error without causing any problems other than heteroskedasticity.

B. Basic Results

The results from estimating our income-difference model are reported in Table 3, employing a variety of specifications of time effects. In column (1), we simply incorporate separate dummy variables for each of the years in the mid-2000s period, along with a time trend. Our basic finding is that there is statistically significant evidence of an increase in the difference in incomes – HMDA minus AHS – in 2005 and 2006. By contrast, the 2004 dummy does not suggest a difference for that year. The estimated coefficient for 2007 is positive, but is of a somewhat lower magnitude than the estimated 2005 and 2006 effects This effect is statistically insignificant (though it is somewhat imprecisely estimated). Given that our 2007 loans are from the early part of that year, this may be indicative of a greater attention being paid to income documentation as the subprime crisis was becoming evident. Column (2) estimates the same specification using HMDA average income calculated with a larger bandwidth -- $5000 – with similar results to that obtained in column (1) using the smaller bandwidth.[16] No evidence of an trend in the HMDA/AHS difference is apparent in columns (1) and (2), and allowing a post-2000 dummy shift in this trend in column (3) does not alter this conclusion.

The specification in column (4) takes the mid-2000 year effects, and breaks them out into possible quarter effects. In this estimation, no loans from the 4th quarter of 2005 or the 1st quarter of 2006 are available (due to the lack of an AHS survey in 2006), and some of the quarter effects are imprecisely estimated (such as the 3rd quarters of 2005 and 2007) due to a small number of loans available for those quarters. The results do suggest that income overstatement had started to increase by the last quarter of 2004 – the coefficient estimate for this quarter is significantly positive, while the earlier quarters in that year are insignificant and in two cases have negative estimated coefficients. Although not individually statistically significant in each case, the coefficient estimates for 2005 and 2006 are all positive and sizeable. The coefficient estimate for the 3rd quarter of 2007 is small and statistically insignificant, as might be expected if concerns about mortgage delinquencies were starting to affect behavior, but the imprecision in this estimate makes it difficult to infer much from this result.

As suggested in Figure 1, the major reason for the growth in income differences between HMDA and AHS was a failure for the AHS incomes in the mid-2000 to increase at the same rate

[16] Expanding the bandwidth to $10,000 also did not change the basic results.

as HMDA and AHS incomes had been increasing since the middle 1990s. This is documented in the final two columns of Table 3, where separate regressions are estimated in which HMDA income or AHS income is the dependent variable. Both income sources have a similar underlying trend – an increase of about $450 per quarter – but only the AHS incomes deviate from that trend in the mid-2000s (there is no evidence of a mid-2000s effect in the HMDA regression). The suggestion, then, is that while true average incomes were falling significantly below trend in 2005 and 2006 – indicative, perhaps, of the increase in subprime lending in this period – it was not reflected in the incomes reported on mortgage loan applications.

It is worth noting that the estimated suggestion of income overstatement in 2005-2006 is quite sizeable. In particular, the suggestion is that the average loan in 2005-2006 had an overstated income of about $15,000-$17,000 (in 2000 dollars), or about a 19-21 percent overstatement of income relative to the average AHS income over the 1995-2007 period. Given that income misrepresentation in prime loans should have been limited, this suggests a large degree of income overstatement in nonprime loans during this period.

B. Borrower Characteristics and Income Overstatement.

While the estimated models in Table 3 are clearly consistent with a tendency to overstate incomes on loan applications in 2005 and 2006, there are several ways of breaking down the data so as to examine the consistency of our estimate with expectations about the prevalence of income overstatement in different segments in the mortgage market. We primarily examine characteristics of the loan that can be measured from the AHS and/or HMDA. In these additional estimations, we take the specification from column (4) of Table 3 and group together the quarters in which the existence of an important deviation from trend was evident – namely, the 4th quarter of 2004 along with all of the available quarters from 2005 and 2006 – and reflect this period with a “suspect quarters” dummy. We then interact this dummy with variables that might be expected to influence income overstatement. In Table 4, we form interactions with characteristics of the loan taken from the AHS (or HMDA and AHS). Column (1) of the table (with no interactions) verifies an income overstatement of roughly $15,000 in the suspect quarters.

The first factor that we consider is a measure of whether or not the loan is a “conventional loan,” that is, a loan not insured through a government agency. Both the AHS and HMDA data provide information on whether the loan was financed through the Federal Housing Authority (FHA), the Veterans Administration (VA), or Farmer Mac, and this was one of the characteristics used in matching HMDA applications to the AHS. Given our focus on loans made within MSAs, only a handful of AHS loans in our sample were made with Farmer Mac guarantees, but we do have a number of loans made with FHA or VA financing (24 percent of loans in general, though this percentage had fallen to only 12 percent during our suspect quarters). These government-guaranteed loans all have stringent income-documentation requirements. For example, FHA requires two years of income tax returns along with income verification with sources; VA also requires income verification, and tax returns for individuals who are self-employed. Given these documentation requirements, we should not expect to see evidence of income overstatement among this group of loans.

The second column of Table 4 includes an interaction of the suspect dummy with a dummy for the loan being a conventional loan, and the results are exactly as we would expect. The suspect dummy reflects income overstatement for the group of government-insured loans, and this dummy is statistically insignificant. The interactive effect, however, is positive and statistically significant, consistent with an estimated income overstatement of roughly $18,000 among the set of conventional loans during the suspect period. Interestingly, the conventional loan dummy coefficient estimate is small, but positive and marginally significant. This suggests that conventional loans might have had a small amount of overstatement even in periods before the mid-2000s boom.

Ideally, we would also include in our specification an interaction with a prime/nonprime loan dummy, but this information is not available from either the HMDA or the AHS. The one exception has to do with loans that are non-conforming because the loan amount exceeds the limit set by the GSEs for potential purchase. As such, income verification requirements of the GSEs are not immediately relevant in this part of the market, and so income overstatement during the boom periods might be expected among this class of loans. On the other hand, jumbo loans are considerably less likely to be securitized than other classes of loans – less than 50% were securitized during 2005 and 2006 (Ashcraft and Scheurmann, 2008) – and so the agency problems associated with securitization may not have been as prevalent with these loans. Including a jumbo loan interaction in column (3), we do see a suggestion of a considerable tendency towards income overstatement in the suspect period among jumbo loans (as well as a general tendency to overstatement in jumbo loans before that period). However, also including an interaction between the suspect dummy and the loan amount (in column 4) significantly reduces the jumbo loan interaction estimate, and leaves it insignificant. Rather, what is evident is a general tendency toward income overstatement (during the suspect period) among larger loans than smaller ones. This evidence of less income overstatement among smaller loan sizes suggests income instatement may not have been characteristic of the subprime market during the mid-2000s.

We also tested some additional hypotheses using interactions with data obtained from the AHS. Column (5) of Table 4 reports a specification that incorporates an interaction of the mortgage rate on the loan with the suspect dummy. Nonprime loans tend to have higher mortgage rates, so the expectation was that we might see a positive coefficient on this interaction. But, in fact, there was no evidence of such an interaction. This may reflect the low mortgage rates that would be reported among some of these loans if they involved teaser rates (in ARMs) or interest-only loans that were common in the Alt-A market. We also considered interactions with income, or with AHS-measured characteristics that should be related to income (such as education, age, and minority status), but failed to find any evidence of an interactive effect.[17] In total, these results provide no evidence to suggest that income overstatement was an important problem in the subprime market.

C. HMDA and MSA Characteristics

Limited information on the lender and loan is available from the HMDA, and in this section we attempt to make use of that information to explore the parts of the mortgage market where income overstatement appears to have been most severe. Unlike in the previous section, however, the match of the information to the underlying AHS loan is less than perfect, so we are only able to assess tendencies for certain lender/loan characteristics across different income groups and geographic locations. The analysis in this section also suffers from the fact that the measured characteristics will tend to predict HMDA income better than AHS income, although (as we argue below) under the assumption that this bias is constant over time our interactive coefficients should still provide useful implications.

[17] There are potential statistical problems with using characteristics that help to predict one income source (AHS) better than the other income source (as we discuss in the next section), though they may not affect the interpretations of the interaction coefficient estimates.

One set of lender characteristics that we use relates to information on the primary government agency that regulates the lending institution from which a HMDA loan is obtained. There are six such regulatory agencies: Housing and Urban Development (HUD); Office of Thrift Supervision (OTS); National Credit Union Association (NCUA); Federal Deposit Insurance Corporation (FDIC); Federal Reserve Board (FRB); and the Office of the Comptroller of the Currency (OCC). Over the period under study, these agencies varied in the type of mortgage lender that they might regulate. The OCC regulates national banks, while the FRB and FDIC are the primary federal regulators of state banks. Credit unions are regulated by the NCUA. OTS primarily regulates thrift institutions, and was the regulator of many of the large subprime and Alt-A lenders -- such as Countrywide, IndyMac, and Washington Mutual -- that experienced severe problems during the housing crisis. Mortgage lenders that fall outside the prudential regulatory framework are classified as regulated by HUD – this would include many nonbank, nonthrift institutions that specialize in mortgages. While we do not observe the regulator of the AHS loan that is part of our matches, we do observe the percentage of matched loans from the HMDA that were regulated by each agency. These percentages do vary with characteristics of the loan – most notably, MSA – which helps in identifying the regulatory effect.[18]

The second column of Table 5 incorporates regulator information by including these percentages along with interactions with the suspect dummy. In this table, we restrict our attention to conventional loans only, given the evidence from Table 4 that overstatement was not prevalent among government-insured loans. Column (1) of Table 5 estimates a similar regression to those in Table 4, suggesting an average overstatement of about $11,000 for a loan at the

[18] Regressions of the matched regulator percentages for each loan in our data on a time trend and a set of MSA dummies revealed R^2‘s from 0.25 to 0.54 for the six regulatory agencies.

average loan amount (the "suspect/loan-amount" interaction is calculated using the deviation of the loan amount from its mean, facilitating the interpretation of the suspect dummy coefficient estimate). In column (2), the regulator variables are added, and these interactive coefficients suggest that both nonbank regulators – OTS and HUD – appeared to be associated with loans in which income overstatement was common (relative to loans from national banks regulated by the OCC). The magnitudes of the coefficient estimates, however, are almost unbelievably large, suggesting going from a 100 percent probability of being an OCC loan to a 100 percent probability of being an OTS loan would increase overstatement in the mid-2000s period by over $100,000. The individual regulator interaction coefficients are imprecisely estimated, so as an alternative we grouped HUD and OTS together as "nonbank regulators" (treating credit unions as banks) in column (3), showing an overstatement impact of nonbanks of just under $60,000 – and providing no evidence of overstatement for an average-sized loan made in the banking sector.

One concern that might arise in interpreting the results in columns (2) and (3) is that the non-interacted coefficient estimates for the regulator variables are often large and statistically significant. For example, in column (2) both the OTS and HUD variables suggest a lower level of income overstatement in institutions regulated by these two agencies in years prior to the mid-2000s. While this could be the case, it also possible that the nonbank lenders may have had a more restrictive definition of allowable income than other lenders. Most likely, however, is that the regulator percentages help to more accurately predict the HMDA income part of the dependent variable than the AHS part. For example, a regression with HMDA average income as the dependent variable shows a strongly negative and significant coefficient on the nonbank percentage, but the same coefficient using AHS income is smaller and statistically insignificant.[19] As well, additional regressions (not reported) show that the major reason for the significant interaction with the nonbank percentage in column (3) of Table 5 is that AHS incomes associated with high nonbank percentages fell during this period. By contrast, the HMDA income difference between banks and nonbanks was stable throughout the period. As the bias from using the regulator percentages has to do with predicting HMDA incomes, not AHS, the implications of this set of findings may not be biased – namely, that nonbank regulators did tend to move towards making loans to households with lower true incomes, although this difference was not reflected in incomes reported on loan applications.

Given the higher underwriting and documentation requirements, we expect loans sold to GSEs to display a lower tendency for income overstatement. HMDA does ask the lending institution whether or not the loan was sold by the end of the year, including whether it was sold to Fannie Mae, Freddie Mac, or Ginnie Mae. As with the bank regulator variables, this is a characteristic of the HMDA loans (and is not available for the AHS loan), but there is evidence of considerable variation in the tendency to sell to a GSE across MSAs. In column (4) of Table 5 we include the GSE percentage along with an interaction with the suspect dummy. The coefficient estimate is as we would expect – negative, suggesting less income overstatement in markets where conforming loans were more common – and is marginally significant. Of course, the GSE variable suffers from the same concerns as the regulator variables, as the GSEs had loan limit and low-income preferences that tended to lower income relative to non-GSE loans (as evidenced by the negative and statistically significant coefficient on the non-interacted GSE percentage). Incorporating the GSE percentage also weakens the evidence for any nonbank institution effect during the suspect period, as the coefficient estimate on this interaction is now

[19] This is from a regression that also has the suspect dummy, and the suspect dummy interacted with the nonbank percentage, as controls (along with a trend and MSA dummies).

statistically insignificant (though the estimate is still quite large). While the GSEs may have started to weaken some of their income documentation requirements during the mid-2000s period, it still appears to be the case that income overstatement was less of a problem in markets with lenders that followed the traditional path of originating conforming loans with the intention to quickly sell loans to the GSEs.

We also studied whether or not income overstatement was related to the degree of house price appreciation in the local market. The tendency to overstate incomes is often thought to have been related to whether or not the MSA market was a “hot market” at that time. For most MSAs in our sample, we can use OFHEO price indices to measure the rate of house price growth in the bubble years. As we measure it, this is an MSA-specific measure that does not vary over time, though we do explore whether the results are sensitive to using house price growth in 2004 or in 2005. Columns (1) and (3) of Table 6 estimate a simple specification that adds the house price growth rate as the only interaction with the suspect dummy, and the suggestion is clear that income overstatement was more characteristic of markets with high house-price growth in these periods. However, when the other interactions from column (4) of Table 5 are added to the regression, the house price interaction is no longer statistically significant. The primary reason for the fall in magnitude of the estimated house-price interactive effect is the addition of the real-income interaction (which remains positive and statistically significant). Income overstatement was more prevalent in markets with higher house-price growth because these markets also tended to have larger loan sizes. Taken at face value, the results would suggest that a $600,000 mortgage loan in San Diego in 2005 would have a similar tendency to income overstatement as a $600,000 loan in Omaha. Income overstatement was likely more prevalent in San Diego than in

Omaha, but this was primarily because high house price growth had led to the average loan amount being higher in San Diego.

D. Comparisons of AHS New Homeowners to Existing Homeowners

The final two columns of Table 3 (as well as Figure 1) suggested that the primary reason for the divergence between our HMDA and AHS sampled incomes was a fall in AHS incomes during the mid-2000s. In this section, we examine whether this fall in AHS incomes also occurred relative to non-moving homeowners over the same period. We can also use the non-moving homeowner population as a frame of reference to examine any changes in the parts of the income distribution from which new homeowners emerge.

The comparison of new homeowners (purchased within the last year) and all other homeowners is complicated by the fact that the AHS samples we use are not nationally representative (some MSAs are over- or under-sampled). A further complication is that there is not a time for non-moving homeowners analogous to the move date that we have used to analyze the timing of new homeowners. On the other hand, we do have interview dates for all homeowners in the AHS, as well as MSA location of the home, so we can make comparisons between new and existing homeowners using samples constructed on a similar basis using these two characteristics. In practice, we use a matching estimator by constructing empirical probabilities of each MSA/interview-date combination from the new homeowner population, and then applying these weights to the existing homeowner distribution to form a comparable expected value from this latter distribution.[20] We make new/existing homeowner comparisons in

[20] This is an estimate of the impact of the "treatment" (new home-owning) on the treated – that is, it reflects how the new homeowner distribution compares to an existing homeowner distribution chosen so as to have similar MSA and interview-date characteristics as the new homeowner distribution (see Angrist, 1998, for a discussion of the estimator we use). As in the earlier analysis, we restrict out attention to single-family households in choosing the reference distribution of existing homeowners. One limitation of this estimator is the possibility of MSA/interview-date combinations occurring among the treated but not the control. In practice, we have only seven such combinations, so we do not see this as an issue.

both the nonsuspect and suspect years, which we accomplish by separating the AHS new homeowners into two sample groups based on their move date, and then forming separate weights to apply to the existing homeowner distribution for these two comparisons. The estimated treatment effects are obtained using these weights in a weighted least squares regression of family income on a move dummy (using robust standard errors assuming clustering at the MSA level).

The top panel of Table 7 reports the comparisons of average income, age, and education between new and existing homeowners, separately for the nonsuspect and suspect years. The difference in average incomes between new and existing homeowners in the nonsuspect period is only around $30. Underlying this is an average age among new homeowners that is about 13 years less than among existing homeowners, and an average education that is about half a year higher.[21] This changes considerably in the suspect quarters, when new homeowners have an estimated average income almost $14,000 less than that of existing homeowners. The average income difference is almost as large as the HMDA/AHS income difference reported in Table 4, suggesting that the changes during the suspect years reflected something particular to the new homeowner distribution, and not some general change in the AHS that might have affected both new and existing homeowners. One implication is that HMDA incomes should not have seemed unusual at the time, as they followed a distribution similar to what one might have expected of new homeowners. Table 7 also suggests that the average age difference is slightly larger in the suspect than in the nonsuspect quarters, while the average education is closer together. Both of these can contribute to a fall in new homeowner income relative to existing homeowners, so to garner the importance of these changes we also estimated regressions for income in which

[21] Age and education are that of the householder when the householder is unmarried, but are the average across the two spouses when the householder is married.

education and age are controlled for.[22] The results suggest that, controlling for age and education, average income was around $6000 less among new homeowners relative to existing homeowners in the nonsuspect years, while the difference was more than $15,000 in the suspect period. While age and education can explain a small portion of this growing difference, most of it remains – there does appear to be something dramatically different about the selection of the new homeowner population from the income distribution during the suspect years.

The matching analysis in the top panel of Table 7 only identifies changes in expected incomes, but says nothing about changes at different points in the distributions. To get some idea of whether the fall in incomes for new homeowners is concentrated more at one end of the distribution, we used our weighting estimator to generate quantile regression comparisons of new and existing homeowners at several different percentiles.[23] For the nonsuspect period, the quantile comparisons suggest that the variance of incomes is lower among new homeowners compared to existing homeowners, as both the bottom quintiles and top quintiles are estimated to be closer to the median among the new homeowners distribution. Similar comparisons for the suspect period suggest some tendency for the bottom quintiles for the new homeowner distribution to fall closer to those of the existing homeowners, but the biggest change is in the upper tail of the distribution, where the 85th percentile for new homeowners has fallen by almost $20,000 during the suspect period relative to what it was during the nonsuspect period. While the apparent increase in the number of new homeowners at the low end of the distribution is

[22] These regressions are WLS estimates with income as the dependent variable, and independent variables including education, age, and age squared along with the new homeowner dummy. Weights are constructed so that the average weights in the new and existing homeowner samples are proportional to the percentage of usable observations from those two subsamples.
[23] We report standard errors using assumptions of homoskedastic and uncorrelated errors. These are merely suggestive, and in practice are likely to understate the appropriate standard deviations.

consistent with a greater prevalence of subprime lending during the suspect years, it is difficult to see how subprime lending can explain the apparent fall in incomes at the top percentiles.

V. Impact of Income Misstatement on Loan Performance

Our regressions in section IV suggest that income overstatement was a characteristic of the general housing market in 2005 and 2006. The occurrence of income misrepresentation leads to the natural question of whether this has had apparent effects on subsequent performance of loans originated in these markets. As noted earlier, loan performance has seriously deteriorated in the years following 2005, so we wish to examine the role that income overstatement and other measures of market characteristics may have played in this deterioration.

We do not have performance measures in either the AHS or HMDA data, and obviously cannot tie performance to income overstatement at the loan level. But we can construct average measures of income overstatement in 2005 by MSA using these data, and relate these to performance measures at the MSA level. In particular, we estimate our basic income-difference model (specification 1 of Table 4), and calculate the average residual in the "suspect period" for each MSA. We then classify MSAs as being in either the top half of the MSAs in the estimated level of income overstatement, or in the bottom half. This measure of income misrepresenatation is then related to our measure of serious delinquencies in each MSA (discussed in section III), where we examine delinquencies in both 2008 and 2009. We can measure income misrepresentation and delinquency rates for 113 MSAs – the average delinquency rate across our sampled MSAs for all loans was 8 percent in 2009 (5 percent in 2008), while for subprime loans it was 39 percent in 2009 (27 percent in 2008).

Several recent papers have estimated models for delinquency or foreclosure rates, with many examining defaults in the recent period of housing-market decline (see Immergluck and

Smith, 2005; Doms, et al., 2007; Grover et al., 2008; Haughwout et al., 2008; Richter, 2008; Sherlund, 2008; Krainer and Laderman, 2009). Two basic characteristics of a market are focused on in these models as potential explanations for an increase in foreclosure rates: one, falling house prices leading to an increase in negative-equity situations for borrowers; and, two, worsening economic conditions leading to problems in borrowers making regular mortgage payments. Bad loans typically go bad in the first few years of the mortgage, so we analyze the extent to which delinquency rates in 2008 and 2009 may depend on initial conditions in the mid-2000s, and changes in those conditions since the mid-2000s. Changes in house prices are measured using percentage changes in the annual average of the OFHEO house-price index for the MSA. The MSA-level unemployment rate is incorporated as a measure of economic conditions, both for 2005 and the later years. In the MSA-level models estimated by Doms et al. (2005), these two measures appear to be the main forces explaining delinquency-rate variation across MSAs, among both prime loans and subprime loans. As a delinquency rate is a probability that can often be small, a logistic functional form was used in estimation. In particular, we assume the true foreclosure rate (r) follows the form

$$r = \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}}$$

and estimate this model with a generalized-linear-model estimator assuming a binomial likelihood function.[24] We report both coefficient estimates and estimated elasticities (in brackets) in Table 8, with elasticities calculated at the sample means of the variables.

The top panel of Table 8 reports estimates of several specifications of this model using delinquency rates for all loans. All models also incorporate the delinquency rate in 2005 for that

[24] The follows the suggestion of Papke and Wooldridge (1996) for estimating fractional-response models. The advantage of this approach over the usual estimator that uses OLS with the log-odds ratio for r as the dependent variable is that it avoids potential retransformation problems in interpreting the coefficient estimates.

MSA as a control for geographic-area effects. A simple model that just includes this lagged delinquency rate and our income overstatement dummy as independent variables provides evidence of a statistically significant relationship between the dummy and the 2009 delinquency rate – the elasticity estimate suggests that 2009 delinquencies rates are about 25 percent higher in MSAs in the top part of the "income overstatement" distribution, compared to those in the bottom half. The bottom panel shows that this correlation is also strong if we focus simply on subprime mortgages. Specification (2) in the table adds unemployment rates in 2009 and 2005 as controls, with strong statistical evidence of effects in the expected direction – higher unemployment rates in 2009 increases delinquency rates in that year, while higher unemployment rates in 2005 (signifying areas where lending may have been more restrictive in that year) are associated with lower 2009 delinquency rates. However, inclusion of these controls reduces the estimated magnitude of the impact of the income-overstatement dummy, and leaves it statistically insignificant when looking at all mortgages. The estimated effect is further reduced when controls for house-price growth are included – in specification (3), there is no evidence pointing to an impact of income overstatement on delinquencies. House price growth in 2009 clearly has a strong impact on delinquency rates, with delinquency rates higher in areas where house prices were falling more (as would be expected). A control for house price growth in 2005 – in hopes of isolating markets where excessive lending was occurring in response to strong price movements – does not suggest growth at the time of the loan was important. The last column of Table 8 uses the 2008 delinquency rate as the dependent variable, with very similar findings to the 2009 regressions.

There are several interpretations that might be given to these estimate income-overstatement impacts. One is that the tendency to overstate incomes in certain MSAs in 2005 did not ultimately affect loan performance, because at the time loan decisions were being made the mortgage industry fully appreciated that this was occurring on most Alt-A loans. It could also be the case that the primary effect of income overstatement on mortgage markets was in driving up prices in the mid-2000s period, and areas where prices increased in the middle of the decade were also areas were prices fell more by the end of the decade. There is indeed a correlation between house price growth and our income-overstatement measure – more overstatement was associated with higher house price growth in 2005, and lower house price growth (greater declines) in 2009[25]. The data do show that house prices fall more in 2009 in MSAs where house price growth was greater back in 2005, so this "market correction" could have been the major avenue through which income overstatement impacted later delinquencies. Finally, there may be a problem with measurement error in our income misstatement variable, as these measures at the MSA level are often based on a very small number of income values from the AHS for many MSAs.[26] The fact that significant correlations can be uncovered in the simpler specifications of Table 8, however, suggests that this problem does not lead to the variable being completely meaningless in its representation.

VI. Conclusions

We examine MSA-level reports on average incomes of buyers of newly-purchased homes in two different data sets – one reflecting income on originated mortgage applications, the other census-surveyed income among newly-purchased-home buyers. We substantiate a considerable difference in income between the two sources at the height of the mid-2000s housing boom. Our

[25] A regression of 2009 house price growth on our income-overstatement dummy provides a coefficient estimate (standard error) of -0.05 (0.01).
[26] Given it is reasonable to argue that this measurement error follows the classical assumptions, our coefficient estimates in the income-overstatement equations in the previous section should still be unbiased (as the error is in the dependent variable in those equations). However, it could cause a bias toward zero when income misstatement is used as an independent variable, as it is in Table 8. If we re-estimate our model only using MSAs where the income overstatement measure is calculated with at least five loans, the coefficient on the income-overstatement dummy does increase from 0.004 to 0.060 (for specification 3, top panel), but remains statistically insignificant.

results suggest a substantial degree of income overstatement on average in 2005 and 2006, one consistent with the average mortgage application overstating income by almost 20 percent. This income overstatement is characterized by a drop in the *actual* incomes of recent buyers, while *reported* incomes on accepted home-mortgage applications continued to increase. This finding is consistent with Shiller's contention that the housing bubble was greatest in the lowest tier of the housing market. We also examined aspects of the loan application that tended to be associated with the tendency to overstate income. Our estimates suggest that the dollar value of income overstatement was larger on higher loan amounts, a result inconsistent with Shiller. The potential sale to a GSE – such as Fannie Mae or Freddie Mac – seemed to play an important role, with applications likely to be sold to a GSE having lower tendencies to overstate income. The most likely explanation for this connection was that requirements for purchase by a GSE generally required full documentation of the borrower's financial situation.

Has the tendency to overstate incomes on mortgage applications led to higher default rates for those loans in subsequent years? While the nature of our data limit the extent to which we can examine this question, we do find a positive simple correlation between the degree of income overstatement in an MSA and the subsequent tendency for mortgages in that MSA to be experiencing a serious delinquency in payment by the end of the decade. However, this correlation appears to be explained by other factors affecting delinquency rates. In particular, we find that both worsening economic conditions in the MSA, and house price declines at the end of the decade, are associated with higher delinquencies. It is possible that the primary relevance of income overstatement to later payment problems works through house price changes, as MSAs with more evidence of income overstatement are also the MSAs with greater price declines by the end of the decade.

References

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica*. 66(2): 249-288

Ashcraft, Adam B. and Til Schuermann. 2008. "Understanding the Securitization of Subprime Mortgage Credit." *Federal Reserve Bank of New York Staff Reports*, No. 318.

Bostic, Raphael W., Kathleen C. Engel, Patricia A. McCoy, Anthony Pennington-Cross, and Susan M. Wachter. 2008. "State and Local Anti-Predatory Lending Laws: The Effect of Legal Enforcement Mechanisms." *Journal of Economics and Business*. 60(1): 47-66.

Calhoun, Charles A. 1996. "OFHEO House Price Indices: HPI Technical Description." Office of Federal Housing Enterprise Oversight Working Paper.

Case, Karl E., Robert J. Shiller and Allan N. Weiss. 1996. "Mortgage Default Risk and Real Estate Prices: The Use of Index-Based Futures and Options in Real Estate." *Journal of Housing Research.* 7(2): 243-258.

Doms, Mark, Fred Furlong, and John Krainer. 2007. "Subprime Mortgage Delinquency Rates." Federal Reserve Bank of San Francisco Working Paper 2007-33.

Foote, Christopher L., Kristopher Gerardi, Lorenz Goette, and Paul S. Willen. 2008. "Just the Facts: An Initial Analysis of Subprime's Role in the Housing Crisis." *Journal of Housing Economics*. 17: 291-305.

Gerardi, Kristopher, Andreas Lehnert, Shane M. Sherlund, and Paul Willen. 2008. "Making Sense of the Subprime Crisis." *Federal Reserve Bank of Boston Public Policy Discussion Paper*, No. 09-1.

Gimein, Mark. 2008. "Inside the Liar's Loan: How the Mortgage Industry Nurtured Deceit." *Slate*. April 24.

Green, Richard K. 2008. "Imperfect Information and the Housing Finance Crisis: A Descriptive Overview." *Journal of Housing Economics*. 17: 262-271.

Grover, Michael, Laura Smith, and Richard M. Todd. 2008. "Targeting Foreclosure Interventions: An Analysis of Neighborhood Characteristics Associated with High Foreclosure Rates in Two Minnesota Counties." *Journal of Economics and Business*. 60: 91-109.

Harney, Kenneth R. 2005. "Agency Sounds Warning on Stated-Income and Interest-Only Mortgages." *Realty Times*. January 10.

Haughwout, Andrew, Richard Peach, and Joseph Tracy. 2008. "Juvenile Delinquent Mortgages: Bad Credit or Bad Economy?" *Journal of Urban Economics*. 64: 246-257.

Immergluck, Dan and Geoff Smith. 2005. “Measuring the Effect of Subprime Lending on Neighborhood Foreclosures: Evidence from Chicago.” *Urban Affairs Review*. 40: 362-289.

Krainer, John, and Elizabeth Laderman. “Mortgage Loan Securitization and Relative Loan Performance.” Federal Reserve Bank of San Francisco Working Paper 2009-22.

Mayer, Christopher and Karen Pence. 2008. “Subprime Mortgages: What, Where, and to Whom?” NBER Working Paper no. 14083.

Mayer, Christopher, Karen Pence, and Shane M. Sherlund. 2009. “The Rise in Mortgage Defaults.” *Journal of Economic Perspectives*. 23: 27-50.

Papke, Leslie E. and Jeffrey M. Wooldridge. 1996. “Econometric Models for Fractional Response Variables with an Application to 401(k) Plan Participation Rates.” *Journal of Applied Econometrics*. 11: 619-632.

Richter, Francisca. 2008. “An Analysis of Foreclosure Rate Differentials in Soft Markets.” Federal Reserve Bank of Cleveland Working Paper No. 08-11.

Rogers, Willam. H. 1993. “Calculation of Quantile Regression Standard Errors.” *Stata Technical Bulletin*. 13: 18-19.

Sanders, Anthony. 2008. “The Subprime Crisis and its Role in the Financial Crisis.” *Journal of Housing Economics*. 17: 254-261.

Sherlund, Shane M. 2008. “The Past, Present, and Future of Subprime Mortgages.” FEDS Working Paper No. 2008-63, Federal Reserve Board, Washington, D.C.

Table 1
Percentage of Origination Value from Mortgage-Loan Types, 2001-2006

Year	*Subprime*	*Alt-A*	*Prime and Jumbo*
2001	9.0	2.8	88.2
2002	8.3	2.5	89.2
2003	8.9	2.3	88.8
2004	20.8	7.7	71.5
2005	22.7	13.8	63.5
2006	23.8	15.9	60.3

Source: Ashcraft and Schuermann (2008)

Table 2
Percent Change in Home Prices within MSAs by Home Price Tier, Jan. 2000 to June 2005

MSA	*Low Tier*	*Middle Tier*	*Upper Tier*
Atlanta	34.1%	25.8%	28.5%
Boston	114.9	86.2	67.7
Chicago	64.1	60.2	47.7
Cleveland	31.9	22.7	21.4
Denver	37.7	35.6	35.0
Las Vegas	118.4	124.4	117.0
Los Angeles	181.9	152.2	119.9
Miami	164.8	143.2	120.9
Minneapolis	80.6	63.5	59.1
New York	128.8	103.5	81.8
Phoenix	84.4	88.4	91.1
Portland	52.7	47.9	47.1
San Diego	188.1	149.5	120.7
San Francisco	156.5	115.2	82.5
Seattle	57.6	53.1	48.4
Tampa	122.2	98.8	91.1
Washington, D.C.	160.8	144.9	117.6

Note: Percent changes are based on the tier classifications available in the November 2008 S&P/Case-Shiller Tiered Price Indices. The three tiers separate houses in the area into three tiers (each with roughly one-third of homes sold) based on the starting price of the houses sold.

Table 3
Estimated Models of Income Differences between HMDA and AHS

Independent Variable	*Dependent Variable*					
	HMDA Income – AHS Income				HMDA Income	AHS Income
	(1)	(2)	(3)	(4)	(5)	(6)
2004 Dummy	2.8 (3.7)	2.8 (3.5)	3.3 (3.8)		-1.6 (3.6)	-4.4 (3.7)
2005 Dummy	14.4** (6.1)	14.4*** (5.4)	15.2** (6.4)		1.5 (6.2)	-12.9*** (4.0)
2006 Dummy	16.7*** (4.5)	18.3*** (5.4)	17.9*** (5.4)		2.1 (4.8)	-14.7*** (4.3)
2007 Dummy	10.0 (8.7)	8.9 (7.3)	11.5 (9.5)		2.4 (5.7)	-7.6 (7.4)
Post-2000 Dummy			3.5 (5.2)			
2004 – 1st Qtr.				1.7 (4.3)		
2004 – 2nd Qtr.				-16.4 (10.4)		
2004 – 3rd Qtr.				-6.2 (7.6)		
2004 – 4th Qtr.				17.5*** (5.1)		
2005 – 1st Qtr.				11.3 (7.0)		
2005 – 2nd Qtr.				16.6* (9.0)		
2005 – 3rd Qtr.				16.8 (13.9)		
2006 – 2nd Qtr.				25.4*** (7.0)		
2006 – 3rd Qtr.				12.1** (4.7)		
2006 – 4th Qtr.				19.0** (7.7)		
2007 – 1st Qtr.				11.9 (10.8)		
2007 – 2nd Qtr.				9.3 (9.1)		
2007 – 3rd Qtr.				1.9 (28.7)		
Trend	-0.00 (.09)	-0.01 (0.08)	-0.12 (0.22)	0.00 (0.09)	0.45*** (0.09)	0.45*** (0.09)
P-value: No mid-2000s Effect	0.000	0.001	0.002	0.000	0.859	0.001
R^2	0.03	0.03	0.03	0.03	0.14	0.06
Mean (S.D.) of Dep. Var.	3.8 (63.8)	3.6 (61.3)	3.8 (63.8)	3.8 (63.8)	83.7 (53.7)	80.0 (62.1)

Note: All models are estimated by OLS. Standard errors (in parentheses) are robust to heteroskedasticity and any arbitrary correlation in the error terms over time within an MSA. A full set of MSA dummies and quarter-of-the-year dummies are included in each specification. The dependent variable is expressed in thousands of year-2000 dollars, corrected for inflation using the personal consumption expenditure deflator. The AHS income measure is family income in all specifications. All matched HMDA incomes are obtained using a $1000 bandwidth, except column (2) where a $5000 bandwidth is used. The sample includes 10,171 observations, except for column (2) where it includes 10,310 observations.
* indicates significance at the 0.10 level, ** at the 0.05 level, *** at the 0.01 level

Table 4
Estimated Models of HMDA-AHS Income Differences, with Interactions with AHS Characteristics

Independent Variables	*Mean (S.D)*	*Specification*				
		(1)	(2)	(3)	(4)	(5)
"Suspect" Quarters Dummy	0.07	15.2*** (3.4)	-5.3 (5.0)	-5.9 (5.1)	-6.2 (5.2)	-7.6 (5.6)
Suspect*Convent-ional Loan			23.1*** (5.2)	17.2*** (5.2)	16.5*** (5.5)	16.6*** (5.6)
Suspect* Jumbo Loan				42.2*** (14.2)	8.6 (16.9)	9.2 (16.8)
Suspect*Real Loan Amount					0.09** (0.04)	0.08** (0.04)
Suspect*Mortgage Rate						-1.1 (2.0)
Conventional Loan	0.76		3.4* (2.0)	2.5 (2.0)	2.6 (2.1)	2.4 (2.1)
Jumbo Loan	0.08			13.5** (6.1)	14.2* (8.1)	13.9* (8.0)
Real Loan Amount	163 (106)				0.03 (0.08)	0.00 (0.03)
Mortgage Rate	7.0 (1.3)					1.2** (0.6)
P-value: joint test of Jumbo and Loan Amount Interactions					0.008	0.011

Note: See the notes to Table 3 A full set of MSA dummies and quarter of the year dummies, and a linear time trend, are included in each specification. The dependent variables is HMDA income minus total family income from the AHS. The sample includes 10,171 observations.

Table 5
Estimated Models of HMDA-AHS Income Differences for Conventional-Loan Mortgages, with Interactions involving HMDA Characteristics

Independent Variables	*Mean (S.D)*	*Specification*			
		(1)	(2)	(3)	(4)
"Suspect" Quarters Dummy	0.08	11.4*** (2.7)	-39.0** (15.7)	-16.5 (13.1)	-4.5 (17.0)
Suspect*HUD			75.4*** (25.6)		
Suspect*OTS			101.5** (49.8)		
Suspect*NCUA			16.5 (49.9)		
Suspect*FDIC			28.3 (58.8)		
Suspect*FRB			65.1* (37.4)		
Suspect*Nonbank				57.2** (25.5)	40.7 (27.4)
Suspect*Gov't Sold					-32.5* (18.9)
Suspect*Real Loan Amount		0.10*** (0.03)	0.09*** (0.03)	0.10*** (0.03)	0.09*** (0.03)
HUD	0.32		-45.2** (16.9)		
OTS	0.20		-44.8*** (13.1)		
NCUA	0.03		42.2 (32.0)		
FDIC	0.08		-30.0 (20.7)		
FRB	0.14		-26.1** (12.2)		
Nonbank	0.51			-33.2*** (10.9)	-31.9*** (10.6)
Gov't Sold	0.37				-30.8*** (5.5)
Real Loan Amount	176 (115)	0.03 (0.03)	0.03 (0.03)	0.03 (0.03)	0.01 (0.03)
Mortgage Rate	7.0 (1.4)	1.6 (0.6)	1.7 (0.6)	1.7 (0.6)	1.4 (0.6)

Note: See the notes to Table 3. A full set of MSA dummies, quarter of the year dummies, and a linear time trend, are included in each specification. The dependent variables is HMDA income minus total family income from the AHS. The sample consists only of conventional loan applications; it includes 7,770 observations.

Table 6
Estimated Models of HMDA-AHS Income Differences for Conventional-Loan Mortgages, with Interactions involving MSA-Level House Price Growth

Independent Variable	*Specification*			
	(1)	(2)	(3)	(4)
Suspect Dummy	15.2*** (2.7)	-1.3 (17.5)	16.0*** (2.7)	-1.1 (17.6)
Suspect Dummy Interacted With:				
House Price Growth Rate, 2004	165.9*** (38.2)	16.6 (33.0		
House Price Growth Rate, 2005			125.6*** (28.9)	33.0 (24.4)
Nonbank		37.3 (28.4)		35.7 (28.6)
Gov't Sold		-35.4* (19.5)		-33.7* (19.2)
Real Loan Amount		0.09*** (0.03)		0.09*** (0.03)

Note: See notes to Table 3. A full set of MSA dummies, quarter of the year dummies, and a linear time trend, are included in each specification. Specifications (2) and (4) also include the additional controls in specification (5) of Table 5. The dependent variable is HMDA income minus total family income from the AHS. The interactions with the house price growth and real loan amount variables use deviations from means for those variables. The sample consists only of conventional loan applications; it includes 7,732 observations.

Table 7
Comparison of AHS Movers to Non-Moving Homeowners

Dependent Variable	***Coefficient on New Homeowner Dummy***	
	Non-Suspect Quarters	Suspect Quarters
OLS Regressions		
Income	0.03 (1.21)	-13.6*** (3.5)
Age	-13.3*** (0.3)	-13.9*** (0.6)
Education	0.64*** (0.04)	0.10 (0.21)
Income (with Age and Education Controls)	-6.3*** (1.1)	-15.2*** (2.7)
Quantile Regressions for Income		
0.15 Percentile	8.6 (0.4)	6.2 (1.8)
0.30 Percentile	7.3 (0.5)	3.7 (1.3)
0.50 Percentile	4.2 (0.6)	-2.6 (2.4)
0.70 Percentile	-0.2 (0.7)	-9.6 (3.8)
0.85 Percentile	-3.2 (1.2)	-22.0 (3.6)

Note: All regressions use weights for non-moving homeowners, where the weight is the empirical probability that an observations with that interview date and SMSA would be in the mover sample for the relevant time period. The sample size for the non-suspect quarters is 184,123, and for the suspect quarters is 17,971. Standard errors in the top panel are clustered at the SMSA level. Standard errors for the quantile regressions are nonrobust, and are calculated using the formulas suggested by Rogers (1993).

Table 8
Logit Models Estimates of Delinquency-Rate Models

Independent Variable	**Mean (S.D.)**	**2009 Delinquency Rate** (1)	(2)	(3)	**2008 Delinquency Rate**
All Mortgages					
"Top Half" in Income Overstatement	0.50	0.292*** (0.099) [0.13]	0.121 (0.074) [0.05]	0.004 (0.063) [0.002]	-0.008 (0.087) [-0.004]
Delinquency Rate in 2005	0.02 (0.03)	0.281 (1.033) [0.05]	6.38*** (0.76) [0.11]	6.47*** (1.33) [0.11]	8.64*** (1.59) [0.15]
Current Unemployment Rate	*2009*: 9.9(2.6) *2008:* 7.2(1.9)		0.182*** (0.015) [1.63]	0.103*** (0.014) [0.92]	0.199*** (0.032) [1.36]
Unemployment Rate in 2005	4.7 (1.3)		-0.223*** (0.036) [-0.97]	-0.145*** (0.019) [-0.63]	-0.229*** (0.034) [-1.03]
Current House Price Growth Rate	*2009:* -0.05(0.06) *2008*: -0.05(0.09)			-4.00*** (1.07) [-0.19]	-1.75** (0.70) [-0.09]
House Price Growth Rate in 2005	0.12 (0.09)			0.17 (0.58) [0.02]	0.02 (0.65) [0.00]
Subprime Mortgages					
"Top Half" in Income Overstatement		0.224*** (0.079) [0.06]	0.126* (0.065) [0.04]	-0.021 (0.052) [-0.006]	-0.014 (0.054) [-0.005]
Delinquency Rate in 2005	0.06 (0.06)	-0.910 (0.698) [-0.03]	1.28* (0.67) [0.05]	1.92*** (0.47) [0.07]	2.59*** (0.58) [0.11]
Current Unemployment Rate			0.133*** (0.015) [0.78]	0.055*** (0.013) [0.33]	0.119*** (0.022) [0.61]
Unemployment Rate in 2005			-0.189*** (0.035) [-0.55]	-0.114*** (0.019) [-0.33]	-0.174*** (0.026) [-0.60]
Current House Price Growth Rate				-4.21*** (0.86) [-0.12]	-1.62*** (0.55) [-0.06]
House Price Growth Rate in 2005				0.51 (0.55) [0.04]	0.69 (0.52) [0.06]
Sample Size		113	111	110	110

Notes. See notes to Table 8. The house-price growth rate measures are the percentage change in the average annual OFFHEO index from the previous year to the next. The numbers in brackets are estimated elasticities.

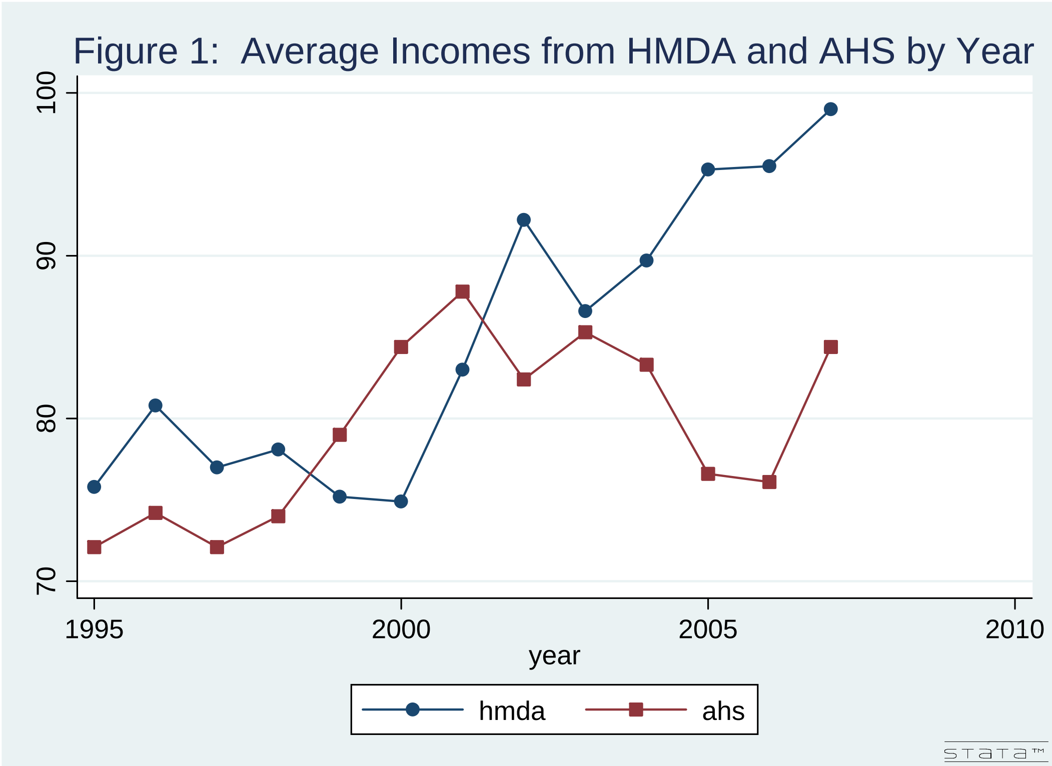

Figure 1: Average Incomes from HMDA and AHS by Year
100
90
80
70
1995
2000
2005
2010
year
hmda
ahs