Missing Race Data in HMDA and the Implications for the Monitoring of Fair Lending Compliance

Jason Dietrich

Economic and Policy Analysis Working Paper 2001-1

March 2001

Missing Race Data in HMDA and the Implications for the Monitoring of Fair Lending Compliance

Jason Dietrich

Office of the Comptroller of the Currency
Economic and Policy Analysis Working Paper

March 2001

Abstract: Home Mortgage Disclosure Act (HMDA) data contains high and increasing percentages of applications that lack race data. As HMDA data are an integral part of current efforts to monitor banks' compliance with fair lending laws, regulators must understand the reasons for, and consequences of, these patterns. Using HMDA data from 1993 to 1999, this study examines trends in missing race data, discusses possible reasons for the findings, and summarizes the salient regulatory issues. The results indicate that race data are missing for systematic reasons and therefore introduce bias and efficiency problems into fair lending exams. Applications that contain race data have higher origination rates than applications without race data, and applications from Blacks and Hispanics may be more likely to be without race data than whites. These findings suggest that denial rate disparities used during early stages of fair lending exams may be understated and that statistically-modeled estimates of racial effects used during latter stages may be overstated.

The views expressed in this paper are those of the author alone and do not necessarily reflect those of the Office of the Comptroller of the Currency or the Department of the Treasury. The author would like to thank Irene Fang, Jeff Brown, Gary Whalen, and Amber Jessup for their insightful comments and editorial assistance.

Please address correspondence to Jason Dietrich, economist, Financial Access and Compliance, Office of the Comptroller of the Currency, 250 E Street, S.W., Washington, DC 20219 (phone: 202-874-5119; e-mail: jason.dietrich@occ.treas.gov)

Introduction

The Home Mortgage Disclosure Act (HMDA) requires financial institutions to collect and report data on applicants for home mortgage product loans. These data allow regulators to monitor a bank's compliance with OCC fair lending laws.[1] Regulators' ability to rely on HMDA data as a monitoring tool may be diminishing, however, as race is not reported for a significant and growing portion of applicants. For example, in 1999, the number of observations in HMDA that lack race data ranged from 19.1 percent for conventional home purchase loans to 48.9 percent for Federal Housing Administration-insured (FHA) home improvement loans.[2] These percentages are up from 11.4 and 34.4 percent in 1993, respectively. In general, all loan products reported in HMDA show relatively large percentages of missing race data, with an upward trend over the period 1993 to 1999.

Most empirical analyses using HMDA data, including current fair lending examinations, merely eliminate from the population applications that lack race data. This approach assumes race data are missing for random reasons and therefore introduce no sample selection problems into the analysis. Huck (2000) presents evidence indicating that this is not necessarily a safe assumption. Using 1997 HMDA data from 10 metropolitan statistical areas (MSA), the author shows that approval rates for applications containing race data consistently overstate true approval rates, especially for refinance

[1] The term "bank" is used generally to refer to various types of regulated financial institutions.
[2] Race is defined as missing if the variable for primary applicant race equals 6 (Other), 7 (Indirect application), 8 (NA) or . (missing). The denominator used to calculate the percentages includes all owner-occupied applications for the specific product.

and home improvement loans. This suggests that race is missing for non-random reasons and that statistical estimators will be biased and inefficient if sample selection is not addressed.

This study takes a more comprehensive approach than Huck, focusing on trends in missing race data at the national level and using HMDA data from 1993 to 1999. Specific attention is placed on potential explanations for these missing race data and their effects on the statistical tools regulators use during fair lending exams. Tests of the hypothesis of equality of variable means for samples that lack and do not lack race data are used to determine whether race data are missing for random or systematic reasons. Weighted logit models using county-level data from the Bureau of Labor Statistics (BLS), Bureau of Economic Analysis (BEA), and Census are then used to indicate the racial composition of applications that lack race data. The identification of racial groups that are more likely to withhold race information will have implications for regulators' interpretation of denial rate disparities, a key indicator used during the early stages of fair lending exams. It will also affect the modeling portion of fair lending exams, since sample selection problems introduced by different approval rates between applications that lack or do not lack race data are compounded, if missing data are also correlated with race.

The remainder of the paper is structured as follows. Section II provides a brief history of fair lending legislation and the role of regulators in assuring compliance. Section III summarizes empirical trends in missing race data in HMDA between 1993 and 1999, by loan type and purpose. Sections IV and V outline potential explanations for missing race data and discuss the salient issues facing regulatory agencies. Section VI

provides a county-level multivariate model to characterize further the pool of applicants that lack race data and section VII concludes the discussion.

Background

Congress passed HMDA in 1975 requiring lenders to collect and make publicly available, data on the number and dollar value of originated home mortgage and home improvement loans on a census-tract-level within MSAs. HMDA provided a means for banking regulators to monitor compliance with the Equal Credit Opportunity Act, the Fair Housing Act, and the Community Reinvestment Act (CRA) legislation passed in the 1960's and 1970's. Redlining (lending discrimination at the community-level) was the main concern at that time and aggregate data was well suited to detect such behavior. Over time, however, discrimination at the individual-level gained attention, providing the impetus for significant modifications to HMDA under the Financial Institutions Reform, Recovery and Enforcement Act (FIRREA) of 1989. FIRREA required banks to collect data at the individual-level for home mortgage product applications, whether approved or denied, within an MSA. In addition to relevant application information, banks were now also required to report data on race, gender, and income.[3] These changes greatly enhanced regulators' efforts at monitoring compliance with fair lending laws, but also increased the importance of data integrity issues, such as missing data.

[3] Additional modifications to HMDA took place in 1988 and 1996. In 1988, HMDA was expanded to cover nondepository institutions, such as savings and loan service corporations and the mortgage banking subsidiaries of bank and thrift holding companies. In 1996, under the interagency Community Reinvestment Act, financial institutions were to begin collecting and reporting data for loan applications in non-MSAs. These changes, as well as others included in the 1988, 1989, and 1996 legislation are summarized in Canner and Smith (1991, 1992) and Avery et. al. (1997). Also see Fishbein (1992) and the 1996 and 1998 Federal Financial Institutions Examination Council (FFIEC) Guides to HMDA Reporting.

HMDA does not require all banks to collect and report data for all loan applications. Because of the costs HMDA imposes on banks, there are size, location, volume, and loan characteristic considerations that exempt certain institutions. For banks, credit unions and savings associations, the institution must have at least $29 million in assets; possess a branch office in an MSA; originate at least one home purchase or refinance loan secured by a first lien on a one-to-four family dwelling; and be federally insured or regulated. For other for-profit institutions, home purchase loan originations must equal or exceed 10 percent of total loan originations; the institution must have a branch office in an MSA or receive at least five applications for loans for property located in an MSA; it must have assets exceeding $10 million; and it must originate at least 100 home purchase loans. These asset and loan cutoffs are based on 1998 standards and are subject to change as the Federal Reserve Board makes annual updates to reflect changes in the Consumer Price Index for Urban Wage Earners and Clerical Workers.

Recent Trends

Table 1 shows the percent of total owner-occupied loan applications in HMDA that contain unreported race data for the primary applicant, by loan purpose and type, from 1993 to 1999. Although HMDA requires collection of race, gender, age, and marital status data, this study focuses only on race since regulators tend to focus more of their resources on explaining racial disparities.[4] There are three items of note in Table 1. First and foremost, considerable percentages of applications reported in HMDA do not

[4] Regulation B prohibits discrimination in any aspect of the credit transaction on the basis of race, color, religion, national origin, gender, marital status, or age.

Table 1: Percent of Total Owner-Occupied Loan Applications in HMDA That Lack Race Data for the Primary Applicant. (race reported as 6, 7, 8 or .)							
	1993	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
Conventional							
Home Purchase	11.4	10.3	10.2	12.3	14.0	17.8	19.1
Home Improvement	16.9	16.8	17.5	25.3	26.9	34.8	31.9
Refinance	12.0	16.5	22.4	27.3	35.4	32.6	38.8
FHA							
Home Purchase	21.7	18.3	19.8	18.8	24.7	25.9	23.2
Home Improvement	34.4	30.1	42.8	57.3	61.4	67.2	48.9
Refinance	25.7	26.8	23.0	27.5	28.2	27.6	29.4
VA							
Home Purchase	22.5	20.3	22.3	24.6	28.0	27.8	27.2
Home Improvement	3.5	88.8	42.1	33.9	44.1	25.1	22.9
Refinance	24.5	28.2	30.6	31.8	36.3	35.3	35.2

contain race data. With the exception of Veterans Administration-insured (VA) home improvement loans in 1993, race data are missing for more than 10 percent of applications for each product/year pairing. Surprisingly, VA home improvement loans in the following year, 1994, have the highest percentage of missing values at 88.8 percent. Second, generally an upward trend exists in the percentages of applications that lack race data. For each product/year pairing, the percentage is higher in 1999 than in 1993. For all conventional loans, this upward trend is quite consistent and pronounced, especially between 1994 and 1999. Other than FHA home improvement loans and VA refinance loans, the percentages for FHA and VA applications show less-pronounced upward trends over the seven-year period. Finally, the percentages of applications that lack race data generally are higher, and more variable, for FHA and VA loans than for conventional loans. Conventional home purchase loans have the lowest percentages of missing data, between 10.2 and 19.1 percent, while FHA home improvement loans have the highest, between 30.1 and 67.2 percent.

Explanations for missing race data

Given these patterns in missing data, and the fact that regulators rely on HMDA data for monitoring purposes, we must understand why those data are missing and why their numbers appear to be increasing over time. There are six main reasons why race data are not reported in a bank's HMDA Loan Application Register (HMDA-LAR).

Customer does not provide information during direct applications

Disclosure of race information by the customer is voluntary, but is required for the bank during direct applications. When a customer declines to provide this information, the loan officer is supposed to record this data based on his or her self-assessment of the customer or the customer's surname (HMDA, Section 203.4 (b1)). If the loan officer chooses not to do this, race is merely coded as not applicable (NA).

Customer does not provide information during indirect applications

There are three possible indirect conduits customers may use when applying for credit: 1) mail, 2) telephone, or 3) Internet. For applications taken entirely by mail, a form is included in the application packet asking for race information. If the applicant declines to volunteer this information, banks are not required to ascertain race based on visual inspection or surname. For telephone applications, banks are not required to ask for race information. Applications taken via the Internet are not explicitly mentioned in the 1998 FFIEC Guide to HMDA reporting. However, it seems safe to assume they would be treated similarly to mail or telephone applications. For all of these indirect conduit applications, banks may mark in the HMDA-LAR race field, "Information not provided by applicant in mail or telephone application," if the customer does not

Table 2: Percent of Total Owner-Occupied Loan Applications in HMDA Taken By Indirect Conduit. (race reported as 7)							
	1993	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
Conventional							
Home Purchase	3.7	3.2	3.0	4.1	5.9	8.9	9.9
Home Improvement	14.3	14.8	15.8	22.9	24.7	32.3	29.8
Refinance	6.0	10.1	16.6	19.3	26.4	23.7	30.2
FHA							
Home Purchase	2.3	3.1	1.6	2.4	6.2	5.8	7.2
Home Improvement	26.6	24.0	38.3	52.2	56.7	52.5	45.9
Refinance	8.3	14.2	10.6	14.6	15.2	11.9	16.2
VA							
Home Purchase	2.9	3.8	3.2	6.9	9.4	8.7	10.9
Home Improvement	2.3	1.6	31.2	24.2	41.6	20.2	12.7
Refinance	7.6	12.3	14.7	15.2	13.6	12.7	15.4

volunteer such information. The introduction of Internet lending in late 1996 may have affected significantly the level of missing race data. The results in Table 2, which show the percentage of total owner-occupied loan applications missing race data because they were taken indirectly, bear this out. Except for FHA and VA home improvement and refinance loans, considerable increases occur in missing race data between 1996 and 1999, because of the use of indirect application conduits. Although acceptance of Internet banking by both the financial industry and consumers has been gradual, more than half of the national banks currently have a web site and many of the larger institutions accept mortgage applications on line.[5] If Internet banking gains popularity, unreported race data most likely will continue to rise.[6]

[5] From mid-1996 to mid-1997, the number of total mortgage-related web sites rose from 60 to more than 3000. Restricting this to lenders only, of the 2,000 lenders with web sites, almost 80 percent provide either pre-application or pre-qualification services [Negroni (1997)]. Restricting this further to only national banks, Furst et.al. (2000) report that 54.2 percent of national banks had web sites as of the 3rd quarter of 1999.

[6] As one example, First Union has pursued Internet lending aggressively and has even stated that it will use the Internet rather than acquisitions and mergers to expand its business (Seidman, 1999). This philosophy is reflected in HMDA where nearly 90 percent of loan applications for First Union were missing race data in 1999.

Loan was purchased

Banks are not required to report race data for loans they purchase and merely mark NA in the HMDA-LAR race field. If a purchased loan were originated by an institution covered by HMDA, the originating institution would be required to report race data, and the loan would be examined as part of the institution's examination. To require the purchaser also to report race would be of little use to regulators for fair lending purposes, since the purchaser did not make the credit decision.[7] Alternatively, if an institution not covered by HMDA originated the loan, the race data for the application would never appear in HMDA. The extent to which this occurs is not known.

Loan was brokered

Currently, approximately half of all mortgages are originated through brokers (Barefoot, 1998). Depending on the stage of data collection that brokers reach before passing the application on to the financial institution, race may not be included in a portion of the applications that banks receive. Given that HMDA does not apply to brokers, brokered loans provide a potential source of missing race data. Banks would mark NA in the HMDA-LAR race field for brokered loans that do not include race information.

The fair lending regulations are vague on data collection issues for brokered applications. HMDA clearly states that the institution making the credit decision must report the loan application data for brokered applications. However, HMDA is not specific about whether the institution needs to track down race information when the broker does not provide it in the application material.

[7] Information for purchased loans is of interest to regulators for CRA reasons.

Data errors

The large volume of data required by HMDA will always invite collection and data entry errors to make their way into the final HMDA-LAR. In most instances, those errors should be few and distributed randomly across races. If not, the bank has Management Information System and data integrity issues, for which examiners routinely check in their standard examination procedures. As one measure of the degree of data error in HMDA, Canner and Smith (1991) report that 4 percent of the loan applications in the 1990 HMDA-LAR records contained errors.[8] Since nearly 6.4 million loan and application records were processed in 1990, approximately 255,000 applications contained errors. Clearly, all of these were not errors in recording race.

Fraud

Some race data may be intentionally unreported or altered to hide discriminatory behavior. Although there is no way to measure the extent that this occurs, it is likely to be rare. Again, examiners would check this as part of their standard examination procedures.

Regulatory issues

Missing race data in HMDA create a number of regulatory issues, since HMDA is an integral part of the regulatory fair lending program. Both the large percentage of missing data and their increasing trend may affect the statistical tools regulators use, as well as the exam procedures they follow. This section focuses mainly on potential

[8] See footnote 6 in Canner and Smith, 1991. The percentage of loan records containing detected errors was 4.4 percent in 1991 and less than 0.5 percent in 1993 [Canner and Passmore (1995)].

sample selection problems, leaving a discussion of other regulatory concerns to Dietrich (1999).[9]

The goal of regulators monitoring fair lending compliance is to ensure that banks are not using prohibited factor information in any aspect of the loan process. The loan process consists of many stages, including referrals, counseling, underwriting, and determination of terms and conditions. Although discrimination can occur in each of these stages, this study restricts its attention for expositional reasons to only the underwriting decision-making process. If race data are missing in HMDA, one could argue that underwriters do not have access to this information either. One could then argue further that applications that lack race data can safely be excluded from examination, since race has already been ruled out as a determining factor of the credit decision. If this characterization is accurate, one must begin to question the efficacy of HMDA as a fair lending monitoring tool. HMDA clearly imposes a cost on banks in data collection and record-keeping efforts. If the recent trends in missing race data persist, banks will continue to incur these costs for all applications, while an ever-diminishing number of applications will actually be used for compliance purposes.[10]

The characterization of missing race data posing no problems for regulators may be inappropriate for both examination and statistical reasons. From an examination perspective, if an applicant's race is unreported in HMDA, but banks have some idea of his or her race, this creates a channel for hiding discriminatory behavior. Banks are

[9] Dietrich (1999) discusses additional regulatory concerns introduced by missing race data, such as adequacy of sample sizes for multivariate analysis, monitoring differential Internet access and price discrimination, selective reporting of race for purchased loans, and the role that brokers play.

[10] HMDA serves more of a role than only to assist regulatory agencies in enforcing compliance with anti-discrimination statutes at the individual level. It is also used to show whether banks are meeting the credit needs of their communities and to determine levels of public and private sector investments. Missing race data at the individual level has less of an effect on both of these roles, since each deals with more aggregate levels of data and aggregate race, gender, and income data are readily available from alternative sources.

likely to have some idea of each applicant's race from surnames, addresses, face-to-face contact, or knowledge of service areas, even when race information is unreported in HMDA. This is especially true for American Indian, Asian, and Hispanic applicants, as well as for smaller, racially-homogenous service areas.[11] There is clearly room for much error in using surnames and other surrogates to ascertain race. However, the FFIEC agencies believe that there is enough merit to this process that banks must use the applicant's surname to help determine race whenever the applicant refuses to provide that information voluntarily during direct applications. If banks can use surnames to help determine race to the satisfaction of regulators for direct applications, they could ascertain the race of all applicants within some margin of error. Therefore, during examinations, regulators should still be concerned about the treatment of applicants with regard to prohibited factors, even if race data are unreported in HMDA.

From a purely statistical perspective, missing race data may also introduce sample selection problems into calculations of denial rate disparities and multivariate estimators. Sample selection issues become relevant when, for some reason, only a portion of the population data is usable. For this study, the sample used for estimations is drawn from only the portion of population observations that contain race data. With a sample selection model, two potential scenarios can occur: 1) the usable subset of the population is distributed randomly across the entire population distribution and unrelated to the outcome of interest, or 2) the usable subset of the population is determined systematically. If the data are missing for random reasons, sample selection introduces no bias into the standard estimators regulators use, and the issue becomes one solely of

[11] This occurred in the Department of Justice's (DOJ) suit against the First National Bank of Gordon in 1993. Even though race data was not included in any of the applications, the Office of the Comptroller of the Currency (OCC) and DOJ determined that the bank unfairly charged American Indians higher rates than similarly-situated white applicants.

statistical precision.[12] A simple strategy for increasing the precision of the estimates is to make observations with missing data usable. This would incorporate any information about the co-variation between explanatory variables with complete data and the dependent variable, which is not used if observations that lack race data are omitted. Replacing missing values with means, and replacing missing values with zeros and including an indicator variable denoting those observations that contain missing values, are two techniques for increasing the number of usable observations. The first approach is unsuitable for fair lending models, since race is an 0/1 indicator variable and replacing missing values of indicator variables with means makes little sense. The second approach does not suffer this drawback, and in addition to gaining information from lost observations, provides an estimate of the effect of missing race.

In the second sample selection scenario, data are missing for systematic reasons. For example, minorities may be less likely to volunteer race information for fear that it will be used against them in the underwriting process. With systematically missing race data, the logit estimator commonly used for fair lending exams will be biased and inefficient if approval rates differ for applications with and without race data. This result is basically an omitted variable problem. If race data are missing for systematic reasons, one can formulate a separate sample selection equation, which models the determinants of an application having non-missing race data. If the errors of the selection equation are correlated with the errors of the underwriting equation, the expected value of the underwriting equation error term conditional on an application having non-missing race data will not equal zero. The new error term will be a composite of an Independently and

[12] All of the bias and precision issues introduced by truncated or censored population distributions also affect judgmental file reviews.

Identically Distributed (IID) error term and an unobserved sample selection effect. This sample selection term can be viewed as an omitted variable if the underwriting equation is estimated using only applications with non-missing race data. With the logit estimator, the direction of bias in the racial estimate depends on the correlation between missing data and the action taken on the loan, and on the correlation of missing data and race conditional on the action taken on the loan.[13]

One indication of whether data are missing for random or systematic reasons is a test of the hypothesis that the mean of a variable in the sample with non-missing race data is equal to the mean of the variable in the sample with missing race data. Tables 3-5 present results for such hypothesis tests for conventional mortgage products from 1993 to 1999, using income, loan amount, and the action taken on the application. Purchased loans are excluded from this, and subsequent analysis since the reporting institution does not make the origination decision on these loans, and action is one of the variables of interest in this study. Except for denials of conventional home purchase applications in 1996, the null hypothesis of equality of means across samples is rejected at the 95 percent significance level for all variables and years. This suggests that race data are missing for systematic reasons, and that sample selection may be an issue for fair lending analyses.

High levels of precision gained from large sample sizes clearly drive these statistical results. Therefore, it is also important to look at the economic meaning of these findings. For home purchase loans, applications that lack race data have higher average incomes and loan amounts, and lower origination rates for all seven years. Excluding

[13] This is different than the typical omitted variable finding for linear models that the direction of bias in the racial estimate depends on the correlation between missing race data and the action taken on the loan, and on the correlation of missing race data and race unconditional on the action taken on the loan (Lee [1992], Maddala [1983,1992]).

Table 3: T-tests of Differences of Means for Samples With and Without Race Data (Conventional Home Purchase)							
	1993	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
N_{race} / $N_{no\ race}$	2543970/ 147810	3527399/ 157975	3741748/ 171838	4384750/ 243630	4719572/ 353245	5539553/ 640863	5717066/ 749431
Income							
Mean with race	62.75	58.15	55.95	56.36	58.34	58.98	62.49
Mean without race	87.55	65.62	67.31	70.16	71.05	66.28	71.59
t-statistic	19.86	13.04	38.17	45.32	41.02	49.20	51.91
Loan Amount							
Mean with race	100.22	99.13	91.34	91.62	93.28	98.56	106.54
Mean without race	129.06	128.99	111.38	111.57	108.02	102.46	118.75
t-statistic	12.84	14.03	48.62	74.92	70.30	22.66	48.85
Originated							
Mean with race	0.69	0.67	0.61	0.57	0.54	0.54	0.53
Mean without race	0.53	0.50	0.50	0.47	0.45	0.40	0.44
t-statistic	-125.84	-135.55	-93.53	-100.64	-101.89	-211.65	-156.85
Approved but NA							
Mean with race	0.05	0.06	0.08	0.09	0.10	0.09	0.10
Mean without race	0.05	0.07	0.08	0.09	0.11	0.11	0.12
t-statistic	-4.89	10.28	-2.92	-1.98	22.87	34.78	48.19
Denied							
Mean with race	0.18	0.18	0.23	0.27	0.30	0.30	0.29
Mean without race	0.24	0.24	0.24	0.27	0.28	0.32	0.28
t-statistic	57.84	54.49	13.51	-0.22	-28.33	22.80	-14.27
Withdrawn							
Mean with race	0.07	0.07	0.07	0.06	0.05	0.06	0.06
Mean without race	0.15	0.15	0.16	0.15	0.13	0.15	0.13
t-statistic	88.44	84.98	101.02	119.55	126.54	198.51	173.48

1993, applications that lack race data have approximately $10,000 more income than applications with race data, with no strong time trends for either sample, or the difference between samples. Loan amounts and origination rates, on the other hand, both show trends over the seven-year period. Average loan amounts for applications that lack race data decline steadily, while loan amounts for applications that contain race data are flat. This produces a declining trend in the difference between samples starting from $29,000 in 1993 and falling to $4,000 in 1998, with a slight increase to $12,000 in 1999. For origination rates, averages for both samples decline at similar rates over the seven-year

Table 4: T-tests of Differences of Means for Samples With and Without Race Data (Conventional Home Improvement)							
	1993	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
N_{race} / $N_{no\ race}$	906030/ 178201	1127326/ 216231	1216967/ 248438	1373508/ 450119	1321051/ 470245	1196013/ 612920	1329904/ 603586
Income							
Mean with race	47.76	48.85	49.50	51.02	52.73	54.92	59.61
Mean without race	50.13	50.95	50.99	52.75	50.66	52.81	57.92
t-statistic	8.43	9.89	8.86	16.31	-21.95	-22.10	-16.30
Loan Amount							
Mean with race	21.49	18.33	15.05	16.59	17.98	19.17	20.18
Mean without race	36.06	29.96	16.13	19.92	20.24	23.46	28.39
t-statistic	12.17	8.65	14.05	39.14	18.64	25.77	76.51
Originated							
Mean with race	0.71	0.70	0.65	0.61	0.59	0.61	0.59
Mean without race	0.41	0.40	0.36	0.32	0.30	0.31	0.31
t-statistic	-235.18	-270.82	-268.32	-361.15	-360.03	-403.18	-388.48
Approved but NA							
Mean with race	0.04	0.05	0.07	0.08	0.10	0.09	0.11
Mean without race	0.07	0.08	0.11	0.09	0.15	0.15	0.12
t-statistic	42.14	59.29	49.14	20.69	91.47	105.16	25.06
Denied							
Mean with race	0.21	0.21	0.24	0.27	0.27	0.27	0.27
Mean without race	0.43	0.42	0.42	0.44	0.46	0.46	0.40
t-statistic	174.30	183.52	171.14	201.41	224.55	252.59	175.76
Withdrawn							
Mean with race	0.03	0.03	0.03	0.04	0.04	0.03	0.03
Mean without race	0.07	0.09	0.09	0.14	0.08	0.08	0.16
t-statistic	59.09	82.44	98.17	186.97	94.99	121.68	266.68

period, such that the difference between samples stays between 9 and 17 percent. Although applications that lack race data have consistently lower origination rates, the denial rates for the two samples are quite similar, especially after 1994. Almost all of the lower origination rates for applications that lack race data are made up by higher withdrawal rates, possibly suggesting that applicants are shopping around for better terms and conditions, and that missing race data are not necessarily correlated with the underwriting decision.

Table 5: T-tests of Differences of Means for Samples With and Without Race Data (Conventional Refinance)							
	1993	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
N_{race} / $N_{no\ race}$	4812663/ 429192	2513438/ 361796	1777884/ 413639	2939586/ 849363	3271422/ 1442092	7251665/ 2790895	5684512/ 2909438
Income							
Mean with race	77.83	68.53	65.81	65.20	66.11	71.39	67.33
Mean without race	79.66	61.53	54.07	53.93	55.54	58.55	56.76
t-statistic	7.43	-23.13	-69.91	-107.34	-67.53	-205.73	-144.00
Loan Amount							
Mean with race	116.47	107.63	103.50	98.29	101.90	114.49	107.54
Mean without race	123.72	94.92	74.65	72.37	78.91	80.86	77.04
t-statistic	10.78	-26.50	-164.94	-184.32	-48.65	-299.79	-284.92
Originated							
Mean with race	0.82	0.72	0.69	0.66	0.62	0.69	0.56
Mean without race	0.56	0.32	0.27	0.27	0.24	0.28	0.24
t-statistic	-330.96	-477.43	-547.19	-707.31	-840.43	-1258.79	-963.89
Approved but NA							
Mean with race	0.02	0.04	0.05	0.06	0.08	0.06	0.09
Mean without race	0.06	0.11	0.13	0.15	0.16	0.16	0.13
t-statistic	102.32	127.38	158.57	219.34	251.98	420.99	209.89
Denied							
Mean with race	0.10	0.13	0.15	0.17	0.19	0.14	0.19
Mean without race	0.21	0.32	0.39	0.35	0.38	0.33	0.34
t-statistic	170.37	233.18	293.43	323.60	420.03	629.43	461.43
Withdrawn							
Mean with race	0.06	0.09	0.09	0.09	0.09	0.09	0.12
Mean without race	0.14	0.19	0.19	0.19	0.17	0.19	0.24
t-statistic	154.68	156.17	149.52	220.20	233.18	392.18	440.56

For home improvement loans, applications that lack race data have higher average loan amounts and lower average origination rates, similar to home purchase loans. Unlike home purchase loans, however, there appears to be little difference in average incomes. Average incomes for both samples trend upward over time, but the differences between samples are consistently near zero with the largest difference being $2,300 in 1993. Average loan amounts for each sample initially decline in 1993 and 1994, but increase over the remaining five years. Differences between samples follow a similar pattern, ranging from $15,000 in 1993 to $1,000 in 1995 and up to $8,000 again in 1999.

For origination rates, both samples trend consistently downward resulting in a fairly constant gap of 29 percent, which is considerably higher than that for home purchase loans. Finally, unlike home purchase loans, applications that lack race data are denied at a much higher rate than applications with race data, ranging from 22 percent in 1993 to 13 percent in 1999.

For refinance loans, average incomes, loan amounts, and origination rates are all lower for applications that lack race data. Excluding 1993 and 1994, applications that lack race data have approximately $11,000 less income than applications that contain race data, almost the exact opposite of the finding for home purchase loans. Neither sample, nor the difference between samples, show any trend over time. Average loan amounts also show no trends, as both samples are fairly flat with differences ranging between $26,000 and $34,000 after 1994. Average origination rates follow similar patterns to home purchase and home improvement loans with rates declining steadily over the seven-year period. The difference between samples is consistently near 40 percent, which is the highest among the three products. Following the results for home improvement loans, denial rates are much higher for applications without race data as well, ranging from 11 percent in 1993 to 24 percent in 1995.

In summary, missing race data are correlated with income, loan amount, and action taken on the loan. Most important for this study is the correlation with action taken on the loan, since this provides direct evidence that sample selection related to missing race data will affect the statistical tools regulators use during fair lending analyses. The fact that denial rates differ for home improvement and refinance loans, but not for home purchase loans, supports Huck's findings that missing race data may be more problematic for those two products.

Multivariate Analysis

This section presents a county-level multivariate analysis to characterize the pool of applications missing race data in HMDA. The main objective is to develop an indication of the actual racial composition of these applicants, which gets at the second contributor to the sample selection bias calculation discussed previously. Achievement of this objective requires creating a link between the applicants that lack race data and some measure with known racial characteristics. A natural proxy for this measure is the racial distribution in the census tracts where applicants reside. This provides a fairly small and economically homogeneous group that can be linked to applicants that lack race data. Unfortunately, census tract data are available only from the decennial census, which was last conducted in 1990 and are fairly dated. Therefore, one higher level of aggregation, county-level data, is used. The use of county-level data removes many of the data availability limitations of census tract data, increasing greatly the range of variables and time periods that can be included in the model. The tradeoff is that much homogeneity is lost within observations, since an aggregate observation is based on an underlying group of heterogeneous units. Further, as the level of aggregation increases, the multivariate analysis becomes more of a search for significant patterns of characteristics, and less of an analysis of causal relationships.

The dependent variable for all of the estimations is the percentage of applications that show missing race data by county. Purchased loans are again excluded from the analysis. Only conventional home purchase, home improvement, and refinance loans are examined, and missing is defined as other, indirect application, NA, or missing. The right-hand-side variables include the following countywide averages, countywide levels, and other miscellaneous measures:

Countywide averages:	age, gender composition, racial composition, education level, loan amount of actual applicants, and income of actual applicants less per capita income;
Countywide levels:	per capita income, unemployment rate, and bank concentration per square mile;
Other variables:	urban/rural status, region, and year.

Annual data are available at the county-level for each of these variables from 1993 to 1998. In addition to year-specific models, pooled models with indicator variables for year to control for unobservable time-variant effects are also presented. To account for the effects of aggregating data over counties of various sizes, a weighted OLS estimator is used with total county population as the weight. Appendix A presents data sources and summary statistics for all of the variables used in the estimations.

Table 6 presents the weighted OLS estimation results for race by year and for the pooled sample.[14] The main result in the table is that counties with higher percentages of Blacks and Hispanics have significantly higher percentages of applications that lack race data, especially for home purchase and refinance loans.[15] There is additional evidence that counties with higher percentages of American Indians also have significantly higher percentages of home purchase applications that lack race data. Those results, as well as those for Asians must be viewed cautiously, however, since most counties have small percentages of each of these racial groups. Combining these results with the tests of differences of means showing applications with race data have higher origination rates suggests denial rate disparities using only applications with non-missing race data may

[14] A full set of estimation results is available upon request.
[15] The estimation assumes the racial composition of applicants who applied for credit generally reflects the racial composition of the entire county.

Table 6: Weighted OLS Estimation Results (standard error estimates in parentheses)
Dependent Variable = percent of non-purchased applications missing race data by county.

	1993	**1994**	**1995**	**1996**	**1997**	**1998**	**Pooled**
N	739	827	829	832	835	835	4897
Conventional Home Purchase							
R-square	.49	.57	.63	.67	.60	.62	.73
American Indian	0.014 (0.075)	0.163** (0.068)	0.142** (0.063)	0.139** (0.066)	0.127 (0.065)	0.400 (0.209)	0.180** (0.046)
Asian	0.044 (0.048)	0.087 (0.055)	0.040 (0.046)	0.001 (0.047)	0.042 (0.050)	0.074 (0.054)	0.055** (0.022)
Black	0.027** (0.012)	0.026** (0.010)	0.024** (0.009)	0.027** (0.009)	0.029** (0.012)	0.058** (0.020)	0.033** (0.007)
Hispanic	0.091** (0.015)	0.060** (0.016)	0.049** (0.016)	0.065** (0.013)	0.034** (0.016)	0.048** (0.017)	0.057** (0.007)
Conventional Home Improvement							
R-square	.54	.61	.52	.60	.51	.42	.61
American Indian	-0.007 (0.330)	0.365 (0.234)	0.213 (0.244)	-0.112 (0.187)	0.294 (0.171)	0.117 (0.200)	0.186 (0.095)
Asian	0.150 (0.181)	0.065 (0.092)	0.057 (0.104)	-0.282 (0.171)	-0.281** (0.129)	-0.153 (0.126)	0.005 (0.046)
Black	0.161** (0.059)	0.067 (0.042)	0.078 (0.047)	0.083 (0.051)	0.155** (0.050)	0.230** (0.054)	0.129** (0.024)
Hispanic	0.162** (0.080)	0.115 (0.059)	0.098 (0.057)	0.072 (0.062)	0.193** (0.051)	0.237** (0.060)	0.155** (0.028)
Conventional Refinance							
R-square	.57	.59	.60	.68	.62	.56	.68
American Indian	0.123 (0.124)	0.076 (0.149)	0.117 (0.233)	0.232 (0.149)	0.541** (0.263)	0.071 (0.240)	0.147 (0.094)
Asian	0.084 (0.055)	0.022 (0.052)	0.095 (0.119)	0.016 (0.128)	0.043 (0.180)	0.439 (0.238)	0.132** (0.057)
Black	0.018 (0.021)	0.142** (0.029)	0.214** (0.052)	0.218** (0.041)	0.200** (0.060)	0.168** (0.060)	0.176** (0.021)
Hispanic	0.137** (0.028)	0.132** (0.027)	0.232** (0.049)	0.178** (0.040)	0.133** (0.057)	0.088 (0.067)	0.162** (0.026)

** Indicates two-tailed significance at the 5 percent level

understate the true disparities. This is in contrast to Huck's finding that applications that lack race data were fairly evenly distributed across racial groups. The results further suggest that the sample selection problems specific to the racial estimates, which stem from different approval rates across applications with and without missing race data may indeed be compounded by correlation between race and whether an application is missing

data. A positive correlation between missing data and the probability of being denied, together with a positive correlation between minority status and missing data suggest racial estimates used in fair lending exams may be biased upward. Finally, if you accept the argument that banks can determine the race of all applicants within some margin of error, the higher percentages of minorities that lack race data in HMDA increase banks' opportunities to discriminate, since regulators never examine these applications.

Clearly, the results of the multivariate models should be viewed with some caution, since they use county-level data and are therefore more of a search for correlation than for causal relationships. In addition, since the models do not control for creditworthiness, the correlation between race and missing data may simply be capturing the fact that less creditworthy applicants may more likely lack race data. However, the results do raise flags for regulators and suggest that further examination of the issues may be warranted.

Conclusion

This study examines recent trends in missing race data in HMDA, summarizes potential reasons for those missing data, and discusses the salient regulatory issues. Three main results are found. First, HMDA contains a surprisingly high percentage of applications that lack race data, and these percentages have trended upward from 1993 to 1999. Second, the percentage of missing race data due to indirect application conduits has been increasing steadily over the last four years. As consumers become more comfortable with banking technologies, missing race data may continue to increase as more applications are taken on-line. This raises interesting regulatory questions as regulators attempt to keep pace with technological changes. Third, race data appear to be

missing for non-random reasons. This point is supported by three findings. First, the mere fact that the percentages of applications that lack race data are large and have trended upward over time suggests that there are patterns to missing data. Second, test results that reject the equality of variable means for samples that lack and do not lack race data provide statistical support to these patterns. Third, multivariate model estimates showing applications that lack race data are not distributed evenly across racial groups indicates racial patterns to missing data. Taken together, these findings suggest missing race data that, if not accounted for, will affect denial rate disparities used during initial stages of fair lending exams and analyses of the determinants of these disparities used in the later stages.

Given these concerns, what can be done? Above all, regulators must recognize and address the problem. Econometrically, regulators can use sample selection estimators, surrogate data as proxies for race, or indicator variables to estimate the effects of missing race. Although use of sample selection estimators may be limited, because of data constraints, surrogates and indicator variables are easy, low cost methods of estimating a general effect of missing race data. In addition, simple hypothesis tests similar to those previously presented should be conducted for every fair lending exam to assess further the existence of potential problems due to missing race data.

Two other suggestions are worth mentioning, even though they are not likely to be feasible options. First, regulators could modify HMDA. Lowering the tolerance for missing race data and forcing banks to track down such information would impose high costs to banks. Alternatively, increasing the tolerance for missing race data and allowing banks to collect and report data only for loan applications with race data would give banks improper incentives to collect and submit loan application data selectively.

Therefore, the current state of HMDA, albeit flawed, may be the best option we have. Second, regulators could explore alternatives and supplements to HMDA. Mystery Shoppers is one such alternative that the Office of Thrift Supervision is currently considering. Regardless of what is currently feasible, regulators should be aware of the existence and consequences of missing race data and should always be searching for improved methods of monitoring compliance with fair lending laws.

References

Avery, Robert B., Patricia E. Beeson, and Paul S. Calem. 1997. Using HMDA Data as a Regulatory Screen for Fair Lending Compliance. *Journal of Financial Services Research* 11:9-42.

Barefoot, Jo Ann. 1998. Stop Focusing on Compliance Exams. *ABA Banking Journal* 90(1):26-9.

Canner, Glenn B., and Wayne Passmore. 1995. Home Purchase Lending in Low-Income Neighborhoods and to Low-Income Borrowers. *Federal Reserve Bulletin* 81(2):71-103.

Canner, Glenn B., and Dolores S. Smith. 1991. Home Mortgage Disclosure Act: Expanded Data on Residential Lending. *Federal Reserve Bulletin* 77(11):859-81.

Canner, Glenn B., and Dolores S. Smith. 1992. Expanded HMDA Data on Residential Lending: One Year Later. *Federal Reserve Bulletin* 78(11):801-24.

Dietrich, Jason. 1999. Missing Race Data in HMDA and the Implications for the Monitoring of Fair Lending Compliance. Unpublished paper. Risk Analysis Division, Office of the Comptroller of the Currency.

Federal Financial Institutions Examination Council. 1996. *A Guide to HMDA Reporting: Getting It Right!*. Washington, DC.

Federal Financial Institutions Examination Council. 1998. *A Guide to HMDA Reporting: Getting It Right!*. Washington, DC.

Fishbein, Allen J. 1992. The Ongoing Experiment with 'Regulation from Below': Expanded Reporting Requirements for HMDA and CRA. *Housing Policy Debate* 3(2):601-36.

Furst, Karen, William W. Lang, and Daniel E. Nolle. 2000. Internet Banking: Developments and Prospects. *Economic and Policy Analysis Working Paper 2000-9*. Washington DC: Office of the Comptroller of the Currency.

Huck, Paul. 2000. Home Mortgage Lending by Applicant Race/Ethnicity: Do HMDA Figures Provide a Distorted Picture. *Consumer Issues Research Series-2000-3*. Chicago: Federal Reserve Bank of Chicago.

Lee, Lung-Fei. 1982. Specification Error in Multinomial Logit Models: Analysis of the Omitted Variable Bias. *Journal of Econometrics* 20:197-209.

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Maddala, G.S. 1992. *Introduction to Econometrics.* 2nd ed. New York: Macmillan Publishing Company.

Negroni, Andrea Lee. 1997. Cyber-regulation. *Mortgage Banking* 58(2):10-5.

Seidman, Ellen. 1999. Testing for CRA Performance. *National Mortgage News* 23(41):4.

Appendix A: Data Tables

Between 1993 and 1998, there were approximately 3,141 counties in the United States. HMDA covered 851 of these counties prior to 1996 and 904 in 1996 and after. After eliminating observations for Puerto Rico, combining sister cities in Virginia, and dropping counties with no reported loans in HMDA, there were 739, 827, 829, 832, 835 and 835 observations for 1993, 1994, 1995, 1996, 1997 and 1998, respectively. This yielded a total of 4,897 observations in the pooled data for the estimations. Looking at the dependent variables first, the overall average of the county averages of missing race data ranges from 5.4 percent for conventional home purchase loans to 18.5 percent for conventional home improvement loans. Looking now at the right-hand-side variables, as expected, whites are the majority with an overall county average of 81.7 percent; Blacks are the second most represented at 10.2 percent. The average age is 35.23, males and females are evenly represented, and a high school diploma is the most prevalent educational attainment, on average, at 32.6 percent. Nearly 64 percent of the counties are urban, the average population and per capita income per county are 2,580,590 and 22,647, respectively, and the unemployment rate averages 5.1 percent. Overall, the summary statistics seem reasonable.

Table A1: Variable Definitions and Summary Statistics (N=4897)				
	Description	**Source**	**Mean**	**S.D.**
Dependent Variables				
Miss11	County % missing race for conventional home purchases	HMDA	0.054	0.034
Miss12	County % missing race for conventional home improvements	HMDA	0.185	0.117
Miss13	County % missing race for conventional refinances	HMDA	0.174	0.113
RHS Variables				
Age	County average age	Census	35.226	2.433
Male	County % who are male	Census	0.490	0.014
Non-Hisp White	County % White, non-Hispanic	Census	0.817	0.164
Non-Hisp Black	County % Black, non-Hispanic	Census	0.102	0.123
Non-Hisp Asians	County % Asian, non-Hispanic	Census	0.017	0.033
Non-Hisp Indians	County % American Indian, non-Hispanic	Census	0.006	0.016
Hispanics	County % Hispanic	Census	0.057	0.106
0-9 years educ	County % 0-9 yrs education	Census	0.092	0.045
9-12 years educ	County % 9-12 yrs education	Census	0.154	0.041
H.S. Degree	County % H.S. diploma	Census	0.326	0.062
Some Col. educ	County % some college	Census	0.206	0.047
Assoc. Degree	County % Associates degree	Census	0.058	0.017
B.S. Degree	County % Bachelors degree	Census	0.110	0.045
Grad. Degree	County % Graduate degree	Census	0.054	0.029
Lamt11	County average loan amount for conventional home purchases	HMDA	87.672	41.086
Lamt12	County average loan amount for conventional home improvements	HMDA	17.227	19.846
Lamt13	County average loan amount for conventional refinances	HMDA	80.886	33.493
Incdif11	County average income for conventional home purchases less pci	HMDA	32.764	14.741
Incdif12	County average income for conventional home improvements less pci	HMDA	26.510	8.904
Incdif13	County average income for conventional refinances less pci	HMDA	36.704	12.756
Population	County population size (100K's)	BEA	0.258	0.050
Pci	County per capita income (thous.)	BEA	22.647	5.389
Unemp. rate	County average unemployment rate	BLS	5.117	2.434
Bank concent.	Number of branches per square mile	Sheshunoff	0.231	0.804
Urban	Urban indicator variable	Census	0.635	0.270
Ntheast	Northeast indicator variable	HMDA	0.193	0.395
Stheast	Southeast indicator variable	HMDA	0.301	0.459
Nthcent	North-central indicator variable	HMDA	0.266	0.442
Sthcent	South-central indicator variable	HMDA	0.128	0.334
West	West indicator variable	HMDA	0.112	0.315
Year93	Year = 1993	HMDA	0.151	0.358
Year94	Year = 1994	HMDA	0.169	0.375
Year95	Year = 1995	HMDA	0.169	0.375
Year96	Year = 1996	HMDA	0.170	0.375
Year97	Year = 1997	HMDA	0.170	0.376
Year98	Year = 1998	HMDA	0.170	0.376