

Home Mortgage Lending by Applicant Race: Do HMDA Figures Provide a Distorted Picture?

Paul Huck
Federal Reserve Bank of Chicago

Abstract

The Home Mortgage Disclosure Act of 1975 (HMDA) was designed to further fair access to mortgage credit and requires lenders to report such information as location, loan amount, income, and race and sex for each application. However, race is missing in a significant proportion of applications taken by mail or phone. Given the widespread use of HMDA data by lenders, community groups, researchers, and regulators and the importance of mortgage lending as a public policy issue, the strengths and shortcomings of these data must be clearly understood.

The main findings are that reported approval rates by race are significantly overstated for refinancing and home improvement loans, while home purchase loans are little affected. A review of trends in how race is reported and in the technology of mortgage lending indicates that missing data on race will become a bigger and bigger problem in the near future.

Keywords: Data; Minorities; Mortgages

Introduction and summary

The Home Mortgage Disclosure Act of 1975 (HMDA) was designed by Congress to spotlight mortgage lending so that the public and regulators could better determine whether or not individuals or specific neighborhoods were being unfairly denied access to credit. HMDA requires lenders to report such information as location, loan amount, income, and race and sex for each applicant.¹ Lenders also report whether the application resulted in an origination, a denial, or some other action. The data generated by HMDA reporting are publicly available and provide a detailed picture of how geographic lending patterns vary depending on the income status or racial makeup of neighborhoods.²

The fact that denial rates are higher for minorities than for nonminorities is clearly evident in the HMDA reports, and this disparity has attracted much attention. Although HMDA data do not contain all of the

¹ The term race in this article is used to refer to both race and ethnicity.

² See Canner and Passmore (1995) for an example of an analysis of the national mortgage market using HMDA data. Not all mortgage applications are reported under HMDA. See Berkovec and Zorn (1996) for an analysis of HMDA coverage.

variables relevant to the credit underwriting decision, these reports have played an important role in research on racial discrimination and redlining—the denial of credit to neighborhoods on the basis of residents’ income or race.³ HMDA reports are also an important source of information for examiners during the course of a Community Reinvestment Act (CRA) or fair lending examination of an institution.

Given the widespread use of HMDA data by lenders, community groups, researchers, and regulators and the importance of mortgage lending as a public policy issue, the strengths and shortcomings of these data must be clearly understood. This article explores the potential for a distorted interpretation of HMDA data as a result of the way race is reported. Race is missing in a significant proportion of applications taken by mail or phone. Since each of these applications really does belong in some racial category, the number of applications for each race is understated by some amount. In addition, applications for which race is missing are less likely to be approved than applications with a known race, suggesting that the approval rates reported for various racial groups are overstated. The key question that this article will address is how much the missing data on race distort the interpretation of HMDA reports.⁴

The main findings are that significant potential exists for skewed interpretation of HMDA data and that the magnitude of the distortion depends on the specific loan product. The problems raised by this issue are more pronounced for refinance and home improvement loans than for home purchase loans. A review of trends in how race is reported and in the technology of mortgage lending indicate that missing data on race will become a bigger and bigger problem in the near future.

Reporting race and action taken

Lenders are required to gather information on the sex and race or ethnicity of applicants for a loan, usually by having them fill out a form.⁵ The categories for reporting race or ethnicity are as follows: (1) American Indian or Alaska Native, (2) Asian or Pacific Islander, (3) black, (4) Hispanic, (5) white, (6) other, (7) information not provided by applicant in mail or telephone application, and (8) not applicable. Applicants are not required to furnish this information, but if they refuse to do so, the

³ See Ladd (1998) for a survey of the research on discrimination in mortgage lending. See Schill and Wachter (1993) for an example of the use of HMDA data for an analysis of redlining in mortgage markets.

⁴ See Dietrich (2001) for a national analysis of missing data on race with some implications for fair lending monitoring that complements the analysis of this article.

⁵ See Federal Financial Institutions Examination Council (1998a) for a detailed discussion of the HMDA reporting process.

lender is required to put down the appropriate category based on visual appearance or surname. Exceptions to this procedure are applications taken entirely by mail or telephone. Lenders who take an application entirely by telephone are not required to ask for this information. Forms for mail applications must include these questions, but if the applicant does not provide the answer, the lender is not required to inquire further. "Not applicable" may be used for purchased loans, which are not originated by the reporter, and should be used for applications made by partnerships or corporations.

HMDA reports also provide information on the actions taken by lenders on the applications they receive. The categories for reporting the action taken are as follows: (1) loan originated, (2) application approved but not accepted, (3) application denied, (4) application withdrawn, (5) file closed for incompleteness, and (6) loan purchased by your institution. For the purposes of this analysis, the first two are combined into a category for approvals, whether originated or not. Denials are left as a category, and four and five are combined into a category for withdrawn/incomplete application. The last category is excluded from analysis.

Although attention usually focuses on racial disparities in denial rates, it is equally valid to look at differences in approval rates. Approval and denial rates are not quite two sides to the same coin because of the third category of withdrawn/incomplete applications. We shall see that applications for which information on race is not reported are much more likely to fall in the denied or withdrawn/incomplete categories as opposed to being approved. Focusing on denials means that we lose track of the withdrawn/incomplete applications, so it will prove more illuminating to look at approval rates in the following discussion.

Although not reporting race is allowed under HMDA reporting rules, some information is obviously lost when this happens. The following question ensues: Does the incidence of missing information on race have an empirically important effect on the interpretation of HMDA data? If race is missing in a purely random fashion, then this reporting practice would not seriously distort our interpretation of HMDA data. However, if this missing information is systematically related to variables of interest, then distortion will be a problem. To see whether this is the case, a first step is to look for systematic patterns in the incidence of missing information on race, using 1997 HMDA data for 10 major metropolitan statistical areas (MSAs) as a case study, as well as some national figures.⁶ To sharpen the focus, only loan applications for one- to four-family houses are included in the analysis. These include applications for purchase, home improvement, and refinance loans. (Loans

⁶ The MSAs in the study are Atlanta, Chicago, Detroit, Houston, Los Angeles, Miami, Milwaukee, New York, Philadelphia, and Washington, DC. These MSAs are clearly not a random sample and are not meant to be representative of the nation.

that were reported as purchased from the originator were excluded from the sample, as were loans with race reported as not applicable.)

Widespread use of information not provided

Given the importance of HMDA data for policy discussions about lending in potentially underserved neighborhoods, variation in the incidence of missing data on race by neighborhood income category is potentially interesting. Prevalence for the 10 MSAs taken together is shown in table 1, where we can see that race is missing for 19.9 percent of all loan applications. Although not reported in the table, this proportion for individual MSAs ranges from 11.9 percent in Houston to 30.3 percent in Philadelphia. The incidence varies by neighborhood income and loan product, however. Across all census tracts, home purchase applications are the least likely to fall into the missing data on race category (6.4 percent), whereas for refinance applications (28.2 percent) and home improvement applications (33.4 percent), the numbers are much higher. The incidence of missing data on race ranges from 4.6 percent in Detroit to 9.5 percent in New York for purchase applications, from 17.9 percent in Detroit to 49.8 percent in New York for refinance applications, and from 19.4 percent in Detroit to 42.0 percent in Los Angeles for home improvement applications.

Table 1. Incidence of Missing Data on Race, 10 MSAs (%)

	Purchase	Refinance	Improvement	All Loans
Low and moderate income	7.1	34.3	35.9	25.7
Middle income	5.7	29.6	33.6	20.5
Upper income	6.9	22.9	30.7	16.2
All tracts	6.4	28.2	33.4	19.9

Source: Federal Financial Institutions Examination Council 1998b and author's calculations.

Note: Neighborhood income categories are defined as follows: Low and moderate income is less than 80 percent of MSA median family income; Middle income is at least 80 percent and less than 120 percent of MSA median family income; and Upper income is equal to or greater than 120 percent of MSA median family income.

National figures provide a similar picture: Race is missing for 17.8 percent of the applications within MSAs in the nation for 1997. The corresponding figures for the various loan products are 5.8 percent for home purchase applications, 26.5 percent for refinance applications, and 29.7 percent for home improvement applications.

Table 1 also shows that race is missing for a higher proportion of applications for all loan products in low- and moderate-income tracts (25.7 percent) relative to middle- (20.5 percent) and upper-income tracts (16.2 percent) in the MSAs taken together. This pattern also holds for each loan product taken separately. If we consider the individual MSAs for

each loan product, New York provides the most extreme example of missing data on race, since 63.8 percent of refinance loans in low- and moderate-income neighborhoods fall into this category. This figure is based on over 21,000 applications, so it is not a result of a small sample. Thus, the incidence of missing data on race is higher in low- and moderate-income neighborhoods in part because this category is used relatively more often for each loan product in these neighborhoods.

Race is also reported less often for applications from low- and moderate-income neighborhoods because a lower proportion of these applications are for home purchase loans, which have a relatively low incidence of missing data on race compared with refinance and home improvement loans. Table 2 shows the proportion of applications for each loan product by neighborhood income. We can see that low- and moderate-income neighborhoods have a lower proportion of purchase applications (32.7 percent) than middle- (41.1 percent) and upper-income neighborhoods (47.8 percent).

Table 2. Distribution of Loan Products, by Tract Income, 10 MSAs (%)

	Purchase	Refinance	Improvement
Low and moderate income	32.7	44.3	23.0
Middle income	41.1	41.5	17.4
Upper income	47.8	39.8	12.4

Source: Federal Financial Institutions Examination Council 1998b and author's calculations.

Note: Neighborhood income categories are defined as follows: Low and moderate income is less than 80 percent of MSA median family income; Middle income is at least 80 percent and less than 120 percent of MSA median family income; and Upper income is equal to or greater than 120 percent of MSA median family income.

Another aspect of an overview of the incidence of missing data on race is that it is increasing over time. I illustrate this increase with national figures for applications made within MSAs using the same criteria for excluding applications that apply to the sample of 10 MSAs. Between 1993 and 1997, the proportion of applications for which race is missing for all loan products in the nation increased from 6.7 percent to 17.8 percent. This proportion rose for purchase loans (from 3.7 to 5.8 percent), for refinance loans (from 6.9 to 26.5 percent), and for improvement loans (from 14.9 to 29.7 percent).⁷

These results demonstrate that the incidence of missing data on race is quite substantial in these selected MSAs and in the nation as a whole, particularly for refinance and home improvement applications, and has

⁷ The figures reported here are not exactly comparable to those reported in Dietrich (2001) because somewhat different definitions of missing data on race are used. The overall trends for both definitions are consistent, and Dietrich shows that the incidence of missing data on race for conventional loans continued to increase from 1997 to 1999.

become more common in recent years. It is also true that applications from low- and moderate-income neighborhoods are more likely not to report race. These findings suggest that there is a systematic component to the incidence and that potential distortions in interpreting lending patterns for various racial groups are more likely to be a problem in low- and moderate-income neighborhoods than in middle- and upper-income neighborhoods. One potential explanation for these findings is that lenders targeting low- and moderate-income and minority neighborhoods market their services so as to promote the use of mail or telephone applications. Thus, product delivery systems may vary systematically by neighborhood, perhaps because of the historical location of branch offices.

Potential explanations for these results remain untested because, despite a considerable search on my part, systematic information on how product delivery systems vary by neighborhood income and minority population is not readily available.⁸ Although anecdotal reports are not a substitute for systematic data, I have heard that some lenders target minority neighborhoods for phone solicitations. Regarding the findings that the New York MSA tends to rank high in the proportion of applications missing data on race, anecdotal reports again support the explanation that the major New York banks are relatively more technologically advanced and thus more likely to use delivery channels with less in-person contact.

Approval rate distortion

The potential for a distorted view of the actions taken on applications for which race is unknown is compounded by the fact that such applications are much less likely to be approved than applications for which the race is known. Approval rates by race and loan product for the aggregated MSAs are shown in table 3,⁹ which illustrates that approval rates for minority applicants are lower than for white applicants, with the exception of purchase applications by Asians.

The table also shows that approval rates for applications with missing data on race are substantially lower than applications for which race is known for all loan products. For example, for all loan products taken

⁸ There are good reasons why this information is not publicly available. Lenders devote considerable resources to tailoring marketing and delivery strategies to neighborhood characteristics and view the results as proprietary. Also, variation in marketing and delivery methods that can be linked to minority populations is an explosive issue for lenders because it raises the possibility of discrimination.

⁹ Applications in the American Indian or Alaska Native and Other categories are omitted from the analysis of loan applications by race because of the small number of applications in these groups.

Table 3. Reported Approval Rates by Race and Loan Product, 10 MSAs (%)

	White	Asian	Black	Hispanic	Known Race/ Ethnicity	Information Not Provided	Total
Purchase	80.0	80.1	69.1	73.5	77.6	60.9	76.5
Refinance	75.3	68.8	59.1	58.8	70.2	37.5	60.8
Improvement	71.1	59.7	53.2	53.6	64.8	38.3	55.7
All loans	77.1	74.8	62.1	66.3	73.1	40.9	66.6

Source: Federal Financial Institutions Examination Council 1998b and author's calculations.

Note: Applications in the American Indian or Alaska Native and Other ethnic categories have been excluded from the analysis.

together, the approval rate for applications for which race is known is 73.1 percent, whereas for applications for which race is missing, it is only 40.9 percent. This pattern holds for each of the various loan products taken separately. It is also the case that the approval rates for each of the racial groups for each loan product are higher than the corresponding figure for the missing race category. A potential explanation for these findings is that some of the people who initially make applications by telephone or mail may at some time have a face-to-face meeting with the lender at which race can be observed. Applicants who are rejected, especially early in the process, are less likely to meet with lenders than those who are accepted. Thus, denied applications are more likely to fall into the missing race category. Another possibility is that less creditworthy applicants may select more anonymous lending channels, such as phone or mail, to reduce embarrassment from rejection.

These findings suggest that ignoring applications with missing data on race results in a distorted picture of approval rates. Table 3 allows us to calculate the magnitude of the distortion. For example, the table shows that the approval rate for purchase applications where race is known is 77.6 percent, while the rate for all purchase applications, including those for which race is missing, is 76.5 percent. Thus, the actual approval rate for total purchase applications is overstated by the difference between these figures: 1.1 percent. This seems to be a minor distortion of the approval rate. This finding is the result of a relatively low incidence of missing race for purchase applications (as shown in table 1) combined with the relatively small difference in approval rates for applications with and without a known race (as shown in table 3).

However, if we calculate the overstatement for the other loan products, we see that the approval rate for refinance loans is overstated by 9.4 percentage points (70.2 percent minus 60.8 percent) and the rate for improvement loans is overstated by 9.1 percentage points (64.8 percent minus 55.7 percent). For all loan products together, the approval rate is overstated by 6.5 percentage points (73.1 percent minus 66.6 percent). These results also hold for the MSAs taken individually. The most extreme example is New York, where the refinance approval rate for appli-

cations with known race is 64.1 percent and the rate for all applications is 47.6 percent, an overstatement of 16.5 percent. Thus, the fairly substantial overstatement of approval rates for refinance and home improvement applications is due to the relatively high incidence of missing data on race for refinance and improvement applications (as shown in table 1), combined with the relatively large difference in approval rates for applications with and without a known race (as shown in table 3).

We can easily calculate the overstatement of the approval rate for all applications combined, but calculating the distortion in the approval rate for separate racial groups is less straightforward. To make this calculation, the applications with missing data on race must be allocated to known racial categories in some way. In other words, the missing information must be somehow filled in, or imputed.

Imputing unknown race

Short of tracking down and surveying applicants for whom race is missing, we can never be certain of their actual race. However, established statistical methods make it possible to impute missing race by using information about the loan and the applicant from HMDA data combined with census data for the tract in which the property is located.

Race is an example of a categorical variable in which the value the variable takes on is limited to a number of discrete outcomes, in this case the categories defined by HMDA. Race is also unordered in the sense that the various categories convey no ranking and could be arranged in any order without affecting the analysis. An example of the opposite case, an ordered categorical variable, is a bond-rating system, for which the order or ranking is clearly important.

The multinomial logit is a standard statistical model for analyzing an unordered categorical variable, and it is used to generate estimated probabilities for the unobserved race category using the known individual loan and neighborhood characteristics.¹⁰ This is an example of a regression imputation, where the missing variable is estimated by the predicted value of a regression on the known variables.¹¹

¹⁰ See Greene (1997) for a technical description of the multinomial logit model.

¹¹ There is a large literature on the analysis of data with missing observations. See Little (1987) and Rubin (1987) for useful introductions to the topic. Another way of filling in missing data is a hot-deck imputation procedure. This method works by matching an observation with a missing variable to another observation similar in terms of the variables that are thought to be useful for prediction. The missing value is then imputed using the variable value of the matched observation. Avery, Beeson, and Sniderman (1999) use a hot-deck procedure in their analysis of mortgage lending. One of the variables they impute is race.

A good deal of information is available for predicting race. HMDA reports provide information about the individual applicant, the lender, the loan characteristics, and the census tract in which the property is located. Tract-level census data, such as information on housing, income, and demographics, can be combined with individual loan data to build a statistical picture of the loan application and the neighborhood in which the property is located. To the extent that these variables are correlated with applicants' race, they will be useful for prediction. Because of the small number of loan applicants who fall into the American Indian or Alaska Native and Other categories, these applications were omitted from the analysis.

To make better predictions across the 10 MSAs and three loan products, separate regressions were run for each MSA and product, resulting in 30 regressions. A fairly large number of applicant- and tract-level variables have been commonly used in the mortgage lending literature.¹² To conserve on computational resources, I began with these variables and then evaluated alternative specifications on the basis of goodness-of-fit measures, such as the log likelihood and cross-tabulations of actual and predicted outcomes. It quickly became apparent that the extra predictive value of adding variables to a fairly parsimonious specification is small. The applicant-level variables used to estimate the model include the log of the loan amount, a dummy variable equal to 1 for lenders who report HMDA data to the Department of Housing and Urban Development (HUD), and a dummy variable equal to 1 if the loan was approved.¹³ The tract-level variables include median family income, the population proportions for Asians, blacks, and Hispanics, and the square of these population proportions. Thus, the regression model consists of 11 variables, including a constant term.

The results are not reported here for two reasons, one practical and the other conceptual. The practical difficulty is that there are too many parameters to report. For marginal effects, the number of parameters in the multinomial logit is the product of the number of variables (11) and the response categories (four), or 44, for each of the 30 separate estimations. The conceptual reason is that this imputation is a pure exercise in prediction, and its success does not depend on the particular values of the parameters. Particular parameter values are simply not important for prediction. Goodness-of-fit results are important and will be discussed shortly. However, some general observations about the results may be of interest. Given widespread housing segregation, it is

¹² See Schill and Wachter (1993) for a representative selection of variables used to describe mortgage lending markets.

¹³ Applicant income and the ratio of loan amount to income were not used as explanatory variables because a significant number of applications in some MSAs did not report income. In practice, income adds little to the results after including loan amount because of the correlation of these two variables.

not surprising that the key variables for predicting race are minority proportions and their squares. There is also a great deal of correlation across the tract-level variables commonly used to describe neighborhood characteristics, so that the particular variables included in the model, apart from minority proportions, make very little difference in terms of prediction.¹⁴

Because the purpose of estimating the model is to impute unknown race, evaluating goodness of fit is important. One method is a cross-tabulation that compares actual and predicted races, where the predicted race for each observation is taken to be the outcome with the maximum probability.¹⁵

Although not reported here because of space constraints, the comparisons between actual and predicted outcomes for each MSA and loan product suggest that overall, the models fit the data reasonably well. Not surprisingly, given that a majority of the applicants are white, a high proportion of these observations are correctly identified. It is also not surprising that the models correctly identify relatively few observations when a racial group makes up a small proportion of the sample and does not stand out in terms of the explanatory variables. This result holds for Asians in all of the MSAs except Los Angeles, where about 20 percent of Asian applications are correctly identified. Thus, the results for Asians discussed later should be taken with a grain of salt. Where blacks and Hispanics make up a more substantial proportion of the observations (and often live in segregated neighborhoods), the models do a reasonable job of correctly categorizing the actual race. For example, black and Hispanic applicants make up about 20 percent and 10 percent, respectively, of all applications in the Chicago MSA. The proportion of correctly identified observations for black applicants ranges from 57.4 percent for purchase loans to 79.2 percent for refinance loans. For Hispanics, this proportion ranges from 27.1 percent for purchase loans to 66.9 for improvement loans.

Results

The predicted probabilities of the racial categories are used as estimates of the unobserved race and provide the foundation for further analysis. At least two approaches could be taken to allocate applications with

¹⁴ Since the number of observations in any given regression ranges into at least the tens of thousands, almost any reasonable tract-level variable is statistically significant, without any necessary relation to its marginal predictive value.

¹⁵ Note that a feature of the logit model is that the estimated probabilities will on average equal the sample proportions for each outcome. Thus, if the predictive variables are utterly useless, the model would simply return the population proportion for each outcome, and the predicted outcome would always be the most common sample outcome.

missing data on race to a particular racial category. One is to allocate the entire application to the race having the highest predicted probability. Another is to, in effect, split the application and allocate a portion to each race equal to the predicted probability for each category. I chose the latter and allocated portions of the unknown applications, on the grounds that information is lost by putting all the weight on the highest predicted probability. Now it is possible to see whether adjusting for omitted race makes a difference in our interpretation of the HMDA information. Specifically, particular attention will be paid to approval rates in light of the importance attached to differences in the outcome of applications across the various racial categories.

The results of the imputation exercise are used to allocate the applications with missing data on race into the known categories. The adjusted figures thus represent estimates of the actions taken on applications in the various racial groups after accounting for missing data. The adjustment is calculated as the residual difference between the adjusted approval rate and the reported rate. Table 4 shows the reported and adjusted approval rates by loan product and race for the aggregated MSAs.

Table 4. Adjustments and Approval Rates by Loan Products and Race, 10 MSAs (%)

		White	Asian	Black	Hispanic
Purchase	Reported	80.0	80.1	69.1	73.5
	Adjustment	-0.9	-0.9	-2.0	-1.0
	Adjusted	79.1	79.2	67.1	72.5
Refinance	Reported	75.3	68.8	59.1	58.8
	Adjustment	-8.5	-8.0	-10.3	-8.0
	Adjusted	66.8	60.8	48.8	50.8
Improvement	Reported	71.1	59.7	53.2	53.6
	Adjustment	-9.0	-9.8	-7.7	-9.0
	Adjusted	62.1	49.9	45.5	44.6
All loans	Reported	77.1	74.8	62.1	66.3
	Adjustment	-5.7	-5.5	-8.4	-6.3
	Adjusted	71.4	69.3	53.7	60.0

Source: Federal Financial Institutions Examination Council 1998b and author's calculations.
Note: Applications in the American Indian or Alaska Native and Other ethnic categories have been excluded from the analysis.

The table suggests that approval rates are overstated for each racial group. For all loan products taken together, the estimated adjustments to the approval rates range from -5.5 percentage points for Asians to -8.4 percentage points for blacks and vary depending on the loan product. The adjustments for home purchase loans are quite small and range from -0.9 percentage points for white and Asian applicants to -2.0 percentage points for black applicants. The adjustments for both refinance and improvement loans are more substantial and range from about

–8 to –10 percentage points. The approval rate adjustments for black applicants tend to be higher (in absolute values) than those of the other racial groups, with the exception of home improvement loans. However, the adjustments tend to be of similar magnitude across the various groups.

The adjustments for some of the individual MSAs are larger than those reported in the table. The most extreme case is refinance loans in New York, where the adjustments range from –13.5 percent for white applicants to –18.2 percent for black applicants. These adjustments are quite substantial, since for example, the reported approval rate for black applicants of 55.9 percent is reduced to an estimated 37.7 percent. In round figures, the reported approval rate of just over one-half is reduced to just over one-third.

Racial disparities in the actions taken on applications have received a good deal of attention as indicators of discrimination in the home mortgage market.¹⁶ Although generally disparities in denial rates have been reported, I have been discussing approval rates. One way to measure differences in outcomes is to calculate the difference between approval rates for white and minority applicants. The adjustments to approval rates for the various loan products in table 4 do not vary markedly across the groups (at most about two percentage points). Since a similar downward adjustment to the approval rate applies to each group, the differences between whites and minorities are not much affected by the adjustment process. The ratios formed by dividing the approval rate for whites by the minority approval rates can also be used to measure differences in outcomes. Although the approval ratio after adjustment tends to show an increase in disparities between white and minority applicants, the increases in approval ratios tend to be fairly small, generally under five percentage points.¹⁷ Again, this reflects the fact that the adjustments to approval rates are fairly similar across racial groups.

Another aspect of the results gets at the issue of how the incidence of missing data on race varies with the applicant's race. After allocating applications with unknown race to a category, we can calculate the estimated proportions of the total applications left out of the HMDA reports in each category. These estimated proportions are shown in table 5. The estimated incidence of omitted race reflects the overall use of the missing race category, since we see relatively low figures for purchase loans and relatively high figures for improvement and refinance loans. The proportions for refinance and improvement applications are quite sub-

¹⁶ For an example of a recent media discussion of discrimination and disparities in denial rates, see "Bias Worsens" (1999).

¹⁷ The biggest change is an increase in the ratio of white to black approval rate from a reported ratio of 1.27 to an adjusted ratio of 1.37.

Table 5. Estimated Proportion of Applications Missing Data on Race, 10 MSAs (%)

	White	Asian	Black	Hispanic	Total
Purchase	6.1	6.3	8.9	6.2	6.5
Refinance	25.5	24.6	38.8	28.9	28.9
Improvement	31.9	37.7	37.7	38.5	34.2
All loans	18.1	16.4	29.4	19.8	20.4

Source: Federal Financial Institutions Examination Council 1998b and author's calculations.
Note: Applications in the American Indian or Alaska Native and Other ethnic categories have been excluded from the analysis.

stantial, ranging from 25.5 percent for white applicants to 38.8 percent for black applicants for refinance loans. If we compare the estimated proportion omitted across racial groups, we see that the proportion for black applicants is higher than the corresponding figures for other groups for purchase and refinance loans. The estimated proportions for improvement loans are similar for minority groups, which in turn are higher than the figure for white applicants. For all loan products taken together, the estimated proportion of applications missing data on race is highest for black applicants at 29.4 percent. The largest estimated proportion across the 10 MSAs is refinance loans in New York, where the proportion ranges from 42.7 percent for white applicants to 65.9 percent for black applicants.

In summary, the results show that reported approval rates are generally overstated for all racial groups. The estimated adjustments to approval rates are small for purchase loans but more substantial for refinance and improvement loans. Since approval rate adjustments are fairly similar across racial groups, the adjustments do not substantially increase disparities in approval rates relative to reported rates. The estimated incidence of missing data on race in each category is quite substantial, particularly for black applicants.

Discussion

One of the most important implications of these results is that the potential for distortion in HMDA reports depends on the loan product in question. The incidence of omitted data on race is relatively small for home purchase loans, both overall and for individual racial groups. Both because of the relatively low incidence and the relative similarity in approval rates, the adjustments are fairly small for purchase loans. Thus, at current levels of the incidence of missing data on race, the interpretation of HMDA reports for purchase loans seems little affected by this reporting issue. However, such is not the case for home improvement and refinance loans.

Approval rates are overstated for all racial groups, substantially so for refinance and improvement loans. Racial disparities in reported approval rates are quite large, and one issue is whether the adjustments widen the disparities. There is some evidence that disparities in approval rates, as measured by differences in this rate between white and minority applicants, as well as the white/minority ratio of approval rates, are somewhat increased by the estimated adjustments, especially for some MSAs. However, because adjustments to approval rates are fairly similar across racial groups for the aggregated MSAs, disparities in approval rates are not affected dramatically by the estimated adjustments.

Although these disparities are little affected, the relatively high incidence of missing data on race for the aggregated MSAs for all loan products (20.4 percent) raises a warning flag for some uses of the data. Table 5 shows that home purchase is the only loan product for which the estimated proportion of applications missing data on race is small. The proportions are much higher for refinance and improvement loans and, although not reported in the table, can be higher yet in selected MSAs.

As one example of problems raised when the incidence of omitted race is high, it is useful to compare minority approval rates for an individual lender with MSA aggregate approval rates, perhaps as part of a fair lending review. Suppose the lender under review fully reports race. Since we have seen that the reported approval rates for an MSA aggregate are overstated, the comparison will be biased against a lender who fully reports race. A better comparison might be made with a control group of lenders that also fully report race.

The widespread incidence of missing data on race also raises problems for the statistical analysis of whether a lender's approve/deny decision depends on an applicant's race. If applications with missing data on race are simply dropped from the analysis, then the sample will obviously be a subsample of the actual population. If the sample selection process is systematically related to the accept/deny process, a statistical bias results. This is clearly the case, because applications that are not approved are much more likely to be reported with missing data on race (see table 3). In other words, race is not missing at random. This sample selection problem is compounded if the selection process depends on an applicant's race. This would be the case, for example, if lenders that marketed phone and mail applications tended to target minority neighborhoods. This selection problem means that statistical tests of discrimination in the loan approval process may be untrustworthy unless selection bias is accounted for.¹⁸

Methods for dealing with sample selection are available. One approach is to statistically model the selection process and include this in the re-

¹⁸ See Phillips, Trost, and Yezer (1994) and Dietrich (2001) for a discussion of selection problems in this context.

gression analysis. Doing so requires a reasonable understanding of the selection process, as well as access to data on the variables in the selection model. However, these data might not be readily available. If the selection process depends on the marketing methods or product delivery systems of individual lenders, for example, then lender-specific data would be required. Another approach is to impute, or fill in, the missing data on race and then proceed as usual with the regression analysis of the accept/deny decision. This is not a simple matter either. The simple regression imputation used in this article is adequate for my purely predictive exercise. However, a statistical test of discrimination requires an evaluation of the unbiased magnitude of particular regression coefficients in a structural econometric model—an entirely different matter from a simple predictive exercise—and more advanced imputation methods would be required.¹⁹ In any case, statistical analysis of discrimination in the accept/deny decision is considerably more difficult in the face of substantial omission of data on race.

As has already been noted, some evidence indicates that the proportion of applications missing data on race has been rising in recent years. This means that reported racial trends in lending will be distorted. For example, if reported home purchase originations by race are compared over time, then an increasing incidence of missing data on race means that the reported number of originations will be increasingly understated for each racial group. Since the number of purchase originations generally has been increasing in recent years, originations by race have actually increased by more than the reported trend,²⁰ although results show that reported approval rates for known race are overstated. Thus, the increasing incidence of missing data on race means that approval rates are increasingly overstated.

We have seen that the current incidence of such missing data raises significant difficulties for interpreting reported lending patterns by race. Will these difficulties become more pronounced in the future? While it is not possible to be certain, a reasonable projection of current mortgage industry trends makes it seem likely that this reporting problem will become more serious.

There can be little doubt that rapidly evolving information technology is making it easier to complete mortgage transactions with less face-to-face contact between lenders and borrowers. Many financial institutions have Internet Web sites that allow customers to apply for a variety of products and services, including mortgage loan products, online.

¹⁹ Multiple imputation is one approach to assigning missing data to hopefully avoid biased statistical inferences. See Rubin (1987) for a discussion of how this approach might be implemented.

²⁰ See Federal Financial Institutions Examination Council (2000) for national lending trends by race in recent years.

Some online banks have no brick-and-mortar branches at all, although it is not clear how successful this strategy will be. A consequence of this investment in online banking is that Internet originations should become more common (“Decisive Shift” 1999).

How the growing presence of online lending will affect HMDA reports depends on the proportion of online applications that fail to report race. Internet application materials, like mail applications, should include a form that gives an applicant a chance to report race, but lenders are not required to follow up if this information is not provided. Thus, whether race is reported for Internet applications depends on how potential borrowers fill out this portion of the form. While we still have little experience with HMDA reporting of online mortgage lending, there is some evidence that applications taken over the Internet will provide little information about applicants’ race.

Although HMDA reports for wholly Internet lenders are still rare, we do have at least one example of how such a lender reports race. Overall, 90.9 percent of applications reported by this lender in 1998 are missing data on race. This high incidence is observed for both home purchase and refinance loans. If this example is at all representative, the increasing use of online lending will entail increasingly less complete HMDA reporting of applicant race.

Another industry trend that results in an increased incidence of missing data on race is the growth of the subprime lending market. It is estimated that firms that specialize in subprime lending took in about 10 percent of the conventional home purchase applications and about 30 percent of the conventional refinance applications in the national HMDA figures in 1998.²¹ Subprime applications are more likely to have race missing than prime applications. For example, subprime specialists in the Chicago MSA in 1998 omitted race for 15.1 percent of home purchase applications, compared with a corresponding figure of 7.7 percent for all lenders in the MSA. The same figure for refinance applications for subprime lenders is 40.0 percent, compared with 17.8 percent for all lenders in the MSA. Thus, the increasing size of the subprime lending market has contributed to the rising incidence of missing data on applicant race in recent years. If the subprime market continues to expand, we can expect to see continued reduction in the completeness of HMDA reporting of applicant race.

Turning to policy options, easy fixes for this problem are not apparent for at least two reasons. First, given that HMDA regulations established a voluntary approach, the reporting of race for applications taken by

²¹ These estimates are taken from Scheessele (1999). Subprime loans are not identified in HMDA reports, but HUD estimates the extent of the subprime lending market by identifying lenders that specialize in this market. A list of subprime specialists can be found in Scheessele (1999).

phone or mail, or on the Internet ultimately depends on the willingness of the applicant to supply the information. It is my personal opinion that compelling applicants to provide race as a condition for accepting an application is not even remotely politically feasible. Second, HMDA reports are meant to provide an accurate picture of loans by individual lenders and for relatively small localities, such as neighborhoods. Thus, statistical measures, such as imputing missing race or surveying nonrespondents, which might be acceptable for establishing a general picture, will not work for specific cases involving a small number of applications.

Conclusion

Imputing missing race allows the calculation of an estimated adjustment to the approval rate and the estimated proportion of a given racial group with missing information for a sample of 10 large MSAs. Results suggest that this omission is not a serious problem for interpreting home purchase lending. The overstatement of the approval rate and the proportion of applications missing data on race for purchase loans are fairly small. However, for refinance and improvement loans, the problem is much more severe, particularly for some MSAs. These findings suggest that the impact of missing information should not be ignored when using HMDA reports to examine racial patterns in lending, particularly for refinance and improvement loans.²²

The rising incidence of missing data on race in recent years and emerging technological trends in the mortgage industry, which facilitate lending without face-to-face contact, make it reasonable to predict that this problem will become more serious in the not-too-distant future. One of HMDA's three stated purposes is to provide loan data that can be used "to assist in identifying possible discriminatory lending patterns and enforcing anti-discrimination statutes" (Board of Governors 1995, 1). If current trends continue, the ability of HMDA reports to further this purpose may be compromised.

Author

Paul Huck is Senior Economist at the Federal Reserve Bank of Chicago.

The views expressed are those of the author and do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System. The author thanks David Shushan for outstanding research assistance.

²² Note that missing information on race is not a major problem for the enforcement of the CRA under the latest procedures because a lender is evaluated on the basis of the distribution of activity by borrower and neighborhood income categories.

References

- Avery, Robert B., Patricia E. Beeson, and Mark S. Sniderman. 1999. Neighborhood Information and Home Mortgage Lending. *Journal of Urban Economics* 45:287–310.
- Berkovec, Jim, and Peter Zorn. 1996. How Complete Is HMDA? HMDA Coverage of Freddie Mac Purchases. *Journal of Real Estate Research* 11:39–55.
- Bias Worsens for Minorities Buying Homes. 1999. *New York Times*, September 16, p. A15.
- Board of Governors of the Federal Reserve System. 1995. *Regulation C Home Mortgage Disclosure*. Washington, DC.
- Canner, Glenn B., and Wayne Passmore. 1995. Home Purchase Lending in Low-Income Neighborhoods and to Low-Income Borrowers. *Federal Reserve Bulletin* 81:71–103.
- Decisive Shift to On-Line Applications Predicted. 1999. *American Banker*, July 27, p. 10.
- Dietrich, Jason. 2001. Missing Race Data in HMDA and the Implications for the Monitoring of Fair Lending Compliance. Economic and Policy Analysis Working Paper 2001–1. Washington, DC: Office of the Comptroller of the Currency.
- Federal Financial Institutions Examination Council. 1998a. *A Guide to HMDA Reporting: Getting It Right!* Washington, DC.
- Federal Financial Institutions Examination Council. 1998b. *1997 Loan Application Register and Transmittal Sheet Raw Data*. Washington, DC.
- Federal Financial Institutions Examination Council. 2000. *August 8, 2000, Press Release*. World Wide Web page <<http://www.ffiec.gov/hmcrpr/hm080800.htm>> (last modified June 22).
- Greene, William H. 1997. *Econometric Analysis*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.
- Ladd, Helen. 1998. Evidence on Discrimination in Mortgage Lending. *Journal of Economic Perspectives* 12:41–62.
- Little, Roderick A. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.
- Phillips, Robert F., Robert P. Trost, and Anthony M. J. Yezer. 1994. Bias in Estimates of Discrimination and Default in Mortgage Lending: The Effects of Simultaneity and Self-Selection. *Journal of Real Estate Finance and Economics* 9:197–215.
- Rubin, Donald B. 1987. *Multiple Imputation of Nonresponse in Surveys*. New York: Wiley.
- Scheessele, Randall M. 1999. 1998 HMDA Highlights. Working Paper No. HF–009. Office of Policy Development and Research, Department of Housing and Urban Development.
- Schill, Michael H., and Susan M. Wachter. 1993. A Tale of Two Cities: Racial and Ethnic Geographic Disparities in Home Mortgage Lending in Boston and Philadelphia. *Journal of Housing Research* 4:245–75.