A Monte Carlo Examination of Bias Tests in Mortgage Lending

by Paul W. Bauer and Brian A. Cromwell

Paul W. Bauer is an economist at the Federal Reserve Bank of Cleveland, and Brian A. Cromwell is a manager at Deloitte & Touche, Washington, D.C. The authors thank Robert Avery, Patricia Beeson, Paul Calem, Fred Furlong, Arden Hall, Gordon Smith, Mark Sniderman, and David Vandre for useful discussions and suggestions. Karen Deangelis provided excellent research assistance. The authors also thank the Federal Reserve Bank of San Francisco for supporting this research while Mr. Cromwell was an economist there.

Introduction

For three years, data released through the Home Mortgage Disclosure Act (HMDA) have documented that blacks are denied loans at a much higher rate than whites.[1] Whether this differential reveals bias by lending institutions, however, is a hotly debated issue. One reason is that HMDA data do not include all of the information contained on loan applications, such as the applicants' job and credit histories and the size of their down payments. Using information from loan applications, the Federal Reserve Bank of Boston (Munnell et al. [1992]) conducted a large statistical study of Boston-area banks (3,062 mortgage applications). The authors found that although the gap in denial rates between blacks and whites narrows when the additional information from the loan file is included, a statistically significant differential remains. Far from settling the issue, however, the Boston Fed study has merely provided the basis for further analysis. Interestingly, the Federal Deposit Insurance Corporation (FDIC) examined the loan applications identified as most likely to have been rejected because of bias, but found no evidence of bias and raised a number of methodological and empirical problems with the Boston Fed study (Horne [1994]).

This *Economic Review* explores the effectiveness of testing procedures in uncovering discrimination by mortgage lenders. After outlining the regulatory issues and the inherent problems in testing for bias, we investigate how well various tests perform under a variety of circumstances using simulated mortgage loan applications. We discuss the problems involved in testing real-world data and demonstrate how these tests perform using artificially generated data in which we can control the degree of bias. It should be emphasized that we cannot answer the question of whether there is bias in mortgage lending, but we have a great deal to say about the performance of bias tests when there are no measurement problems with the data (a condition for which all researchers strive) and when the degree of bias is known beforehand.

We find that tests employing all of the information included in our simulated loan files perform much better than those using only the HMDA subset of data. This result is not unexpected, yet the

■ 1 Under HMDA, lending institutions are required to record and report data on applicants' race, sex, income, type of loan, loan amount, and whether the loan was approved or denied.

BOX 1

Legal Definitions of Discrimination and Current Testing Methods

Legal Definitions

Several regulatory agencies are responsible for ensuring that discrimination does not occur in lending institutions. Each agency has developed a means of testing for discrimination in the bank class it is responsible for overseeing. Three general types of discrimination are recognized: overt discrimination, disparate treatment, and disparate impact.[a]

Overt discrimination occurs "when a lender openly discriminates on a prohibited basis." In addition, overt discrimination exists even when a lender expresses, but does not act on, a discriminatory preference. For example, a lender may offer two equally qualified applicants of different races different credit limits. Regulatory agencies would classify this action as overt discrimination.

There is evidence of *disparate treatment* when "a lender treats applicants differently based on prohibited factors." Disparate treatment may range from overt discrimination to subtle differences in treatment. For example, a lender may provide a nonminority applicant with more assistance in the application process than it would a minority applicant.

Disparate impact occurs when a lender "applies a practice uniformly to all applicants but the practice has a discriminatory effect on a prohibited basis and is not justified by business necessity." For example, a high-minimum-loan requirement may prevent low-income housing applicants, who are typically minorities, from being granted a loan.

Testing Methods

In order to test for lending discrimination, three procedures have been developed: testing, matched pairs, and statistical analyses. Each regulatory agency's method utilizes one or a combination of these general procedures.

Testing is a means of measuring differences in treatment among loan applicants. It involves sending "testers" disguised as loan applicants into an institution where they attempt to apply for a loan. Treatment of the "applicants" is then compared by the regulatory agency and determinations are made concerning the existence of discrimination.

Matched pairs are a means of grouping minority and nonminority applicants in a manner that will allow for accurate comparison of treatment between groups. Matched pairs are determined by comparing loan-to-value and debt-to-income ratios among applicants. Matched pairs are compared in a manner similar to that used for testers.

Statistical analyses may also be used to test for discrimination. An institution's lending data from previous years are collected, entered into a statistical program, and then analyzed for evidence of discrimination.

a. These definitions are outlined in "Interagency Task Force on Fair Lending," *Federal Register*, vol. 59, no. 73 (April 15, 1994), p. 18268.

difficulty of detecting low levels of bias even with large sample sizes is somewhat surprising. With the bias parameter set to the level that results in rejection rates similar to the actual HMDA data, sample size is crucial. At this level of bias, tests with sample sizes under 50 almost always fail to detect bias, whereas econometric tests with sample sizes greater than 200 perform well. Finally, we conduct nonparametric tests that have their roots in the procedures employed by examiners. Although these tests work very well in small samples, they also tend to find bias even in simulations when it is not present.

I. Methods of Testing for Bias

The Equal Credit Opportunity Act prohibits discrimination with respect to any aspect of a credit transaction based on race, color, religion, national origin, sex, marital status, age (provided the applicant has the capacity to contract), receipt of income from public assistance programs, and good-faith exercise of any rights under the Consumer Credit Protection Act. Regulatory institutions such as the FDIC, the Office of the Comptroller of the Currency, and the Federal Reserve Board are charged with enforcing this Act and uncovering discriminatory credit approval processes (see box 1). Of the thousands of consumer examinations conducted each year, few indicate credit discrimination on the basis of race.[2] In 1992, about 90 percent of the 5,602 banking institutions in the United States received outstanding or satisfactory ratings on their consumer exams.

On the other hand, over the last three years, large disparities between credit approval rates of white and minority applicants have been revealed by the revised HMDA data. Even controlling for income, minority applicants were rejected for credit at rates two to three times those of white applicants. This result potentially indicates widespread discrimination on the basis of race. A competing explanation for this credit-approval disparity, though, is that minority populations are commonly found to be less creditworthy (for example, because of lower asset levels) than the nonminority population. The revised HMDA data, however, do not include relevant financial information on credit

■ 2 A consumer exam is conducted to ensure that the regulated financial institution is in compliance with the various statutes relating to the treatment of consumers, such as the Equal Credit Opportunity Act and the Community Reinvestment Act.

applicants (such as assets and credit history) that is available to decisionmaking institutions.[3]

The issue of bias in mortgage lending is a broad one, and some researchers have raised the concern that simple comparisons of lenders' denial rates are not sufficient for grasping the complexities surrounding community-oriented lending.[4] Our purpose here is to explore the narrower issue of looking at the performance of tests that examiners could use to detect bias in the course of their regulatory duties.

Examiners have access to the complete loan files for both approved and rejected credit applications and consequently are able to look at the financial information missing from the HMDA data. In the past, applicant profiles were constructed for a sample of white and minority applicants (both acceptances and rejections). No formal statistical test was conducted, but the examiner looked for evidence that applicants are treated according to the articulated lending criteria of the institution. Financial institutions are required to maintain such criteria, and inspection of them is part of the exam process.

Currently, the Federal Reserve is implementing a testing procedure that estimates a logit model using the HMDA data for an institution over roughly the previous three years. If the race variable is found to be significant, then a random sample of loan files is selected and the model is reestimated after adding pertinent non-HMDA variables, such as employment and credit histories, net worth, and amount of other debt obtained from the loan application. If the race variable continues to be significant, the examiners pull the loan applications that the estimated model predicts were influenced by bias and seek out the bank's management for an explanation.

The sample size of applications actually examined, however, is constrained by the time (and number) of examiners that the agency is able to devote to the procedure. Fed guidelines suggest constructing a matched sample of 100 white and 100 minority applications. Constrained regulatory resources thus potentially undermine the effectiveness of a consumer exam in uncovering bias. Clearly, two important questions are whether pretests employing the HMDA data (which are relatively costless to the examiners, but not to the lenders) provide useful information, and how large a sample is required to determine whether lending bias exists. After developing a simulation model that allows us to vary the amount of bias against minorities, we use it to see how well various testing procedures can identify an institution that discriminates.

II. Simulation Model

Before going into the details of our simulation method, a brief overview is useful to highlight the key parts of the Monte Carlo process. Throughout the discussion of the model and consequently of our findings, it must be remembered that this is a simulation model and thus cannot answer the question of whether lending institutions are really subject to bias. Our goal is to explore how well various tests for bias perform when the level of bias is known beforehand. To accomplish this task, it is not necessary to mimic the underlying real-world process precisely. The key qualitative characteristics we wish to simulate are 1) that lenders base their mortgage approval decision on a larger set of variables than is included in the HMDA data, 2) that some of these omitted variables are correlated with race, and 3) that we can control the degree to which our simulated lender allows race to influence the loan approval process.

The first step is to generate a pool of loan applicants to simulate the actual population of both nonminorities and minorities in terms of income, net worth, debt payments, and credit history. Wherever possible, the variables are calibrated using the results of actual consumer surveys. These generated applicants then apply for loans in a credit approval model that is representative of actual approval processes used by financial institutions. The credit approval model allows for the possibility of bias against minorities, with the level of discrimination able to be varied from zero (in which case credit decisions are made solely on the basis of financial characteristics) to a level that results in a significantly higher level of rejections for minority applications. The results from the credit approval model are a set of loan files

■ 3 Avery, Beeson, and Sniderman (1993) cite results from an extremely large regression on the national HMDA data set, controlling for institutional and neighborhood characteristics and available individual information, and find a 7 to 10 percent unexplained differential linked to race. Munnell et al. (1992) explore the importance of the missing financial information in evaluating lending decisions in the Boston metropolitan statistical area. They find that differences in financial characteristics explain 9.9 percentage points of the observed 17.8 percentage-point discrepancy in denial rates of whites and minorities. The remaining 7.9 percentage points are considered to be linked to race. A discussion of a number of systematic problems present in HMDA data can be found in Horne (1994).

■ 4 See Avery, Beeson, and Sniderman (1993).

with applicant information and a 0/1 variable indicating whether or not the loan was granted. Although we have tried to benchmark our generated applicants to nationally reported data, this was not possible in all cases. Our numerical results will be sensitive to changes in the applicant generation process, but the qualitative import of our results will not.

At this point, our simulated examiner extracts a sample from the set of loan files and tests for discrimination. Several tests are possible. A "bank examiner" approach could search for evidence that whites and minorities with similar characteristics are treated differently, perhaps through matching rejected minority applications with approved white applications. Various levels of sophistication are possible. Alternatively, an "econometric" approach would estimate an equation and test for a significant coefficient on the variable representing race. In either the "examiner" or "econometric" approach, we will take repeated draws from the loan-file population and measure the proportion of times the test indicates a positive result for discrimination. By running the tests on loan files generated from a discriminatory credit approval process, we are able to explore the sensitivity of various tests for discrimination.

Generation of Applicant Data

The applicant sample is generated with the following characteristics: income, net worth, loan amount, other debt payments, credit history, and race. Actual loan applications would contain many more variables, but in our model these are the only ones the bank considers. More variables could be incorporated into the simulation model, but their addition would be unlikely to alter the basic thrust of our findings. Where possible, we have initially calibrated the means and correlations of these variables to those from consumer financial surveys and other sources.[5]

To generate the samples, we first created a matrix of the variances and covariances of the financial variables for the white and minority populations. The covariances of the loan amount and income (in log form) were identified from the 1990 national set of HMDA data for both nonminority and minority populations. We do not have information on the correlation of loan amount and the other financial variables, so we set these to plausible values. The means of the sample for income and loan amount were also determined using the HMDA data. We set the means for other financial variables using information from the 1989 Survey of Consumer Finance (SCF), a nationally representative wealth survey.[6] In particular, the mean of net assets in the sample was established by multiplying the mean income in the HMDA data by the ratio of assets to income in the SCF (for white and minority populations, respectively). The variance of assets and the correlation between assets and income for white and minority populations were also derived from this survey. We determined the mean of "other debt" payments using the ratio of other debt payments to income.

Our information on credit history for real loan applicants is limited. We used the answers to the SCF question on the timely loan and credit card payments to establish the sign of the correlation between bad credit history and the other financial variables, and modeled the tendency to have credit problems as an underlying normal random variable (larger values of credit history are considered bad) that correlates negatively with income and net worth.

Given the means, variances, and correlations, the applicant sample was generated by 1) multiplying the draws from the log-normal distribution by the Cholesky decomposition of the desired covariance matrix, and 2) rescaling the resulting series to match the desired means. This procedure ensures that the generated sample exhibits the desired correlations across variables. Credit history is rescaled into a categorical variable as follows: For whites, about 5 percent will have serious credit problems, 25 percent minor problems, and 70 percent no credit problems. For minorities, the corresponding percentages are roughly 7, 31, and 62, so that by construction they have a higher incidence of credit problems. These thresholds are arbitrarily chosen to give minorities more credit history problems in order to match the qualitative characteristics of real-world data.

The people surveyed in the SCF do not necessarily represent the population of potential mortgage applicants, because potential homeowners tend to be more affluent than the population as a whole. For our initial set of

■ **5** We focus on race in this paper, but a similar approach could be used to determine the power of tests for lending bias related to sex, age, and marital status.

■ **6** See Kennickell and Shack-Marquez (1992) for further information on the SCF.

TABLE 1

Sample Means of Generated Population (Sample size of 5,000, 50% minority)

	White	Minority
Income (annual)	$63,728	$36,029
Net worth	$291,682	$36,455
Loan amount	$73,744	$45,561
Loan payment (monthly)	$647	$400
Other debt payment (monthly)	$452	$202
Minor credit problems (percent)	25	31
Major credit problems (percent)	5	7

SOURCE: Authors' calculations.

variables, however, we found that rejection rates for both whites and blacks from the credit model (presented in the next section) were "too high" in comparison to those found in actual HMDA data, due in part to the low level of assets of minority families seen in the SCF. To adjust for this, we marginally increased the income and net assets of minority families, and marginally reduced the loan amount for both blacks and whites. The resulting generated samples should be viewed as broadly representative of the financial characteristics seen in the actual white and minority population, but as only partially calibrated due to lack of information on the financial characteristics of mortgage applicants.[7] Changing the financial characteristics, of course, does affect the probability of acceptance or rejection, but is unlikely to change the qualitative characteristics of our results.

The sample means of a draw of 5,000 applicants (half minorities) from our samples are reported in table 1.[8] Corresponding sample correlations are reported in table 2. Our white applicants (for this draw) have significantly larger incomes and net worths than do minorities, consistent with SCF data. Correspondingly, average loan amounts are higher for whites than for minorities. Our sample was generated so that positive correlations would be observed between income, net worth, loan amount, and other debt, and a negative correlation seen between financial variables and credit history. Again, we view this generated sample as only partially (and imprecisely) calibrated, but as reflecting broad relationships observed in the financial characteristics of populations in the real world.

Credit Approval Model

Once our applicant pool is generated, the "forms" are fed into our credit approval model that determines whether or not the financial institution makes the loan. The process is modeled so that "good" applications are almost always approved, and "bad" applications are almost always rejected. Borderline applications are approved or denied with a probability determined by the number of problems in the application, and by race in the case of a discriminatory bank.

We assume that the application is initially for a 30-year loan at a 10 percent interest rate with monthly payments and a 20 percent down payment.[9] The loan amount is initially determined through the applicant generation model. However, an applicant is unlikely to apply for a 20 percent down payment loan if he lacks the necessary assets. We model this down payment decision process in the following way: First, if the 20 percent down payment is greater than the applicant's net worth (plus two monthly payments), the applicant shifts to a 10 percent down payment. The loan amount and monthly payments are recalculated accordingly. Second, if net worth still falls short of the down payment, the applicant shifts to a loan with a 5 percent down payment.[10] Setting the loan amount and down payment in this sequential fashion is somewhat arbitrary, but it allows marginal applicants to apply for appropriate loans. Imposing a strong positive correlation between loan amount and net worth further tends to prevent paupers from applying for million-dollar mortgages.

Loan applications are scored according to four standard criteria: 1) the ratio of loan payment to income, 2) the ratio of total debt payment to income, 3) the percentage of the down payment, and 4) credit history. Any of the first three criteria can result in automatic rejection if

■ 7 Supplementing the SCF data with information in Munnell et al (1992) is one possible strategy for correcting this shortcoming.

■ 8 Our reason for oversampling is discussed later.

■ 9 Varying the interest rate and the term of the loan would introduce diversionary complications into the simulation model.

■ 10 Not addressed in this version of the model is the decision of the private mortgage insurer for down payments below 20 percent, or the effect of government insurance programs on loans with 5 percent down payments. Again, these factors are unlikely to affect the basic thrust of our results.

TABLE 2

Sample Correlations of Generated Population (Sample size of 5,000, 50% minority)

	Income (Annual)	Net Worth	Loan Amount	Other Debt	Minor Credit Problems	Major Credit Problems	White
Income (annual)	1.000	0.341	0.354	0.076	−0.406	−0.237	0.349
Net worth	0.341	1.000	0.336	0.048	−0.078	−0.047	0.180
Loan amount	0.354	0.336	1.000	0.095	−0.170	−0.123	0.247
Other debt payment	0.076	0.048	0.095	1.000	−0.009	−0.020	0.301
Minor credit problems	−0.406	−0.078	−0.170	−0.009	1.000	−0.156	−0.069
Major credit problems	−0.237	−0.047	−0.123	−0.020	−0.156	1.000	−0.030
White	0.349	0.180	0.247	0.301	−0.069	−0.030	1.000

SOURCE: Authors' calculations.

it is violated. Each of the criteria also has a "borderline" gray area (called GRAY1, GRAY2, GRAY3, and GRAY4, respectively) that results in a positive probability of rejection. Since the fourth criterion is a qualitative variable, possibly subject to varying interpretations of its severity, an "autoreject" here means that failing this criterion, by itself, results in a 50 percent chance of denial.[11] If all four criteria meet approval, then the application is almost always automatically accepted. With real loan applications, several other criteria (such as employment history and an appraisal) are considered during the underwriting process, but we focus on just these four standard criteria in an attempt to make our model more tractable.

The regions for the four criteria are as follows:

1) **Loan payment to income (PMT/Y):**
If PMT/Y > 0.40, then reject the application;
if 0.40 > PMT/Y > 0.28, then GRAY1;
if PMT/Y < 0.28, then the application passes this criterion.

2) **Total debt payment to income (TPMT/Y):**
If TPMT/Y > 0.48, then reject the application;
if 0.48 > TPMT/Y > 0.36, then GRAY2;
if TPMT/Y < 0.36, then the application passes this criterion.

3) **Net worth (NW):**
If down payment + 2 x PMT < net worth, then reject the application;
if 5 or 10 percent down payment, then GRAY3;
if 20 percent down payment, then the application passes this criterion.

4) **Credit history:**
If there are major credit problems, then randomly reject the application half the time;
if there are minor credit problems, then GRAY4;
if there are no credit problems, then the application passes this criterion.

These credit rules are motivated by actual credit processes. The financial ratios of 28 and 36 percent in rules 1 and 2, respectively, mirror actual tests used by financial institutions and suggested by secondary market purchasers, such as the Federal National Mortgage Association (FNMA).[12] Rule 3 checks for the level of down payment and a minimum net worth. Rule 4 seeks evidence of major and minor credit problems.

We allow for gray areas, however, in order to mimic the judgment that goes into the credit process for borderline applications. For example, the financial ratios suggested by FNMA are considered guidelines subject to the discretion of the lender. The down payment requirement reflects the bank's adjustment for an increased likelihood of default on low-down-payment loans. Finally, allowing for major and minor

■ **11** An additional consideration is that logit regressions cannot handle an independent variable that is too highly correlated with the dependent variable.

■ **12** See Federal National Mortgage Association (1992), pp. 601–94.

TABLE 3

Loan Scoring of Generated Population (Percent of sample, sample size of 5,000, 50% minority)

	White	Minority
AUTO APPROVE (Meets all criteria)	65	49
BORDERLINE (Violates some criteria)		
GRAY1 Payment/income between 28 and 40 percent	6	7
GRAY2 (Payment + other debt)/income between 36 and 48 percent	9	9
GRAY3 Down payment below 20 percent	6	26
GRAY4 Minor credit problems	25	31
AUTOREJECTS (Serious problems)		
AUTOR1 Payment/income above 40 percent	6	8
AUTOR2 (Payment + other debt)/income above 48 percent	14	13
AUTOR3 Net worth below down payment plus 2 PMTs	3	15
AUTOR4 Major credit problems	5	7

SOURCE: Authors' calculations.

credit problems allows for the distinction between recent bankruptcies versus a couple of late payments. Past credit problems may also be the result of unusual circumstances. The more GRAY areas that an application hits, however, the more likely that it will be rejected. In addition, we include a small probability of rejection of "clean" applications with no GRAYs to reflect some randomness in the decision process. The probability of approval contingent on the total number of GRAY areas is modeled as follows:

If TOTAL GRAYs = 0, then 3 percent rejection rate;

If TOTAL GRAYs = 1, then 20 percent rejection rate;

If TOTAL GRAYs = 2, then 30 percent rejection rate;

If TOTAL GRAYs = 3, then 40 percent rejection rate; and

If TOTAL GRAYs = 4, then 50 percent rejection rate.

These rates were chosen in order to generate a plausible number of rejections corresponding to the severity of credit problems.

We also use this process for borderline applications to introduce discrimination against minority applicants. Given this modeling, discrimination occurs because minorities are more likely than nonminorities to be turned down for a loan when there are blemishes in their loan applications. In general, we multiply the vector of approval probabilities by a bias parameter (BIAS) to increase the probability of rejection of minority applications.

If TOTAL GRAYs = 0, then 1/(1 – BIAS) percent rejection rate;

If TOTAL GRAYs = 1, then 20/(1 – BIAS) percent rejection rate;

If TOTAL GRAYs = 2, then 30/(1 – BIAS) percent rejection rate;

If TOTAL GRAYs = 3, then 40/(1 – BIAS) percent rejection rate;

If TOTAL GRAYs = 4, then 50/(1 – BIAS) percent rejection rate.

There are many ways to introduce bias into the loan approval process. This approach has the advantage of employing only a single parameter that can easily be varied from no bias (BIAS = 0) to the point where no minorities ever receive loans (BIAS = 1).

For example, if the bias parameter is set to 0.5, then minority applicants with a single GRAY will be rejected 40 percent of the time, applicants with two GRAYs will be rejected 60 percent of the time, and applicants with three or four GRAYs will always be rejected at 80 and 100 percent rates, respectively. We use this simple model so that we can easily test (by varying one parameter) the sensitivity of the results to varying levels of discrimination. Although more complicated models of discrimination can be used, we believe this model adequately captures the flavor of a discriminatory process where minority applicants are less likely to be approved in borderline cases.[13]

The sample statistics for GRAY1 – GRAY4 for whites and minorities are shown in table 3 (the sample includes 2,500 whites and 2,500 minorities).

■ **13** This result is implied by the findings of Munnell et al. (1992).

TABLE 4

Credit Application Decisions (5,000 draws, 50% minority)

	Discrimination Parameter[a]				
	0.0	0.2	0.4	0.6	0.8
Percent approved					
Total	71.3	70.7	67.1	59.0	54.0
White	76.1	76.1	76.1	76.1	76.1
Minority	66.4	65.2	58.0	41.9	31.9
Percent denied					
Total	28.7	29.3	32.9	41.0	46.0
White	23.9	23.9	23.9	23.9	23.9
Minority	33.6	34.8	42.0	58.1	68.1
Minority/white					
Percentage point difference	9.7	10.9	18.1	34.2	44.2
Due to financial characteristics	9.7	9.7	9.7	9.7	9.7
Due to discrimination	0.0	1.2	8.4	24.5	34.5

a. If zero, no bias. If one, no loans made to minorities.
SOURCE: Authors' calculations.

The proportion automatically rejected is also shown along with the four reasons for rejection, AUTOR1 – AUTOR4. (Applicants may have multiple reasons for automatic rejection.) By construction, whites are more likely than minorities to have clean applications that are automatically approved and are less likely to be automatically rejected. In our model, the most common GRAY areas hit are the credit history rule for both races and the down payment rule for minorities. Forty percent of the loans to minorities are for down payments below 20 percent, and of these, 15 percent are still automatically rejected for not having the necessary net worth for a mortgage with a 5 percent down payment. In addition, 13 percent are rejected for a high ratio of total debt to income. For whites, this financial ratio is also the most common reason for automatic rejection (14 percent).

Approval rates for varying levels of bias are shown in table 4. Whites are approved about 76.1 percent of the time in our model. With no discrimination, the minority approval rate is 66.4 percent, a difference of 9.7 percent. This compares with a national approval rate of 75.5 percent for whites and 55.7 percent for minorities (a 19.8 percentage-point difference) observed in 1990 national HMDA statistics. In our model, this difference is due solely to the financial characteristics of the applicant, and it is also close to that attributed to financial characteristics in Munnell et al. (1992). By varying our BIAS parameter, however, we can generate approval rates for minorities that mimic the observed approval rates in the HMDA data.

A small level of bias (0.2) results in only a slight increase in the disparity between whites and minorities. The next level of bias reported (0.4) raises the disparity to 18.1 percentage points, close to the observed 19.8 percentage-point difference in the national statistics. Raising the level of bias to 0.8 results in a rejection rate for minorities that is almost three times that of whites. Thus, our credit approval/discrimination model can generate the range of credit approvals observed in the HMDA data, allows for easy variation of the level of discrimination, and generates loan files that can be used to test for discrimination through a bank examination.

TABLE 5

Proportion Passing Examination (1,000 repetitions, logit test)

Sample size	Discrimination Parameter—Full Set of Variables[a] 0.0	0.2	0.4	0.6	0.8
50	0.990	0.981	0.976	0.875	0.555
75	0.970	0.932	0.892	0.701	0.238
100	0.971	0.937	0.856	0.572	0.068
150	0.971	0.929	0.824	0.441	0.005
200	0.971	0.936	0.777	0.323	0.003
250	0.972	0.925	0.720	0.184	0.000
300	0.971	0.923	0.695	0.161	0.000
350	0.966	0.914	0.656	0.103	0.000
400	0.970	0.902	0.584	0.062	0.000
450	0.974	0.908	0.568	0.040	0.000
500	0.967	0.883	0.505	0.027	0.000
600	0.969	0.870	0.456	0.019	0.000
800	0.973	0.850	0.339	0.001	0.000
Sample size	Discrimination Parameter—HMDA Variables[a] 0.0	0.2	0.4	0.6	0.8
50	0.926	0.905	0.835	0.693	0.265
75	0.876	0.845	0.734	0.497	0.112
100	0.879	0.790	0.676	0.406	0.038
150	0.828	0.728	0.527	0.203	0.007
200	0.760	0.671	0.412	0.109	0.000
250	0.744	0.564	0.341	0.040	0.000
300	0.648	0.506	0.272	0.027	0.000
350	0.635	0.446	0.182	0.015	0.000
400	0.586	0.383	0.136	0.005	0.000
450	0.572	0.333	0.110	0.003	0.000
500	0.521	0.291	0.068	0.001	0.000
600	0.423	0.214	0.041	0.000	0.000
800	0.316	0.109	0.016	0.000	0.000

a. If zero, no bias. If one, no loans made to minorities.
SOURCE. Authors' calculations.

III. Analysis of Econometric Tests

In this section, we test the statistical power of econometric exam tools through Monte Carlo simulation. We vary both the sample size and level of bias to test the sensitivity of the exams to these factors. Repeated draws (of our preset sample size) from bank loan files (with our preset level of bias) are used in a logit regression. The dependent variable is a 0/1 variable indicating credit approval. Independent variables include those corresponding to the credit approval process: income, net worth, payment/income, total debt payments/income, down payment/net worth, CREDIT1 (0/1 dummy for minor credit problems), and CREDIT2 (0/1 dummy for major credit problems). In addition, a 0/1 dummy variable indicating a minority is included. A significant (negative) estimated coefficient is taken as a positive test for discrimination.

Logit is the preferred estimator on theoretical grounds for such a model because it allows for the 0/1 nature of the dependent variable (determining whether the loan was approved) and for slightly more outliers than the Probit model.[14] These advantages have also led to its use in the Boston Fed study and in the current Federal Reserve testing procedure for investigating possible lending bias. We conducted 1,000 repetitions for each setting of the model, oversampling minorities so that they compose 50 percent of the sample.[15]

Table 5 reports the proportion of examinations that "passed" (failed to find statistically significant evidence of discrimination) at the standard 5 percent significance level. Figure 1 plots the same data, but is useful for illustrating how the performance of the test improves as the sample size increases. In the first column, there is no bias, yet some banks fail to pass. The level of "false positives" is an important factor in evaluating the usefulness of a test. False positives represent the risk of erroneously accusing a bank of discriminating when it in fact does not. An ideal test would always find bias when it is present, but would never find it when it is absent. For logit, this rate is typically in the 1 to 3 percent range over the sample sizes studied and tends to be slightly better than the 5 percent we allowed for in our selection of the significance level.

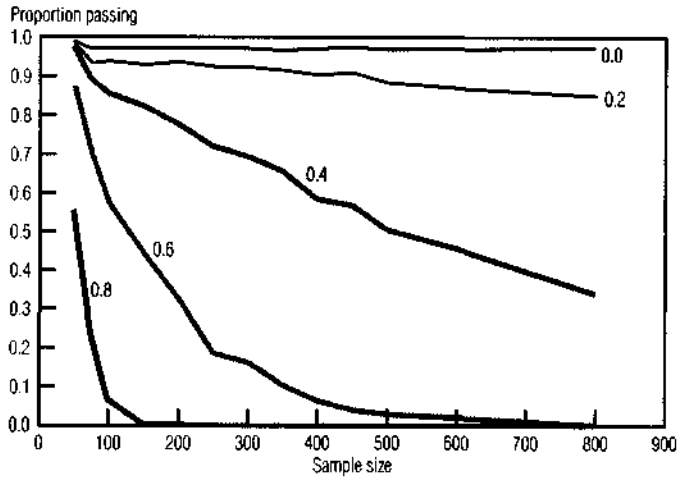
In the second column, we set our bias parameter to 0.2, introducing a low-level bias that, as seen in table 4, increases the rejection rate for minorities only slightly. For small sample sizes, we rarely find evidence of this discrimination. Even for a sample size of 800, our tests successfully uncover discrimination only 15 percent of the time. A small level of discrimination can go undetected by statistical methods even with very large samples.

14 We know that banks do not use an equation like this to make their decisions, but we use it in our model to approximate their decision-making process.

15 In earlier work, we explored the importance of oversampling. We found that it increases the statistical power of the exam by a very small amount, but reduces the incidence of false positives in small sample sizes. Oversampling also avoids anomalous results in small sample sizes, such as having no minority acceptances.

FIGURE 1

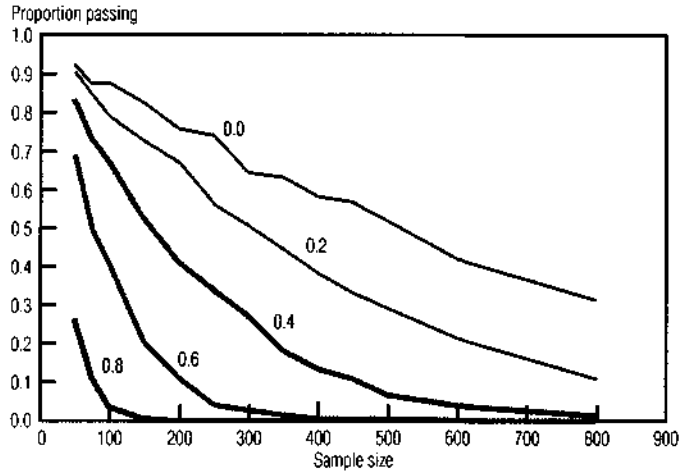
**Power of Logit Test—
Full Variable Set**



SOURCE: Authors' calculations.

FIGURE 2

**Power of Logit Test—
HMDA Variables**



SOURCE: Authors' calculations.

At moderate levels of discrimination (with bias equal to 0.4, raising the lending differential between whites and minorities significantly, as seen in table 4), sample size plays a critical role in the power of the exam. We find discrimination less than 15 percent of the time for sample sizes of 100 or less. Raising the sample size from 100 to 200 increases the power of the exam to about 23 percent, and a sample size of 800 results in a statistical power of 66 percent.

At larger levels of bias, the logit test is better able to detect discrimination. For a bias of 0.6, a sample of 50 uncovered discrimination less than 15 percent of the time, but the power increases sharply with sample size. For a sample of 200, the power is nearly 67 percent; for a sample of 400, the power is 94 percent. For our highest level of bias, our smallest sample size found discrimination 45 percent of the time, and samples over 250 uncovered discrimination every time.

Employing only the variables available in the HMDA data significantly lowered the chance of passing the examination (see bottom of table 5 and figure 2). One criticism of a test using this incomplete data is that it suffers from omitted variable bias. In our model, differences in credit history and assets result in a higher rejection rate for minorities, but the regression results attribute this to race even in the absence of discrimination. While this tends to make it easier to find bias when it exists, it also makes it easier to find bias when it does not exist. In the case of no bias, the power of the test plummets as the sample size increases, so that with sample sizes of 800, nondiscriminatory banks would pass less than a third of the time.

Given that one perceived advantage of using the available HMDA data is that large samples can be put together at low cost, this result suggests that indications of discrimination that rely solely on HMDA data should be treated with caution. The usefulness of employing the readily available HMDA data to pretest banks in order to direct scarce regulatory resources more effectively is a possible extension of our analysis.

Of course, the problem of false positives arises partly because of the built-in correlations between race and the other variables. If race and these variables were uncorrelated, the problem of false positives would be reduced. The degree of correlation between the two is an empirical question.

Initially, we ran ordinary least squares (OLS) instead of the more sophisticated logit model to save time in our simulations. While crude because it fails to adjust for the 0/1 nature of the dependent variable, the OLS test performs nearly as well as the logit estimator (see table 6). It also never fails to achieve convergence as the logit sometimes does with small sample sizes and a high degree of bias. Consequently, the OLS test using the HMDA data may be useful as a pretest when logit fails.

TABLE 6

Proportion Passing Examination (1,000 repetitions, DLS test)

Sample size	Discrimination Parameter—Full Set of Variables[a] 0.0	0.2	0.4	0.6	0.8
50	0.974	0.942	0.923	0.806	0.358
75	0.968	0.953	0.886	0.694	0.173
100	0.970	0.928	0.851	0.635	0.089
150	0.969	0.922	0.832	0.453	0.014
200	0.959	0.900	0.776	0.361	0.005
250	0.957	0.895	0.736	0.225	0.000
300	0.942	0.887	0.684	0.156	0.000
350	0.951	0.863	0.634	0.111	0.000
400	0.932	0.859	0.562	0.097	0.000
450	0.935	0.834	0.528	0.055	0.000
500	0.924	0.825	0.490	0.038	0.000
600	0.909	0.785	0.415	0.021	0.000
800	0.898	0.748	0.306	0.003	0.000
Sample size	Discrimination Parameter—HMDA Variables[a] 0.0	0.2	0.4	0.6	0.8
50	0.904	0.864	0.791	0.657	0.245
75	0.873	0.829	0.710	0.479	0.117
100	0.860	0.782	0.664	0.399	0.039
150	0.819	0.720	0.540	0.208	0.008
200	0.758	0.674	0.425	0.131	0.000
250	0.731	0.567	0.336	0.043	0.000
300	0.650	0.512	0.269	0.030	0.000
350	0.641	0.443	0.191	0.012	0.000
400	0.582	0.384	0.140	0.006	0.000
450	0.568	0.336	0.113	0.002	0.000
500	0.519	0.301	0.072	0.000	0.000
600	0.430	0.226	0.046	0.000	0.000
800	0.317	0.112	0.017	0.000	0.000

a. If zero, no bias. If one, no loans made to minorities.
SOURCE: Authors' calculations.

IV. Analysis of Alternative Exam Procedures

A goal of this research is to test the power of actual examination techniques used in consumer exam settings. The *Federal Reserve System Consumer Compliance Handbook*, published by the Federal Reserve Board of Governors (1989), provides guidance on how to model a consumer exam. In addition, we have met with consumer examiners and one of us went on an actual consumer examination to observe procedures firsthand.

Applicant profile worksheets are the main tool used by consumer examiners to test for discrimination against protected classes such as minorities or females. Examiners complete these forms from a sample of both accepted and rejected loan files. They list the applicant's class characteristics along with his or her length of employment, length of residence, and monthly debt/income ratio. The forms also include the date and terms of the requested credit. If the application is rejected, reasons for rejection are noted.

The examiner then uses the profiles to compare the characteristics of applicants who receive credit with those who do not, and to make comparisons between protected classes. As a first check, the examiner sees whether those who are accepted or rejected are treated in accordance with the bank's articulated lending criteria. Any instances of credit decisions that fall outside the criteria are flagged for further investigation. The examiner then has considerable flexibility in how the files are selected and segregated for analysis between protected classes. Various comparisons suggested by the handbook include accepted minority versus accepted nonminority, rejected minority versus accepted nonminority, and rejected minority versus rejected nonminority. While not conducting a formal statistical test, the examiner then makes a judgment as to whether the classes have received equal treatment. With respect to sample size, the *Consumer Compliance Handbook* notes:

> Since statistical validity is not a key issue, the ideal size of the judgmental sample cannot be stated in terms of numbers. Enough items should be selected in order to draw a reasonable conclusion. Again, the examiner should exercise careful discretion based upon experience and related examination findings. (p. 1.B.25)

Discussion with experienced examiners suggests that the examiner starts off with a fairly small sample size of perhaps 40 acceptances and 40 rejections. This small sample size is due in part to the limited amount of time available to conduct the examination. In addition, for many other of the regulations tested on an examination — such as truth in lending — compliance can be adequately ascertained through a small sample. If the examiner finds any evidence of discrimination — for example, a rejected minority whose characteristics dominated those of an accepted white — then the sample size is expanded and a more intensive investigation is conducted.

TABLE 7

Proportion Passing Examination (1,000 repetitions, NP I)

Sample size	Discrimination Parameter—Full Set of Variables[a] 0.0	0.2	0.4	0.6	0.8
50	0.484	0.466	0.390	0.288	0.058
75	0.653	0.578	0.495	0.328	0.043
100	0.665	0.561	0.485	0.316	0.020
150	0.658	0.588	0.458	0.235	0.006
200	0.656	0.561	0.421	0.200	0.002
250	0.617	0.548	0.406	0.154	0.001
300	0.585	0.477	0.377	0.135	0.000
350	0.591	0.480	0.335	0.098	0.000
400	0.558	0.457	0.319	0.084	0.000
450	0.530	0.406	0.260	0.077	0.000
500	0.495	0.372	0.227	0.060	0.000
600	0.426	0.354	0.196	0.031	0.000
800	0.318	0.227	0.125	0.020	0.000
Sample size	Discrimination Parameter—HMDA Variables[a] 0.0	0.2	0.4	0.6	0.8
50	0.044	0.058	0.044	0.033	0.015
75	0.027	0.021	0.016	0.011	0.005
100	0.009	0.010	0.008	0.001	0.001
150	0.001	0.002	0.002	0.000	0.000
200	0.001	0.000	0.002	0.000	0.000
250	0.000	0.000	0.000	0.000	0.000
300	0.000	0.000	0.000	0.000	0.000
350	0.000	0.000	0.000	0.000	0.000
400	0.000	0.000	0.000	0.000	0.000
450	0.000	0.000	0.000	0.000	0.000
500	0.000	0.000	0.000	0.000	0.000
600	0.000	0.000	0.000	0.000	0.000
800	0.000	0.000	0.000	0.000	0.000

a. If zero, no bias. If one, no loans made to minorities.
SOURCE: Authors' calculations.

As an approximation of actual exam procedures, we tested the power of two potential exam procedures on our generated loan files. Our first test (NP 0) looked for *any* instance of a rejected minority with financial characteristics that dominated those of an accepted white. Domination was defined as having more favorable characteristics for all four of the criteria used in the loan scoring procedures. The second test (NP I) compared the proportion of minority rejections that dominated white acceptances with the proportion of minority acceptances that dominated white rejections. A third test attempted to see whether differences in the latter proportions are statistically significant.

NP 0 proved to create a high degree of false positives. With BIAS set equal to zero and a sample size of 50, the exam indicated discrimination 39 percent of the time (with 100 repetitions). For a sample size of 100, discrimination was found 72 percent of the time. Finally, for sample sizes above 200, we almost always found an instance of a minority rejected applicant who had more favorable financial characteristics than an accepted white. This high degree of false positives suggests that the test was overly stringent given the degree of randomness we introduced in the loan files. Whenever there is even a small amount of randomness in the approval process, the probability of finding a rejected minority applicant who dominated an approved white applicant approaches one, so this test actually performs worse as the sample size expands in the case where there is no bias. On the other hand, if there is no randomness and all of the variables are measured without error, then this test would perform flawlessly. These are conditions that are unlikely to be met with actual data.

We used NP I to account for the underlying uncertainty in the credit approval process. We compared the proportion of minority rejections that were dominated by white acceptances with the proportion of white rejections that were dominated by minority acceptances. If the first proportion was larger, we took this as a flag for discrimination. The power of this exam is shown in table 7 for sample sizes of 50 through 800.

For BIAS set equal to 0, we find that this test reports a large proportion of false positives, better than NP 0, but much worse than the logit and OLS tests. Using only the HMDA variables resulted in false positives almost all of the time regardless of the sample size. When bias is introduced, this test outperforms logit and OLS in small samples, but not in large samples. Unfortunately, the large proportion of false positives in the case of no bias makes this test less than ideal.

In table 8, we report results for a modified version of this test (NP II) that attempts to determine whether differences between the proportion of minority rejections that dominated some whites and white rejections that were dominated by some minorities was statistically significant using a chi-squared test. When there is no bias, this test has fewer false positives in small sample sizes than NP I, but has more when the sample size is large. Like NP I, when there is bias, the test is much better at detecting it than are logit and OLS in small samples, but the test is not as good with large samples.

TABLE 8

Proportion Passing Examination (1,000 repetitions, NP II)

Sample size	Discrimination Parameter—Full Set of Variables[a] 0.0	0.2	0.4	0.6	0.8
50	0.703	0.684	0.685	0.640	0.538
75	0.629	0.626	0.570	0.550	0.403
100	0.545	0.537	0.479	0.458	0.280
150	0.420	0.406	0.395	0.346	0.139
200	0.405	0.344	0.361	0.309	0.078
250	0.339	0.331	0.329	0.252	0.042
300	0.337	0.322	0.275	0.256	0.025
350	0.262	0.283	0.289	0.236	0.022
400	0.285	0.270	0.260	0.197	0.017
450	0.279	0.275	0.249	0.190	0.004
500	0.237	0.227	0.224	0.184	0.005
600	0.211	0.210	0.198	0.163	0.005
800	0.175	0.184	0.179	0.164	0.001
Sample size	Discrimination Parameter—HMDA Variables[a] 0.0	0.2	0.4	0.6	0.8
50	0.749	0.798	0.753	0.728	0.742
75	0.741	0.727	0.707	0.689	0.640
100	0.671	0.651	0.679	0.636	0.612
150	0.584	0.615	0.608	0.539	0.509
200	0.555	0.554	0.556	0.503	0.490
250	0.511	0.538	0.532	0.502	0.385
300	0.480	0.493	0.469	0.470	0.324
350	0.466	0.452	0.471	0.465	0.331
400	0.420	0.432	0.450	0.446	0.314
450	0.422	0.422	0.441	0.449	0.284
500	0.400	0.409	0.399	0.413	0.282
600	0.350	0.366	0.376	0.402	0.239
800	0.318	0.330	0.338	0.382	0.212

a. If zero, no bias. If one, no loans made to minorities.
SOURCE: Authors' calculations.

V. Conclusion

Using a simulation model, we have examined several approaches to testing whether a financial institution discriminates. Because we employ a simulation model, the degree of bias can be varied from no bias to the point where no minorities are given loans.

Tests that employ all of the information included in our simulated loan files perform much better than those that use only the HMDA subset of data. For example, using the logit test, a nondiscriminating bank with 800 applications has less chance of passing than a smaller discriminating bank (bias = 0.4) with only 250 applications (see table 5). More surprisingly, low levels of bias can be difficult to detect even with large sample sizes. With levels of apparent bias found in actual HMDA data, sample size is very important. Tests with sample sizes under 50 almost always fail to detect bias, whereas tests with sample sizes greater than 200 perform well. Our test that attempts to mimic the procedures employed by examiners suggests that they work well in small samples, but also tend to find bias even in simulations when it is not present.

The qualitative characteristics of these findings are unlikely to be affected by either better calibration of the data or more elaborate modeling of the approval process. Detecting bias, particularly a small degree of bias at an institution, is likely to be a difficult endeavor. Even examiners, who have access to the applicants' loan files, are apt to face problems. Statistical methods require large sample sizes for low bias levels, which may require a great deal of regulatory resources. Examiner-inspired methods work well in small samples, but have a tendency to find bias even when it is not present. In particular, any randomness in lending decisions makes simple match-pair tests (such as NP 0) yield a high degree of false positives. More sophisticated versions (such as NP I and NP II) perform better because they allow for some underlying randomness.

Future research will look at the usefulness of employing the HMDA variables as a pretest to direct regulatory resources. By construction, this paper cannot say whether there is discrimination in mortgage lending, but by laying out the issues and problems involved in testing for discrimination and by exploring the robustness of the various approaches to testing for bias, it allows a more informed debate to proceed.

References

Avery, Robert B., Patricia E. Beeson, and Mark S. Sniderman. "Lender Consistency in Housing Credit Markets," Federal Reserve Bank of Cleveland, Working Paper 9309, December 1993.

Board of Governors of the Federal Reserve System, Division of Consumer and Community Affairs. *Federal Reserve System Consumer Compliance Handbook*, Washington, D.C., June 1989.

Federal National Mortgage Association, *Fannie Mae Guides*, vol. 1 (Selling Guide), Washington, D.C., 1992.

Horne, David K. "Evaluating the Role of Race in Mortgage Lending," *FDIC Banking Review*, vol. 7, no. 1 (Spring/Summer 1994), pp. 1–15.

Kennickell, Arthur, and Janice Shack-Marquez. "Changes in Family Finances from 1983 to 1989: Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin*, vol. 78, no. 1 (January 1992), pp. 1–18.

Munnell, Alicia H., Lynn E. Browne, James McEneaney, and Geoffrey M.B. Tootell. "Mortgage Lending in Boston: Interpreting HMDA Data," Federal Reserve Bank of Boston, Working Paper No. 92-7, October 1992.