How Costly is Noise?

Data and Disparities in the US Mortgage Market*

Laura Blattner Scott Nelson

Stanford University Chicago Booth

March 31, 2020

Abstract

We quantify the economic cost of the differential informativeness of credit scores across different consumer groups in the US mortgage market. We provide evidence that widely used credit scores are poorer risk predictors for minority applicants but play equal weight in lenders' approval and pricing decisions. Motivated by these facts, we build a structural model of a lender making loan approval and pricing decisions based on a credit score signal. We estimate the model exploiting novel empirical moments from credit bureau data combined with instrumental variable estimates of marginally approved borrowers' default rates, revealing the distribution of unobserved risk across different parts of the borrower and score distribution. Our estimation results suggest that commonly used credit scores feature 47% higher standard deviation of signal noise in credit scores for minority mortgage applicants, suggesting substantially lower precision for minorities. In preliminary results, our model suggests that this unequal precision accounts for roughly half of the gap in mortgage approval rates between minority and non-minority applicants. Finally, we translate our estimates into a dollar estimate of loan mis-pricing for mortgages purchased by the government-sponsored enterprises (GSEs).

*We thank Mario Curiki, David Du, Pranav Garg, and Xuechen Hong for excellent research assistance. We thank Juliane Begenau, Greg Buchak, Claudia Robles Garcia, David Schoenherr, Amit Seru, Ken Singleton, and seminar participants at Stanford University for helpful discussions and comments, and the Fama-Miller center at Chicago Booth as well as the Becker Friedman Institute for generous support. Any errors or omissions are the responsibility of the authors. This draft is preliminary and incomplete; please do not circulate or cite without permission.

1 Introduction

Automated decision-making is increasingly replacing human discretion in US consumer credit markets. Credit scores, such as FICO, drive underwriting and pricing for at least USD 1 trillion in mortgages annually, and advanced machine learning algorithms are already in use for unsecured consumer lending. This shift toward algorithmic lending has benefits: it reduces the inefficiencies and disparities in credit access that arise from human frictions or biases (Bartlett et al., 2019; Dobbie et al., 2019). This shift also has costs. A growing machine-learning literature emphasizes how algorithms may reinforce or increase disparities in credit access (e.g. Barocas et al., 2019). However, these costs have been difficult to quantify in economic terms.

In this paper, we quantify the economic costs of a particular source of disparities in algorithmic lending: the differential precision and accuracy of credit scores across different consumers. We build a parsimonious structural model of a lender approving, and pricing, loans based on a credit score signal that can differ in precision (signal noise) and accuracy (expected deviation from true risk). Focusing on protected classes that are emphasized in US consumer protection law, we find that commonly used credit scores feature 47% higher standard deviation of signal noise for minority mortgage applicants. These statistical differences have economic costs: we estimate that roughly half of the gap in approval rates between minority and non-minority applicants is attributable to credit score noise alone. In ongoing work, we also quantify what share of minority relative to non-minority mortgage applicants are assigned to higher-cost loans under standard GSE underwriting criteria as a result of credit score noise, and among mortgages that ultimately are originated, what the differences are between minority and non-minority borrowers in total interest costs due to (mis)pricing driven by noise in credit scores.

Prior research has emphasized several reasons why statistical risk prediction models can be less informative for minority groups. Minority groups are, almost by definition, represented less numerously in the data used to train algorithms such as credit scoring models. Such under-representation can be compounded by historical biases, for example redlining, which cause even fewer minority

applicants to be approved and thus represented in the training data. This results in poorer model fit whenever the true relationships in the data differ across groups but models are not allowed to condition on group membership, a phenomenon referred to as aggregation bias (Suresh and Guttag, 2019). Numerous other reasons exist as well.[1] In this paper, we remain agnostic to the specific source of differential informativeness but rather express the resulting differences in parsimonious mean-variance terms in order to quantify the economic effects.

We establish two stylized facts that motivate our analysis. First, we show that widely used credit scores are less informative for minority groups. We find that the relationship between default and credit score is less strong for minority applicants, consistent with variability, or noise, in the score attenuating the relationship towards zero. We confirm this finding using receiver operating characteristic (ROC) curves, a widely used tool in machine learning to evaluate the performance of a classification algorithm. The "area under the curve" (AUC) – a simple summary measure of the difference in predictive power exhibited in ROC curves – suggests a 5% lower fit for minority relative to non-minority applicants. A unique strength of our data is the ability to verify that these patterns are robust to including rejected applicants in the analysis – an empirical challenge often referred to as reject inference. Second, we provide evidence that mortgage lenders put equal weight on credit scores in both approval and pricing decisions despite the differences in informativeness. In line with existing evidence, we find that approval rates are lower and interest rates higher for minority applicants across the entire range of credit scores. However, we find that the dependence of the approval rates and interest rates vary identically for both groups. Put differently, lenders do not correct for statistical differences in credit score performance during underwriting and pricing decisions. This result is consistent with lenders following US fair lending regulation in this respect, as differential treatment of credit score information across groups would be prohibited.

[1] In addition to under-representation, training data might suffer from *mis-representation*. For example, if minority applicants are more likely to be given predatory loan terms and the model does not account for loan terms, minority borrowers will default more often given identical credit histories and ex-ante default risk. Separately, if default outcomes are assigned differently to minority and non-minority groups – for example, if minorities use auto lenders that are more likely to repossess cars after loan default, resulting in severely negative information on a credit report – then algorithmic decision-making can develop biases against one group or another (Rambachan and Roth, 2019; Barocas and Selbst, 2016).

Our analysis is based on two datasets that cover much of the US mortgage market. The first dataset is a panel of credit report data on 50m individuals from 2009-2016 combined with socio-economic characteristics (from a marketing dataset), information on mortgage originations (from public deed records), and lender information (from deed records and public bank data). The second dataset combines loan-level information at origination (from CoreLogic LLMA and HMDA), borrower socio-economic characteristics (from a marketing dataset) and subsequent loan performance (from CoreLogic LLMA). The dataset has several crucial advantages over existing data. First we are able to combine detailed information at origination (including credit scores, interest rates, and loan-to-value ratios) with loan performance data. Second, we are able to merge in socio-economic characters on the borrower (including race/ethnicity, age, gender). Finally, the ability to merge our data with lender-level information allows us to utilize lender-based instrumental variables, which play a key role in our empirical strategy.

To identify the precision and accuracy of credit score information, we build a model in which a risk-neutral lender decides whether to approve, and how to price, a loan based on two signals of the applicant's default risk. The first signal is the credit score, while the second signal summarizes all other information the lender may have access to. We model the differential accuracy of the credit score by allowing the signal distribution for each group to deviate from the true population mean (statistical mean bias) and to feature different variance of signal noise (imprecision). The lender is constrained by fair-lending laws to use the same decision rules for all groups as a function of these signals, and the lender's first-order condition implies an optimal signal cut-off for approval. This parsimonious model allows us to estimate the distribution of true default risk, credit score signal noise, and the signals contained in other information, and to do so across groups.

Identification stems from several key empirical facts that our merged data make available. First, we observe how the level and dispersion of credit scores differ across minority and non-minority individuals and across approved and rejected mortgage applications. The difference in these distributions, and in particular the relatively high dispersion of credit scores among accepted, minority

applicants, helps identify differences in signal noise across groups.[2] Second, we observe how credit scores differentially predict default across different borrower types and different segments of the mortgage market. Finally, using the lender-specific information in our merged data, an instrumental variables strategy identifies the characteristics of *marginally* approved mortgage applicants. Intuitively, the more noise there is in credit scores, the more our model predicts that marginal borrowers have similar credit scores to infra-marginal borrowers, and that marginal default rates are uncorrelated with marginal credit scores. Model estimation uses these empirical patterns to recover a rich set of parameters that describe consumer heterogeneity and signal precision and accuracy in the US mortgage market.

We use an instrumental variable approach to identify marginal default rates by minority group status. Our instrument exploits Community Reinvestment Act (CRA) exams following Agarwal et al. (2012). The Community Reinvestment Act, passed in 1977, mandates US financial regulators "to encourage insured depository institutions to help meet credit needs of all segments of their local communities" (FED 2018). Agarwal et al. (2012) show that banks increase lending, in particular in CRA-eligible locations, around CRA exams. This plausibly exogenous variation in approval leniency of mortgage lenders allows us to identify marginal loan applicants and their default outcomes.

The instrumental variables strategy reveals that marginally approved minority loan applicants are higher-risk than marginally approved non-minority applicants, with roughly twice as a high a probability of default within a two-year horizon. Together with our reduced-form results that minority applicants face lower approval rates, these results are consistent with two features of primitives in the market: (1) lenders are constrained to treat credit score data equally from different groups, (2) credit scores provide less informative signals for minority relative to non-minority applicants. Estimating the model on these and the other moments described above, this is indeed what we quantify: we find that credit scores feature 47% higher standard deviation in signal noise for minority mortgage applicants. Within the simple framework of the model, we estimate that roughly half of

[2]Accepted minority applicants have higher dispersion in credit scores than accepted non-minority applicants (a standard deviation of 70 rather than 63), whereas rejected applicants have nearly identical credit score dispersion for minorities and non-minorities, and in the population overall, minority credit score dispersion is in fact lower than non-minority credit score dispersion (a standard deviation of 103 rather than 108).

the gap in approval rates between minority and non-minority applicants is attributable to credit score noise alone.

Turning from the model back to the institutional details of the mortgage market, we pursue two exercises to bound the economic costs of these statistical discrepancies in dollar terms, focusing on the GSE market for reasons related to both our data and related institutional details. First, we look across the universe of US mortgage applicants, and we assess what pricing that applicant would face on a GSE-backed mortgage loan if the applicant were approved; we then re-generate credit scores for all applicants while equalizing signal noise between minority and non-minority applicants, and we calculate what share of applicants are assigned to higher or lower pricing based on GSE underwriting using these scores with newly equalized informativeness across groups. Second, we look at actual loan pricing among all approved, GSE-backed mortgage loans, and we estimate how this pricing varies with credit score; we then simulate the total dollar value in interest costs due to price changes that would result from re-generated scores that equalize signal noise between minority and non-minority applicants. These exercises are partial equilibrium in nature and do no take into account potential market-wide adjustments from changes in lender behavior, which is a broader agenda we are pursuing in companion work. While these two exercises are still in progress as of this writing, we view them as promising avenues for translating our baseline estimates into economic costs in dollar terms.

Literature Our paper is related to a large literature on discrimination in lending markets. Most of the existing literature has focused on disparities in approval rates or pricing conditional on credit scores, either on average (Bartlett et al., 2019; Bayer et al., 2016), on the margin using instrumental variables strategies (Dobbie et al., 2019), or in randomized audit studies focused on loan officer behavior (e.g., Ross et al., 2008, Hanson et al., 2016). We offer a different and complementary emphasis, showing how disparities can emerge *due to* features of credit scoring, rather than being in spite of or conditional on credit scores. Related, we also identify an important channel through which disparities can persist even when lenders transition from human to automated decision-making.

Furthermore, we illustrate a way to measure disparities across groups that is robust to concerns about infra-marginality (Ayres, 2002) or group composition differences (Kearns and Roth, 2019).

Our paper is also related to a literature on the effects of statistical technology in credit predictions. Fuster et al. (2018) study the effect of more precise default prediction technology on the number of approvals for different social groups. In contrast, our focus is on incorrect classifications of risk driving disparities across groups in the number of loans *incorrectly* approved or priced. Berg et al. (2019) and Khandani et al. (2010) study how alternative data sources can be used to predict consumer default, for example "digital footprints" or credit card transactions data.

We also contribute to a rapidly growing literature on Fairness, Accountability, and Transparency (FAT) in computer science (for summaries see e.g. Barocas et al., 2019; Kleinberg et al., 2018). This literature has largely focused on different concepts of fairness in the context of algorithmic decision-making (Corbett-Davies and Goel, 2018) and their relative merits and compatibility (Kleinberg et al., 2016). Instead of adjudicating between different hypothetical concepts of fairness, we instead focus on quantifying the economic effects of disparities driven by credit scores currently in use. Such quantification exercises, which are key for welfare considerations, have largely been absent from the axiomatic approach taken by the machine learning literature (e.g. demanding equal error rates across demographic subgroups).

Unlike the machine learning/statistics learning literature that abstracts from how lenders use the information present in credit scores, there is a large literature in economics devoted to this question. For the credit card market, Agarwal et al. (2017) show how FICO guides banks' decisions to extend credit limits, sometimes using inefficiently coarse decision rules, and Keys et al. (2012) illustrate how reliance on credit scores may disincentivize lenders from gathering other soft information that is predictive default. Related, Rajan et al. (2015) show that the increased reliance on credit scores and other 'hard' information in subprime mortgage market can lead to the approval of borrowers that are worse on unobserved dimensions. Their findings are consistent with our evidence that credit scores are less reliable signals for certain parts of the population. Our analysis of how GSE loan pricing generates differential impacts across groups also recalls evidence on regional redistribution

due to GSE pricing in Hurst et al. (2016). Similar to their approach, we show that the coarse GSE pricing leads to redistribution across borrowers although we focus on the redistribution across race/ethnicity groups and a different mechanism of redistribution.

More broadly, our work uses insights from the labor economics literature on statistical discrimination to emphasize the importance of difference in signal noise in determining outcomes across groups (Phelps, 1972; Aigner and Cain, 1977; Lundberg and Startz, 1983; Charles and Guryan, 2011; Bartik and Nelson (2019)). Whereas Bartik and Nelson (2019) provide general results about how the effects of removing information from a market affect different groups differently as a function of which groups have access to relatively precise signals (information) in the market, we emphasize here a setting where estimates of differential noise across groups are particularly important, and we also illustrate how to recover estimates of this noise without having access to empirical variation in which one signal or another is removed from a market.

This paper is organized as follows: Section 2 briefly provides describes background and institutional details. Section 3 presents motivational evidence and Section 4 develops the theoretical framework. Section 5 describes the data, the main empirical strategy, and results. Section 6 concludes.

2 Background

Since the mid-1990s, the most widely used credit score in US mortgage market is FICO, which is calculated by a proprietary algorithm licensed by the Fair Isaac Corporation to the three major credit bureaus in the US.[3] FICO uses a borrower's credit history to predict a default risk but does not use any information about a borrower's income or assets. FICO ranges from 300 to 850 or 250 to 900, depending on the scoring model. There are four requirements for a consumer to be scored: data must exist at a credit bureau, at least one trade line must be six or more months old, at least one trade line must have had activity reported within the past six months, and there must be no

[3]Competing credit scores exist, notably VantageScore, which is developed and maintained by the three main credit bureaus.

indication that the consumer is deceased.

Mortgage lenders in the US rely heavily on credit scores, in particular FICO scores, both for loan approval and pricing decisions. FICO scores are frequently used to define approval cut-offs by mortgage lenders (Laufer and Paciorek (2016); Keys et al. (2010); Bubb and Kaufman (2014)). Historically, the most prominent cut-off has been 620. Originally established as a minimum threshold by the Government-Sponsored Enterprises (GSEs) in their underwriting guidelines (Avery et al. (1996)), the cut-off emerged as a de-facto rule of thumb even among non-GSE securitized loans (Keys et al. (2010)). After the crisis, the cut-off became most relevant for the FHA market. While the Department of Housing and Urban Development, which administers the FHA program, has set minimum score requirements of 500 (580 with a down payment below 10 percent) since 2010, in practice lenders set cut-offs above those official minimums. For example, Laufer and Paciorek (2016) show that in the FHA market the effective FICO cut-off increased from 620 to 640 in 2012.

Credit scores also play an important role in pricing decisions. For example, the GSEs charge a guarantee fee ("G-fee") that covers projected credit losses from borrower defaults over the life of the loans, administrative costs, and a return on capital. These fees depend on FICO credit scores along with loan-to-value ratios, product type (e.g. fixed-rate/adjustable-rate; 15-year/30-year term), property type, occupancy status, loan purpose (e.g.purchase, rate-refi, cash-out refi) and level of documentation. The exact dependence on observables has changed over time. In particular, in 2008 GSEs introduced loan-level pricing adjustment (LLPA) to introduce more risk-based pricing of mortgages to more closely tie the G-fee to a borrower's FICO score and loan-to-value ratio at origination. While LLPAs do not apply to Ginnie Mae (FHA, VA lending), FICO scores still matter for FHA pricing.[4]

An important feature of credit scoring models is that disparate treatment statutes effectively bar the use of protected characteristics, like race/ethnicity and gender, in credit scoring models in the US. Similarly, disparate treatment statutes ban lenders from putting different weights on credit scores across different protected classes in underwriting or pricing decisions. In practice, these rules

[4]See for example: http://www.fhahandbook.com/blog/how-fha-rates-are-determined/

mean that FICO provides a single universal model and lenders are required to apply the same FICO-based decision rule to all loan applicants.[5]

3 Data and Motivational Evidence

In this section, we describe our data and we document two key empirical facts that motivate our analysis.

3.1 Data

We compile a novel, large-scale dataset of US mortgage lending from several data sources. Our first data source are loan-level origination and performance data from CoreLogic. We merge this data with borrower information from a marketing company, with a panel of consumer credit reports, and with publicly available lender data. Our second data source is a credit bureau data on a panel of 50m individuals.

CoreLogic data CoreLogic provides loan-level information on about 60% of mortgages originated in the US. The data include detailed information at origination, including FICO scores, loan-to-value (LTV), debt-to-income (DTI) ratio, loan documentation level, loan type (insured by the GSE or insured by the FHA), and interest rate. The performance data provides information on mortgage payments and default. We merge additional information on the mortgage holder from a marketing company, which provides name, age, and address histories of a large number of individuals residing in the US. Our merge keys are the property address and the date of the loan origination. The marketing dataset allows us to impute race/ethnicity and gender based on a commonly-used algorithm that combines name and geographic information.

The merged data contains almost 20m loans and spans 1998-2017. For comparison, HMDA records around 100m loans in the same time period (see Table 3 for annual coverage statistics).

[5]These laws include the Community Reinvestment Act, the Fair Housing Act, the Equal Credit Opportunity Act, and implementing regulation (see e.g. 12 CFR 1002.4).

Table 1 shows that our loan characteristics are broadly comparable: Loan sizes, average interest rate, LTV and the share of FHA loans are generally comparable. Relative to HMDA, our dataset diverges in one important regard. CoreLogic does not collect information from all US lenders tilting our sample to loans securitized and sold either to the GSEs or to Ginnie Mae (FHA loans). Our results should therefore be interpreted as statements about the originate-to-distribute mortgage market. Table 2 presents summary statistics on the demographics of our sample. Relative to HMDA, we get a lower share of whites and African-Americans. This is driven by the fact that our race/ethnicity imputation, described more below and in a future appendix, only assigns a race/ethnicity in cases where either a person's name or geography (or both) suggest a high probability – over 95% – of particular group membership. Since white and African-Americans are slightly more likely to have ambiguous results in this imputation exercise, we have a higher share of non-classified mortgage holders.

We also merge in information on the lender originating the mortgage exploiting that lender names are recorded on public deed records for each recorded mortgage. We add information on the timing of lenders' Community Reinvestment Act examinations, which are publicly available on the CRA website. Our sample includes an average of 1000 CRA exams per year for 3,949 distinct mortgage lenders. On average, banks undergo 3 CRA exams in our sample period.

Credit bureau data Our second dataset contains credit report data on 50 m individuals between 2009-2016. Similar to our CoreLogic sample, we merge in additional information from a marketing dataset and public information on mortgage lenders. The credit report data allows us to observe rejected loan applications. We infer a rejection in any instance where we see an application for a new mortgage and no new mortgage origination in the subsequent 6 months. We infer applications from a hard credit inquiry by a mortgage lender. The credit report data also provide a supplementary measure of subsequent loan performance. The availability of credit scores in the credit report data also allows us to measure how the mean and variance of credit scores differs across accepted and rejected applicants, which is informative about the amount of noise in credit scores relative to other

information sources used in approval decisions.

Race and Ethnicity Imputation We use a two-step procedure to impute race and ethnicity in the marketing dataset that provides names and addresses, which we merge to both the CoreLogic and credit bureau data. First, we follow an approach developed by Sood and Laohaprapanon (2018).[6] which models the relationship between the sequence of characters in a name and race and ethnicity using Florida Voter Registration data. After implementing their approach, we update each individual's baseline racial/ethnic probabilities with the racial and ethnic characteristics of the census block associated with her or his place of residence in 2000 using Bayes' Rule, and we then update the Bayesian posterior again using an individual's 2010 address and the 2010 census data. An individual is assigned to a racial/ethnic category if this category has the highest posterior probability for that individual. This two-step method is similar to methods used by the CFPB to construct race and ethnicity in fair lending analysis. Work by the CFPB (2014) and Elliott et al. (2009) shows that combining geographic and name-based information outperforms methods using either of these sources of information alone. In ongoing work, we validate our results in a sample of originated loans matched to HMDA records which contain self-reported race and ethnicity.

3.2 Empirical Facts

We document two key empirical facts that motivate our analysis, First, we show that widely used credit scores have lower classification accuracy for minority groups. Second, we show that lenders do not correct for statistical differences in scores.

Fact 1: Credit Scores Have Different Informativeness Figure 1 plots realized default outcomes against credit score at origination by minority status for both purchase and refinance mortgages. In line with existing research from data in the 2000s (Avery et al. (2009), Avery et al. (2012)), we find that default rates are higher for minority mortgage borrowers for all levels of credit scores. For example, at FICO 700, non-minority borrowers have a 2% chance of being 90 days (or more)

[6]Their approach is implemented in a Python package available at https://github.com/appeler/ethnicolr.

delinquent on a mortgage payment over a 24 month period whereas minority borrowers have 5% change of being similarly delinquent.

We add to this existing evidence by showing that the slope between default and credit score is flatter for minorities, consistent with higher noise – intuitively, measurement error – in minority scores attenuating the relationship towards zero. To understand this result, consider regressing the log odds ratio of default on a signal a of default risk (credit score) for each group g.

$$\log(\mathrm{pd}_g/(1-\mathrm{pd}_g)) = \alpha + \beta \mathrm{a}_g + \epsilon_g \tag{1}$$

where pd_g is probability of default for individuals in group g, conditional on a realization of the signal or score a. Assume the signal is measured with noise $\mathrm{a}_g = \theta_g + v_g$. If the signal has more measurement error for the minority group, classical errors-in-variable theory indicates that the coefficient for that group is more heavily attenuated towards zero. Table A1 in Appendix A confirms that the differences in slopes are statistically significant.

Figure 2 presents a different way of presenting this stylized fact using receiver operating characteristic (ROC) curves, which are a popular tool in machine learning to assess model fit. ROC curves show the ability of a prediction algorithm to predict "goods" (non-defaulting loans) separately from "bads" (defaulting loans) as the approval threshold is varied. A good predictor maximizes the fraction of "goods" approved (referred to as a true positive rate) and minimizes the number of "bads" mistakenly approved (referred to as a false positive rate). While a typical ROC curve plots true positive rates against false positive rates for a curve traced out by different approval thresholds, Figure 2 simplifies this exposition by showing what fraction of "goods" and what fraction of "bads" are approved at each possible credit score cut-off. For example, if a lender approved every applicant with a credit score of 700 or above, she would admit 70% of "goods" for the non-minority population but only 50% of the "goods" in the minority population.

The key take-away from Figure 2 is that both true positive and false positive rates for minority borrowers lie below the respective rates for non-minority borrowers. This difference implies that

credit scores are poorer predictors for minority borrowers. Based on Area Under the Curve (AUC) measures which summarize the differential model fit across all possible credit scores, fit is 5% higher for non-minority applicants (an AUC of 0.78 for non-minority versus 0.74 for minority applicants).

A common challenge with assessing the informativeness of credit scores arises because we do not observe default outcomes for rejected applicants – referred to as the problem of reject inference (Thomas, 2009). We address this challenge by using several ways of imputing outcomes for rejected applicants. In Figure 2, we show default on other loans (e.g. auto, credit cards, personal loans) which exploits that many rejected applicants have other types of loan applications.

Fact 2: Lenders Put Equal Weight on Credit Scores The second key stylized fact is that lenders do not correct for differences in statistical accuracy in their mortgage underwriting and pricing decisions. Figure 3 shows rejection rates for mortgage applicants as a function of their credit score at the time of application in a given county-time period. In line with existing research (e.g. Bartlett et al., 2019), we find lower approval rates for minority applicants of about 7 pp.[7] However, our key finding is that the approval rates vary in the same way across credit scores for minority and non-minority applicants. In other words, even though moving from a lower to a higher credit score is less informative for minority applications, a higher credit score has symmetric effect on the lender's decision to approve a loan. In the appendix, we show that this fact holds if we additionally control for income, presence of co-applicants, and the estimated loan-to-value and debt-to-income ratios on their loan application.

We now document that the same fact also holds for pricing decisions. Figure 4 shows the relationship between interest rates and credit score at origination residualized on rate fixation, LTV, original loan balance, loan type, term, DTI, year and lender. We again confirm existing research that finds higher interest rates for minority borrowers for purchase loans.[8] However, our key finding is that interest rates vary symmetrically with credit scores for both non-minority and minority groups. Unlike with approvals, we can now condition on the same mortgage lender making loans

[7] These numbers are similar to results in Bartlett et al., 2019 who find mortgage rejection rates that are between 9.6 and 7.3 pp higher for minorities.

[8] However, for cautionary evidence on this point, see citetbhutta2019minorities.

in the same county. This result suggests that our results are driven by within-lender as opposed to between-lender variation.

Together these facts suggest that lenders do not treat credit score signals differently across groups, despite differences in scores' informativeness. This equal treatment is likely motivated by fair-lending law that explicitly requires equal underwriting processes for all applicants regardless of group membership. As we develop more next, a key implication of these patterns is that inequality in information across groups can undermine the goals of fair-lending restrictions and can exacerbate between-group disparities.

4 A Model of Information in Mortgage Lending

We develop a model to distinguish algorithmic noise from bias, to differentiate both of these from human animus, and to illustrate the potential costs of signal noise in particular. We then use the model to design an estimation strategy to quantify these different informational frictions empirically.

4.1 Setup

We model a lender who receives algorithmic signals $a_j \in \mathbb{R}^J$ from algorithms $j = 1 \dots J$ and receives subjective signals $s_k \in \mathbb{R}^K$ from sources $k = 1 \dots K$; the latter may include, for example, an impression formed in an interview or private information gathered through relationship lending. These signals provide information about a loan applicant's true type $\theta \in \mathbb{R}$.

We define noise and bias in a standard way relative to the true type θ. For each signal a_j or s_k, the signal error v_j or u_k is the gap between the signal value and the true type:

$$a_j = \theta + v_j$$

$$s_k = \theta + u_k$$

Then the bias and noise in an algorithmic signal are respectively $\mu_j = \mathbb{E}[v_j]$ and $\sigma_j^2 = \text{Var}(v_j)$. We

define μ_k and σ_k^2 for subjective information analogously.

Our goal will be to show how to estimate bias and noise terms such as μ_j, σ_j^2, μ_k, and σ_k^2 from data on lenders' approval decisions and and data on loan performance (default). Naturally, this depends on the process governing loan default and the approval decisions that lenders make.

We specify that approved loans default with probability $\delta(\theta)$, which depends on the borrower's type. Meanwhile lenders make approval decisions based on their expected profit $\hat{\Pi}$, which, for any given realization of signals $\mathbf{a}$ and $\mathbf{s}$, depends on default risk as well as revenue and other costs. Specifically, we model expected profits as,

$$\hat{\Pi}(\mathbf{a}, \mathbf{s}) = \alpha + \beta R(\mathbf{a}, \mathbf{s}) - \gamma \mathbb{E}[\delta(\theta) \,|\, \mathbf{a}, \mathbf{s}] - c \tag{2}$$

Here the parameters α, β, and γ respectively capture non-interest revenue, the interest revenue retained by the lender given net interest rate R, and the lender's exposure to losses from loan default. Meanwhile the per-loan marginal cost c captures other variable costs not driven by default. This is a relatively general profit function because it describes both portfolio lending, where the lender fully internalizes interest revenue and default losses, as well as originate-to-distribute and GSE-funded business models, where lenders share revenue and risk externally.

A risk-neutral lender approves all loans for which $\hat{\Pi}(\mathbf{a}, \mathbf{s}) > 0$. How do lenders decide which combinations of signals $\mathbf{a}$ and $\mathbf{s}$ imply positive expected profits? We focus on cases where a one-dimensional posterior belief about an applicant's type θ is sufficient for optimal pricing and for expected default. Such cases include, as we develop further in the next subsection, "normal-normal" settings where both true types and signal noise are normally distributed.

In such a setting, lenders form a posterior $x = x(\mathbf{a}, \mathbf{s})$ based on the signals they observe about an applicant's true type. We suppose lenders know the true distribution of signal noise except for the mean, and lenders *believe* the signal noise is mean zero. In other words, whenever signal biases μ_j or μ_k are nonzero, lenders are unable to "de-bias" the signals, which allows our sense of bias in a signal to still have empirical content. For a signal a_j with true CDF denoted F_j, let the distribution

$F_j^0(t) = F(t - \mu_j)$ denote the noise-mean-zero translation of F_j, and let the joint distribution across all signals $F_{\mathbf{a},\mathbf{s}}^0$ be defined analogously. Lenders then form subjective Bayesian posteriors about θ as,

$$g(t \,|\, \mathbf{a}, \mathbf{s}) = \frac{f_{\mathbf{a},\mathbf{s}}^0(\mathbf{a}, \mathbf{s})}{f_{\mathbf{a},\mathbf{s}}^0(\mathbf{a}, \mathbf{s} \,|\, t) h(t)} \tag{3}$$

where H is the CDF of true types θ. The posterior belief x about θ is then simply $x(\mathbf{a}, \mathbf{s}) = \int_t t \, dG(t \,|\, \mathbf{a}, \mathbf{s})$.

The sufficiency of these posteriors implies lenders' approval rules can be described by a set $\mathbb{X} \in \mathbb{R}$ of values for x that yield positive expected profits. Consistent with the institutional details in much of the mortgage market, we further assume that $\mathbb{X}$ is a half-line $[\hat{x}, \infty)$. Given this form for the approval region,[9] we summarize lender approval strategies using the cutoff value $\hat{x}$, and we write lender approval decisions as $d(x) = \mathbf{1}\{x > \hat{x}\}$.

When lenders face applicants from observably different groups, for example applicants from racial or ethnic majorities or minorities, these groups may differ both in their signal noise and biases, and in lenders' propensity to exhibit animus towards these different groups. In our empirical implementation, we consider various versions of how lenders may or may not do so. In general, any of the above features in our model setup may admit a group-specific subscript m to signify either that primitives may differ across groups or that lender behaviors or beliefs differ across groups.

4.2 Bias vs. Noise

Before turning to our empirical implementation, we first present a theoretical result highlighting why canonical empirical tools used to identify human animus towards some group – distinct from those tools in our later empirical strategy – may be difficult to interpret in an era of algorithmic noise and bias.

Proposition: Suppose two groups m and m' face different signal noise from an algorithmic signal, with $\sigma_{jm}^2 < \sigma_{jm'}^2$, and suppose lenders are constrained by regulation to weight this algorithmic

[9] In a future appendix we explore conditions on optimal pricing $R^\star(x)$ relative to expected default rates $\mathbb{E}[\delta(\theta) \,|\, x]$ that are sufficient for the approval region $\mathbb{X}$ takes this form.

signal equally in loan approval decisions, such that when lenders follow a cutoff rule $d(x) = \mathbf{1}\{x > \hat{x}\}$, the relevant posterior is equalized across groups whenever other signals are equal, $x_m(a_j, \mathbf{a}_{-j}, \mathbf{s}) = x_{m'}(a_j, \mathbf{a}_{-j}, \mathbf{s}) \, \forall a_j$. Then marginal applicants from group m' have lower default than marginal applicants from group m,

$$\mathbb{E}[\delta(\theta_{im}) \,|\, x_{im} = \hat{x}] > \mathbb{E}[\delta(\theta_{im'}) \,|\, x_{im'} = \hat{x}]$$

Remark: Bias in such a signal also easily generates the same result, without signal noise.

The proposition says that algorithmic noise can generate *Becker test failures* in the sense of Becker (1957), whereby differences in loan performance between two groups' marginally approved applicants is seen as evidence for discrimination against one group. While this test has been highlighted in recent work as potentially also rejecting equality across groups because of misalignment in incentives between banks and their loan officers (Dobbie et al., 2019), to our knowledge this is the first result to illustrate the potential for Becker test failures due purely to differences in algorithmic signal noise across groups.

The second premise in the proposition focuses on regulation that constrains algorithmic signals to be weighted equally across groups in underwriting decisions. This premise reflects the effect of *disparate-treatment* regulation in US fair lending standards: two groups with the same observable characteristics cannot be treated systematically differently by lenders. To this end, it is important that algorithmic inputs in underwriting by *verifiable*, and this is why we emphasize algorithmic inputs in the proposition's first premise. As a mnemonic, recall from the preceding subsection that errors in algorithmic or verifiable signals were denoted v, and errors in subjective or unverifiable signals were denoted u.

To see the intuition for this result, consider the case where $J = 1$ and $K = 0$, so that $x(\mathbf{a}, \mathbf{s}) = a_j$, and for this one algorithmic signal, suppose the signal is completely uninformative for one group and perfectly informative for another group. Both groups have the same expected profit of zero, and suppose there are some types the lender would like to reject. Given the signal for one group

is completely uninformative and that group is on average positive-payoff, the lender would like to set $\hat{x} = -\infty$ for that one group and $\hat{x} = 0$ for the other group. However given the constraint of equalizing the treatment of a_j across groups, the lender chooses a common threshold across groups, and this generates a Becker test failure: worse-than-average applicants are rejected from one group, and while rejected applicants from the other group are all no different from average.

This claim motivates our empirical work to quantify the role of algorithmic noise and bias relative to human informational frictions and animus.

4.3 Estimating the Model

We study a mortgage lender's problem of which applications to approve given noisy information about default, using an empirical analog to the setup introduced earlier in this section.

To make empirical headway, we suppose that true types for individuals i from each applicant group m are normally distributed with potentially group-specific mean and inverse variance,

$$\theta_{im} \sim \mathcal{N}\left(\mu_m, \frac{1}{h_{0,m}}\right) \tag{4}$$

Likewise signal noise is normally distributed with potentially group-specific mean and variance,

$$v_{ijm} \sim \mathcal{N}\left(\mu_{jm}, \frac{1}{h_{0,j,m}}\right), \;\; u_{ikm} \sim \mathcal{N}\left(\mu_{km}, \frac{1}{h_{0,k,m}}\right) \tag{5}$$

This structure implies the lender's Bayesian posterior of each applicant's type is,

$$x_{im} = \frac{\sum_{j=1}^{J} h_{i,j,m} a_{ij} + \sum_{k=1}^{K} h_{i,k,m} s_{ik} + h_{0,m}\mu_m}{\sum_{j=1}^{J} h_{i,j,m} + \sum_{k=1}^{K} h_{i,k,m} + h_{0,m}} \tag{6}$$

We also map from types, which are unobserved, to observable default, via default probabilities parameterized as,

$$\delta = \frac{\exp(\theta)}{1 + \exp(\theta)} \tag{7}$$

We also follow the standard parameterization in the credit scoring industry (Thomas, 2009) and suppose that observed credit scores are an affine transformation of a credit scoring agency's predicted log odds of default,

$$FICO_i = a_1 + a_2 * log\Big(\frac{\delta(a_{i,j})}{1-\delta(a_{i,j})}\Big)$$

With these parameterizations in hand, we use a simulated method of moments strategy to estimate the key primitives of the environment: two parameters for each group describing the distribution of true types for that group's loan applicants, μ_m and h_m; signal bias and noise for each group and for each signal, μ_j and h_j (for algorithmic signals) and μ_k and h_k (for subjective signals); the scalar parameters of the affine mapping from credit report signals to FICO scores, a_1 and a_2; and finally the approval threshold $\hat{x}_m$ for each group.[10] Note that the existence of an optimal cutoff strategy for loan approval $\hat{x}$, as described in the setup earlier in this section, implies we do not need to estimate the profit function parameters directly in order to recover these primitives on signal bias and noise.

To estimate these parameters, we simulate the following moments based on the setup and the parameterizations above, and we match these to corresponding empirical moments: loan approval rates for each group of applicants; the mean and variance of FICO scores among accepted applicants in each group; the mean and variance of FICO scores among *rejected* applicants in each group; the correlation of default probabilities with FICO scores among approved loans; the average default rate among approved loans; and the default rate of *marginally* approved loans. This final moment corresponds to the expectation $\mathbb{E}[\delta(\theta_{im}) \,|\, x_{im} = \hat{x}]$ highlighted in the prior subsection.

As an illustration of the model identification, in Figure 6 we present graphs similar to Figure 1 from our earlier motivating evidence in section 3.2 where we now estimate estimate the relationship between ex-post default and credit scores using simulated data. The simulated data include

[10] Note that the optimal approval threshold can differ across groups even if lenders do not exhibit animus towards one group and even if other profit function parameters are the same across groups, because of non-linearity in δ will make groups with noisier signals exhibit higher average default for the same posterior x.

cases with and without between-group differences in signal noise, as well as cases with and without between-group differences in the mean-bias of signals. While these graphs do not account for potential differences in approval thresholds across groups, or differences in the underlying distribution of default risk across groups, these point to patterns in the data that help identify some of the key model moments. Interestingly, the simulated data that most closely match the empirical patterns seen previously in Figure 1 come from panel (d) of the new Figure 6, which corresponds to the case where signal noise is greater for minority applicants, and yet mean-bias *increases* the scores of minority applicants relative to what would be consistent with ex-post default rates.

The full estimation strategy involves estimating $2 + 3 \times M + 2 \times M \times J + 2 \times M \times K$ model parameters when there are M groups, J algorithmic signals, and K subjective signals, using $8 \times M$ empirical moments. So for $J = 1$ and $K = 1$, which corresponds to lenders using FICO scores together with a single subjective signal, the parameters are either identified or over-identified for any number of groups $M \geq 2$; in our main empirical application we focus on the case with $J = 1$, $K = 1$, and $M = 2$, such that we recover 16 model parameters from 16 empirical moments.

5 Empirical Strategy and Preliminary Results

5.1 Construction of Instrument

Important in our empirical strategy is the ability to identify the characteristics of *marginally* approved mortgage applicants. We use CRA exams as an instrumental variable to obtain exogenous variation in the approval leniency of banks.

The Community Reinvestment Act, passed in 1977, directs federal banking regulators "to encourage insured depository institutions to help meet credit needs of all segments of their local communities" (FRB, 2018). The Federal Reserve, FDIC and the OCC periodically conduct exams to evaluate whether lenders comply with the CRA lending requirements. Unsatisfactory CRA exams can be grounds for the regulator to deny applications for merger/acquisitions or the opening of new branches.

Compliance is measured in loan quantity (not in terms of the interest rates changed on the loans). Two types of loans satisfy the CRA requirements. First, any loans made in low- and moderate-income areas where the bank has a physical branch and takes deposits, or made in moderate-income areas deemed depressed or under-served by banking regulators ("CRA-eligible census tracts"). Second, lending to low- and moderate-income individuals in any area regardless of average income.

We construct the CRA instrument following Agarwal et al. (2012) and consider a six-quarter window around a lender's CRA exam. These exams occur at fixed intervals – every five (resp. two) years for banks with assets under (resp. over) $250 million – so as argued by Agarwal et al. (2012), the timing of these exams is likely unrelated to local economic conditions and, in particular for our setting, credit demand and credit risk.

To establish that our instrument affects origination, we estimate the following first stage in a quarter (t) - tract (g) - bank (b) panel, separately by observable groups m:

$$\begin{aligned} Q_{bgmt} = \beta_m \text{CRA exam}_{bt} \times \text{CRA tract}_{gt} + \gamma_m \text{CRA exam}_{bt} \times \text{CRA ineligible tract}_{gt} \\ + \alpha_{bgm} + \alpha_{gmt} + \epsilon_{bgmt} \end{aligned} \tag{8}$$

where Q_{bgmt} is the quantity of loans originated, α_{bgm} is a bank-tract fixed effect for group m and α_{gmt} is a tract-quarter fixed effect for group m. We weight by the bank's origination share in a given tract. Standard errors are clustered at the tract level.

To capture potential differences between CRA-eligible and CRA-ineligible regions, we estimate separate coefficients for CRA eligible and ineligible tracts. For example, lenders may substitute lending from CRA-ineligible areas to CRA-eligible areas in the time period around an exam, so it is important to allow the two first-stage coefficients to have different sizes or signs. We define a CRA-eligible tract using low- and moderate-income status relative to area median income as defined by the FFIEC.[11] These two coefficients are identified off of *within*-tract-time variation by comparing lenders that are and are not subject to a CRA exam at a particular time.

[11] We focus on CRA-eligible tracts that are eligible by virtue of their income level, rather than higher-income tracts deemed distressed or under-served in a given year; these counties include the vast majority of loans made in CRA-eligible tracts.

We restrict the estimating sample in several ways. First, we include only county-bank combinations in which banks make a substantial number of loans per year. Our goal in doing so is both to approximate the CRA assessment areas in which banks are incentivized to expand CRA lending – these include areas where the bank has branches, takes deposits, or has issued a "substantial" portion of its loans – and also to ensure the bank-county-specific and bank-time-specific fixed effects in our baseline regression are well identified. Because of ambiguity about what makes a number of loans "substantial" for sake of defining assessment areas, we select a cutoff number of loans per year at the county-bank level so that banks on average about 80% of loans are included in these bank-county combinations, which matches the approximate share nationally covered by bank assessment areas. This results in us keeping markets with at least 100 refinance loans or at least 20 purchase loans per bank-county per year.

Second, we also restrict the sample to years prior to 2011, as existing evidence suggests the CRA may have been most binding during this period (Bhutta, 2011) and we divide the sample by purchase and refinance loans separately. Because refinance loans constitute the bulk of all mortgage loans, we focus on refinance loans in our main results and present results for purchase loans in the appendix. The institutional details of the refinancing market also make our instrument and estimates more straightforward to interpret for refinance loans than for purchase loans: the much smaller market share of FHA and VA loans among refinances means that there are fewer margins of adjustment for lenders in response to the instrument, and smaller effective choice sets of borrowers; we discuss this choice in the context of the standard IV "monotonicity" assumption more in the following section.

5.2 Instrument validity

Our identification assumption is that a CRA exam affects origination only through the increased leniency of the lender. The exclusion restriction would be violated if there are other time trends coinciding with the CRA examinations driving loan demand at the examined lenders. Including county-time fixed helps to address these concerns. Additional loan-level controls are only required if we believe that that the instrument is only valid conditional on these additional controls; we view

this as unlikely, given how the timing of CRA exams is set independently of local market conditions.

We see three reasons why a bank may choose to defer CRA-eligible lending until just before an exam. First, CRA-induced lending is, almost by definition, lower-NPV loans than other loans for the bank: these are the marginal loans the bank is only induced to make because of the CRA restrictions. Making these lower-NPV loans as late as possible – just before an exam – therefore maximizes overall value for the bank. Second, there may be option value in waiting as long as possible to increase CRA lending, as the bank potentially will issue sufficient CRA-eligible loans without needing to change leniency at all, for example if there is a house price boom in particular areas or if an interest rate decrease induces high refinancing activity. Third, there may be behavioral or agency frictions that make it difficult to incentivize loan officers to make CRA-eligible loans, which results in under-production of CRA loans that the bank has to correct for just before its exam.

To further validate the instrument, we examined CRA exam reports to find evidence of institutions responding to CRA incentives with expanded lending just before its exam date. An apt example comes from Fifth Third Bank's 2016 CRA exam (emphasis added):

> In 2015, Fifth Third launched the "*Community Reinvestment* Mortgage Special," whereby all lender fees are waived for borrowers purchasing properties located in a *low-income tract*. Fifth Third's use of these programs was evaluated when there was sufficient volume necessary to conduct a meaningful analysis. Overall, Fifth Third made extensive use of flexible lending practices in serving *low- and moderate-income* needs within its assessment areas.

Both the use of "community reinvestment" in the program name and the emphasis on low- and moderate-income tracts point to this expanded lending being directly in response to CRA incentives; the regulator's report also notes how the the bank made "extensive use" of such incentives more broadly.

To illustrate the validity of the instrument in a statistical rather than anecdotal sense, in Table 4 we present the results of estimating the first stage (equation 8) on refinance loans in our data.

We do so separately for the two prominent subgroups in our analysis, minority and non-minority mortgage applicants. Analogous results for purchase loans are presented in Appendix Table A2.

Instrument validity also requires two assumptions that we view as reasonable in our setting – an exclusion restriction that the timing of CRA exams is independent of credit demand and credit risk, and a monotonicity condition that the CRA exam does not cause any applicants in CRA-eligible tracts to be denied credit when they otherwise would be approved (i.e., no "defiers" in the language of Angrist and Pischke (2008)). Our focus on refinance loans helps support the validity of the monotonicity condition in our setting, because borrowers' smaller choice sets for refinance loans implies that lenders' response to the instrument is less likely to change the mix or characteristics of multiple products available to borrowers, and rather, lenders' response is likely to entail an across-the-board credit expansion.

5.3 Empirical specification

We use the first-stage variation in loan volume from the previous section to estimate the default rate of marginal borrowers. We can recover these marginal default rates in a second stage regression of loan default volume on total (instrumented) origination volume. Specifically, the second stage at the quarter (t) - tract (g) - bank (b) level is, separately by observable groups m:,

$$Y_{bgmt} = \beta^{IV} Q_{bgmt} + \alpha_{bgm} + \alpha_{gmt} + \eta_{bgmt} \tag{9}$$

where Y_{bgct} is the volume of defaulted loans of group g (minority or non-minority), Q_{bgct} is the (instrumented) volume of loans originated, and α_{bgm} and α_{gmt} are tract-time and bank-tract fixed effects respectively. Intuitively this regression captures in the coefficient β^{IV} the relationship $\partial Y/\partial Q$, or a marginal default rate, evaluated at equilibrium quantities in the absence of an exam. For example, if the instrument leads banks to originate 10 additional loans and we observe 3 additional defaults, then the 2SLS estimator would estimate a default rate of 0.3 or 30% among the marginal loan applicants. As in the first stage, we weight by the bank's origination share in a given tract,

and standard errors are clustered at the tract level.

We focus on default at a two-year horizon, where we define default as non-payment for at least ninety days. This is the same default outcome that standard credit scores are designed to predict (Federal Reserve Board (2007)).

5.4 Results

Table 5 shows our second stage estimates for refinance loans, and Appendix Table A3 shows analogous results for purchase loans. In each table we show results for minority and non-minority borrowers separately. We present the 2SLS estimates of *marginal* default rates alongside standard OLS estimates of *average* default rates to show both moments of the default-probability distribution that we later target in the structural model estimation.

Our baseline estimate of the default rate for marginally approved minority applicants is roughly twice as high as our estimate for marginally approved non-minority applicants, at 11.3% and 5.2% respectively. Meanwhile both groups feature, as would be consistent with lenders using approval strategies summarized by cutoff rules, substantially higher marginal than average default rates. For minorities, the marginally approved applicant is roughly four times as likely to default as the average minority borrower, and for non-minorities, the marginally approved applicant is roughly eight times as likely to default as the average non-minority borrower.

We use these four coefficients and other moments of the data to estimate our model of mortgage lenders screening applicants using noisy credit scores and other signals. Specifically, we use the moments outlined in section 4.3, which we summarize in Table 6. These moments also include the slopes and intercepts of the linear relationships in the motivating evidence in Figure 1, as well as the mean and variance of FICO scores among accepted applicants in each group, and the mean and variance of FICO scores among *rejected* applicants in each group. We estimate the sixteen model parameters described in section 4.3, including measures of human animus, first-moment differences across groups in credit-score and non-credit-score signals, and the variance of noise in credit-score and non-credit-score signals.

Figure 7 presents a summary of some of our current parameter estimates. In reviewing our current results, we focus on estimates of credit score noise, where we find that credit scores feature 47% higher standard deviation of signal noise for minority mortgage applicants. Fuller estimation that allows full flexibility across the different model parameters while using all available data moments is still in progress. However, one way to interpret the magnitude of this estimate of credit score noise is already available: we keep other model parameters at their baseline value and allow minority applicants access to equally informative credit scores as non-minority applicants have, and then we re-estimate the model to find whether and how the gap decreases between minority and non-minority approval probabilities. In preliminary results, we estimate that equalizing credit score signal noise across groups reduces roughly half of the gap in approval rates between minority and non-minority applicants.

5.5 Ongoing Work: Pricing Errors and Credit Score Noise

In ongoing work, we quantify how differential noise in credit score signals leads to systematic mispricing of GSE-backed mortgage loans. We focus on GSE loans for two reasons. First, they are prevalent: they cover the vast majority of the refinance market and, as was shown earlier in Figure 5, between 25% and 50% of the purchase-mortgage market depending on credit score. Second, focusing on GSE loans makes it straightforward to measure how noise in a borrower's FICO score can change the guarantee fee, or G-fee, based on which the borrower's loan is priced, leveraging the relationship discussed in section 2 between these fees and FICO scores.

We pursue two exercises to quantify such mispricing. First, looking across the universe of US mortgage applicants, we assess what pricing each applicant would face on a GSE-backed mortgage loan if she were approved. To do so, we assume that each applicant is seeking a loan with the same average characteristics (e.g., LTV) as are chosen by approved applicants with her same FICO score. Essentially this relationship between FICO and pricing is the same as the pricing gradient shown in Figure 4. For all applicants, we then probabilistically re-generate credit scores while equalizing signal noise between minority and non-minority applicants. Finally, we calculate what share of applicants

are assigned to higher or lower pricing based on the guarantee fees assigned by GSE underwriting, under these new counterfactual scores relative to their actual score.

There are two key steps in generating these counterfactual credit scores. First, using our model parameter estimates, we map each observed FICO score probabilistically to a true underlying default risk, effectively "inverting" the realized measurement error in observed FICO scores. Second, we then simulate new credit scores by drawing from the counterfactual signal distribution that equalizes noise across groups.[12]

Our second exercise is similar to the first in that we again simulate counterfactual credit scores after equalizing noise distributions across groups. However in this exercise, we restrict our focus to actual loan pricing among all approved, GSE-backed mortgage loans – that is, borrowers rather than applicants – and we estimate the total dollar value of how pricing would change under the counterfactual scores. Whereas the counterfactual scores naturally imply a change in interest rates (rather than in dollar terms), we dollarize these interest rate changes under two plausible scenarios: in the first scenario, borrowers choose the same LTV and and loan amount, and have the same effective (i.e. prepayment-risk adjusted) loan duration, as they chose given their actual FICO score; in the second scenario, borrowers choose the same location in the FICO-conditional distribution of LTV and loan amount at their counterfactual FICO score as they chose at their actual FICO score, and they face the same effective loan duration as does the average borrower at their counterfactual FICO score.[13] These two scenarios span two extreme cases of how product choice and prepayment behavior covaries with credit score noise, and accordingly, we view these two scenarios as providing reasonable bounds for each individual on how credit score noise leads to systematic mispricing given actual mortgage product choices.

These exercises are partial equilibrium in nature and do no take into account potential market-

[12]In order to provide a conservative estimate about how much score reassignment this generates, we impose that an individual's true-type-conditional (i.e. θ-conditional) rank in the the counterfactual noise distribution is the same as the individual's true-type-conditional rank in the actual noise distribution. In other words, for individuals for whom actual FICO scores are high (low) relative to their true type, we impose that counterfactual credit scores are also high (low) relative to their true type – the counterfactual scores are just *less* high (low) in cases when the counterfactual score is less noisy than the actual score.

[13]We follow guidance from the Securities Industry and Financial Markets Association (SIFMA) in calculating effective loan durations and in translating these into borrowers' expected payments.

wide adjustments from changes in lender behavior, which is a broader agenda we are pursuing in companion work. While these two exercises are still in progress as of this writing, we view them as promising avenues for translating our baseline estimates into economic costs in dollar terms.

6 Conclusion

With the rapid growth of automated decision-making in US credit markets, there are newly important questions about how algorithmic underwriting can reduce – or fail to reduce – inefficiencies and disparities in credit access. Focusing on widely used credit scores in the US mortgage market, we show that credit scores are differentially informative for different groups that are emphasized in US consumer protection law, in particular for racial and ethnic minorities, and that this leads to substantial disparities in loan approval and pricing. First using reduced-form evidence that has an intuitive errors-in-variables interpretation, and next using a structural model that allows for more economically meaningful quantification, we estimate that commonly used credit scores feature 47% higher standard deviation of signal noise for minority mortgage applicants than for non-minorities. Identification stems from novel moments in credit report data and from an instrumental variables strategy that recovers the unobserved default risk of marginally approved mortgage applicants. We preliminarily estimate that this inequality in signal noise accounts for roughly 50% of the gap in approval rates for minority relative to non-minority mortgage applicants, and we work towards quantifying in dollar terms how this leads to systematic mispricing of GSE-backed mortgages.

As lenders move to more complex, machine-learning based scoring models, disparities due to differential accuracy or precision will continue to be an important concern. While diagnosing the source of differential informativeness is beyond the scope of this paper, many of the potential sources of differential model fit – e.g. historical bias, unequal representation in training data, and legal requirements not to condition on protected class – are likely to remain with us for some time to come. Understanding and quantifying the source of differential informativeness and developing potential solutions is a fruitful avenue for future research that we are pursuing in companion work.

Tables and Figures

Table 1: Summary Statistics

	Purchase Loans + Refinancing (1990-2017)			**Purchase Loans (2005-2017)**
	CL + Infutor	**CL + Infutor + Banks**	**CL + Infutor + Banks + LLMA**	**CL + Infutor + Banks + HMDA**
Mortgage Amount	159,586.43	162,186.43	213,171.12	197,029.91
	(145,327.19)	(155,611.89)	(180,851.79)	(116,989.58)
Interest Rate	5.90	5.25	4.96	5.09
	(2.06)	(1.78)	(1.28)	(1.45)
LTV	82.93	82.84	76.02	84.49
	(14.96)	(15.11)	(21.25)	(14.21)
FICO	-	-	735.48	-
			(59.48)	
DTI	-	-	36.06	-
			(15.94)	
Default (Delinquency)	-	-	0.10	-
			(0.86)	
FHA share	10.15%	8.23%	14.41%	23.22%
Portfolio share	-	-	14.97%	17.74%
N (total count)	201,665,930	80,898,786	19,467,279	9,780,865

Note: This table reports descriptive statistics on three different loan-level (purchase and refinancing) samples over the period 1990-2017, and an additional sample restricted to purchase loans from 2005-2017. The first column refers to observations in Corelogic (CL) Deed for which we have matched Infutor address and name. The second column further restricts the sample to those observations with non-missing bank information. The third column adds the condition of a match in the CL Loan-Level Market Analytics dataset. The last column refers to HMDA-matched CL Deed observations with Infutor address and non-missing bank information. We report sample averages (first six statistics) and shares (last two). Corelogic statistics exclude outliers (distance to mean larger than 2 std. deviations). LTV (derived from mortgage and sale amounts) in CL Deed is estimated only for observations with sale amount as large as mortgage amount. Average default refers to the average count of 90+ day delinquencies within 24 months post-origination. Standard deviations are reported in parentheses.

Table 2: Summary Statistics (Demographics)

	LLMA	**Deed**	**Matched**	**HMDA**
White	-	-	64.03%	73.58%
Hispanic	-	-	4.95%	4.48%
African-American	-	-	3.08%	6.02%
Female Only	-	-	12.29%	11.60%
Male Only	-	-	14.50%	17.49%
Mixed Household	-	-	39.82%	46.80%
Age	-	-	47.44	-
N (total count)	-	-	21,601,490	112,104,289

Note: This table reports descriptive statistics on four different loan-level samples over the period 1990-2017. Matched sample demographic information comes from the Infutor dataset. We report sample shares for each racial and gender-composition groups. Mixed household refers to mixed gender households. Age refers to the average age in each sample.

Table 3: Summary Statistics by Year

	Purchase Loans			Refinance Loans		
Year	**Nr Loans**	**Default Rate**	**FICO**	**Nr Loans**	**Default Rate**	**FICO**
1998	36,879	3.6%	705.4	35,425	1.8%	719.7
1999	49,847	3.7%	697.5	24,143	3.3%	700.6
2000	109,872	3.3%	699.1	24,149	4.6%	668.8
2001	170,308	3.5%	703.2	217,515	1.7%	714.3
2002	205,858	3.4%	705.5	334,203	1.0%	729.4
2003	206,524	3.4%	707.9	790,644	0.7%	733.5
2004	256,471	4.1%	711.2	320,908	1.9%	718.3
2005	353,878	8.9%	718.8	345,386	5.4%	717.3
2006	405,715	17.4%	716.1	280,952	15.1%	710.1
2007	489,328	21.1%	714.5	410,928	20.1%	710.4
2008	563,420	13.1%	714.2	559,609	12.0%	722.5
2009	570,869	6.8%	723.8	1,165,914	4.0%	751.6
2010	533,249	5.4%	727.4	1,126,689	2.4%	755.2
2011	504,028	4.0%	730.5	1,009,256	1.6%	756.9
2012	481,404	3.2%	732.7	1,347,570	1.4%	753.0
2013	467,451	2.6%	737.1	959,710	1.3%	744.2
2014	435,152	2.7%	737.6	329,070	1.4%	739.3
2015	454,500	2.5%	740.4	420,194	1.0%	748.4
2016	437,509	1.6%	744.5	470,385	0.6%	751.8

Note: This table reports descriptive statistics of the Corelogic LLMA-Deed matched sample by loan type and origination year. Default rate refers to the share of loans with at least one 90+ day delinquency within 48 months post-origination.

Table 4: First Stage Regressions (Refinance Loans)

	Non-Minority	**Minority**
Bank Exam X Eligible Tract	0.143***	0.0808*
	(0.0332)	(0.0484)
Bank Exam X Non-eligible Tract	0.103***	0.161***
	(0.00937)	(0.0189)
Constant	1.832***	0.625***
	(0.00246)	(0.00696)
Controls		
Bank x Tract FE	X	X
Tract x Time FE	X	X
First-stage F-stat	248.323	81.32
Clusters	16514	622
N	963231	47320
R-sq	0.832	0.599

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Note: The table shows results from estimating the first-stage regression of loan origination volumes on CRA exam activity, as described in equation 8. The sample in this table corresponds to refinance loans for years prior to 2011, for bank-market pairs with substantial loan volume as described in section 5.1.

Table 5: Second Stage Regressions (Refinance Loans)

	Non-Minority		Minority	
	Marginal	**Average**	**Marginal**	**Average**
Originated Loans	0.0516***	0.00649***	0.113***	0.0311***
	(0.00616)	(0.000233)	(0.0231)	(0.00171)
Controls				
Bank x Tract FE	X		X	
Tract x Time FE	X		X	
Clusters	16514	16514	622	622
N	963231	963231	47320	47320
R-sq	-0.440	0.031	-0.103	0.052

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

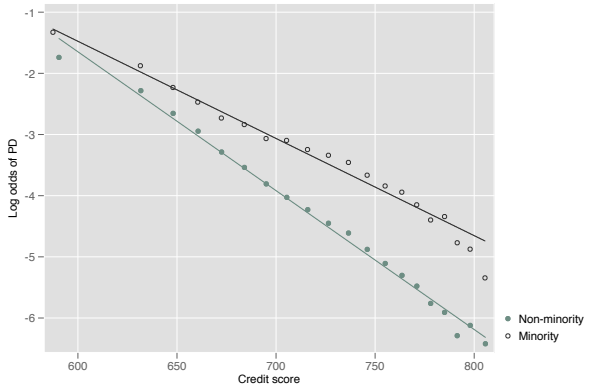
Note: The table shows results from estimating the second-stage regression of loan default volumes on loan origination volumes instrumented by variation in CRA exam activity, as described in equation 9. The sample in this table corresponds to refinance loans for years prior to 2011, for bank-market pairs with substantial loan volume as described in section 5.1.

Table 6: Model Estimation: Target Moments

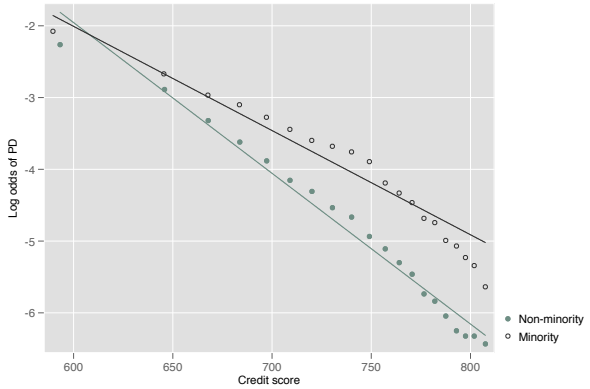
	Non-Minority	**Minority**
Mortgage Approval Rate	0.547	0.721
Mean(Score) — Approved	741.41	711.83
SD(Score) — Approved	63.42	69.61
Mean(Score) — Rejected	682.18	634.74
SD(Score) — Rejected	93.71	93.97
Correlation of Default and Score	-0.021	-0.015
Average Default Rate	0.0065	0.031
Marginal Default Rate	0.052	0.11

Note: This table reports the empirical moments we target in the full model estimation; as discussed in section 5.4, full estimation using all moments and parameters is still in progress. The first row, approval rates, are taken from HMDA data; approval and rejection credit score statistics are taken from the credit report data; the correlations between default and score are the coefficients of the log odds of default regressed on score as reported in appendix table A1; average and marginal default rates are estimated via OLS and 2SLS and are taken from the results reported in table 5.

Figure 1: Relationship between Probability of Default and Credit Score



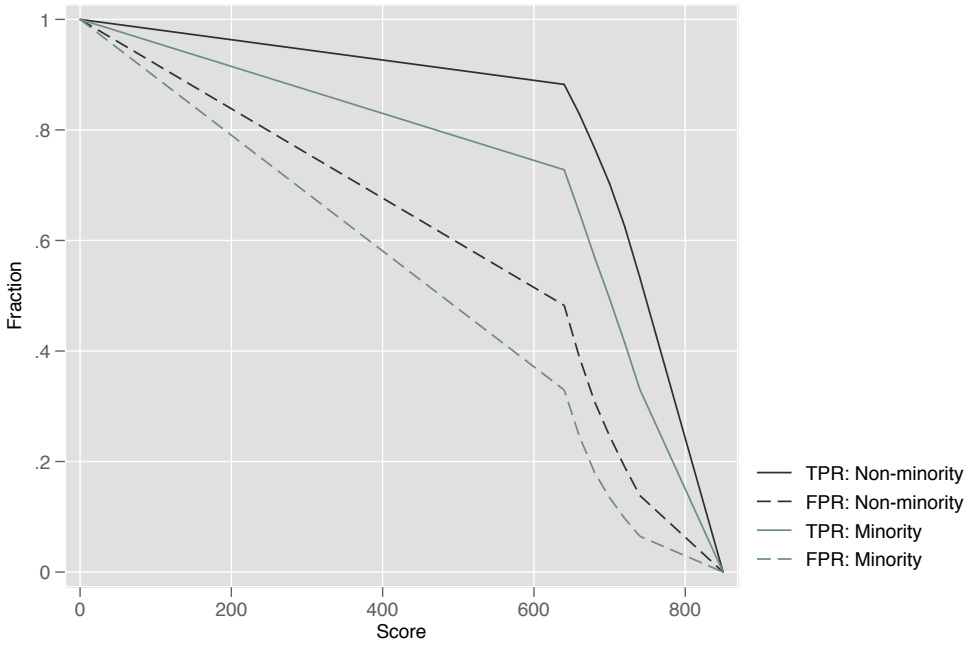
(a) Purchase loans



(b) Refinance loans

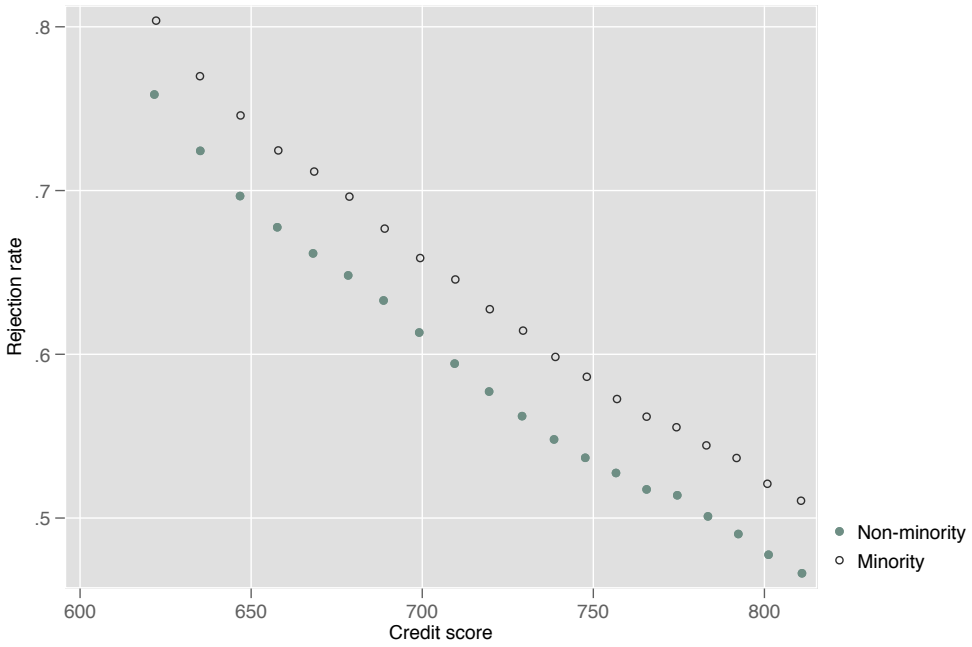
Note: This figure shows the relationship between the log odds of default and credit score at origination. Panel a shows the relationship for the sample of purchase loans originated between 2009 and 2016 while panel b shows refinance loans in the same period. Log odds is defined as log(pd/(1-p)) where the probability of default (d) is define as the fraction borrowers who are ever 90+ days overdue on their mortgage 24 months after origination. Credit score is FICO. The graph is a binned scatter plot.

Figure 2: Deconstructed Receiver Operating Characteristic Curves



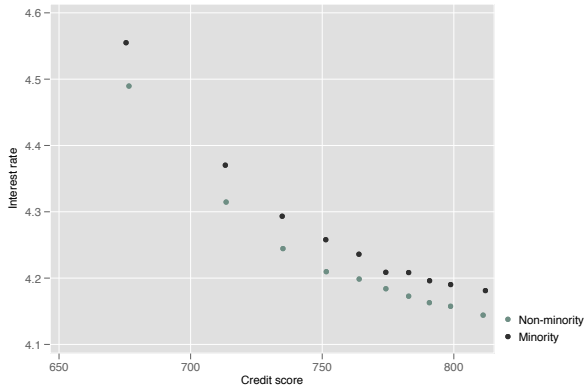
Note: This figure shows the two components of receiver operating Characteristic (ROC) curves for the sample of mortgage applicants. TPR (true positive rate) refers to the fraction of borrower who do not default who would be approved at a given credit score cut-off. FPR (false positive rates) refers to the fraction of borrowers who default who would be approved at a given credit score cut-off. By definition, TPR and FPR are 1 when all inquiries are approved (cutoff = 0) and TPR and FPR are equal to zero when no inquiries are approve (cutoff = 850). Minority refers to Hispanic or African-American applicants. Applications are calculated based on hard mortgage inquiries in credit bureau data.

Figure 3: Relationship between Rejection Rates and Credit Score

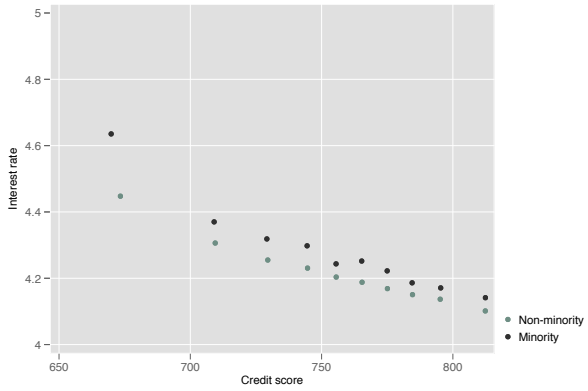


Note: This figure shows the relationship between mortgage rejection rates and credit score at origination. Rejection rates are calculated based on hard mortgage inquiries in credit bureau panel data. Minority is defined as Hispanic or African-American applicants. Credit score is Vantage. The graph is a binned scatter plot.

Figure 4: Relationship between Interest Rates and Credit Score



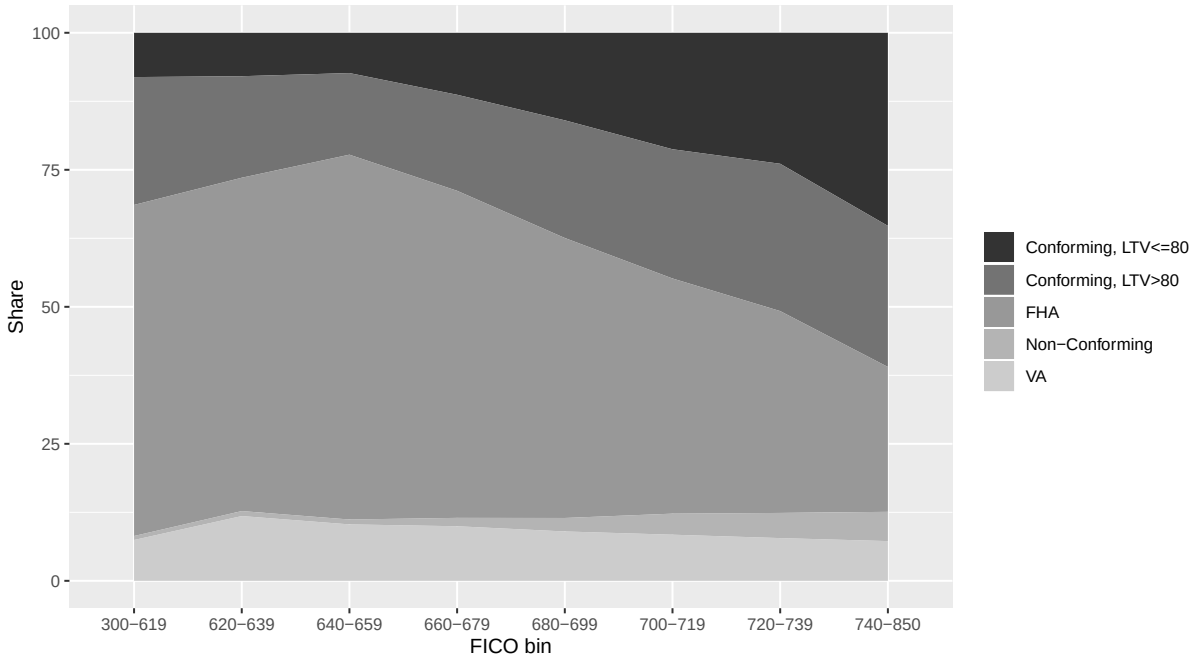
(a) Purchase loans



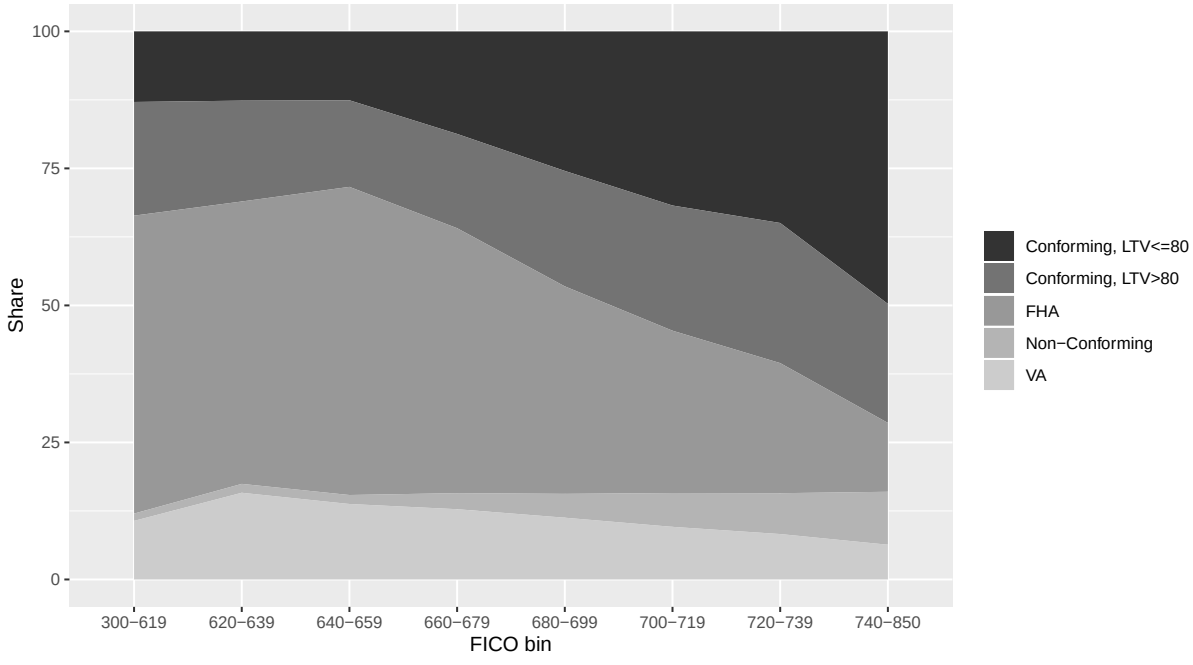
(b) Refinance loans

Note: This figure shows the relationship between the residualized interest rate and credit score at origination. Interest rates are residualized on rate fixation, LTV, original loan balance, loan type, term, DTI, year and lender. Panel a shows the relationship for the sample of purchase loans originated between 2009 and 2016 while panel b shows rate refinance loans in the same period. Credit score is FICO. The graph is a binned scatter plot.

Figure 5: Purchase Mortgage Type Distribution by Credit Score



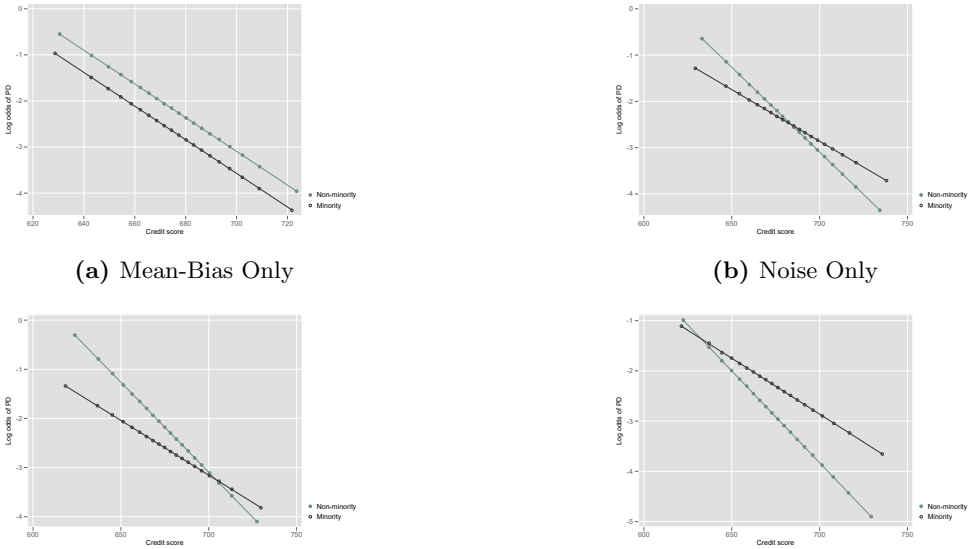
(a) Minority



(b) Non-Minority

Note: These figures shows the distribution of mortgage types across different credit score (FICO) levels for minorities and non-minorities.

Figure 6: Model Identification: Noise and Mean-Bias



(a) Mean-Bias Only

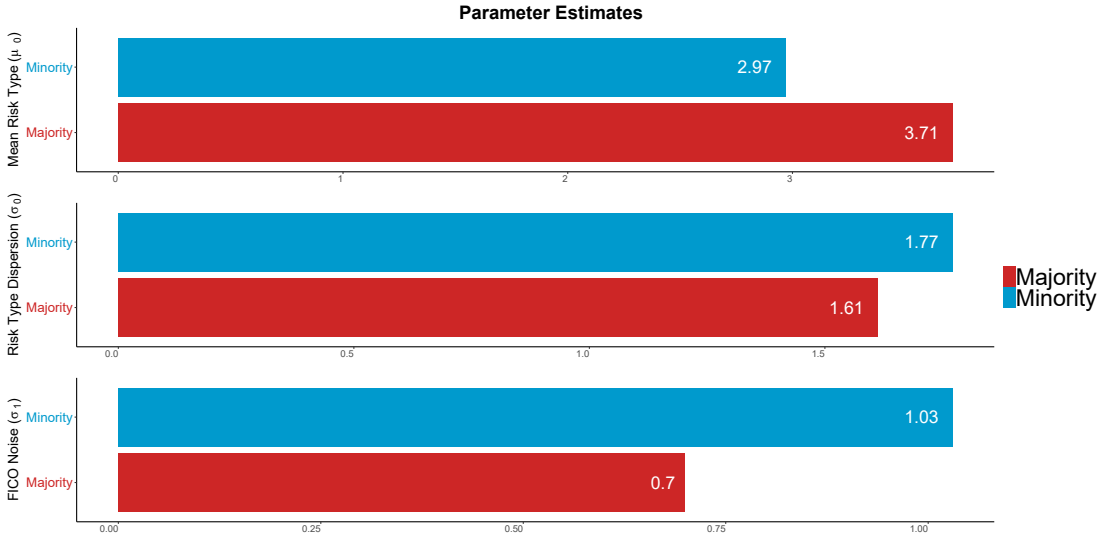
(b) Noise Only

(c) Noise and Reverse Bias

(d) Noise and Bias

Note: The four panels of the figure show simulated credit scores and default rates for two identical populations, labeled minority and non-minority applicants, that face differing noise and/or mean-bias in their credit scores. While mean bias introduces a level-shift in default outcomes, differences in noise lead to differences in slopes because of differential measurement error in credit scores. Panel (d), "Noise and Bias," shows the case where credit scores under-estimate default risk for minorities but also include more noise for minorities; note that this closely matches the empirical analog shown in Figure 1. Panel (b) is the same as (d) but with no bias; panel (c) is the same as (d) but with the bias reversed; panel (a) is the same as (d) but with no measurement error.

Figure 7: Model Parameter Estimates



Note: The figure shows a subset of preliminary parameter estimates from the model described in section 4.3. In our current estimates, we emphasize the difference across groups in the standard deviation of signal noise in credit scores, σ_1, where we find that credit scores feature 47% higher standard deviation of signal noise for minority mortgage applicants as compared to non-minority mortgage applicants.

References

Agarwal, S., E. Benmelech, N. Bergman, and A. Seru (2012): "Did the Community Reinvestment Act (CRA) Lead to Risky Lending?" *NBER Working Paper No., 18609*.

Agarwal, S., S. Chomsisengphet, N. Mahoney, and J. Stroebel (2017): "Do Banks Pass through Credit Expansions to Consumers Who want to Borrow?" *The Quarterly Journal of Economics*, 133, 129–190.

Aigner, D. J. and G. G. Cain (1977): "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relations Review*, 30, 175–187.

Angrist, J. D. and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricistís Companion*, Princeton, NJ: Princeton Univ. Press.

Avery, R., R. Bostic, P. Calem, and G. Canner (1996): "Credit Risk, Credit Scoring and the Performance of Home Mortgages," *Federal Reserve Bulletin*, 82, 621–648.

Avery, R. B., K. P. Brevoort, and G. Canner (2012): "Does Credit Scoring Produce a Disparate Impact?" *Real Estate Economics*, 40, S65–S114.

Avery, R. B., K. P. Brevoort, and G. B. Canner (2009): "Credit scoring and its effects on the availability and affordability of credit," *Journal of Consumer Affairs*, 43, 516–537.

Ayres, I. (2002): "Outcome Tests of Racial Disparities in Police Practices," *Justice Research and Policy*, 4, 131–142.

Barocas, S., M. Hardt, and A. Narayanan (2019): *Fairness and Machine Learning*, fairmlbook.org, http://www.fairmlbook.org.

Barocas, S. and A. D. Selbst (2016): "Big Data's Disparate Impact," *California Law Review*, 104, 671–732.

Bartik, A. and S. Nelson (2019): "Deleting a Signal: Evidence from Pre-Employment Credit Checks," *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

Bartlett, R., A. Morse, R. Stanton, and N. Wallace (2019): "Consumer-Lending Discrimination in the FinTech Era," NBER Working Papers 25943, National Bureau of Economic Research, Inc.

Bayer, P., F. Ferreira, and S. L. Ross (2016): "What Drives Racial and Ethnic Differences in High Cost Mortgages? The Role of High Risk Lenders," Working Paper 22004, National Bureau of Economic Research.

Becker, G. S. (1957): *The Economics of Discrimination*, Chicago: University of Chicago Press.

Berg, T., V. Burg, A. Gombović, and M. Puri (2019): "On the Rise of FinTechs: Credit Scoring Using Digital Footprints," *The Review of Financial Studies*, hhz099.

Bhutta, N. (2011): "The Community Reinvestment Act and mortgage lending to lower income borrowers and neighborhoods," *The Journal of Law and Economics*, 54, 953–983.

Bubb, R. and A. Kaufman (2014): "Securitization and moral hazard: Evidence from credit score cutoff rules," *Journal of Monetary Economics*, 63, 1–18.

CFPB (2014): "Using publicly available information to proxy for unidentified race and ethnicity," .

Charles, K. K. and J. Guryan (2011): "Studying Discrimination: Fundamental Challenges and Recent Progress," Working Paper 17156, National Bureau of Economic Research.

Corbett-Davies, S. and S. Goel (2018): "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," *CoRR*, abs/1808.00023.

Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2019): "Measuring Bias in Consumer Lending," Working Paper 24953, National Bureau of Economic Research.

Elliott, M., P. P. Peter Morrison, Allen Fremont, and N. Luri (2009): "Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities." *Health Serv Outcomes Res Methodn*, 9, 69–83.

Federal Reserve Board (2007): *Report to the Congress on Credit Scoring and its Effects on the Availability and Affordability of Credit*, Federal Reserve Board of Governors.

FRB (2018): "The Community Reinvestment Act," .

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2018): "Predictably Unequal? The Effects of Machine Learning on Credit Markets," .

Hanson, A., Z. Hawley, H. Martin, and B. Liu (2016): "Discrimination in mortgage lending: Evidence from a correspondence experiment," *Journal of Urban Economics*, 92, 48 – 65.

Hurst, E., B. J. Keys, A. Seru, and J. Vavra (2016): "Regional Redistribution through the US Mortgage Market," *American Economic Review*, 106, 2982–3028.

Kearns, M. and A. Roth (2019): *The Ethical Algorithm*, Oxford, UK: Oxford University Press.

Keys, B., T. Mukherjee, V. Vig, and A. Seru (2010): "Did Securitization Lead to Lax Screening? Evidence From Subprime Loans," *Quarterly Journal of Economics*, 125, 307–362.

Keys, B. J., A. Seru, and V. Vig (2012): "Lender screening and the role of securitization: evidence from prime and subprime mortgage markets," *The Review of Financial Studies*, 25, 2071–2108.

Khandani, A. E., A. J. Kim, and A. Lo (2010): "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking Finance*, 34, 2767–2787.

Kleinberg, J., J. Ludwig, S. Mullainathan, and A. Rambachan (2018): "Algorithmic Fairness," *AEA Papers and Proceedings*, 108, 22–27.

Kleinberg, J. M., S. Mullainathan, and M. Raghavan (2016): "Inherent Trade-Offs in the Fair Determination of Risk Scores," *CoRR*, abs/1609.05807.

Laufer, S. and A. Paciorek (2016): "The Effects of Mortgage Credit Availability: Evidence from Minimum Credit Score Lending Rules," .

Lundberg, S. J. and R. Startz (1983): "Private Discrimination and Social Intervention in Competitive Labor Market," *The American Economic Review*, 73, 340–347.

Phelps, E. S. (1972): "The Statistical Theory of Racism and Sexism," *The American Economic Review*, 62, 659–661.

Rajan, U., A. Seru, and V. Vig (2015): "The failure of models that predict failure: Distance, incentives, and defaults," *Journal of Financial Economics*, 115, 237–260.

Rambachan, A. and J. Roth (2019): "Bias In, Bias Out? Evaluating the Folk Wisdom," .

Ross, S. L., M. A. Turner, E. Godfrey, and R. R. Smith (2008): "Mortgage lending in Chicago and Los Angeles: A paired testing study of the pre-application process," *Journal of Urban Economics*, 63, 902 – 919.

Sood, G. and S. Laohaprapanon (2018): "Predicting Race and Ethnicity From the Sequence of Characters in a Nam," .

Suresh, H. and J. V. Guttag (2019): "A Framework for Understanding Unintended Consequences of Machine Learning," *CoRR*, abs/1901.10002.

Thomas, L. C. (2009): *Consumer credit models: pricing, profit and portfolios: pricing, profit and portfolios*, OUP Oxford.

A Appendix Tables and Figures

Table A1: Regression Test for Attenuation Bias

	Purchase Loans	Refinance Loans
FICO x Non-Minority	-0.023***	-0.021***
	0.000	0.000
FICO x Minority	-0.016***	-0.015***
	0.000	0.000
Minority	-3.92***	-3.04***
	0.003	0.003
N	3,110,617	6,046,602
R2	0.983	0.965
H0: Equal slopes		
p-value	0.000	0.000

Note: The table shows regression results corresponding to the binned scatter plot in Figure 1. The dependent variable is the log odds of default. Log odds is defined as log(pd/(1-p)) where the probability of default (pd) is defined as the fraction borrowers who are ever 90+ days overdue on their mortgage at any point within 24 months after origination. Credit score is FICO. The bottom row of table shows that we can strongly reject the null that the coefficients of default log odds regressed on FICO are equal for minority and non-minority borrowers (p<.0001).

Table A2: First Stage Regressions (Purchase Loans)

	Non-Minority	**Minority**
Bank Exam X Eligible Tract	-0.140	0.620
	(0.187)	(0.569)
Bank Exam X Non-eligible Tract	0.0997***	0.242**
	(0.0344)	(0.107)
Constant	1.052***	1.320***
	(0.00570)	(0.0147)
Controls		
Bank x Tract FE	X	X
Tract x Time FE	X	X
First-stage F-stat	18.230	7.249
Clusters	2530	156
N	131,360	8,747
R-sq	0.737	0.834

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Note: The table shows results from estimating the first-stage regression of loan origination volumes on CRA exam activity, as described in equation 8. The sample in this table corresponds to purchase loans for years prior to 2011, for bank-market pairs with substantial loan volume as described in section 5.1.

Table A3: Second Stage Regressions (Purchase Loans)

	Non-Minority		**Minority**	
	Marginal	**Average**	**Marginal**	**Average**
Originated Loans	0.0827*	0.0368***	-0.105	0.136***
	(0.0476)	(0.00223)	(0.159)	(0.0104)
Controls				
Bank x Tract FE	X		X	
Tract x Time FE	X		X	
Clusters	2530	2530	156	156
N	131,360	131,360	8,747	8,747
R-sq	-0.042	0.091	-0.356	0.283

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Note: The table shows results from estimating the second-stage regression of loan default volumes on loan origination volumes instrumented by variation in CRA exam activity, as described in equation 9. The sample in this table corresponds to purchase loans for years prior to 2011, for bank-market pairs with substantial loan volume as described in section 5.1.