Measuring Bias in Consumer Lending*

Will Dobbie† Andres Liberman‡ Daniel Paravisini§ Vikram Pathania¶

May 2020

Abstract

This paper tests for bias in consumer lending using administrative data from a high-cost lender in the United Kingdom. We motivate our analysis using a new principal-agent model of bias where loan examiners maximize a short-term outcome, not long-term profits, leading to bias against illiquid applicants at the margin of loan decisions. We identify the profitability of marginal applicants using the quasi-random assignment of loan examiners. Consistent with our model, we find significant bias against immigrant and older applicants when using the firm's preferred measure of long-run profits, but not when using the short-run measure used to evaluate examiner performance.

Keywords: Discrimination, Consumer Credit

JEL codes: G41, J15, J16

*First version: August 2018. We are extremely grateful to the Lender for providing the data used in this analysis. We also thank Leah Boustan, Hank Farber, Alex Mas, Crystal Yang, and numerous seminar participants for helpful comments and suggestions. Emily Battaglia, Nicole Gandre, Jared Grogan, Ashley Litwin, Alexia Olaizola, Bailey Palmer, and James Reeves provided excellent research assistance. All errors and omissions are ours alone.

†Harvard Kennedy School and NBER. Email: will_dobbie@hks.harvard.edu

‡New York University. Email: aliberma@stern.nyu.edu

§London School of Economics and CEPR. Email: d.paravisini@lse.ac.uk

¶University of Sussex. Email: v.pathania@sussex.ac.uk

I. Introduction

There are large disparities in the availability and cost of credit across different demographic groups within many developed countries. In the United States, for example, blacks pay higher interest rates and are more likely to be rejected for mortgage loans compared to whites, even after accounting for observable differences in credit history and earnings (e.g., Charles and Hurst 2002; Bayer, Ferreira and Ross 2017). There are also large disparities in interest rates and credit usage across ethnic and gender lines within many European countries that cannot be explained by observable differences in creditworthiness (e.g., Alesina, Lotti and Mistrulli 2013; Deku, Kara and Molyneux 2016).

These unexplained disparities have fueled concerns that lenders may be biased against minorities and women. Yet, there are both statistical and economic reasons that these disparities may not be driven by ethnic- and gender-related bias. Lenders may, for example, use variables that are not observed by the econometrician and are correlated with both creditworthiness and group traits when making lending decisions, such as an applicant's expected future income or job prospects, leading to omitted variable bias when estimating credit disparities across these groups. Lenders may also use observable group traits such as ethnicity or gender to form accurate beliefs about the unobservable characteristics of different applicants, a practice commonly known as statistical discrimination (e.g., Phelps 1972; Arrow 1973). Many economists are also skeptical that bias against minorities and women can survive in competitive lending markets, as such behavior typically implies that lenders are not profit-maximizing (e.g., Arrow 1972).

In this paper, we test for bias in consumer lending decisions using detailed administrative data on loan outcomes from a high-cost lender in the United Kingdom (hereafter, "the Lender"). We motivate our analysis using a new principal-agent model of bias that explains why bias against some groups can survive even in competitive lending markets. In the model, bias arises because loan examiners are encouraged to maximize a short-term outcome, not long-term profits, due to principal-agent concerns.[1] The focus on short-run outcomes leads to bias in lending against illiquid

[1]Our model is based on the classic multitasking model of Holmstrom and Milgrom (1991). The multitasking problem we explore is a widely recognized concern for loan officers (Agarwal and Ben-David 2018), as well as teachers (Carrell and West 2010), executives (Healy 1985), sales-people (Oyer 1998), and doctors (Gravelle, Sutton and Ma 2010). Healy (1985) and Oyer (1998) show, for example, that multitasking problems lead both executives and sales-people to increase their compensation at the expense of long-run firm value. We extend this work by showing how the incidence of such multitasking problems can be unequal across consumers and lead to bias in outcomes against vulnerable populations.

subpopulations with volatile income or unpredictable consumption needs for whom short-term and long-term outcomes diverge.[2]

Our incentive-based model also yields the now familiar Becker "outcome test" for bias that compares the success or failure of decisions across groups at the margin (Becker 1957, 1993). In the context of consumer lending, the outcome test is based on the idea that long-run profits should be identical for marginal applicants from all groups if loan examiners are unbiased and the disparities across groups are solely due to omitted variables or statistical discrimination. In contrast, marginal applicants from a targeted group (e.g., ethnic minorities) will yield higher profits than marginal applicants from the non-targeted reference group (e.g., non-minorities) if loan examiners are biased against the targeted group. The outcome test has been difficult to implement in practice, however, as comparisons based on average borrower outcomes are biased if there are unobserved differences in creditworthiness across groups – the well-known infra-marginality problem (e.g., Ayres 2002).

We then show that we can identify the differences in profitability at the margin required for the Becker outcome test using variation in the approval tendencies of the quasi-randomly assigned loan examiners. Using the assigned loan examiner as an instrumental variable (IV) for loan take-up, we can recover the causal effect of loan take-up on both short-run default and long-run profits for applicants at the margin across groups. Though IV estimators are often criticized for the local nature of the estimates, we exploit the fact that the outcome test relies on the difference between exactly these kinds of local treatment effects to test for bias in consumer lending.[3]

Our empirical setting offers an ideal laboratory to implement the Becker test for bias and distinguish our incentive-based model of bias from alternative behavioral models (e.g., ethnic- or gender-related animus or inaccurate stereotypes). First, we observe detailed data on cash flows that allow us to construct individual-level measures of profitability for the Lender. Second, the Lender uses a blind rotation system to assign first-time loan applicants to examiners of the same nationality,

[2]The idea that short- and long-run outcomes can diverge in credit markets when borrowers are illiquid is not new, although past work has largely focused on developing countries. Field et al. (2013), for example, show that requiring early repayment can decrease short-run default, but at the cost of long-run profits. They conclude that the focus on short-run outcomes in microfinance contracts can inhibit investment by high-return but illiquid entrepreneurs.

[3]Our empirical strategy builds on Arnold, Dobbie and Yang (2018), who test for bias in bail decisions using the quasi-random assignment of bail judges to identify outcomes for marginal white and marginal black defendants, and Marx (2018), who tests for bias in police stops using police officers of different races to identify bounds on the outcomes for marginal white, Hispanic, and black drivers. Our IV strategy is also related to research designs used by Liberman, Paravisini and Pathania (2016) to study the effects of high-cost credit and Dobbie and Song (2015) and Dobbie, Goldsmith-Pinkham and Yang (2017) to estimate the impact of bankruptcy protection.

effectively randomizing new applicants to examiners within each branch and nationality. Third, the Lender's loan examiners make on-the-spot, discretionary judgments about whether to accept a loan application with only the standard credit information and limited interaction with applicants, making their decisions particularly prone to the kind of inaccurate stereotypes or categorical heuristics that can lead to bias. The loan examiners can also only make discretionary judgments on whether to accept a loan application or not, with no ability to affect the loan amount, interest rate, maturity date, the number of installments, or the amount of each installment. Finally, examiners are evaluated based on a short-term outcome that we observe in the data. This setup allows us to develop empirical tests that distinguish the incentive-based model of bias from taste-based and stereotypes-based models of bias.

In our empirical analysis, we find significant bias against both immigrant and older loan applicants when using a measure of long-term profits that includes all the future cash flows from the borrower as an outcome. Following the initial loan decision, we find that marginal immigrant applicants yield long-term profits that are £566 larger than marginal native-born applicants, or nearly four times larger. Marginal older applicants also yield profits that are £348 larger than marginal younger applicants, or more than two times larger. Conversely, marginal female and male applicants yield statistically identical profits, suggesting no bias against (or in favor of) female applicants. We show that these results cannot be explained by other ethnic or age-related differences in baseline characteristics, differences in the level of systematic risk across groups, or the way that the IV estimator averages the level of bias across different examiners. In contrast to our IV estimates, however, naïve OLS estimates indicate much more modest levels of bias across all three groups, highlighting the importance of accounting for both infra-marginality and omitted variables when estimating bias in consumer credit decisions.

Three additional results support our proposed incentive-based model of bias. First, there is no evidence of bias against immigrant or older applicants when we implement the Becker outcome test using the short-run default outcome used to evaluate examiner performance. Second, the decisions made by loan examiners are strikingly consistent with a data-based decision rule minimizing this short-term default outcome, but inconsistent with a decision rule maximizing long-term profits. Finally, immigrant and older applicants are more likely to default in the short run compared to native-born and younger applicants with the same level of expected long-run profits, while no such

differences exist for female and male applicants. Taken together, these three results suggest that examiners are equalizing the private returns of lending across groups at the margin, just as predicted by our incentive-based model of bias. In contrast, none of these findings can be easily explained by models based on prejudice and inaccurate stereotypes. We also investigate several more suggestive tests of the taste- and stereotypes-based models.

We conclude by showing that a decision rule based on machine learning (ML) predictions of long-run profits could simultaneously increase profits and eliminate bias. Following Kleinberg et al. (2018), we use the quasi-random assignment of loan examiners to identify the implicit rankings of applicants by loan examiners, which we then compare to the rankings produced by a standard ML algorithm. Our approach uses the loan applicants approved by the most lenient examiner to construct hypothetical outcomes for stricter loan examiners. The quasi-random assignment of loan applicants across loan examiners (combined with a monotonicity assumption) means that we can directly compare the hypothetical outcomes generated by the ML predictions to the actual outcomes generated by strict loan examiners. Consistent with our earlier results, we find that loan examiners systematically misrank loan applicants at the margin of loan take-up, particularly immigrant and older applicants. The Lender would earn approximately 58 percent more per applicant if marginal lending decisions were made using the ML algorithm rather than loan examiners. Including all applicants, not just those at the margin of loan take-up, the Lender would earn over 30 percent more per applicant if lending decisions were made using the ML algorithm.

Our paper is related to an important literature documenting disparities in the availability and cost of credit by ethnicity and gender. There is considerable evidence that minorities have either less access to credit or are forced to pay more for credit compared to observably similar non-minorities for mortgage loans (e.g., Charles and Hurst 2002; Bayer, Ferreira and Ross 2017; Bartlett et al. 2018), auto loans (e.g., Charles, Hurst and Stephens 2008), small business loans (e.g., Cavalluzzo and Cavalluzzo 1998; Cavalluzzo, Cavalluzzo and Wolken 2002; Blanchflower, Levine and Zimmerman 2003), and consumer loans (e.g., Cohen-Cole 2011). There is also evidence that women pay more for both consumer credit (e.g., Alesina, Lotti and Mistrulli 2013) and small business loans (e.g., Bellucci, Borisov and Zazzaro 2010) than observably similar men, and that blacks are more likely to be rejected for peer-to-peer loans than observably similar whites (e.g., Pope and Sydnor 2011). However, none of these papers have been able to determine whether these disparities are due to

omitted variables, statistical discrimination, or bias.

Our results also contribute to a much smaller important literature testing for bias in consumer credit decisions. Outcome tests based on standard OLS estimates show that black mortgage borrowers in the United States have, if anything, slightly higher default rates compared to observably similar white mortgage borrowers (e.g., Van Order, Lekkas and Quigley 1993; Berkovec et al. 1994; Han 2004). However, these OLS-based comparisons will only recover the true level of black-white bias if there are no unobserved differences between black and white mortgage borrowers. In contrast to these OLS-based tests, recent work using both in-person and correspondence audit studies shows that loan officers treat fictitious black and Hispanic mortgage applicants worse than identical fictitious white applicants (e.g., Ross et al. 2008; Hanson et al. 2016). These audit-based tests provide convincing evidence of discriminatory behavior among mortgage lenders, but are generally unable to distinguish between statistical discrimination and bias. Our paper complements this work by providing an empirical test of bias in consumer credit decisions. Our IV-test of bias also has the advantage of being based on actual borrowers who have fully optimized their loan application strategy, not fictitious borrowers applying to a randomly selected lender (e.g., Heckman 1998).

Finally, our paper contributes to an emerging literature examining the underlying reasons for bias against women and minorities. Economists have traditionally viewed bias as rooted in preferences based on ethnic- or gender-related prejudice (e.g., Becker 1957, 1993). In recent work, Bordalo et al. (2016) show that bias can also be driven by inaccurate stereotypes based on low probability but highly representative outcomes. These inaccurate stereotypes can then give rise to both self-stereotyping and biased behavior against women and minorities, as evidenced in both laboratory experiments (Coffman 2014; Bohren, Imas and Rosenberg 2018; Bordalo et al. 2019; Coffman, Exley and Niederle 2018) and real-world settings (Arnold, Dobbie and Yang 2018). Our paper contributes to this literature by showing that bias can also be driven by misaligned incentives, even when prejudice and inaccurate stereotyping are absent. This new incentive-based explanation for bias is important given that incentive problems are prevalent in settings where market forces should, in principle, drive out both prejudice and inaccurate stereotyping (e.g., Becker 1957; Peterson 1981).[4]

[4]The misalignment of firm and examiner incentives is likely widespread in credit markets (e.g., Heider and Inderst 2012). Keys et al. (2010) show, for example, that the securitization of subprime mortgage loans before the financial crisis reduced the incentives of financial intermediaries to carefully screen borrowers. Berg, Puri and Rocholl (2013) and Agarwal and Ben-David (2018) similarly show that volume incentives distort the incentives of loan examiners, leading to higher default risk, while Cole, Kanz and Klapper (2015) show that aligning examiner and firm incentives

The rest of the paper is structured as follows. Section II describes the theoretical model underlying our analysis and develops our empirical test for bias. Section III describes our institutional setting, the data used in our analysis, and the construction of our instrument. Section IV presents the main results. Section V explores potential mechanisms, and Section VI concludes. The Online Appendix provides additional results, details on two alternative models of bias, and information on the outcomes used in our analysis.

II. An Empirical Test of Bias in Consumer Lending

In this section, we motivate and develop our empirical test for bias in consumer lending decisions. We begin by briefly reviewing standard models of bias based on prejudice and inaccurate stereotypes, with the details of these models available in Online Appendix B. We then develop a new principal-agent model of bias that explains why bias against some groups can survive even in competitive lending markets. We develop several testable implications that differentiate our incentive-based model of bias from the standard models based on prejudice and inaccurate stereotypes. We conclude by showing how to estimate the group-specific treatment effects necessary for the bias test using the quasi-random assignment of loan applications to examiners.

A. Prejudice and Inaccurate Stereotypes Models of Bias

To set the stage and motivate our modeling choices, we briefly discuss the standard models of bias from the previous literature based on prejudice and inaccurate stereotypes. Details of both models are available in the Online Appendix.

In taste-based models of bias, bias emerges because there is a direct utility cost of lending to applicants in the target group relative to applicants in the reference group (e.g., Becker 1957, 1993). Risk neutral loan examiners are assumed to maximize long-run profits, leading to a simple decision rule where examiners approve the loan if the expected profit is larger than the perceived cost of making the loan. As a result, long-run profits to the lender should be identical for marginal applicants from all groups if loan examiners are unbiased. In contrast, marginal applicants from

leads to better lending decisions. Hertzberg, Liberti and Paravisini (2010) show that loan officers under-report bad news due to reputational concerns. The Wells Fargo account fraud scandal, where millions of fraudulent savings and checking accounts were created without customers' consent, is also widely thought to be the result of poorly designed sales incentives among branch employees.

the target group will yield higher long-run profits to the lender than marginal applicants from the reference group if loan examiners are biased against the target group.

Stereotypes-based models of bias instead assume that loan examiners may systematically underestimate the long-run profits of lending to applicants in the target group relative to applicants in the reference group (e.g., Bordalo et al. 2016). These stereotypes-based models nevertheless yield an identical set of empirical predictions, where long-run profits to the lender should be identical for marginal applicants from all groups if loan examiners are unbiased, but marginal applicants from the target group will yield higher long-run profits to the lender than marginal applicants from the reference group if loan examiners are biased. Bias can therefore be best understood as unequal outcomes at the margin of a decision, as opposed to one particular behavioral model.

One important shortcoming of existing taste-based and stereotypes-based models of bias is that neither model can easily explain why such bias would survive in a competitive lending market (e.g., Arrow 1972). Bias, by definition, means that lenders are not maximizing long-run profits. Standard economic models imply that new, less-biased lenders should enter these markets and extend credit to target group applicants at the margin (e.g., Becker 1957; Peterson 1981). Bias against target group applicants should therefore decrease with the level of competition in a market, with no bias in a perfectly competitive market.[5]

B. A Principal-Agent Model of Bias

We develop a new principal-agent model of bias that explains why bias, in the sense of unequal long-run outcomes at the margin of a decision, can survive even in competitive lending markets. In the model, such bias arises because loan examiners maximize a short-term outcome instead of long-term profits due to principal-agent concerns. There is ample empirical evidence that agents that make loan decisions inside a financial institution have conflicting objectives with those of the principal (e.g., Hertzberg, Liberti and Paravisini 2010; Berg, Puri and Rocholl 2013; Qian, Strahan

[5]See Berkovec et al. (1998) and Buchak and Jørring (2016) for empirical work estimating the effects of competition on lending disparities. The existing evidence largely supports the idea that lending disparities are decreasing with market pressure, although this literature has been unable to isolate changes in bias per se. We tentatively investigate the relationship between bias and market competition in our setting in Appendix Table A1, which estimates the level of bias in branches located in London where there is considerable market competition in subprime lending and branches located in smaller cities where there is potentially less market competition in this space. We find similar levels of bias against immigrants in branches located in London and in other areas, and slightly lower (but statistically indistinguishable) evidence of bias against older applicants in London stores. These results are consistent with the idea that incentive-based bias is unaffected by market competition.

and Yang 2015). There is also little reason to believe that principal-agent problems inside financial institutions will decrease with the level of competition in a market. We show that the focus on short-run outcomes leads to bias against illiquid subpopulations with volatile income or unpredictable consumption needs for whom short-term and long-term outcomes diverge. The model also delivers testable implications that distinguish incentive-based bias from taste-based or stereotypes-based forms of bias.

B.1. Setup

Let i denote loan applicants and $\mathbf{V}_i$ denote all applicant characteristics excluding group identity, g_i, such as ethnicity or gender. Individuals receiving a loan generate a discounted sum of cash flows from all the future loans, α_i^{LR}, which we refer to as long-run profits. These individuals also generate a short-run outcome α_i^{SR} that is a known function of long-run profits α_i^{LR} and group identity, $g_i \in \{T, R\}$, which indexes target and reference group applicants, respectively. As a result, $\alpha_i^{SR} = \beta_g \alpha_i^{LR} + \eta_i$, where β_g is the correlation between short- and long-run outcomes for each group and η_i is noise. $\beta_R - \beta_T$ indicates the relative misalignment of short- versus long-run outcomes for the target group relative to the reference group.

The perceived cost of lending to applicant i assigned to examiner e is denoted by t^e, which includes both the firm's opportunity cost of making a loan and the personal benefits to examiner e from any direct utility or disutility from being known as either a lenient or tough loan examiner, respectively. We assume these costs are independent of the group identity of the applicant, unlike taste-based models of bias that allow for subjective examiner preferences to differ for applicants from the target group and the reference group. See the Online Appendix for an example of such a model.

For simplicity, we assume that both the lender and loan examiners are risk neutral and maximize the perceived net benefit of approving a loan. We also assume that the loan examiner's sole task is to decide whether to approve or reject a loan application given that, in practice, this is the only decision margin in our setting.

We allow for the potential misalignment of examiner and lender incentives by assuming that the lender cannot contract with the loan examiner on long-run profits α_i^{LR} directly, a standard assumption in the multitasking literature (e.g., Holmstrom and Milgrom, 1991). This assumption

can be rationalized in our setting by assuming that loan examiners have a high discount rate, so that delaying compensation until long-run profits are observed weakens the effectiveness of the contract. We further assume that the unconditional expectation of long-term profits is negative, such that the lender cannot profitably make loans to all applicants without examiners. Loan examiners can observe the information set $\mathbf{V}_i$ at a given effort cost to form accurate predictions $\mathbb{E}[\alpha_i^{LR}|\mathbf{V}_i, g_i]$ and $\mathbb{E}[\alpha_i^{SR}|\mathbf{V}_i, g_i]$, with these predictions allowing the lender to profitably make loans to a subset of applicants. Finally, we assume that lenders cannot offer different compensation contracts for applicants from different groups, reflecting the fact that consumer protection and anti-discrimination laws ban the use of race, gender, or ethnicity in any aspect of the loan decision process. This setup yields the standard result that the examiner's optimal compensation contract includes a fixed salary ω and a bonus equal to $b\alpha_i^{SR}$ for every loan that is approved, with $b > 0$.[6]

The expected benefit from lending to individual i from the perspective of the loan examiner is $b\mathbb{E}[\alpha_i^{SR}|\mathbf{V}_i, g_i]$, or, written as a function of the long term outcome:

$$b\left(\beta_g \mathbb{E}[\alpha_i^{LR}|\mathbf{V}_i, g_i]\right) \tag{1}$$

where b captures examiners' systematic undervaluation of long-run profitability across all groups due to misaligned incentives, a standard result in the multitasking literature. Our contribution comes from the term β_g, which captures examiners' undervaluation of long-run profitability for the target group relative to the reference group due to the misalignment of short- versus long-run outcomes.

Loan examiner e will lend to applicant i if and only if examiner e's perceived benefit expressed in (1) is weakly greater than the cost of issuing the loan:

$$b\beta_g \mathbb{E}[\alpha_i^{LR}|\mathbf{V}_i, g_i] \geq t^e \tag{2}$$

We use the above decision rule to define the *marginal* applicant for examiner e and group g as the applicant i for whom the expected benefit is exactly equal to the perceived cost: $b\beta_g \mathbb{E}[\alpha_i^{LR}|\mathbf{V}_i, g_i] = t^e$. Rearranging in terms of long-run profit, we see the decision rule from the lender's perspective

[6]This result follows directly from the assumption that the lender's unconditional expected profits are negative. A contract with a fixed wage is not incentive compatible due to the cost of effort, and will lead to the examiner approving all loans with negative profits for the lender. So long as the short-term outcome α^{SR} and long-term outcome α^{LR} are related ($\beta_g \neq 0$), the lender can always increase profits by offering the examiner a contract that incentivizes her to screen on the short-term outcome α^{SR}.

which mirrors the familiar Becker outcome-based test:

$$\begin{aligned}\mathbb{E}[\alpha_i^{LR}|\mathbf{V}_i, g_i] &= \frac{1}{\beta_g}\left(\frac{t^e}{b}\right) \\ &= t_g^e\end{aligned} \tag{3}$$

where we define $t_g^e \equiv \frac{1}{\beta_g}\left(\frac{t^e}{b}\right)$ to represent examiner e's threshold for loan approval for applicants from group g based on long-run profit. Moving forward, we let α_g^{SR} and α_g^{LR} represent expected profit $\mathbb{E}[\alpha_i^{SR}|\mathbf{V}_i, g_i]$ and $\mathbb{E}[\alpha_i^{LR}|\mathbf{V}_i, g_i]$, respectively.

Equation (3) shows that bias in the sense of unequal long-run outcomes at the margin of a decision arises when there is a relative misalignment of short- versus long-run outcomes for the target group relative to the reference group. When target and reference group applicants have the same misalignment of short- and long-run outcomes so that $\beta_T = \beta_R$ and $t_T^e = t_R^e = t^e$, loan examiners will apply the same decision threshold to both groups and long-run outcomes will be equalized at the margin across groups. But when target group applicants have relatively worse short-run outcomes compared to reference group applicants with the same expected level of long-run profitability so that $\beta_T < \beta_R$ and $t_T^e > t_R^e$ for a given α_g^{LR}, loan examiners will apply a stricter decision threshold for target group applicants and long-run profits will be higher for marginal applicants from this group. Importantly, such incentive-based bias exists even though there is no taste-based or stereotypes-based bias in the model.

Our model shows that incentive-based bias is driven by the relative relationship between short- and long-run outcomes across different groups. Changes in the market structure of the banking sector, outside options for examiners, the cost of providing credit, and so on are unlikely to change the relative relationship between short- and long-run outcomes across different groups, and are therefore unlikely to change the level of bias. The model also clarifies that principal-agent problems do not always lead to bias. Examiners can undervalue the long-run profitability of all groups due to misaligned incentives ($b < 1$) without bias in the sense of unequal outcomes at the margin. Principal-agent problems only lead to such bias when examiners systematically undervalue the long-run profitability of the target group relative to the reference group ($\beta_T < \beta_R$).

C. Testable Implications of the Model

The above framework demonstrates that principal-agent conflicts may lead to a bias in lending that is indistinguishable from the one that arises from alternative behavioral models such as the taste-based and stereotypes-based models. In this section, we develop three testable implications that are unique to the incentive-based model and can be tested in our data.

The first testable implication of the incentive-based model comes from the fact that examiners are compensated based on short-run outcomes such as loan volume and first-loan default. Examiners will therefore make lending decisions based on these short-run outcomes alone, meaning that these outcomes will be equalized across groups at the margin in the absence of any other forms of bias. The incentive-based model of bias therefore yields the following testable prediction.

PREDICTION 1. If examiners are evaluated using only the short-run outcome α^{SR}, then the expected short-run outcome will be identical for the marginal target group and marginal reference group applicants: $\alpha_T^{SR} = \alpha_R^{SR}$. In contrast, the taste-based and stereotypes-based models generally predict that the expected short-run outcome will be higher for the marginal target group applicants compared to the marginal reference group applicants when there is bias against the target group: $\alpha_T^{SR} > \alpha_R^{SR}$.

Testing Prediction 1 simply requires implementing the Becker outcome test using the short-run outcome used to evaluate the examiner. The incentive-based model predicts that the short-run outcome will be identical for the target and reference groups at the margin, as examiners have no incentive to treat applicants differently at the margin. In contrast, the taste-based and stereotypes-based models both predict better short-run outcomes for the marginal target group applicants compared to marginal reference group applicants when there is bias, as there is a positive correlation between the short- and long-run outcomes.

The second testable implication of our model comes from the fact that examiners make lending decisions based on the short-run outcome in our incentive-based model, but based on long-run profits in the taste-based and stereotypes-based models. As a result, the incentive-based model should only predict bias when examiners' rankings of applicants are different when using the short- and long-run outcomes. We can formalize this observation through the following testable prediction.

PREDICTION 2: The incentive-based model of bias predicts that examiners will make lending decisions as though they are ranking applicants based on the short-run outcome, α_g^{SR}. In contrast, the taste-based and stereotypes-based models predict that examiners will make lending decisions as though they are ranking applicants based on the long-run outcome, α_g^{LR}.

Testing Prediction 2 requires recovering examiners' implicit ranking of loan applicants. In Section V, we show that we can implement this test by comparing examiners' decisions to those generated by a machine learning (ML) algorithm trained using either the short- or long-run outcome.

The final testable implication of our model comes from the observation that incentive-based bias will only emerge when short- and long-run outcomes diverge for the target group relative to the reference group ($\beta_T < \beta_R$). The divergence of short- and long-run outcomes could emerge if, for example, target group applicants are more likely to have a binding liquidity constraint compared to reference group applicants, increasing the probability of short-run default without changing the long-run profitability of lending to that group. In contrast, the taste-based and stereotypes-based models yield no predictions on the relationship between short- and long-run outcomes. We can formalize this observation through the following testable prediction.

PREDICTION 3. The incentive-based model predicts that bias will emerge only if, conditional on expected long-term profits, the expected short-run outcome is lower for the marginal target group applicants compared to the marginal reference group applicants: $\beta_T < \beta_R$. In contrast, the taste-based and stereotypes-based models yield no predictions on the relationship between short- and long-run outcomes.

Testing Prediction 3 requires a comparison of short- and long-run outcomes for different, mutually exclusive subpopulations. The incentive-based model of bias requires a divergence between short- and long-run outcomes whenever we observe bias, while the taste-based and stereotypes-based models do not.

There are also other testable implications of the incentive-based model that cannot be examined in our data. Due to confidentiality reasons, for example, we cannot share the exact details of the examiners' compensation contract or explore heterogeneity in the results along these dimensions. We also do not have information on examiners' expected tenure with the Lender, which may also

impact the examiners' incentives.

D. Empirical Test of Bias in Consumer Lending

This section explains how we identify the differences in profitability at the margin required for the Becker outcome test using variation in the approval tendencies of quasi-randomly assigned loan examiners, building on work by Arnold, Dobbie and Yang (2018) in the context of bail decisions. We begin with a definition of the target parameter and a series of simple graphical examples that illustrate our approach. We then formally describe the conditions under which our examiner IV strategy yields consistent estimates of bias in consumer lending decisions and discuss the interpretation of the estimates.

Overview: Following the theory model, let the average long-run profitability for applicants from group g at the margin for examiner e, $\alpha_g^{LR,e}$, for some weighting scheme, w^e, across all loan examiners, $e = 1...E$, be given by:

$$\begin{aligned} \alpha_g^{LR^*,w} &= \sum_{e=1}^{E} w^e \alpha_g^{LR,e} \\ &= \sum_{e=1}^{E} w^e t_g^e \end{aligned} \tag{4}$$

where w^e are non-negative weights which sum to one that will be discussed in further detail below. As seen in Equation (3), $\alpha_g^{LR,e} = t_g^e$, where t_g^e represents examiner e's threshold for loan approval for applicants from group g. In our context, applicants who do not take up a loan yield exactly zero profit and so we can identify profitability as the treatment effect of loan take-up on long-run profits. Thus, $\alpha_g^{LR^*,w}$ represents a weighted average of the treatment effects for applicants of group g at the margin of loan take-up across all examiners.

Following this notation, the average level of bias among loan examiners, $B^{LR^*,w}$, for the weighting

scheme w^e is given by:

$$\begin{aligned} B^{LR^*,w} &= \sum_{e=1}^{E} w^e \left(t_T^e - t_R^e\right) \\ &= \sum_{e=1}^{E} w^e t_T^e - \sum_{e=1}^{E} w^e t_R^e \\ &= \alpha_T^{LR^*,w} - \alpha_R^{LR^*,w} \end{aligned} \tag{5}$$

Equation (5) generalizes the outcome test to the case where there are many examiners and the level of bias across examiners may vary. Following Equation (4), we can then express the target parameter, $B^{LR^*,w}$, as a weighted average across all examiners of bias in lending decisions, measured by the difference in treatment effects for target and reference group applicants at the margin of loan take-up.

Recall that standard OLS estimates will typically not recover unbiased estimates of the weighted average of bias, $B^{LR^*,w}$, for two reasons. The first is that characteristics observable to the loan examiner but not the econometrician may be correlated with loan take-up, resulting in omitted variable bias when estimating the treatment effects for different types of loan applicants. The second, and more important, reason OLS estimates will not recover unbiased estimates of bias is that the average treatment effect identified by OLS will not equal the treatment effect at the margin required by the outcome test unless there is either an identical distribution of potential profits for loan applicants from different groups or constant treatment effects across the entire distribution of loan applicants — the well-known infra-marginality problem (e.g., Ayres 2002).

Following Arnold, Dobbie and Yang (2018), we estimate the differences in profitability at the margin required for the Becker outcome test, $B^{LR^*,w}$, using variation in the approval tendencies of quasi-randomly assigned loan examiners. Our estimator uses the standard IV framework to identify the difference in local average treatment effects (LATEs) for reference group and target group applicants near the margin of loan take-up. Though IV estimators are often criticized for the local nature of the estimates, we exploit the fact that the outcome test relies on the difference between exactly these kinds of local treatment effects to test for bias. This empirical design allows us to recover a weighted average of the long-run profitability of different groups near the margin, where the weights are equal to the standard IV weights described in further detail below.

Figure 1 provides a series of simple graphical examples to illustrate the intuition of our approach. In Panel A, we consider the case where there is a single unbiased examiner to illustrate the potential for infra-marginality bias when using a standard OLS estimator. The examiner perfectly observes expected profitability and chooses the same approval threshold for all loan applicants, but the distributions of profitability differ by group identity such that reference group applicants, on average, yield higher profits than target group applicants. Letting the vertical lines denote the examiner's approval threshold, standard OLS estimates of $\alpha_T^{LR^*,w}$ and $\alpha_R^{LR^*,w}$ measure the average profitability for target and reference group applicants who take up a loan, respectively. In the case illustrated in Panel A, the standard OLS estimator indicates that the examiner is biased against reference group applicants, when, in reality, the examiner is unbiased. Panel B illustrates a similar case where the standard OLS estimator indicates that the examiner is unbiased, when, in reality, the examiner is biased against target group applicants.

To illustrate how our IV estimator identifies the profitability of marginal applicants, the last two panels of Figure 1 consider a case where there are two loan examiners, one that is lenient and one that is strict. In Panel C, we consider the case where the two examiners are unbiased, while in panel D we consider the case where the two examiners are both biased against target group applicants. In both cases, an IV estimator using examiner leniency as an instrument for loan take-up will measure the average profitability of "compliers," or applicants who take up a loan when assigned to the lenient examiner but not when assigned to the strict examiner. In other words, the IV estimator only measures the profitability of applicants between the two examiner thresholds, ignoring applicants that are either above or below both examiner thresholds and eliminating the need to observe the hypothetical profits of rejected applicants.[7] When the two examiners in our example are "close enough" in leniency, the IV estimator will therefore measure the profitability of applicants only at the margin of loan take-up, allowing us to correctly conclude that the examiners are unbiased in the example illustrated in Panel C and biased against target group applicants in Panel D.

Consistency of the IV Estimator: We now briefly review the conditions under which our examiner

[7]In our empirical implementation we assign individuals who do not borrow (whose profits are not observed) a level of zero profits. Note that this choice is irrelevant because individuals who do not borrow are never right above their assigned examiner's threshold. In the language of the LATE literature, these individuals are "never-takers" and do not affect the IV estimate.

IV strategy yields consistent estimates of bias in consumer lending decisions. See Arnold, Dobbie and Yang (2018) for formal proofs.

Let Z_i be a scalar measure of the assigned examiner's propensity for loan take-up for applicant i that takes on values ordered $\{z_0, ..., z_E\}$, where $E + 1$ is the total number of examiners. For example, a value of $z_e = 0.7$ indicates that 70 percent of all applicants assigned to examiner e take up a loan. We construct Z_i using a standard leave-out procedure that captures the approval tendencies of examiners. We calculate a single Z_i for all groups to minimize measurement error in our instrument, but we show in robustness checks that our results are similar (if less precise) if we allow the instrument to vary by group.

Following Imbens and Angrist (1994), an estimator using Z_i as an instrumental variable for loan take-up is valid and well-defined under the following three assumptions:

Assumption 1. (Existence) $Cov(TakeUp_i, Z_i) \neq 0$

Assumption 2. (Exclusion) $Cov(Z_i, \mathbf{v}_i) = 0$

Assumption 3. (Monotonicity) $TakeUp_i(Z_e) - TakeUp_i(Z_{e-1}) \geq 0$

where $\mathbf{v}_i = \mathbf{U}_i + \varepsilon_i$ consists of characteristics unobserved by the econometrician but observed by the examiner, $\mathbf{U}_i$, and idiosyncratic variation unobserved by both the econometrician and examiner, ε_i. Assumption 1 requires the instrument Z_i to increase the probability of loan take-up $TakeUp_i$. Assumption 2 requires the instrument Z_i to be as good as randomly assigned and to only influence profitability through the channel of loan take-up. In other words, Assumption 2 ensures that our instrument is orthogonal to characteristics unobserved by the econometrician, $\mathbf{v}_i$. Assumption 3 requires the instrument Z_i to weakly increase the probability of loan take up $TakeUp_i$ for all individuals.

Taking Assumptions 1–3 as given, let the true IV-weighted level of bias, $B^{LR^*,IV}$ be defined as:

$$\begin{aligned} B^{LR^*,IV} &= \sum_{e=1}^{E} w^e \left(t_T^e - t_R^e\right) \\ &= \sum_{e=1}^{E} \lambda^e \left(t_T^e - t_R^e\right) \end{aligned} \tag{6}$$

where $w^e = \lambda^e$, the standard IV weights defined in Imbens and Angrist (1994).

Let our IV estimator that uses examiner leniency as an instrumental variable for loan take-up be defined as:

$$\begin{aligned} B^{IV} &= \alpha_T^{IV} - \alpha_R^{IV} \\ &= \sum_{e=1}^{E} \lambda_T^e \alpha_T^{e,e-1} - \sum_{e=1}^{E} \lambda_R^e \alpha_R^{e,e-1} \end{aligned} \tag{7}$$

where λ_g^e are again the standard IV weights and each pairwise treatment effect $\alpha_g^{e,e-1}$ captures the treatment effects of compliers within each e, $e-1$ pair.

Our IV estimator B^{IV} provides a consistent estimate of the true level of bias $B^{LR^*,IV}$ if two conditions hold: (1) the instrument Z_i is continuously distributed over some interval $[\underline{z}, \bar{z}]$, and (2) the weights on the pairwise LATEs λ_g^e are identical across groups. The first condition ensures that the group-specific IV estimates are equal to the true IV-weighted average of treatment effects for applicants at the margin of loan take-up.[8] The second condition for consistency ensures that any difference in the group-specific IV estimates is driven by differences in the true group-specific treatment effects, not differences in the IV weights applied to those treatment effects. This equal weights condition holds if there is a linear first stage across groups, as is true in our data (see Figure 2). We also find that the distributions of IV weights by nationality, gender, and age are visually indistinguishable from each other (see Appendix Figure A1) and that the IV weights for each examiner are highly correlated across groups (see Appendix Figure A2), indicating that the equal weights condition is unlikely to be violated in our setting.

Interpretation of IV Weights: We conclude this section by discussing the economic interpretation of our IV-weighted estimate of bias, B^{IV}. Appendix Table A2 presents OLS estimates of IV weights in each examiner-by-branch cell and observable examiner characteristics. We find that our IV weights are uncorrelated with the number of applications, examiner experience, examiner leniency, and examiner gender. We also find that our IV weights are largely uncorrelated with examiner-level estimates of bias against each group obtained from our MTE specification described below, indicat-

[8]The maximum estimation bias from using a discrete instrument, as we do in this paper, can be calculated using the empirical distribution of examiner leniency and the worst-case treatment effect heterogeneity among compliers. Using the 10th and 90th percentiles of observed profits as the worst-case treatment effect heterogeneity among compliers, we find that the maximum estimation bias when using a discrete instrument is only £18, indicating that this issue is unlikely to be a significant problem in our setting.

ing that our IV-weighted estimates of bias are likely to be very similar to estimates based on other weighting schemes. In robustness checks, we also report estimates from an MTE specification that allows us to impose equal weights when calculating the average level of bias across loan examiners at the cost of statistical precision and additional auxiliary assumptions.

III. Background, Data, and Instrument Construction

This section summarizes the most relevant information regarding our institutional setting and data, describes the construction of our examiner leniency measure, and provides support for the baseline assumptions required for our IV estimator. Further details on the cleaning and coding of variables are contained in Online Appendix C.

A. Institutional Setting

We test for bias in consumer lending decisions using information from a large subprime lender in the United Kingdom. The Lender offers short-term, uncollateralized, high-cost loans to subprime borrowers. Loan maturities are typically less than six months, and can be as short as a few weeks. Loan amounts range from £200 to £2,000, with an average first-loan amount of just under £300. All loans require weekly payments starting soon after the loan is disbursed, with interest rates that average about 600 percent. By comparison, the typical payday loan in the United States is below $300 with an APR of 400 to 1,000 percent and a seven- to thirty-day maturity (Stegman 2007). The Lender also allows applicants who remain in good standing after one month the option of "topping up" their initial loan, or increasing their outstanding balance back to the initially approved loan amount. In other words, applicants can convert their initial loan to a line of credit up to the original loan amount after one month. The Lender's profits are largely driven by the use of these loan top-ups over the next one to two years, with only about 25 percent of the variation in long-run profits coming from the repayment of the original loan amount. In contrast, the number of loan top-ups explains 34 percent of the variation in long-run profits among individuals taking up a loan, while the number of loans explains 41 percent of the variation in long-run profits in this sample. The repayment of the original loan amount also explains very little additional variation in long-run profits once these longer-run measures are included.

The Lender operates 24 branches throughout the United Kingdom to handle all in-person applications, and a virtual branch to handle all online and phone applications. In the physical branches maintained by the Lender, loan applicants are first greeted by a receptionist, who gathers basic information such as the applicant's name, address, phone number, and nationality. Loan applicants are then randomly assigned to one of the loan examiners working in the branch that day using a blind rotation system. The blind rotation system randomly assigns native-born applicants to the full set of loan examiners working in the branch that day, but only randomly assigns foreign-born and non-English speaking applicants to the set of the loan examiners with the same ethnic background to put these applicants at ease and improve the accuracy of the screening process. Next, the assigned loan examiner reviews the applicant's credit history (e.g., credit score, outstanding debt, past repayment behavior) and inquires about the applicant's income and employment status, as well as any other relevant information, during the initial interview. The examiner uses this information to make an on-the-spot decision as to whether to approve or reject the application. The assigned examiner only makes discretionary judgments on whether to accept a loan application or not; the examiner has no discretion to affect the loan amount, interest rate, maturity date, the number of installments, or the amount of each installment, all of which are determined by the Lender. Following the examiner's approval decision, approved loan applicants decide whether to take up the loan or not, as well as the total amount to borrow from the maximum allowable credit line. Loans are then disbursed to approved applicants before leaving the store. The process is broadly similar for online and phone applications, although applicants are typically not randomly assigned to loan examiners and, for that reason, we do not include these applicants in our analysis.

The assigned loan examiner has complete discretion to approve or reject first-time loan applicants whose credit scores exceed a minimum threshold established by the Lender. The loan examiners' compensation contract includes a fixed salary and a bonus amount that increases with the number of loans issued and decreases with the number of loans that are in default within the first few months after origination.[9] Loan examiners are not compensated for long-run profits. The short-run nature of this compensation contract can be explained by the high turnover among examiners, who

[9]Due to confidentiality reasons, we cannot share the exact details of the bonus contract. There were also several small changes to the bonus contract during our sample period, all related to the amount of time that defaults decreased the bonus amount. In robustness checks, we show that our results are robust to using different time horizons for the short-run default outcome.

have a 10 percent monthly turnover rate and a 62 percent six-month turnover rate. Measuring and compensating long-run performance in such high-turnover jobs is likely infeasible. Due to the short-run nature of the compensation contract, we expect examiners to focus on the probability of default in the first few months, disregarding what happens after the horizon of their incentive structure. We explore the potential importance of this compensation contract when explaining our results in Section V.

B. Data Sources and Descriptive Statistics

We use administrative data on all loan applications and loan outcomes at the Lender between May 2012 and February 2015. The loan-level data contain detailed information pulled from a private credit registry at the time of application, including credit scores and information on outstanding debts and past repayment behavior. The data also contain information gathered by the examiner during the interview, including the applicant's nationality, age, gender, earnings and employment, marriage status, number of dependents, months at his or her current residence, and the stated reason for the loan. Finally, for individuals who take up at least one loan, the data contain information on loan disbursal amounts, interest rates, maturities, payments, top-ups, and defaults for all loans during our sample period. The data are high-quality and complete with one important exception: information on earnings and employment is only collected when examiners believe the application is likely to be approved, meaning that it is missing for a relatively large part of our sample. We therefore do not include earnings and employment controls in our baseline results, as the availability of these controls is mechanically correlated with examiner leniency. None of our results are significantly changed if we include these controls, however.

We measure long-run profits using the sum of all payments made by the applicant minus all disbursements from the Lender for both the first loan and all subsequent loans during our sample period. We are unable to measure the direct costs of lending at the applicant level, meaning that our long-run profit measure reflects the net revenue generated by each applicant. Our estimates will be systematically biased to the extent that the marginal cost of making or servicing a loan varies by group. In practice, we view our estimates as far too large to be plausibly explained by such marginal cost differences. In robustness checks, we investigate this concern more formally by estimating results using the net present value of long-run profits for a variety of discount rates that

are allowed to vary across groups, therefore accounting for possible differences in the opportunity cost of capital by group.

We make five restrictions to the estimation sample. First, we drop repeat applications, as these applications are not randomly assigned to examiners and have a nearly 100 percent approval rate. Second, we drop all online and phone applications, as these individuals are also not randomly assigned to examiners during our sample period. Third, we drop loans assigned to loan examiners with fewer than 50 applications, and loan applications where there is only a single applicant in a branch by nationality cell. Fourth, we drop a handful of applications where applicants are younger than 18 years old, older than 75 years old, or where the credit check information is missing. Finally, we drop all applications after December 2014 to ensure that we observe loan outcomes over a reasonable period. The final sample contains 45,507 first-time loan applications assigned to 254 loan examiners between May 2012 and December 2014.

Table 1 reports summary statistics for our estimation sample separately by loan take-up, with approximately 12.1 percent of approved applicants not taking up a loan. Forty percent of first-time applicants are immigrants, 56 percent are female, 73 percent have lived at least one year at their current residence, and 42 percent are married. The average age of first-time applicants in our sample is 33.9 years old, with the typical applicant having just under one dependent. Over 91 percent of first-time applicants have a bank account and 29 percent have other loan payments. Twenty-seven percent of loans are for emergency expenses, 11 percent are for a large one-time expense, 5 percent are to avoid an overdraft, and 23 percent are for shopping or a holiday.

For the 66 percent of first-time applicants who take out a loan, the average amount is about £290, with an APR of 663 percent and a maturity of 5.5 months. For these first loans, 35 percent end in default, 44 percent result in a top-up, with the remainder ending in the full repayment of the original balance. The high default rate is in line with the market for high-cost credit in the UK, where only 64 percent of payday loans issued in 2012 were repaid in full, either early or on time (Competition and Markets Authority, 2015).The average long-run profit for individuals taking out a loan, defined as the sum of all payments made by the applicant minus all disbursements from the Lender for both the first loan and all subsequent loans, is equal to £267. By definition, applicants who do not take out a loan have a 0 percent default rate, 0 percent top-up rate, 0 percent repayment rate, and yield profits of exactly £0.

C. Construction of the Instrumental Variable

We estimate the causal impact of loan take-up for the marginal loan applicant using a leave-out measure of loan examiner leniency as an instrumental variable for loan take-up. As discussed above, first-time loan applicants are assigned to loan examiners of the same nationality using a blind rotation system, effectively randomizing applicants to a subset of examiners within each branch. For example, Polish loan applicants visiting a particular branch on a particular day are randomly assigned to one of the Polish-speaking loan examiners working in that branch on that day. In contrast, native-born loan applicants are randomly assigned to the full set of examiners within each branch, including the Polish-speaking loan examiners. Importantly, the assigned loan examiner is given complete discretion to approve or reject these first-time loan applicants, leading to significant variation in approval rates across examiners.

Our empirical design relies on the fact that individuals assigned to a more lenient loan examiner are more likely to take up a loan, or, in other words, relies on a significant first stage relationship between loan take-up and examiner assignment. There are several potential reasons that examiners may differ in their leniency. Loan examiners could, for example, have different risk preferences, leading to different application decisions for marginal loan applicants. Loan examiners could also vary in their impatience or liquidity constraints, again leading to application decisions at the margin. A final possibility is that loan examiners differ in how they evaluate the same set of information, a potential concern if these differences lead to violations of the monotonicity assumption discussed below. We return to this issue in Section V, where we show suggestive evidence that variation in leniency is unlikely to be driven by differences in how examiners evaluate the same information, but rather by differences in examiner preferences.

We measure examiner leniency using a leave-out, residualized measure that accounts for the assignment process used by the Lender following Dahl, Kostøl and Mogstad (2014) and Arnold, Dobbie and Yang (2018). To construct this residualized examiner leniency measure, we first regress loan take-up on an exhaustive set of branch-by-month-by-nationality fixed effects, the level at which loan applicants are randomly assigned to loan examiners. We then use the residuals from this regression to calculate the leave-out mean examiner-by-branch take-up rate for each loan applicant. We calculate our instrument across all applicants assigned to an examiner within a branch to

increase the precision of our leniency measure. In robustness checks, we present results that use an instrument that is allowed to vary by nationality, age, and gender, as well as results that use an instrument based on loan approval instead of loan take-up.

Appendix Figure A3 presents the distribution of our residualized examiner leniency measure for loan take-up separately by nationality, age, and gender. Our sample includes 254 examiners, with the typical examiner-by-branch cell including 179 first-time loan applications. Controlling for branch-by-month-by-nationality fixed effects, our examiner leniency measure ranges from -0.165 to 0.195 with a standard deviation of 0.047. In other words, moving from the least to most lenient loan examiner increases the probability of loan take-up by 36.0 percentage points, a 55 percent change from the mean take-up rate of 66.1 percentage points.

D. Instrument Validity

Existence of First Stage: The first baseline assumption needed for our IV estimator is that examiner assignment is associated with loan take-up. To examine the first-stage relationship between examiner leniency (Z_{ite}) and loan take-up ($TakeUp_{ite}$), we estimate the following specification for applicant i assigned to examiner e at time t using a linear probability model:

$$TakeUp_{ite} = \gamma Z_{ite} + \pi \mathbf{X}_{it} + \mathbf{v}_{ite} \tag{8}$$

where, as described previously, Z_{ite} are leave-out (jackknife) measures of examiner leniency. The vector $\mathbf{X}_{it}$ includes branch-by-month-by-nationality fixed effects and the baseline controls in Table 1. The error term $\mathbf{v}_{ite}$ is composed of characteristics unobserved by the econometrician but observed by the examiner, as well as idiosyncratic variation unobserved to both the examiner and econometrician. Robust standard errors are clustered at the examiner level.

Figure 2 provides a graphical representation of the first stage relationship between our residualized measure of examiner leniency and the residualized probability of loan take-up that accounts for our exhaustive set of branch-by-month-by-nationality fixed effects, overlaid with the distribution of examiner leniency. Appendix Figure A3 presents the same results separately by nationality, age, and gender. Figure 2 and Appendix Figure A3 are a flexible analog to Equation (8), where we plot a local linear regression of residualized loan take-up against examiner leniency. The individual rate

of residualized loan take-up is monotonically increasing in our leniency measure for all groups. The first stage relationship between loan take-up and examiner leniency is also linear over nearly the entire distribution of our examiner leniency measure, consistent with the identifying assumptions discussed in Section II. These results are also consistent with our first baseline assumption that loan examiners exert significant discretionary judgment on the extensive margin.

Column 1 of Table 2 presents formal first stage results from Equation (8) for all applicants. Columns 1-6 of Appendix Table A3 present results separately by nationality, age, and gender. Consistent with the graphical results in Figure 2 and Appendix Figure A3, we find that our residualized examiner instrument is highly predictive of whether an individual receives a loan in both the full sample and within each subgroup. Table 2 shows, for example, that an applicant assigned to a loan examiner that is 10 percentage points more likely to approve a loan is 7.2 percentage points more likely to receive a loan in the full sample. Several other applicant characteristics are highly predictive of loan take-up. For example, women are 2.5 percentage points more likely to receive a loan compared to male applicants in the full sample, while applicants who are ten years older are 0.9 percentage points less likely to receive a loan compared to younger applicants.

Appendix Figure A4 shows the distribution of the number of advisors in a given store by applicant nationality by month of application cell. On average, each cell has 2.8 advisors and the median cell has 2 advisors. The branch-by-month-by-nationality fixed effects drop observations where there is only one advisor per cell, meaning that our estimates are obtained from cells with more than one advisor.

Exclusion Restriction: The second baseline assumption needed for our IV estimator is that examiner assignment only impacts applicant outcomes through the probability of receiving a loan. This assumption would be violated if examiner leniency is correlated with any unobservable determinants of future outcomes. Column 2 of Table 2 and columns 1-6 of Appendix Table A4 present a series of randomization checks to partially assess the validity of this exclusion restriction. Following the first stage results, we control for branch-by-month-by-nationality fixed effects and cluster standard errors at the examiner level. We find that examiners with differing leniencies are assigned observably identical applicants, both in the full sample and within each subgroup. None of the results suggest that there is systematic non-random assignment of applications to examiners.

The exclusion restriction could also be violated if examiner assignment impacts the profitability of a loan through channels other than loan take-up. The assumption that examiners only systematically affect loan outcomes through loan take-up is fundamentally untestable, but we argue that the exclusion restriction assumption is reasonable in our setting. Loan examiners only meet with applicants one time, and are forbidden, by law, to give advice or counsel applicants, leaving relatively little scope through which the assigned examiner could influence outcomes other than through loan take-up. Loan examiners are also only allowed to make discretionary judgments on whether to accept a loan application or not; the examiner has no discretion to affect the loan amount, interest rate, maturity date, the number of installments, or the amount of each installment, all of which are determined by the Lender. Thus, it seems unlikely that loan examiners would significantly impact loan applicants other than through the loan approval decision.

Monotonicity: The final baseline assumption needed for our IV estimator is that the impact of examiner assignment on loan take-up is monotonic across loan applicants. In our setting, the monotonicity assumption requires that applicants who receive a loan when assigned to a strict examiner would also receive a loan when assigned to a more lenient examiner, and that applicants not receiving a loan when assigned to a lenient examiner would also not receive a loan when assigned to a stricter examiner. To partially test the monotonicity assumption, Appendix Figure A5 plots examiner leniency measures that are calculated separately for each examiner by nationality, age, and gender. Consistent with our monotonicity assumption, examiners exhibit similar tendencies across observably different types of applicants. We also find a strong first-stage relationship across various applicant types in Appendix Table A3. None of the results suggest that the monotonicity assumption is invalid in our setting. In robustness checks, we also relax the monotonicity assumption by letting our leave-out measure of examiner leniency differ across applicant characteristics.

IV. Results

In this section, we present our main results applying our empirical test for bias in consumer lending. We then show the robustness of our results to alternative specifications, before comparing the results from our empirical test with standard tests based on OLS specifications.

A. Empirical Test for Bias

We estimate the profitability of marginal loan applicants using the following two-stage least squares specification for applicant i assigned to examiner e at time t:

$$Y_{ite} = \alpha_R^{IV} TakeUp_{ite} + B^{IV} TakeUp_{ite} \times TargetGroup_i + \phi \mathbf{X}_{it} + \mathbf{v}_{ite} \tag{9}$$

where Y_{ite} is the long-run profitability of the loan, as measured by the difference between total loan payments minus total loan disbursals. α_R^{IV} measures the profitability of the marginal loan to the reference group. B^{IV} is our measure of bias, or the difference in profitability between the reference and target group applicants. The vector $\mathbf{X}_{it}$ includes branch-by-month-by-nationality fixed effects and the baseline controls listed in Table 1. As described previously, the error term $\mathbf{v}_{ite} = \mathbf{U}_{ite} + \varepsilon_{ite}$ consists of characteristics unobserved by the econometrician but observed by the loan examiner, $\mathbf{U}_{ite}$, and idiosyncratic variation unobserved by both the econometrician and examiner, ε_{ite}. We instrument for loan take-up, $TakeUp_{ite}$, with our measure of examiner leniency, Z_{ite}. We similarly instrument for the interaction of loan take-up and target group status, $TakeUp_{ite} \times TargetGroup_i$, with the interaction of our examiner leniency measure and group status, $Z_{ite} \times TargetGroup_i$. Robust standard errors are clustered at the examiner level.

Estimates from Equation (9) are presented in Table 3. Column 1 reports results pooling across all applicants. Columns 2-4 report results with interactions for applicant nationality, age, and gender, respectively. Column 5 reports results with all interactions simultaneously. For completeness, Appendix Figure A3 provides a graphical representation of our reduced form results separately by nationality, age, and gender. Following the first stage results, we plot the reduced form relationship between our examiner leniency measure and the residualized profitability of loan take-up.

We find convincing evidence of bias against immigrants and older applicants using our IV estimator. We find that marginal loan applicants yield a profit of £331 following the initial loan decision (column 1), 24 percent larger than the mean profit level of £267. Marginal native-born applicants yield a profit of only £195, however, £568 less than marginal immigrant applicants (column 2), consistent with bias against immigrant applicants. We similarly find that marginal younger applicants yield a profit of £337, £348 less than marginal older applicants (column 3), consistent with

bias against older applicants as well. These estimates imply that marginal immigrant applicants are almost four times more profitable than marginal native-born applicants, while marginal older applicants are more than two times as profitable as marginal younger applicants. In contrast, we find that marginal female and male applicants yield similar profits, suggesting no bias against (or in favor of) female applicants.

B. Robustness

Threats to Interpretation: Our test for bias assumes that there are no differences in the true cost of lending to different groups. This assumption would be violated if, for example, there are differences in the systematic risk of lending to different groups. We explore this concern in Appendix Table A5, where we calculate long-run profits using a 10 percent discount rate for applicants from the reference group and a variety of higher discount rates for applicants from the target group. Our test is motivated by a standard CAPM model, where the higher discount rate for target group applicants captures the additional risk of lending to these applicants. We continue to find evidence of bias against immigrants and older applicants in Appendix Table A5 even when we assume that the target group applicants have a 50 percent higher discount rate, equivalent to assuming that these target group applicants are more than seven times riskier than reference group applicants at a market risk premium of 5 percent. In unreported results, we find that the bias against older and immigrant applicants is no longer statistically significant at a discount rate differential of 70 and 110 percent, respectively. These estimates indicate that our results are unlikely to be driven by differences in the systematic risk of lending to different groups.

Differences in Loan Take-Up by Group: In our model, we abstract away from the fact that loan examiners only approve or reject loan applicants. Loan applicants must then decide whether to take up the loan, which also impacts long-run profits. Extending our model to incorporate these institutional details means that bias could also be driven by, for example, examiners encouraging certain groups to take up loans conditional on loan approval.

We explore the empirical relevance of differences in loan take-up by group in two ways. First, we test whether loan approval has a larger impact on the probability of taking up a loan for marginal applicants from any particular group, which could occur if examiners encourage those groups to

take up the loans. To test this idea, Appendix Table A6 presents two-stage least squares estimates of the impact of loan approval on loan take-up using a leave-out measure based on loan approval as an instrumental variable. We find that loan approval has a nearly identical impact on loan take-up rates for all groups at the margin. Second, we directly estimate bias in the setting of loan approval versus loan rejection to incorporate any additional bias stemming from this margin. We estimate these effects using a two-stage least squares regression of long-run profits on loan approval, again using a leave-out measure based on loan approvals as an instrumental variable. Appendix Table A7 presents these estimates. We find similar estimates of bias when focusing on the loan approval versus loan rejection margin when we scale the estimated treatment effects by the "first stage" effect of loan approval on loan take-up from Appendix Table A6.

Alternative Specifications: Appendix Tables A8-A11 explore the sensitivity of our main results to a number of different specifications. Appendix Table A8 presents results where we use a net present value measure of long-run profits for a variety of different discount rates. Appendix Table A9 presents re-weighted estimates with the weights chosen to match the distribution of observable characteristics for target group loan applicants to explore whether differences in characteristics such as credit history or earnings can explain our results.[10] Appendix Table A10 presents results from an MTE estimator that puts equal weight on each examiner in our sample, instead of the IV weights as in our preferred specification.[11] Finally, Appendix Table A11 presents estimates where the instrument is calculated separately for each subgroup in the data, relaxing the monotonicity assumption. Results are generally similar to our preferred specification across all alternative specifications, although some of our estimates lose statistical significance. In particular, our MTE estimates in

[10] Arnold, Dobbie and Yang (2018) show that it is possible to test for bias holding fixed other group differences using a re-weighting procedure that weights the distribution of observables of the target group to match observables of the reference group in the spirit of DiNardo, Fortin and Lemieux (1996) and Angrist and Fernández-Val (2013). This narrower test for bias relies on the assumption that examiner preferences vary only by observable characteristics. See Appendix Table A12 for the complier characteristics used to construct the weights and Arnold, Dobbie and Yang (2018) for additional details.

[11] While the MTE estimator has the advantage of allowing the researcher to choose any weighting scheme across examiners, it comes at the cost of statistical precision and the additional functional form assumptions needed to interpolate estimates between observed values of the instrument. Following Arnold, Dobbie and Yang (2018), we estimate these MTE results using a two-step procedure. In the first step, we estimate the entire distribution of MTEs using the derivative of residualized profits with respect to variation in the propensity score provided by our instrument. To do this, we regress the residualized profit variable on the residualized examiner leniency measure to calculate the group-specific propensity score. We then compute the numerical derivative of a local quadratic estimator to estimate group-specific MTEs (see Appendix Figure A6). In the second step, we use the group-specific MTEs to calculate the level of bias for each individual examiner, and the simple average of these examiner-specific estimates. We calculate standard errors by bootstrapping this two-step procedure at the examiner level.

Appendix Table A10 are particularly noisy due to the increased weight put on a handful of imprecise examiner-level estimates compared to our preferred specification. There is also considerably more noise when using smaller cells to calculate the leave-out examiner leniency measure in Appendix Table A11. The estimates are not economically or statistically different across specifications, however, and none of the results suggest that our preferred estimates are invalid.

C. Comparison to OLS Estimates

Appendix Table A13 replicates the outcome tests from the prior literature (e.g., Han 2004) that rely on standard OLS estimates of Equation (9). In contrast to our IV test for bias, standard OLS estimates suggest much lower levels of bias against immigrant and older applicants. We find, for example, that the gap between the average native-born and immigrant applicant is only £102 (column 2), 82 percent lower than our IV estimate for marginal applicants in Table 3. The gap between the average younger and older applicant is also only £88 (column 3), 74 percent lower than our IV estimate for marginal applicants. Standard OLS estimates also suggest, incorrectly, that there is bias against female applicants (column 4). Taken together, these results highlight the importance of accounting for both infra-marginality and omitted variables when testing for bias in consumer lending.

V. The Misalignment of Examiner and Lender Incentives

In this section, we attempt to differentiate bias due to the misalignment of firm and examiner incentives from explanations based on prejudice or inaccurate stereotypes. The model of bias based on misaligned examiner incentives developed in Section II has three testable implications. First, that there should be no evidence of bias when using the short-run default measure used to evaluate loan examiners' performance. Second, loan examiners should make decisions as though they are ranking applicants based on the short-run default measure used to evaluate their performance, even when these rankings diverge from a ranking based on long-run profits. Finally, incentive-based bias should only emerge among groups where short- and long-run outcomes diverge. We discuss each of these tests below, before providing more suggestive evidence against alternative models based on prejudice and inaccurate stereotypes.

A. Short-Run Default

The first testable hypothesis from our incentive-based model of bias is that if examiners make lending decisions to minimize short-run default probability, but their decisions are not distorted either by prejudice nor by inaccurate beliefs, then the Becker outcome test applied to short-run outcomes should show no bias. In terms of the model developed in Section II, we should therefore find that $\alpha_T^{SR} = \alpha_R^{SR}$.

To test whether there is bias when using short-run default as the outcome, we estimate the default risk of marginal loan applicants using our IV estimator in Table 4. Consistent with the misalignment of examiner incentives, there are no statistically significant differences in the first-loan default risk of marginal loan applicants by nationality, age, or gender. Marginal first-loan loan applicants, in general, default on 44.7 percent of loans, 28 percent more than the average default rate of 35.0 percent. Marginal immigrant applicants, however, are only 2.5 percentage points less likely to default on the first loan than marginal native-born applicants. Marginal older applicants are 14.4 percentage points less likely to default on the first loan than marginal younger applicants, and marginal female applicants are 8.3 percentage points less likely to default than marginal male applicants, with none of these differences being statistically significant.[12]

These results confirm that examiners are making unbiased decisions based on the short-term default outcome used to evaluate their performance. These findings are not only consistent with our incentive-based model of bias, but they also suggest that animus and inaccurate stereotypes are not affecting examiners' decisions at the margin. A general implication of our incentive-based model of bias is that it is not always possible to infer examiners' utility or beliefs using long-run profits. Learning about examiners' utility or beliefs instead requires using the short-run measure used to evaluate their performance, just as we have done here.

B. Examiner Decision Rule

The second testable hypothesis from our incentive-based model of bias is that loan examiners should make decisions as though they are ranking applicants based on the short-run default measure used to measure their performance, α_i^{SR}, even when these rankings diverge from a ranking based on long-

[12]Appendix Table A14 estimates bias using default in the first month of the loan. We again see no evidence of bias against immigrants or older applicants using this even shorter-run measure of default.

run profits, α_i. We can evaluate examiners' objective function by contrasting examiners' decisions with two different data-based decision rules, one based on short-run default and the other based on long-run profits.

To implement this test for misaligned examiner incentives, we first estimate predicted short- and long-run outcomes using an ML algorithm that efficiently uses all observable applicant characteristics. In short, we use a randomly-selected subset of the data to train the model using all individuals who receive a loan. In training the model, we must choose the shrinkage, the number of trees, and the depth of each tree. Following common practice (e.g., Kleinberg et al. 2018), we choose the smallest shrinkage parameter (i.e., 0.005) that allows the training process to run in a reasonable time frame. We use a five-fold cross-validation on the training sample to choose the optimal number of trees for the predictions. The interaction depth is set to four, which allows each tree to use at most four variables. Using the optimal number of trees from the cross-validation step, predicted outcomes are then created for the full sample.[13,14]

One important challenge is that we only observe outcomes for applicants who receive a loan, not those who do not. This missing data problem makes it hard to evaluate counterfactual decision rules based on algorithmic predictions or to identify the implicit decision-rule used by loan examiners. To overcome this missing data problem, we follow Kleinberg et al. (2018) and start with the set of loan applicants receiving a loan from the most lenient examiners. From this set of applicants, we then choose additional applicants to hypothetically reject according to the predicted outcomes calculated using our ML algorithm. For each additional hypothetical rejection, this allows us to calculate the hypothetical change in profitability or default risk for the now smaller set of applicants that would have received a loan. Importantly, the hypothetical change in profitability and default risk can be compared to the outcomes produced by the stricter examiners because applicant characteristics are, on average, similar across examiners due to the quasi-random assignment of applicants to examiners.

[13]Appendix Table A15 presents the correlates of our predicted profitability measure. Predicted profitability is increasing in the credit score used by the lender. Predicted profitability is also higher for female applicants, older applicants, and applicants with more dependents.

[14]One potential concern is that our measures of predicted profitability may be biased if loan examiners base their decisions on variables that are not observed by the econometrician (e.g., demeanor during the loan application). Following Kleinberg et al. (2018), we test for the importance of unobservables in loan decisions by splitting our sample into a training set to generate the profitability predictions and a test set to test those predictions. We find that our measure of predicted profitability from the training set is a strong predictor of true profitability in the test set, indicating that our measure of predicted profitability is not systematically biased by unobservables (see Appendix Figure A7).

Figure 3 presents these results for both long-run profits and short-run default. The solid black curve calculates the change in profitability or default that would have resulted if additional applicants had been rejected in order of the algorithm's predicted profitability and default rates. Each of the points denotes the different examiner leniency quintiles. We also plot the change in profitability and default rates that would have resulted if we used a decision rule based on an ML algorithm that does not include nationality, age, and gender, a decision rule based on the baseline credit scores used to screen loan applicants, and a decision rule based on a random number generator.

We find that the decisions made by loan examiners are strikingly consistent with a data-based decision rule minimizing short-run default, but inconsistent with a decision rule maximizing long-run profits. Long-run profits decrease as we move from the most lenient to most strict examiners, worse than a decision rule based on a random number generator and far worse than a decision rule based on our ML algorithm.[15] In contrast, loan examiners are nearly as effective as our ML algorithm in decreasing default rates. The second quintile of examiners, for example, reduce the default rate by 2.3 percentage points relative to the most lenient quintile examiners by increasing the rejection rate by 8.5 percentage points. Our ML algorithm using all characteristics could have decreased the default rate by 4.8 percentage points with the same 8.5 percentage point increase in the rejection rate, or just 2.5 percentage points better than the loan examiners. The ML algorithm using only allowable characteristics could have similarly decreased the default rate by 4.1 percentage points, or 1.8 percentage points more than the loan examiners.[16,17]

Figure 3 also shows that examiners are almost as good as the ML algorithm at predicting default *across* the leniency distribution. This finding suggests that the variation in examiner leniency is

[15]For example, the second quintile of examiners reduce profitability by £29 per applicant relative to the most lenient quintile examiners by increasing the rejection rate by 8.5 percentage points. In contrast, the ML algorithm using all characteristics could have increased profits by £38 per applicant with the same 8.5 percentage point increase in the rejection rate. The ML algorithm using only allowable characteristics could have increased profits by £33 per applicant, just £6 less than the full algorithm, while the rule based on baseline credit scores could have only increased profits by £9 per applicant.

[16]In contrast, the credit score used by the Lender could have decreased the default rate by only 0.5 percentage points, 1.8 percentage points less than the loan examiners and only 0.6 percentage points more than a random decision rule. The poor performance of the credit score variable is likely driven by the fact that the credit score variable purchased by the Lender is calibrated to the entire credit market, not the subprime market that the Lender operates in.

[17]Appendix Figures A9 and A10 show a comparison of examiners' decisions with the ML algorithm for examiners with longer and shorter tenures at the Lender. More experienced examiners perform relatively worse when the predicted outcome is long-run profits, but slightly better when the outcome is short-run default. These findings are consistent with examiners becoming slightly better at ranking loan applicants over time, at least with respect to short-run default. We view these results as further evidence in support of our incentives-based model of bias.

unlikely to be driven by differences in how examiners evaluate the same information, but is instead driven by differences in, for example, risk preferences across examiners.

C. Misalignment of Short- and Long-Run Outcomes

The final testable hypothesis from our incentive-based model is that bias should only emerge among groups where short- and long-run outcomes diverge. If the observed bias is due to incentive misalignment, we should therefore observe that immigrant and older applicants have a higher short-run default probability for a given long-run profitability relative to native-born and younger applicants. No such pattern should exist for female applicants compared to male applicants if our model is correct.

To test whether immigrant and older applicants are both high-profit and high-default, we plot the distributions of predicted long-run profits and predicted short-run default by nationality, age, and gender in Figure 4. We calculate predicted profits and predicted default risk using the ML algorithm described in the previous subsection. Consistent with the misalignment of examiner incentives, both immigrant and older applicants are visually more likely to default in the short-run for a given level of long-run profits compared to native-born and younger applicants. In contrast, female and male applicants are equally as likely to default for a given level of long-run profits.

To provide a more formal test of this hypothesis, Appendix Table A16 presents OLS results regressing predicted long-run profits on applicant characteristics and a quadratic in predicted short-run default. Controlling for predicted default, immigrant applicants have predicted profits that are £36 larger than native applicants (column 1), while older applicants have predicted profits that are £31 larger than younger applicants. In contrast, male and female applicants have statistically identical predicted profits for a given level of predicted default.

A final question is why immigrant and older applicants are more likely to default in the short-run compared to native-born and younger applicants. One possible explanation is that immigrant and older applicants are less liquid than native-born and younger applicants, and, as a result, more susceptible to the kinds of unanticipated income or expense shocks that lead to default. Consistent with this explanation, we find that immigrant and older applicants have lower credit scores at a given level of income compared to native-born and younger applicants. Unfortunately, we do not have the necessary panel data on income and expenditures to further test this hypothesis or explore

alternative explanations.

D. Models Based on Prejudice and Inaccurate Stereotypes

We conclude this section by discussing more suggestive evidence against alternative models that could explain our main results.

First, we consider the possibility that loan examiners either knowingly or unknowingly discriminate against immigrant and older applicants at the margin. Loan examiners could, for example, harbor explicit biases against immigrant and older applicants that leads them to exaggerate the cost of lending to these individuals (e.g., Becker 1957, 1993). Loan examiners could also harbor implicit biases against immigrant and older applicants, leading to biased lending decisions despite the lack of any explicit prejudice (e.g., Greenwald et al. 2009). However, immigrant applicants are typically matched to loan examiners from the same ethnic background and all loan examiners tend to be older themselves, institutional features that are inconsistent with most models of ethnic or age-related prejudice. We also find no bias against female applicants even among male examiners, another finding that is inconsistent with the simplest models of prejudice, although we note that these results are very imprecise (see Appendix Table A18).[18] These results suggest that either prejudice is not driving our results or that loan examiners are prejudiced against immigrant and older applicants despite sharing those same characteristics.

Second, we consider the possibility that loan examiners are making biased prediction errors, potentially due to inaccurate stereotypes against immigrant and older applicants. Bordalo et al. (2016) show, for example, that representativeness heuristics—probability judgments based on the most distinctive differences between groups—can exaggerate perceived differences between groups. In our setting, these kinds of group-based heuristics or inaccurate stereotypes could lead loan examiners to systematically underestimate the potential profitability of lending to immigrant and older applicants relative to native-born and younger applicants at the margin.

Following Arnold, Dobbie and Yang (2018), we first explore whether our data are consistent with the formation of negative stereotypes that could lead to these kinds of biased prediction errors. Extending Bordalo et al. (2016) to our setting, negative stereotypes against immigrant and older

[18] We cannot estimate results separately by examiner ethnicity, as immigrant applicants are typically matched to loan examiners from a similar ethnic background. We also cannot estimate results separately by examiner age, as our data do not include this variable.

loan applicants should only be present if these applicants are over-represented in the left tail of the predicted profit distribution compared to native-born and younger loan applicants. Appendix Figure A11 presents the distribution of the predicted long-run profits by nationality, age, and gender, where the predicted profits are calculated using the ML algorithm described below and the full set of baseline applicant characteristics in our data. Results for each individual characteristic in our predicted risk measure are also presented in Appendix Table A12. In stark contrast to the predictions of the Bordalo et al. (2016) model, we find that both immigrant and older loan applicants are significantly over-represented in the right tail of the predicted profit distribution. For example, immigrant applicants are 2.1 times more likely than native-born applicants to be represented among the top 25 percent of the predicted profit distribution, while older loan applicants are 2.6 times more likely than younger applicants to be represented among the top 25 percent.

We can also test for biased prediction errors by examining situations where prediction errors of any kind are more likely to occur. One such test for biased prediction errors uses a comparison of experienced and inexperienced loan examiners, as examiners may be less likely to rely on inaccurate group stereotypes as they acquire greater on-the-job experience, at least in settings with limited information and contact. To test this idea, Appendix Table A19 presents subsample results for more and less experienced examiners, where we measure experience using an indicator for being employed by the Lender when our sample period begins. There are no systematic patterns by examiner experience and, if anything, the estimates suggest more bias against immigrants among more experienced loan examiners. Taken together with Appendix Figure A11, these results suggest that inaccurate stereotypes are unlikely to be driving our results.

VI. Conclusion

In this paper, we test for bias in consumer lending using the quasi-random assignment of loan examiners to identify the profitability of marginal loan applicants. We find that there is substantial bias against immigrant and older loan applicants, ruling out statistical discrimination and omitted variable bias as the sole explanations for the disparities in credit availability for these groups. Our results also cannot be explained by other ethnic or age-related differences in baseline characteristics, systematic differences in the level of risk across groups, or the way that the IV estimator averages

the level of bias across different examiners.

Three additional findings are consistent with our results being driven by the misalignment between firm and examiner incentives. First, there is no evidence of bias against immigrant or older applicants when we use the short-run default outcome used to evaluate examiner performance. Second, the decisions made by loan examiners are strikingly consistent with a data-based decision rule minimizing short-term default, but inconsistent with a decision rule maximizing long-term profits. Finally, immigrant and older applicants are more likely to default in the short run compared to native-born and younger applicants with the same level of expected long-run profits, while no such differences exist for female and male applicants. Taken together, these three results suggest that examiners are equalizing the private returns of lending across groups at the margin, just as predicted by our incentive-based model of bias. In contrast, none of these findings can be easily explained by models based on prejudice and inaccurate stereotypes.

An important implication of our analysis is that incentive-based bias is likely to emerge whenever a group is disproportionately exposed to the kinds of income and expenditure shocks that lead to short-run default. Our findings therefore suggest that policymakers should consider the incentives created by examiner contracts when addressing bias in consumer lending markets, not just the possibility of prejudice or inaccurate stereotypes. Another implication of our analysis is that a data-based decision rule based on long-run profits could simultaneously increase profits and eliminate bias compared to examiner-based decisions. In our setting, for example, the Lender would earn approximately £157 more in profit per applicant if marginal lending decisions were made using the machine learning algorithm rather than loan examiners, a 58 percent increase from the mean. Including all applicants, not just those at the margin of loan take-up, the Lender would earn over £53 more per applicant if lending decisions were made using the machine learning algorithm, a 30 percent increase from the mean.[19,20]

[19]Following Kleinberg et al. (2018), we compute the gains from the machine learning algorithm relative to the average examiner in our test sample using a three-step process. First, we impute long-run profits for applicants who never took up a loan under the assumption that our algorithm perfectly predicts true long-run profits and that all selection is on observable characteristics. Second, we re-rank all applicants, both rejected and approved, and select a hypothetical group to approve using predicted long-run profits. Finally, we compare the predicted long-run profitability of applicants for both the algorithm and average examiner, both at the margin of loan approval and aggregating across all loan applicants.

[20]We can also test whether a data-based decision rule is biased by comparing the observed and predicted profitability of applicants from different groups. Following the logic of our outcome test developed above, the observed profitability at each level of predicted profitability will be identical across groups if our algorithm is unbiased. In contrast, if our algorithm is biased against, say, non-native applicants, the observed profitability of non-natives at a

There are three important caveats to our analysis. First, while misaligned incentives are commonplace in most consumer credit markets, the exact magnitude of the bias may differ in these other credit markets. Second, given that cultural proximity between borrowers and lenders has been shown to reduce the incidence of bias in other settings (see Fisman, Paravisini and Vig 2017), prejudice and inaccurate stereotypes may be more important factors in other credit markets where, for example, loan examiners are not matched to culturally and ethnically similar loan applicants. Finally, the welfare effects of credit availability, particularly for high-cost credit, are largely unknown (e.g., Agarwal, Skiba and Tobacman 2009, Melzer 2011, Morgan, Strain and Seblani 2012, Morse 2011, Gathergood, Guttman-Kenney and Hunt 2019, Liberman, Paravisini and Pathania 2016, Zaki 2016). Given these concerns, we are unable to estimate the welfare effects of bias in consumer lending decisions using our data and research design. Developing a framework to assess the precise welfare effects of bias in consumer credit decisions is an important area of future work.

given level of predicted profitability will be higher than the observed profitability of natives. There is no evidence of bias by nationality, age, or gender (see Appendix Figure A12).

References

Agarwal, Sumit, and Itzhak Ben-David. 2018. "Loan Prospecting and the Loss of Soft Information." *Journal of Financial Economics*, 129(3): 608–628.

Agarwal, Sumit, Paige Marta Skiba, and Jeremy Tobacman. 2009. "Payday Loans and Credit Cards: New Liquidity and Credit Scoring Puzzles?" *American Economic Review*, 99(2): 412–417.

Alesina, Alberto F., Francesca Lotti, and Paolo Emilio Mistrulli. 2013. "Do Women Pay More for Credit? Evidence from Italy." *Journal of the European Economic Association*, 11(1): 45–66.

Angrist, Joshua, and Ivan Fernández-Val. 2013. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." *Advances in Economics and Econometrics: Volume 3, Econometrics: Tenth World Congress*, , ed. D. Acemoglu, M. Arellano and E. Dekel Vol. 51, 401–434. Cambridge University Press.

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics*, 133(4): 1885–1932.

Arrow, Kenneth. 1972. "Some Mathematical Models of Race Discrimination in the Labor Market." *Racial Discrimination in Economic Life*, 187–204. Lexington, Mass.: Lexington Books.

Arrow, Kenneth. 1973. "The Theory of Discrimination." *Discrimination in Labor Markets*, , ed. Orley Ashenfelter and Albert Rees, 3–33. Princeton University Press.

Ayres, Ian. 2002. "Outcome Tests of Racial Disparities in Police Practices." *Justice Research and Policy*, 4(1-2): 131–142.

Bartlett, Richard, Adair Morse, Richard Stanton, and Nany Wallace. 2018. "Consumer Lending Discrimination in the FinTech Era." *Unpublished Working Paper.*

Bayer, Patrick, Fernando Ferreira, and Stephen L. Ross. 2017. "What Drives Racial and Ethnic Differences in High-Cost Mortgages? The Role of High-Risk Lenders." *The Review of Financial Studies*, 31(1): 175–205.

Becker, Gary S. 1957. *The Economics of Discrimination.* University of Chicago Press.

Becker, Gary S. 1993. "Nobel Lecture: The Economic Way of Looking at Behavior." *Journal of Political Economy*, 101(3): 385–409.

Bellucci, Andrea, Alexander Borisov, and Alberto Zazzaro. 2010. "Does Gender Matter in Bank-Firm Relationships? Evidence from Small Business Lending." *Journal of Banking & Finance*, 34(12): 2968–2984.

Berg, Tobias, Manju Puri, and Jorg Rocholl. 2013. "Loan Officer Incentives and the Limits of Hard Information." *NBER Working Paper No. 19051.*

Berkovec, James A., Glenn B. Canner, Stuart A. Gabriel, and Timothy H. Hannan. 1994. "Race, Redlining, and Residential Mortgage Loan Performance." *The Journal of Real Estate Finance and Economics*, 9(3): 263–294.

Berkovec, James A., Glenn B. Canner, Stuart A. Gabriel, and Timothy H. Hannan. 1998. "Discrimination, Competition, and Loan Performance in FHA Mortgage Lending." *Review of Economics and Statistics*, 80(2): 241–250.

Blanchflower, David G., Phillip B. Levine, and David J. Zimmerman. 2003. "Discrimination in the Small-Business Credit Market." *Review of Economics and Statistics*, 85(4): 930–943.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2018. "The Dynamics of Discrimination: Theory and Evidence." *PIER Working Paper No. 18-016.*

Bordalo, Pedro, Katherine B. Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review*, 109(3): 739–773.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics*, 131(4): 1753–1794.

Buchak, Greg, and Adam Jørring. 2016. "Does Competition Reduce Racial Discrimination in Lending?" *Unpublished Working Paper.*

Carrell, Scott E., and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy*, 118(3): 409–432.

Cavalluzzo, Ken S., and Linda C. Cavalluzzo. 1998. "Market Structure and Discrimination: The Case of Small Businesses." *Journal of Money, Credit and Banking*, 30(4): 771–792.

Cavalluzzo, Ken S., Linda C. Cavalluzzo, and John D. Wolken. 2002. "Competition, Small Business Financing, and Discrimination: Evidence from a New Survey." *The Journal of Business*, 75(4): 641–679.

Charles, Kerwin Kofi, and Erik Hurst. 2002. "The Transition to Home Ownership and the Black-White Wealth Gap." *Review of Economics and Statistics*, 84(2): 281–297.

Charles, Kerwin Kofi, Erik Hurst, and Melvin Stephens. 2008. "Rates for Vehicle Loans: Race and Loan Source." *American Economic Review*, 98(2): 315–320.

Coffman, Katherine B. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *The Quarterly Journal of Economics*, 129(4): 1625–1660.

Coffman, Katherine B., Christine L. Exley, and Muriel Niederle. 2018. "When Gender Discrimination is Not About Gender?" *Harvard Business School Working Paper No. 18-054*.

Cohen-Cole, Ethan. 2011. "Credit Card Redlining." *Review of Economics and Statistics*, 93(2): 700–713.

Cole, Shawn, Martin Kanz, and Leora Klapper. 2015. "Incentivizing Calculated Risk-Taking: Evidence from an Experiment with Commercial Bank Loan Officers." *Journal of Finance*, 52(2): 537–575.

Competition and Markets Authority. 2015. "Payday Lending Market Investigation: Final Report." `https://assets.publishing.service.gov.uk/media/54ebb03bed915d0cf7000014/Payday_investigation_Final_report.pdf` (accessed May 5, 2020).

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2016. "From LATE to MTE: Alternative Methods for the Evaluation of Policy Interventions." *Labour Economics*, 41: 47–60.

Dahl, Gordon B., Andreas Ravndal Kostøl, and Magne Mogstad. 2014. "Family Welfare Cultures." *Quarterly Journal of Economics*, 129(4): 1711–1752.

Deku, Solomon Y., Alper Kara, and Philip Molyneux. 2016. "Access to Consumer Credit in the UK." *The European Journal of Finance*, 22(10): 941–964.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica*, 64(5): 1001–1044.

Dobbie, Will, and Jae Song. 2015. "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection." *American Economic Review*, 105(3): 1272–1311.

Dobbie, Will, Paul Goldsmith-Pinkham, and Crystal S. Yang. 2017. "Consumer Bankruptcy and Financial Health." *Review of Economics and Statistics*, 99(5): 853–869.

Field, Erica, Rohini Pande, John Papp, and Natalia Rigol. 2013. "Does the Classic Microfinance Model Discourage Entrepreneurship among the Poor? Experimental Evidence from India." *American Economic Review*, 103(6): 2196–2226.

Fisman, Raymond, Daniel Paravisini, and Vikrant Vig. 2017. "Cultural Proximity and Loan Outcomes." *American Economic Review*, 107(2): 457–92.

Gathergood, John, Ben Guttman-Kenney, and Stefan Hunt. 2019. "How Do Payday Loans Affect Borrowers? Evidence from the U.K. Market." *The Review of Financial Studies*, 32(2): 496–523.

Gravelle, Hugh, Matt Sutton, and Ada Ma. 2010. "Doctor Behaviour Under a Pay for Performance Contract: Treating, Cheating and Case Finding?" *The Economic Journal*, 120(542): F129–F156.

Greenwald, Anthony G., T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji. 2009. "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity." *Journal of Personality and Social Psychology*, 97(1): 17–41.

Hanson, Andrew, Zackary Hawley, Hal Martin, and Bo Liu. 2016. "Discrimination in Mortgage Lending: Evidence from a Correspondence Experiment." *Journal of Urban Economics*, 92: 48–65.

Han, Song. 2004. "Discrimination in Lending: Theory and Evidence." *The Journal of Real Estate Finance and Economics*, 29(1): 5–46.

Healy, Paul M. 1985. "The Effect of Bonus Schemes on Accounting Decisions." *Journal of Accounting and Economics*, 7(1): 85 – 107.

Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives*, 12(2): 101–116.

Heider, Florian, and Roman Inderst. 2012. "Loan Prospecting." *The Review of Financial Studies*, 25(8): 2381–2415.

Hertzberg, Andrew, Jose Maria Liberti, and Daniel Paravisini. 2010. "Information and Inventives Inside the Firm: Evidence from Loan Officer Rotation." *Journal of Finance*, 65(3): 795–828.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*, 7: 24–52.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Keys, Benjamin J, Tanmoy Mukherjee, Amit Seru, and Vikrant Vig. 2010. "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans." *Quarterly Journal of Economics*, 125(1): 307–362.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

Liberman, Andres, Daniel Paravisini, and Vikram Pathania. 2016. "High-Cost Debt and Borrower Reputation: Evidence from the U.K." *Unpublished Working Paper.*

Marx, Philip. 2018. "An Absolute Test of Racial Prejudice." *Unpublished Working Paper.*

Melzer, Brian T. 2011. "The Real Costs of Credit Access: Evidence from the Payday Lending Market." *Quarterly Journal of Economics*, 126(1): 517–555.

Morgan, Donald P., Michael R. Strain, and Ihab Seblani. 2012. "How Payday Credit Access Affects Overdrafts and Other Outcomes." *Journal of Money, Credit and Banking*, 44(2-3): 519–531.

Morse, Adair. 2011. "Payday Lenders: Heroes or Villains?" *Journal of Financial Economics*, 102(1): 28–44.

Oyer, Paul. 1998. "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality." *The Quarterly Journal of Economics*, 113(1): 149–185.

Peterson, Richard L. 1981. "An Investigation of Sex Discrimination in Commercial Banks' Direct Consumer Lending." *The Bell Journal of Economics*, 12(2): 547–561.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review*, 62(4): 659–661.

Pope, Devin G., and Justin R. Sydnor. 2011. "What's in a Picture? Evidence of Discrimination from Prosper.com." *Journal of Human Resources*, 46(1): 53–92.

Qian, Jun, Philip E. Strahan, and Zhishu Yang. 2015. "The Impact of Incentives and Communication Costs on Information Production and Use: Evidence from Bank Lending." *The Journal of Finance*, 70(4): 1457–1493.

Ross, Stephen L., Margery Austin Turner, Erin Godfrey, and Robert R. Smith. 2008. "Mortgage Lending in Chicago and Los Angeles: A Paired Testing Study of the Pre-Application Process." *Journal of Urban Economics*, 63(3): 902–919.

Stegman, Michael A. 2007. "Payday Lending." *Journal of Economic Perspectives*, 21(1): 169–190.

Van Order, Robert, Vassilis Lekkas, and John M. Quigley. 1993. "Loan Loss Severity and Optimal Mortgage Default." *Real Estate Economics*, 21(4): 353–371.

Zaki, Mary. 2016. "Access to Short-Term Credit and Consumption Smoothing Within the Paycycle." *Unpublished Working Paper.*

Table 1: Descriptive Statistics

	All Loans	Loan Taken-Up	Loan Not Taken-Up
Panel A: Applicant Characteristics	(1)	(2)	(3)
Immigrant	0.402	0.409	0.388
Age	33.883	33.631	34.381
Female	0.559	0.570	0.539
One-Plus Years at Residence	0.731	0.772	0.651
Married	0.415	0.373	0.496
Number of Dependents	0.973	1.054	0.813
Credit Score	538.174	545.893	522.264
Has Bank Account	0.912	0.938	0.861
Has Other Loan Payments	0.288	0.349	0.167
Loan Amount Requested (£)	409.039	396.519	433.721
Loan for Emergency	0.266	0.267	0.265
Loan for Large One-Time Expense	0.106	0.109	0.100
Loan for Overdraft Avoidance	0.052	0.053	0.052
Loan for Shopping or Holiday	0.231	0.233	0.227
Panel B: Loan Characteristics			
Loan APR (%)	—	663.139	—
Loan Duration (Months)	—	5.498	—
Loan Amount Net of Fees (£)	—	289.897	—
Panel C: Loan Outcomes			
Loan Take-up	0.663	1.000	0.000
Loan Approved	0.755	1.000	0.271
Loan Default	0.232	0.350	0.000
Loan Top-Up	0.290	0.438	0.000
Long-Run Profits (£)	177.229	267.114	0.000
Observations	45,507	30,192	15,315

Note: This table reports descriptive statistics. The sample consists of first-time loan applicants assigned to a loan examiner between 2012 and 2014. We drop online and phone applicants, applicants younger than 18 or older than 75 years old, applicants assigned to examiners with fewer than 50 observations, and applicants that are unique to a store-by-month-by-nationality cell. Loan uses are self-reported at the time of application. Long-run profits are the sum of all payments made from the applicant to the Lender over the course of their entire relationship minus all disbursements from the Lender to the applicant. Loan default, top-up, and profits are all equal to zero for applicants not taking out a loan. See the data appendix for additional details on the variable construction.

Table 2: First Stage and Balance Tests

	Loan Take-Up	Examiner Leniency
	(1)	(2)
Examiner Leniency	0.71983***	
	(0.07351)	
Age	−0.00089***	−0.00001
	(0.00024)	(0.00002)
Female	0.02451***	−0.00022
	(0.00487)	(0.00063)
One-Plus Years at Residence	0.10186***	−0.00071
	(0.00868)	(0.00110)
Married	−0.08338***	−0.00000
	(0.00693)	(0.00106)
Number of Dependents	0.03315***	0.00006
	(0.00255)	(0.00022)
Credit Score (/1000)	1.55005***	0.00761*
	(0.05059)	(0.00421)
Has Bank Account	0.13770***	0.00370**
	(0.01203)	(0.00160)
Has Other Loan Payments	0.21633***	−0.00027
	(0.00874)	(0.00087)
Loan Amount Requested (£/1000)	−0.09962***	−0.00046
	(0.00626)	(0.00071)
Loan for Emergency	0.00373	−0.00125
	(0.00556)	(0.00341)
Loan for Large One-Time Expense	0.02052**	0.00091
	(0.00873)	(0.00389)
Loan for Overdraft Avoidance	−0.00150	−0.00134
	(0.00868)	(0.00382)
Loan for Shopping or Holiday	0.00662	−0.00108
	(0.00572)	(0.00338)
Dep. Variable Mean	0.663	0.000
Observations	45,507	45,507
p-value on Joint F-test	[0.000]	[0.426]
Clusters	254	254

Note: This table reports first stage results and balance tests. The regressions are estimated on the sample described in the notes to Table 1. Examiner leniency is estimated using data from other loan applicants assigned to an examiner following the procedure described in Section III. Column 1 reports estimates from an OLS regression of loan take-up on the variables listed. Column 2 reports estimates from an OLS regression of examiner leniency on the variables listed. The p-value reported at the bottom of the columns is for an F-test of the joint significance of the variables listed in the rows. All specifications control for store-by-month-by-nationality fixed effects. Standard errors clustered at the examiner level are reported in parentheses. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Table 3: Loan Take-Up and Long-Run Profits

	2SLS Estimates				
	(1)	(2)	(3)	(4)	(5)
Loan Take-Up	331.193***	194.851***	337.473***	168.296*	−46.970
	(60.338)	(63.480)	(98.566)	(97.381)	(146.413)
Take-Up x Immigrant Applicant		567.911***			605.517***
		(173.167)			(188.001)
Take-Up x Older Applicant			348.469**		350.255**
			(163.583)		(163.504)
Take-Up x Female Applicant				−10.735	118.053
				(128.474)	(133.994)
Dep. Variable Mean	177.229	177.229	177.229	177.229	177.229
Observations	45,507	45,507	45,507	45,507	45,507
Clusters	254	254	254	254	254

Note: This table reports IV estimates of bias in consumer lending decisions using long-run profits as an outcome. Each column reports two-stage least squares estimates of the impact of loan take-up on long-run profits using the sample described in the notes to Table 1. We instrument for loan take-up using the leave-out examiner leniency measure constructed using the procedure described in Section III, and for the interaction of loan take-up and applicant characteristics using the interaction of leave-out leniency and the same characteristic. All specifications control for store-by-month-by-nationality fixed effects and the baseline characteristics listed in Panel A of Table 1. Standard errors clustered at the examiner level are reported in parentheses. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

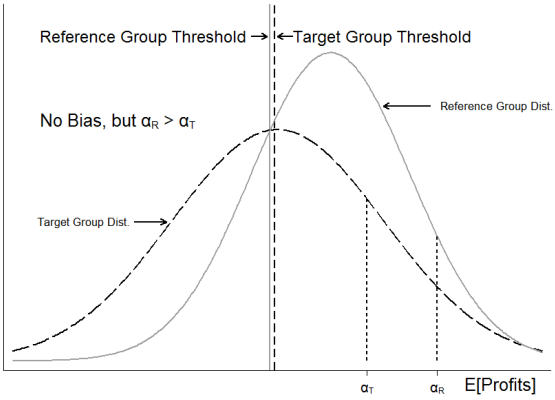
Table 4: Loan Take-Up and Short-Run Default

	2SLS Estimates				
	(1)	(2)	(3)	(4)	(5)
Loan Take-Up	0.447***	0.453***	0.495***	0.514***	0.586***
	(0.050)	(0.057)	(0.086)	(0.072)	(0.113)
Take-Up x Immigrant Applicant		−0.025			−0.058
		(0.114)			(0.123)
Take-Up x Older Applicant			−0.144		−0.148
			(0.104)		(0.104)
Take-Up x Female Applicant				−0.083	−0.097
				(0.119)	(0.124)
Dep. Variable Mean	0.232	0.232	0.232	0.232	0.232
Observations	45,507	45,507	45,507	45,507	45,507
Clusters	254	254	254	254	254

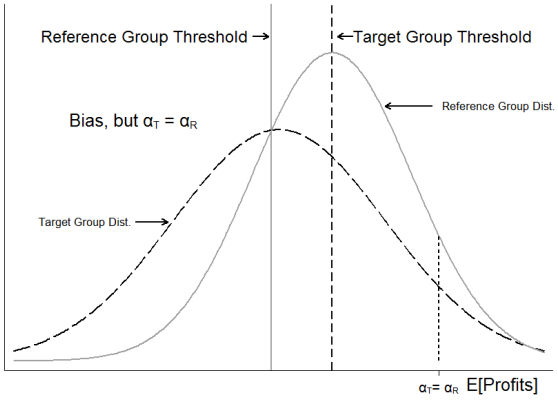
Note: This table reports IV estimates of bias in consumer lending using short-run default as an outcome. Each column reports two-stage least squares estimates using the sample described in the notes to Table 1. We instrument for loan take-up using the leave-out examiner leniency measure constructed using the procedure described in Section III, and for the interaction of loan take-up and applicant characteristics using the interaction of leave-out leniency and the same characteristic. All specifications control for store-by-month-by-nationality fixed effects and the baseline characteristics listed in Panel A of Table 1. Standard errors clustered at the examiner level are reported in parentheses. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Figure 1: Illustration of Estimation Problem

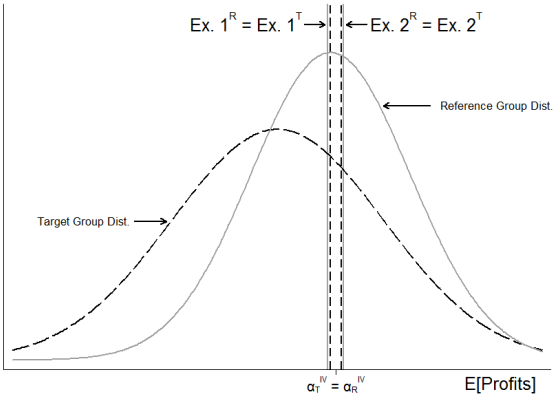
(a) Case 1: OLS with Unbiased Examiner



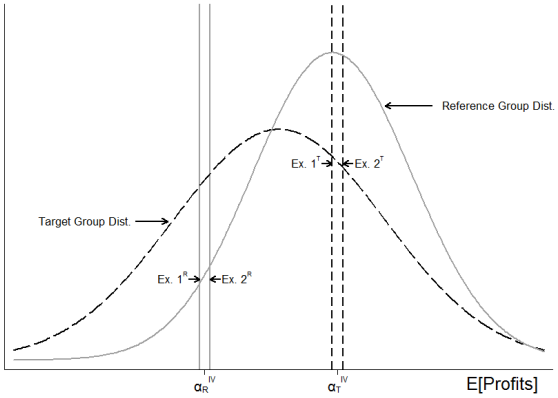
(b) Case 2: OLS with Biased Examiner



(c) Case 3: IV with Unbiased Examiners



(d) Case 4: IV with Biased Examiners

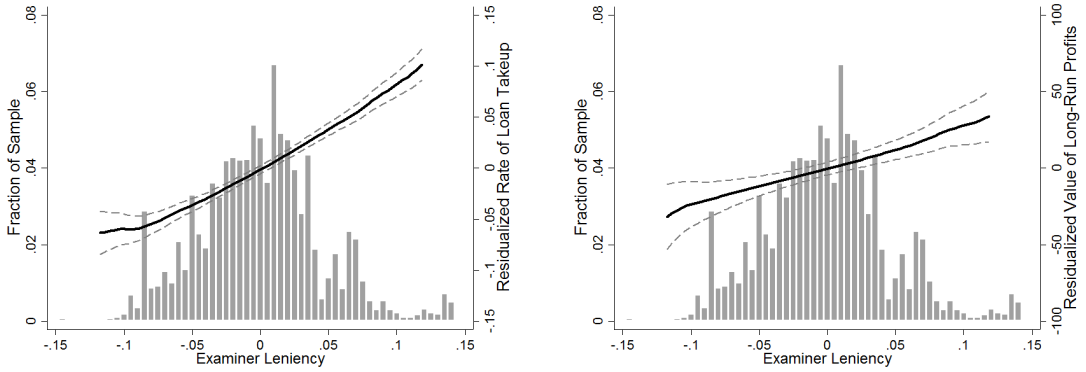


Note: These figures report hypothetical risk distributions for reference and target group applicants. Panel A illustrates the OLS estimator for an unbiased examiner who chooses the same threshold for take-up for both reference and target applicants. Panel B illustrates the OLS estimator for a biased examiner who chooses a higher threshold for loan take-up for target applicants than reference applicants. Panel C illustrates the IV estimator for two unbiased examiners. Panel D illustrates the IV estimator with two biased examiners. See the text for additional details.

Figure 2: First Stage and Reduced Form Results

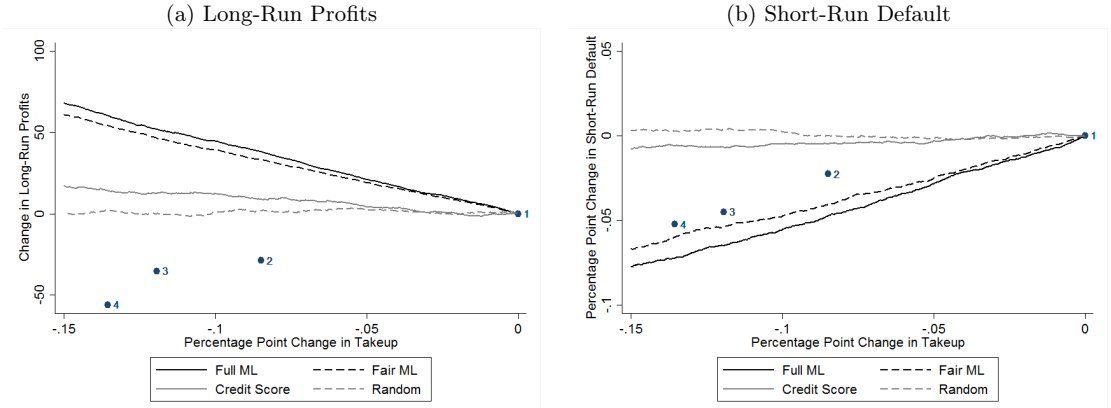
(a) First Stage for All Applicants

(b) Reduced Form for All Applicants



Note: These figures report the first stage and reduced form relationships between applicant outcomes and examiner leniency. The regressions are estimated on the sample described in the notes to Table 1. Examiner leniency is estimated using data from other applicants assigned to a loan examiner following the procedure described in Section III. In the first stage regression, the solid line represents a local linear regression of loan take-up on examiner leniency. In the reduced form regression, the solid line represents a local linear regression of long-run profits on examiner leniency. Loan take-up and long-run profits are residualized using store-by-month-by-nationality fixed effects. Standard errors are clustered at the examiner level.

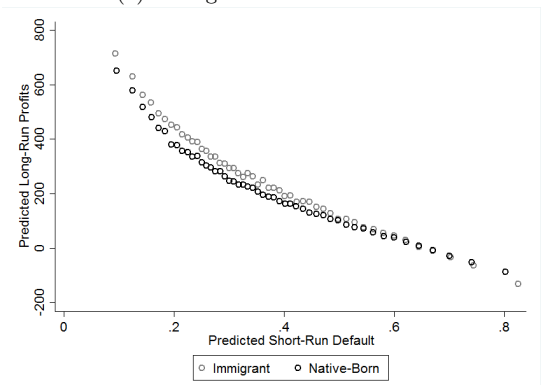
Figure 3: Comparing Additional Profits and Defaults by Ranking Method



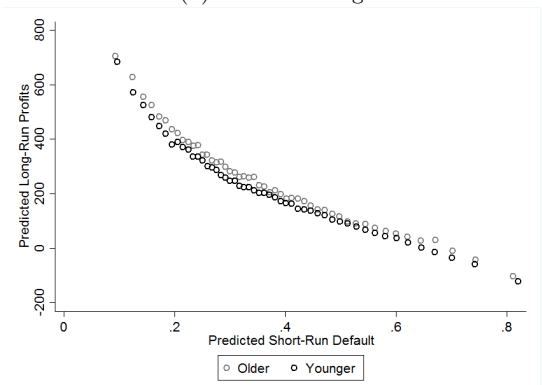
Note: These figures examine the performance of different data-based decision rules versus the actual decisions made by stricter loan examiners. The rightmost point in the graph represents the loan outcomes and loan take-up rate of the most lenient bin of examiners. The additional three points on the graph show loan outcomes and take-up rates for the actual decisions made by the second through fourth most lenient bins of examiners. Each line shows the loan outcome and take-up trade-off that comes from denying additional applicants within the most lenient bin of examiners' approval set using different data-based decision rules. The solid black line shows the trade-off when using the machine learning algorithm described in Section V trained using all available variables; the dashed black line for the same machine learning algorithm omitting nationality, gender, and age; the solid gray line for the credit score used to screen applicants; and the dashed gray line for randomly rejecting applicants. Panel A presents these results for long-run profits. Panel B presents these results for short-run default. See the text for additional details.

Figure 4: Joint Distributions of Machine Learning Predictions

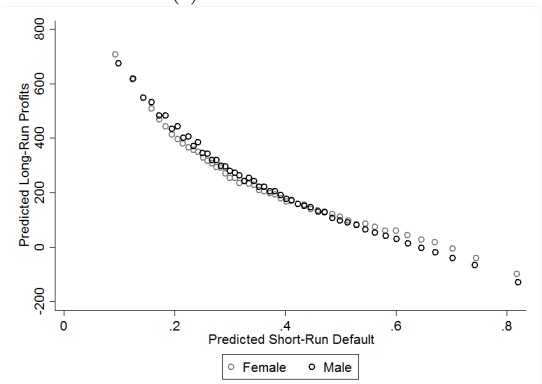
(a) Immigrant vs. Native-Born



(b) Old vs. Young



(c) Female vs. Male



Note: These figures report the relationship between predicted long-run profits and predicted short-run default separately by group. Predicted long-run profits and predicted short-run default are obtained using the machine learning algorithm described in Section V. See the text for additional details.

Appendix Table A1: IV Estimates of Loan Take-Up and Long-Run Profits by Store Location

	London Store Location				Non-London Store Location			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Loan Take-Up	144.795**	177.787	338.549**	−11.488	269.437**	168.175	363.796***	−60.858
	(70.037)	(127.980)	(167.624)	(216.329)	(107.236)	(161.661)	(120.919)	(209.874)
Take-Up x Immigrant Applicant	706.758**			743.615**	427.141*			496.955**
	(280.364)			(308.628)	(218.195)			(226.979)
Take-Up x Older Applicant		234.050		191.790		427.761*		462.484*
		(251.218)		(234.604)		(232.851)		(241.970)
Take-Up x Female Applicant			−100.037	109.058			66.838	118.565
			(192.784)	(198.845)			(175.020)	(184.199)
Dep. Variable Mean	141.948	141.948	141.948	141.948	219.153	219.153	219.153	219.153
Observations	24,711	24,711	24,711	24,711	20,796	20,796	20,796	20,796
Clusters	117	117	117	117	137	137	137	137

Note: This table reports IV estimates of bias in consumer lending separately by store location. See the notes to Table 3 for details on the sample and empirical specification. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A2: Correlation Between IV Weights and Examiner Observables

	Applicant Nationality		Applicant Age		Applicant Gender	
	Immigrant IV Weights x 100	Native-Born IV Weights x 100	Old IV Weights x 100	Young IV Weights x 100	Female IV Weights x 100	Male IV Weights x 100
	(1)	(2)	(3)	(4)	(5)	(6)
Examiner Discrimination (/1000)	−0.024	−0.005	0.045	0.029	−0.030	−0.039
	(0.043)	(0.039)	(0.036)	(0.036)	(0.096)	(0.100)
Examiner Case Load (/100)	0.007	0.001	0.002	0.003	0.006*	0.008*
	(0.005)	(0.003)	(0.005)	(0.004)	(0.004)	(0.005)
Examiner Average Leniency	−0.069	−0.136	0.098	0.014	−0.051	−0.042
	(0.206)	(0.189)	(0.189)	(0.190)	(0.185)	(0.193)
Experienced Examiner	0.014	0.015	0.013	0.016	0.013	0.016
	(0.023)	(0.021)	(0.021)	(0.022)	(0.021)	(0.022)
Male Examiner	−0.026	−0.024	−0.025	−0.025	−0.025	−0.025
	(0.022)	(0.021)	(0.021)	(0.021)	(0.021)	(0.022)
Dep. Variable Mean	0.394	0.394	0.394	0.394	0.394	0.394
Observations	254	254	254	254	254	254

Note: This table reports the relationship between instrumental variable weights assigned to a specific examiner cell and observables of the examiner cell by group. To ease readability, the weights are multiplied by 100. To compute the IV weights assigned to an examiner cell, we compute continuous IV weights following the procedure described in Cornelissen et al. (2016). We then discretize these continuous weights by assigning each examiner the weight associated with his or her average leniency. Finally, we divide the discretized examiner weights by the sum total to ensure the discretized weights sum to one. Examiner discrimination is calculated using the MTE procedure described in the text. Examiner case load is measured over the entire sample period. Examiner experience is an indicator for being employed by the Lender at the start of the sample period. See the data appendix for additional information on the construction of the variables. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A3: First Stage Results by Applicant Characteristics

	Applicant Nationality		Applicant Age		Applicant Gender	
	Immigrant	Native-Born	Old	Young	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)
Loan Take-Up	0.553***	0.793***	0.719***	0.717***	0.750***	0.637***
	(0.095)	(0.082)	(0.080)	(0.090)	(0.074)	(0.113)
Dep. Variable Mean	0.675	0.656	0.650	0.675	0.676	0.648
Observations	18,160	27,226	20,394	23,945	24,881	19,507
Clusters	254	254	254	254	254	254

Note: This table reports first stage results of examiner leniency on loan take-up estimated separately by group. The regressions are estimated on the sample as described in the notes to Table 1. Examiner leniency is estimated using data from other loan applicants assigned to an examiner following the procedure described in Section III. Observations in subgroups do not always add to the full sample size due to dropping of singleton observations. All specifications control for store-by-month-by-nationality fixed effects and the baseline characteristics listed in Panel A of Table 1. Standard errors clustered at the examiner level are reported in parentheses. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A4: Balance Tests by Subgroup

	Applicant Nationality		Applicant Age		Applicant Gender	
	Immigrant	Native-Born	Old	Young	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)
Age	0.00001	−0.00002	0.00001	−0.00002	−0.00001	−0.00002
	(0.00003)	(0.00002)	(0.00004)	(0.00008)	(0.00003)	(0.00003)
Female	0.00000	−0.00031	−0.00025	−0.00040		
	(0.00066)	(0.00084)	(0.00082)	(0.00082)		
One-Plus Years at Residence	0.00056	−0.00186	−0.00147	−0.00029	−0.00125	0.00014
	(0.00087)	(0.00155)	(0.00143)	(0.00114)	(0.00143)	(0.00097)
Married	0.00112	−0.00079	−0.00046	0.00026	−0.00014	0.00024
	(0.00113)	(0.00126)	(0.00101)	(0.00141)	(0.00109)	(0.00142)
Number of Dependents	0.00022	−0.00005	0.00025	−0.00003	0.00012	−0.00008
	(0.00035)	(0.00023)	(0.00028)	(0.00030)	(0.00022)	(0.00039)
Credit Score	−0.00000	0.00001**	0.00001	0.00001*	0.00001	0.00001**
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
Has Bank Account	0.00338*	0.00387**	0.00281	0.00453***	0.00311*	0.00469***
	(0.00193)	(0.00179)	(0.00184)	(0.00167)	(0.00174)	(0.00178)
Has Other Loan Payments	0.00039	−0.00060	−0.00021	−0.00024	0.00009	−0.00082
	(0.00120)	(0.00099)	(0.00109)	(0.00098)	(0.00100)	(0.00105)
Loan Amount Requested (£)	−0.00000	−0.00000	0.00000	−0.00000	0.00000	−0.00000**
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Loan for Emergency	−0.00323	−0.00006	−0.00163	−0.00120	−0.00095	−0.00139
	(0.00274)	(0.00411)	(0.00348)	(0.00360)	(0.00376)	(0.00320)
Loan for Large One-Time Expense	−0.00139	0.00317	0.00149	0.00035	0.00291	−0.00036
	(0.00277)	(0.00526)	(0.00441)	(0.00386)	(0.00458)	(0.00347)
Loan for Overdraft Avoidance	−0.00115	−0.00121	−0.00142	−0.00178	0.00018	−0.00304
	(0.00356)	(0.00427)	(0.00403)	(0.00389)	(0.00433)	(0.00346)
Loan for Shopping or Holiday	−0.00176	−0.00071	−0.00227	−0.00037	−0.00087	−0.00092
	(0.00254)	(0.00412)	(0.00343)	(0.00355)	(0.00362)	(0.00329)
Dep. Variable Mean	0.001	-0.000	0.000	0.000	0.000	0.001
Observations	18,160	27,226	20,394	23,945	24,881	19,507
p-value on Joint F-test	[0.656]	[0.225]	[0.834]	[0.324]	[0.857]	[0.033]
Clusters	254	254	254	254	254	254

Note: This table reports balance tests separately by group. The regressions are estimated on the sample as described in the notes to Table 1. The dependent variable is examiner leniency, which is estimated using data from other loan applicants assigned to an examiner following the procedure described in Section III. The p-value reported at the bottom of the columns is for a F-test of the joint significance of the variables listed in the rows. Observations in the subgroups do not always add to the full sample size due to dropping of singleton observations. All specifications control for store-by-month-by-nationality fixed effects. Standard errors clustered at the examiner level are reported in parentheses. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A5: IV Estimates of Loan Take-Up and Subgroup-Specific Discounted Long-Run Profits

	Additional Target Group Discount Rate					
	5 Percent	10 Percent	20 Percent	30 Percent	40 Percent	50 Percent
	(1)	(2)	(3)	(4)	(5)	(6)
Loan Take-Up	−0.703	8.153	24.422	40.292	54.789	67.078
	(119.673)	(117.303)	(114.004)	(112.071)	(111.241)	(111.117)
Take-Up x Immigrant Applicant	436.515***	418.979***	386.738**	354.412**	327.842**	305.414**
	(159.560)	(155.900)	(150.215)	(146.136)	(143.567)	(141.943)
Take-Up x Older Applicant	256.638**	247.373*	232.513*	218.988*	207.693*	198.102*
	(130.344)	(127.222)	(122.662)	(119.206)	(116.786)	(115.080)
Take-Up x Female Applicant	61.656	52.913	36.074	19.513	4.166	−8.848
	(111.313)	(109.658)	(107.365)	(106.317)	(106.315)	(106.870)
Dep. Variable Mean	168.927	166.699	162.772	159.396	156.607	154.203
Observations	45,507	45,507	45,507	45,507	45,507	45,507
Clusters	254	254	254	254	254	254

Note: This table reports IV estimates of bias in consumer lending using discounted long-run profits as the outcome. Reference group long-run profits are discounted using a discount rate of 10 percent. Target group long-run profits are additionally discounted by the amount listed in the column header. See the notes to Table 3 for additional details. Standard errors clustered at the examiner level are reported in parentheses. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A6: IV Estimates of Loan Approval and Loan Take-Up

	2SLS Estimates				
	(1)	(2)	(3)	(4)	(5)
Loan Approved	1.254***	1.228***	1.357***	1.302***	1.376***
	(0.077)	(0.071)	(0.116)	(0.106)	(0.141)
Approved x Immigrant Applicant		0.116			0.065
		(0.164)			(0.186)
Approved x Older Applicant			−0.101		−0.095
			(0.131)		(0.133)
Approved x Female Applicant				−0.162	−0.145
				(0.127)	(0.141)
Dep. Variable Mean	0.661	0.661	0.661	0.661	0.661
Observations	45,687	45,687	45,687	45,687	45,687
Clusters	254	254	254	254	254

Note: This table reports two-stage least squares estimates of loan approval on loan take-up. Loan approval is instrumented for using the residuals from a regression of loan approval on store x month x applicant nationality fixed effects. See the notes to Table 3 for additional details on the sample and specification. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A7: IV Estimates of Loan Approval and Long-Run Profits

	2SLS Estimates				
	(1)	(2)	(3)	(4)	(5)
Loan Approved	351.621***	212.873**	366.809**	170.857	−7.521
	(90.247)	(90.108)	(145.200)	(138.874)	(194.790)
Approved x Immigrant Applicant		625.982**			619.523**
		(252.961)			(280.825)
Approved x Older Applicant			378.979		358.552
			(230.964)		(226.412)
Approved x Female Applicant				−24.047	80.442
				(177.016)	(177.322)
Dep. Variable Mean	177.229	177.229	177.229	177.229	177.229
Observations	45,507	45,507	45,507	45,507	45,507
Clusters	254	254	254	254	254

Note: This table reports IV estimates of bias in consumer lending using loan approvals instead of loan take-up. See the notes to Table 3 for details on the sample and empirical specification. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A8: IV Estimates of Loan Take-Up and Discounted Long-Run Profits

	Discount Rates					
	5 Percent	10 Percent	20 Percent	30 Percent	40 Percent	50 Percent
	(1)	(2)	(3)	(4)	(5)	(6)
Loan Take-Up	−6.141	−10.504	−19.631	−27.925	−35.050	−41.235
	(132.336)	(122.257)	(105.128)	(91.091)	(79.583)	(70.162)
Take-Up x Immigrant Applicant	493.987***	454.416***	388.566***	333.469***	285.608***	245.721***
	(176.894)	(163.422)	(140.582)	(121.753)	(106.265)	(93.825)
Take-Up x Older Applicant	291.016**	266.600**	224.364**	188.100*	157.779*	132.628*
	(145.503)	(133.750)	(113.656)	(96.845)	(82.992)	(71.723)
Take-Up x Female Applicant	72.406	71.130	69.515	68.219	66.016	63.714
	(122.460)	(113.124)	(97.369)	(84.576)	(74.180)	(65.724)
Dep. Variable Mean	186.609	171.194	144.073	120.846	100.867	83.624
Observations	45,507	45,507	45,507	45,507	45,507	45,507
Clusters	254	254	254	254	254	254

Note: This table reports IV estimates of bias in consumer lending using discounted long-run profits as the outcome. Long-run profits are discounted using the discount rate listed in the column header. See the notes to Table 3 for additional details. Standard errors clustered at the examiner level are reported in parentheses. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A9: Re-Weighted IV Estimates of Loan Take-Up and Long-Run Profits

	2SLS Estimates				
	(1)	(2)	(3)	(4)	(5)
Loan Take-Up	331.193***	192.756***	381.996***	169.339*	−28.076
	(60.338)	(64.770)	(97.945)	(98.378)	(143.947)
Take-Up x Immigrant Applicant		366.358*			553.833***
		(192.899)			(191.709)
Take-Up x Older Applicant			313.718*		398.575**
			(170.915)		(170.992)
Take-Up x Female Applicant				101.937	135.741
				(133.555)	(130.320)
Dep. Variable Mean	177.229	177.229	177.229	177.229	177.229
Observations	45,507	45,507	45,507	45,507	45,507
Clusters	254	254	254	254	254

Note: This table reports robustness results testing for bias when the estimates are re-weighted so that the observable characteristics of target group applicants match the distribution of observable characteristics of reference group applicants. See Appendix Table A12 for the complier characteristics used to construct the weights and Arnold, Dobbie and Yang (2018) for additional details. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A10: Marginal Treatment Effect Estimates of Loan Take-Up and Long-Run Profits

	MTE Estimates			
	(1)	(2)	(3)	(4)
Loan Take-Up	406.713**	270.838*	237.288*	382.745**
	(171.668)	(162.620)	(138.041)	(157.429)
Loan Take-Up x Immigrant Applicant		357.697		
		(301.596)		
Loan Take-Up x Older Applicant			363.226*	
			(207.811)	
Loan Take-Up x Female Applicant				45.268
				(205.548)
Dep. Variable Mean	177.229	177.229	177.229	177.229
Observations	45,507	45,507	45,507	45,507

Note: This table reports robustness results testing for bias using an MTE estimator that puts equal weight on each examiner. We estimate these MTE results using a two-step procedure. In the first step, we estimate the entire distribution of MTEs using the derivative of residualized profits with respect to variation in the propensity score provided by our instrument. To do this, we regress the residualized profit variable on the residualized examiner leniency measure to calculate the group-specific propensity score. We then compute the numerical derivative of a local quadratic estimator to estimate group-specific MTEs that are presented in Appendix Figure A6. In the second step, we use these group-specific MTEs to calculate the level of bias for each examiner and the average level of bias across all loan examiners. We calculate standard errors by bootstrapping this two-step procedure at the examiner level. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A11: IV Estimates of Loan Take-Up and Long-Run Profits Using Group-Specific Leniencies

	2SLS Estimates				
	(1)	(2)	(3)	(4)	(5)
Loan Take-Up	331.193***	168.221**	447.493***	77.285	−72.608
	(60.338)	(69.172)	(158.974)	(159.295)	(468.296)
Take-Up x Immigrant Applicant		695.455			634.798
		(440.707)			(893.513)
Take-Up x Older Applicant			473.528*		384.599
			(260.099)		(385.881)
Take-Up x Female Applicant				−86.557	130.602
				(182.530)	(445.535)
Dep. Variable Mean	177.229	177.229	177.229	177.229	177.229
Observations	45,507	45,507	45,507	45,507	45,507
Clusters	254	254	254	254	254

Note: This table reports robustness results testing for bias using an examiner leniency measure that is estimated separately by group. See the notes to Table 3 for details on the sample and empirical specification. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A12: Representativeness Statistics

	$\mathbb{E}$(x|Immigrant)/ $\mathbb{E}$(x|Native-Born)	$\mathbb{E}$(x|Old)/ $\mathbb{E}$(x|Young)	$\mathbb{E}$(x|Female)/ $\mathbb{E}$(x|Male)
	(1)	(2)	(3)
Immigrant	–	1.100	0.607
Age	1.007	–	0.986
Female	0.677	0.964	–
One Plus Years at Residence	0.803	1.094	1.056
Married	1.403	1.723	0.852
Number of Dependents	0.741	1.218	2.027
Credit Score	1.051	0.991	0.967
Has Bank Account	1.028	0.990	0.986
Has Other Loan Payments	0.685	1.270	1.375
Loan Amount Requested (£)	1.300	1.163	0.860
Loan for Emergency	1.031	1.003	0.887
Loan for Large One-Time Expense	2.118	0.911	0.653
Loan for Overdraft Avoidance	0.716	1.189	0.974
Loan for Shopping or Holiday	0.886	0.908	1.308
Observations	45,507	45,507	45,507

Note: This table reports the mean of the listed variable conditional on target group status, divided by the mean of the listed variable, conditional on reference group status. The means are estimated for the sample as described in the notes to Table 1.

Appendix Table A13: OLS Estimates of Loan Take-Up and Long-Run Profits

	OLS Estimates				
	(1)	(2)	(3)	(4)	(5)
Loan Take-Up	209.393***	176.350***	154.663***	169.103***	60.083***
	(8.270)	(8.256)	(8.796)	(8.312)	(8.536)
Take-Up x Immigrant Applicant		101.921***			126.945***
		(11.931)			(12.030)
Take-Up x Older Applicant			88.257***		89.059***
			(7.478)		(7.405)
Take-Up x Female Applicant				99.889***	123.195***
				(8.405)	(8.287)
Dep. Variable Mean	177.229	177.229	177.229	177.229	177.229
Observations	45,507	45,507	45,507	45,507	45,507
Clusters	254	254	254	254	254

Note: This table reports OLS estimates of bias in consumer lending decisions based on long-run profits. The regressions are estimated on the sample as described in the notes to Table 1. All specifications control for store-by-month-by-nationality fixed effects and the baseline characteristics listed in Panel A of Table 1. Standard errors clustered at the examiner level are reported in parentheses. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A14: IV Estimates of Loan Take-Up and 30-Day Default

	2SLS Estimates				
	(1)	(2)	(3)	(4)	(5)
Loan Take-Up	0.271***	0.263***	0.279***	0.239***	0.235***
	(0.048)	(0.052)	(0.072)	(0.059)	(0.091)
Take-Up x Immigrant Applicant		0.033			0.030
		(0.106)			(0.115)
Take-Up x Older Applicant			0.070		0.069
			(0.078)		(0.078)
Take-Up x Female Applicant				−0.013	−0.006
				(0.090)	(0.095)
Dep. Variable Mean	0.124	0.124	0.124	0.124	0.124
Observations	45,507	45,507	45,507	45,507	45,507
Clusters	254	254	254	254	254

Note: This table reports IV estimates of bias in consumer lending using an indicator for no payments in the first 30 days as the dependent variable. The dependent variable is one if the applicant makes no payments in the first 30 days after the loan application and is zero otherwise. See the notes to Table 3 for details on the sample and empirical specification. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A15: Correlates of the Machine Learning Prediction for Long-Run Profits

	OLS Estimates			
	(1)	(2)	(3)	(4)
Credit Score (/1000)	431.651***	457.684***	364.420***	184.493***
	(14.303)	(12.278)	(9.751)	(5.301)
Immigrant	36.443***	94.785***	45.903***	−14.904***
	(5.643)	(5.468)	(4.949)	(5.027)
Age		7.132***	7.368***	6.797***
		(0.146)	(0.120)	(0.091)
Female	−4.039	120.344***	112.911***	116.683***
	(3.115)	(3.174)	(2.727)	(2.392)
Married		18.431***	6.945***	10.772***
		(2.579)	(1.902)	(1.016)
Disposable Income (£/1000)			−30.161***	12.915***
			(9.595)	(2.804)
Months in UK (/1000)			−128.951***	−77.004***
			(8.220)	(4.207)
Number of Dependents			13.291***	20.891***
			(1.051)	(0.708)
Loan for Emergency			−9.092***	−7.283***
			(2.603)	(1.308)
Customer was Referred			113.429***	89.342***
			(3.079)	(2.106)
Loan Amount Requested (£/1000)			144.377***	127.876***
			(5.497)	(4.443)
Total Income (£/1000)			62.365***	−1.873
			(6.431)	(1.302)
Salary (£/1000)				14.382***
				(1.222)
Other Loan Payments (£/1000)				235.630***
				(9.499)
Debt to Income Ratio (/1000)				1,428.603***
				(265.997)
Total Credit Outstanding (£/10^6)				37.244
				(71.338)
Number Open Lines of Credit				−0.333***
				(0.124)
Months at Current Residence (/1000)				87.657***
				(6.010)
Credit Arrears from Other Lenders (£/10^6)				63.501
				(51.705)
Dep. Variable Mean	229.189	229.189	229.189	229.189
Observations	45,507	45,507	45,507	45,507
Adjusted R-Squared	0.216	0.458	0.659	0.858

Note: This table reports selected coefficients from an OLS regression of predicted total profits on loan and applicant characteristics. The regressions are estimated on the sample as described in the notes to Table 1. All regressions include store and time fixed effects. Predicted total profits are obtained using the machine learning algorithm described in Section V. Standard errors clustered at the examiner level are reported in parentheses. See the text for additional details. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A16: Predicted Long-Run Profits and Applicant Characteristics

	OLS Estimates			
	(1)	(2)	(3)	(4)
Immigrant Applicant	34.309***			37.313***
	(4.842)			(4.604)
Older Applicant		29.507***		31.428***
		(2.188)		(2.200)
Female Applicant			−4.344	9.519***
			(2.703)	(2.054)
Predicted Default (x100)	−19.000***	−19.587***	−20.169***	−18.226***
	(0.447)	(0.482)	(0.484)	(0.465)
Predicted Default Squared (x100)	0.109***	0.118***	0.121***	0.104***
	(0.447)	(0.004)	(0.004)	(0.004)
Dep. Variable Mean	229.189	229.189	229.189	229.189
Observations	45,507	45,507	45,507	45,507
Clusters	254	254	254	254

Note: This table reports the correlation between predicted long-run profits and applicant characteristics. Predicted long-run profits and predicted short-run default are obtained using the machine learning algorithm described in Section V. Standard errors clustered at the examiner level are reported in parentheses. See the text for additional details. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A17: Total Monthly Income and Applicant Characteristics

	OLS Estimates			
	(1)	(2)	(3)	(4)
Immigrant Applicant	96.024***			74.563***
	(7.090)			(7.498)
Older Applicant		148.265***		142.438***
		(5.988)		(6.084)
Female Applicant			−86.642***	−69.182***
			(6.641)	(7.186)
Credit Score × (-1)	−2.102**	−0.863	−0.896	−1.696**
	(0.874)	(0.854)	(0.880)	(0.842)
Credit Score Squared	−0.003***	−0.001	−0.001*	−0.002***
	(0.874)	(0.001)	(0.001)	(0.001)
Dep. Variable Mean	1,110.815	1,110.815	1,110.815	1,110.815
Observations	37,332	37,332	37,332	37,332
Clusters	254	254	254	254

Note: This table reports the correlation between applicant total monthly income and applicant characteristics. The sample excludes applicants with a missing credit score or missing income. Standard errors clustered at the examiner level are reported in parentheses. See the text for additional details. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Table A18: IV Estimates of Loan Take-Up and Long-Run Profits by Examiner Gender

	Female Examiner				Male Examiner			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Loan Take-Up	246.375***	53.643	388.215***	−215.785	28.287	−185.160	−304.144	−195.856
	(95.372)	(409.203)	(120.359)	(428.846)	(237.768)	(231.626)	(461.093)	(440.809)
Take-Up x Immigrant Applicant	902.355**			894.045**	−1,038.901			−1,091.074
	(439.780)			(422.518)	(1,693.207)			(1,794.075)
Take-Up x Older Applicant		690.198		718.549		152.022		155.291
		(767.067)		(797.304)		(118.520)		(128.874)
Take-Up x Female Applicant			60.308	147.461			223.382	206.955
			(144.901)	(166.348)			(397.751)	(421.641)
Dep. Variable Mean	177.059	177.059	177.059	177.059	177.502	177.502	177.502	177.502
Observations	28,085	28,085	28,085	28,085	16,450	16,450	16,450	16,450
Clusters	157	157	157	157	95	95	95	95

Note: This table reports IV estimates of bias in consumer lending separately by examiner gender. See the notes to Table 3 for details on the sample and empirical specification. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

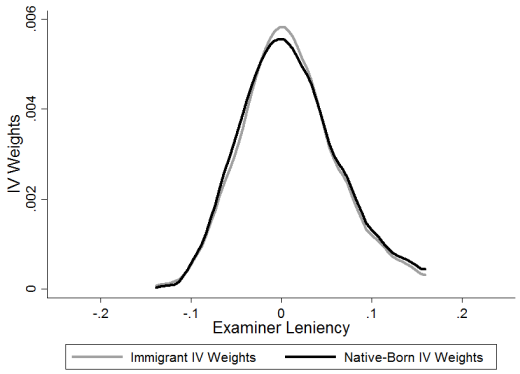
Appendix Table A19: IV Estimates of Loan Take-Up and Long-Run Profits by Examiner Experience

	Inexperienced Examiner				Experienced Examiner			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Loan Take-Up	105.673	39.144	216.032**	−97.586	362.133*	664.084***	744.224***	−78.405
	(64.253)	(110.320)	(101.188)	(145.665)	(191.969)	(249.264)	(231.154)	(437.368)
Take-Up x Immigrant Applicant	557.631***			535.883***	1,426.386*			1,629.382*
	(181.942)			(191.177)	(769.458)			(872.736)
Take-Up x Older Applicant		381.287**		386.453**		194.684		239.952
		(183.749)		(187.942)		(368.000)		(362.622)
Take-Up x Female Applicant			−11.684	58.447			31.324	417.320
			(133.661)	(141.250)			(310.282)	(409.073)
Dep. Variable Mean	145.982	145.982	145.982	145.982	232.355	232.355	232.355	232.355
Observations	28,798	28,798	28,798	28,798	16,208	16,208	16,208	16,208
Clusters	165	165	165	165	89	89	89	89

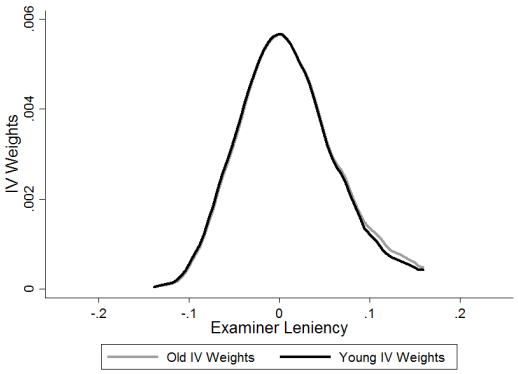
Note: This table reports IV estimates of bias in consumer lending separately by examiner experience. Experienced examiners are those employed by the Lender at the start of the sample period. See the notes to Table 3 for details on the sample and empirical specification. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Appendix Figure A1: Distribution of IV weights

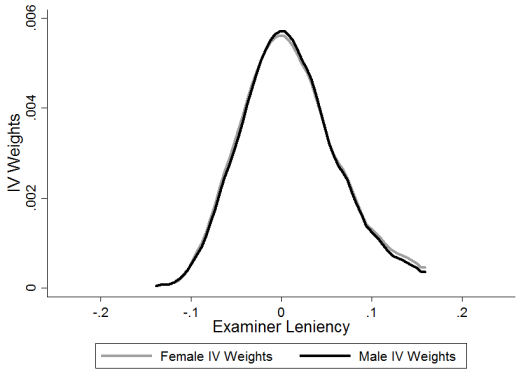
(a) Immigrant vs. Native-Born



(b) Old vs. Young



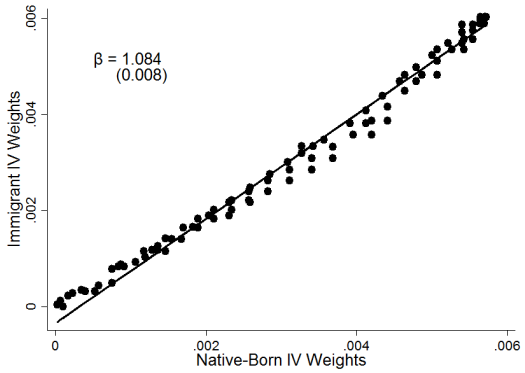
(c) Female vs. Male



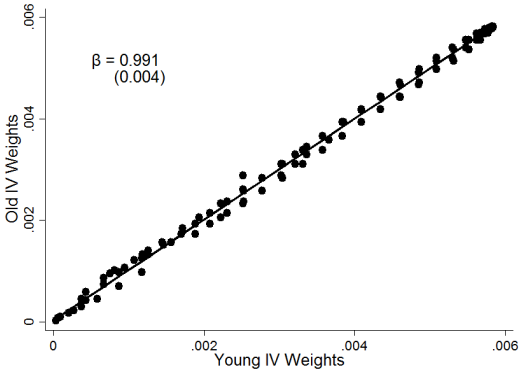
Note: These figures report the distribution of IV weights separately by group. To compute the IV weights assigned to an examiner cell, we compute continuous IV weights following the procedure described in Cornelissen et al. (2016). We then discretize these continuous weights by assigning each examiner the weight associated with his or her average leniency. Finally, we divide the discretized examiner weights by the sum total to ensure the discretized weights sum to one. See the text for additional details.

Appendix Figure A2: Correlation Between Subgroup-Specific IV Weights

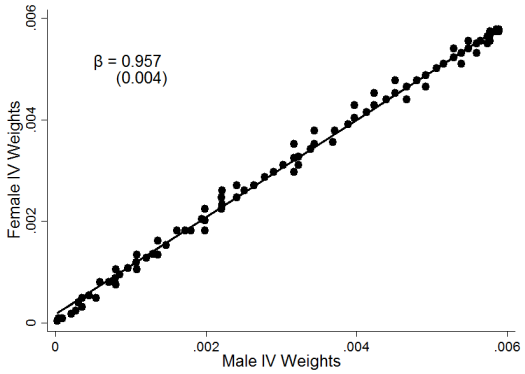
(a) Immigrant vs. Native-Born



(b) Old vs. Young



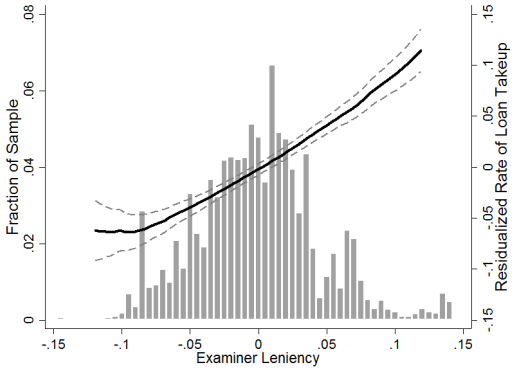
(c) Female vs. Male



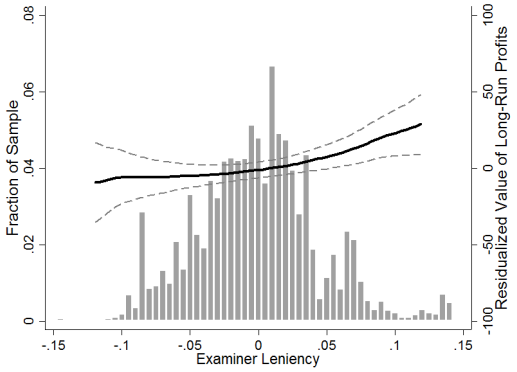
Note: These figures report the correlation between group-specific IV weights. To compute the IV weights assigned to an examiner cell, we compute continuous IV weights following the procedure described in Cornelissen et al. (2016). We then discretize these continuous weights by assigning each examiner the weight associated with his or her average leniency. Finally, we divide the discretized examiner weights by the sum total to ensure the discretized weights sum to one. The best fit line is estimated using OLS. See the text for additional details.

Appendix Figure A3: First Stage and Reduced Form Results by Group

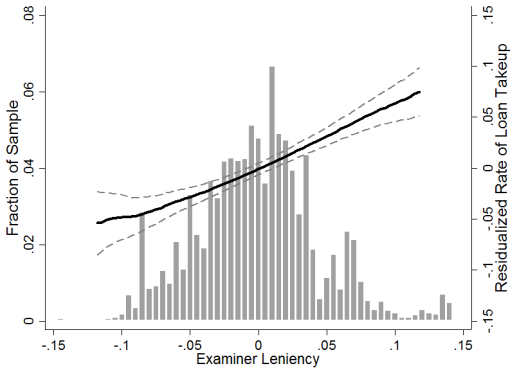
(a) First-Stage for Native-Born Applicants



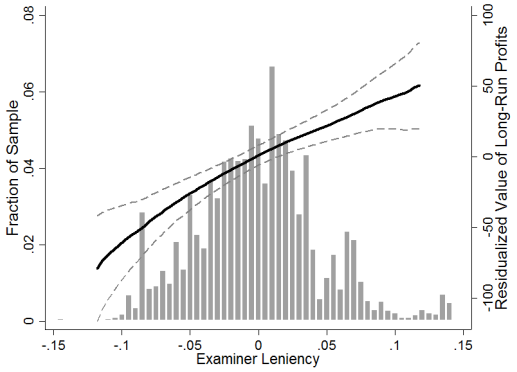
(b) Reduced Form for Native-Born Applicants



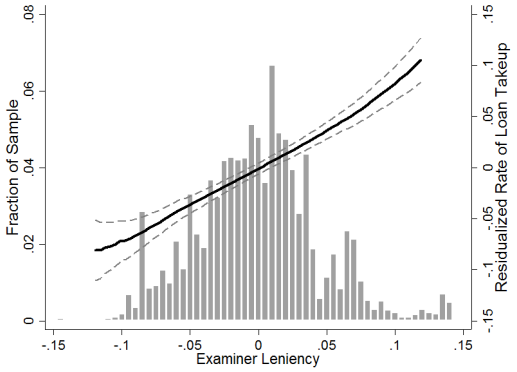
(c) First-Stage for Immigrant Applicants



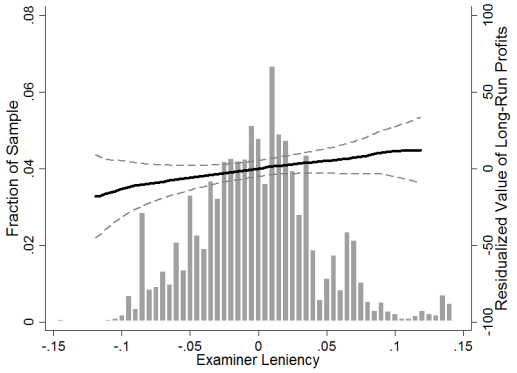
(d) Reduced Form for Immigrant Applicants



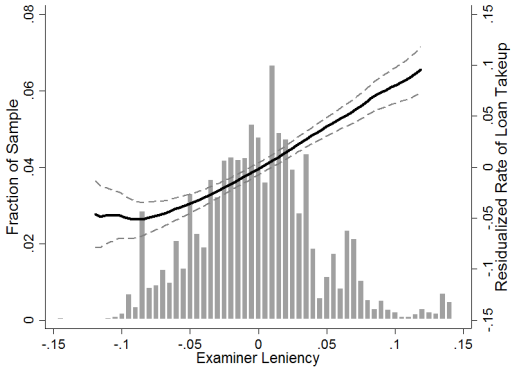
(e) First-Stage for Young Applicants



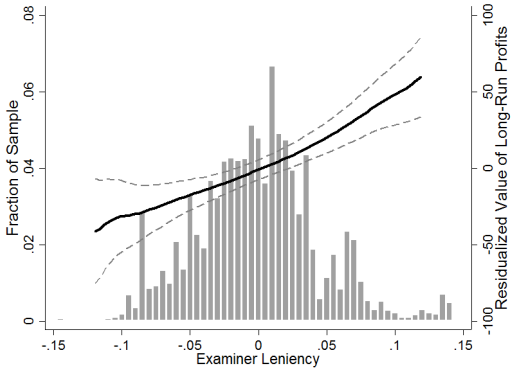
(f) Reduced Form for Young Applicants



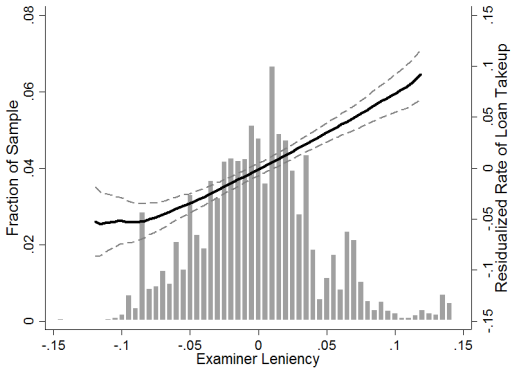
(g) First-Stage for Old Applicants



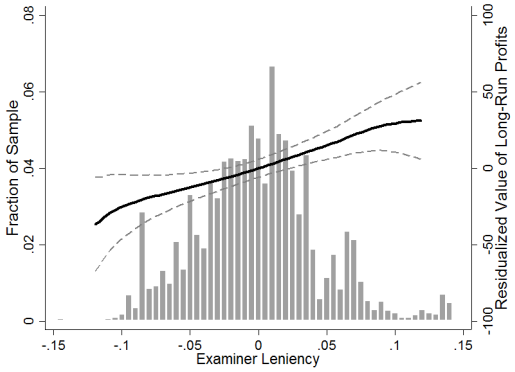
(h) Reduced Form for Old Applicants



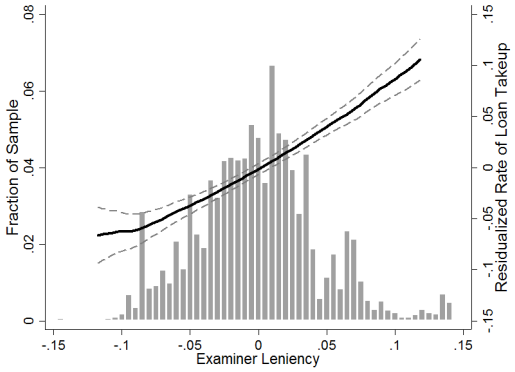
(i) First-Stage for Male Applicants



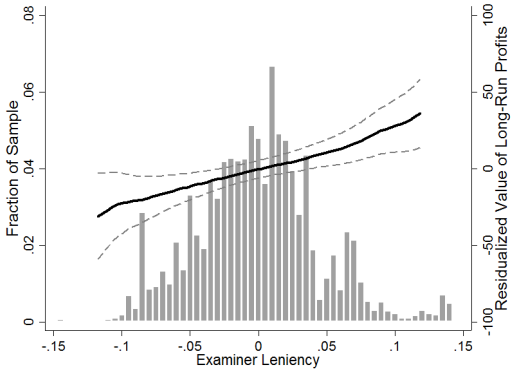
(j) Reduced Form for Male Applicants



(k) First-Stage for Female Applicants

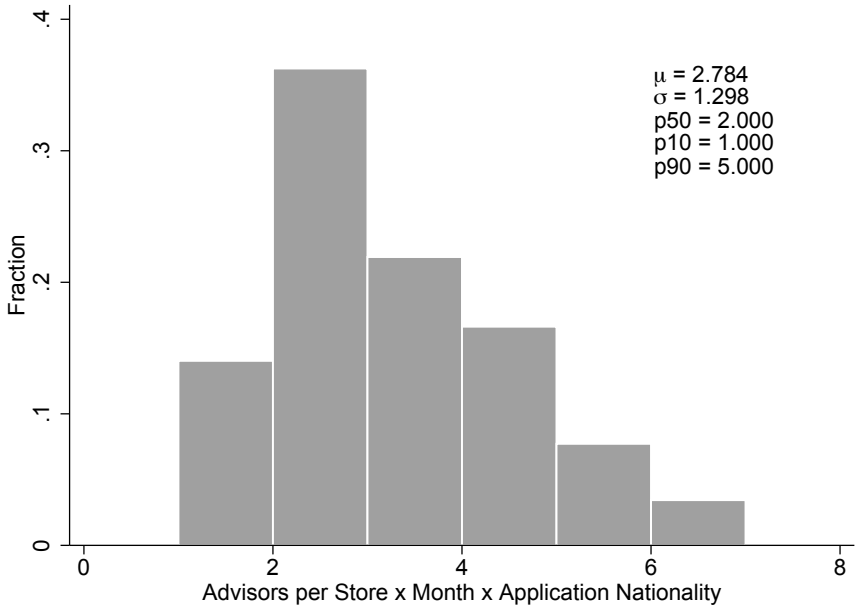


(l) Reduced Form for Female Applicants



Note: These figures report the first stage and reduced form relationships between applicant outcomes and examiner leniency by group. The regressions are estimated on the sample described in the notes to Table 1. Examiner leniency is estimated using data from other applicants assigned to a loan examiner following the procedure described in Section III. In the first stage regressions, the solid line represents a local linear regression of loan take-up on examiner leniency. In the reduced form regressions, the solid line represents a local linear regression of long-run profits on examiner leniency. Loan take-up and long-run profits are residualized using store-by-month-by-nationality fixed effects. Standard errors are clustered at the examiner level.

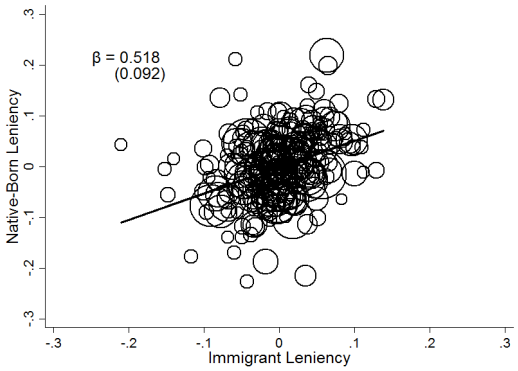
Appendix Figure A4: Number of Loan Examiners in Each Store x Month x Nationality Cell



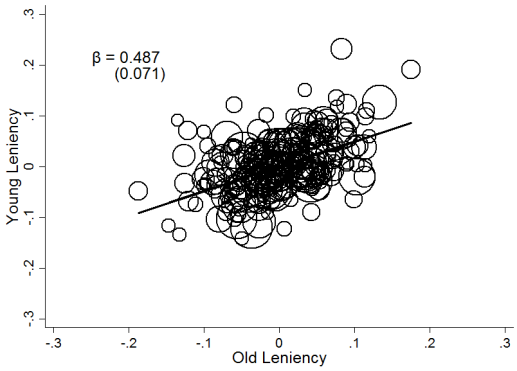
Note: This figure reports the number of loan examiners in each store x month x applicant nationality cell. The number of loan examiners in each cell are calculated using the sample described in the notes to Table 1.

Appendix Figure A5: Correlation Between Subgroup-Specific Examiner Leniency Measures

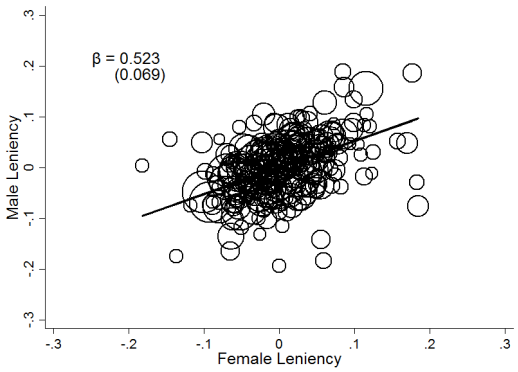
(a) Immigrant vs. Native-Born



(b) Old vs. Young



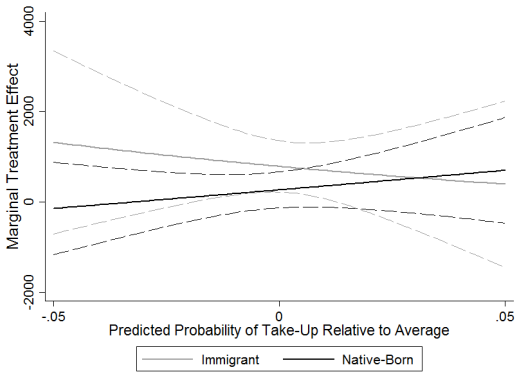
(c) Female vs. Male



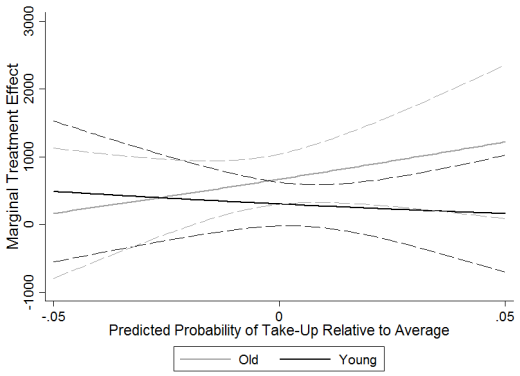
Note: These figures report the correlation between group-specific examiner leniency measures. Examiner leniency by group is estimated using data from other applicants from the same group assigned to a loan examiner following the procedure described in Section III. The best line fit is estimated using OLS.

Appendix Figure A6: Marginal Treatment Effects

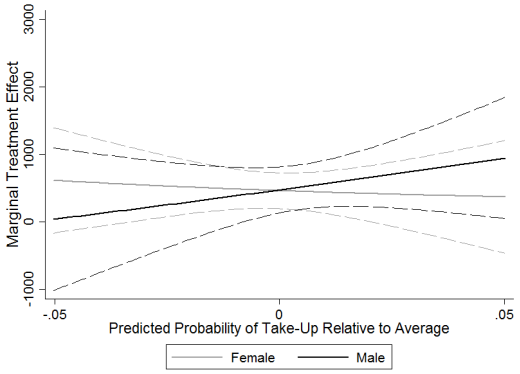
(a) Immigrant vs. Native-Born



(b) Old vs. Young

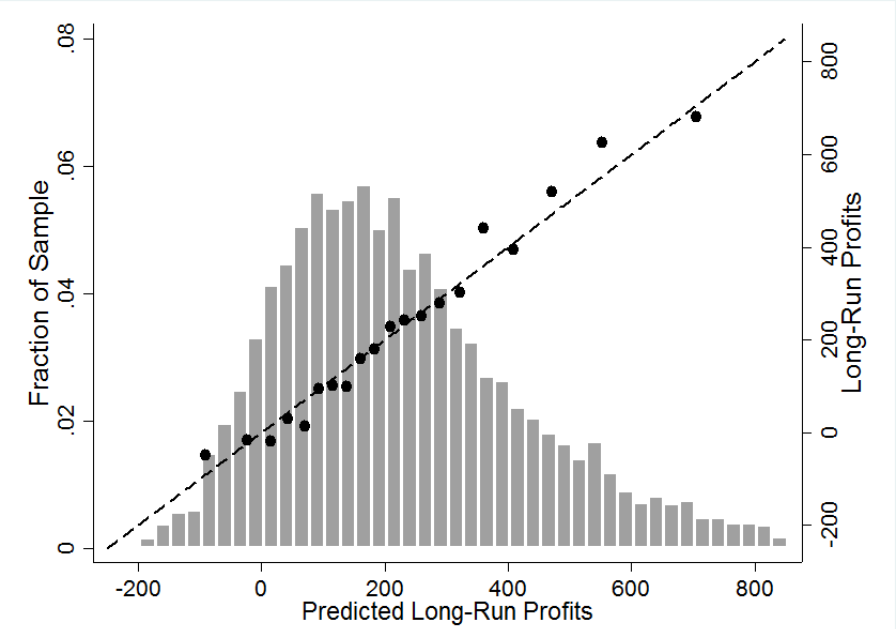


(c) Female vs. Male



Note: These figures report estimated marginal treatment effects of loan take-up on total profits separately by group. To compute the marginal treatment effects, we first estimate the predicted probability of take-up using only variation in examiner leniency. We then estimate the relationship between predicted probability of take-up and total profits using a local quadratic estimator. We calculate the numerical derivative to estimate the marginal treatment effect at each point in the distribution. The solid lines represent the estimated marginal treatment effects separately for each subgroup, while the dashed lines represent 90 percent confidence intervals, with standard errors that are computed using 500 bootstrap replications clustered at the examiner level. See the text for additional details.

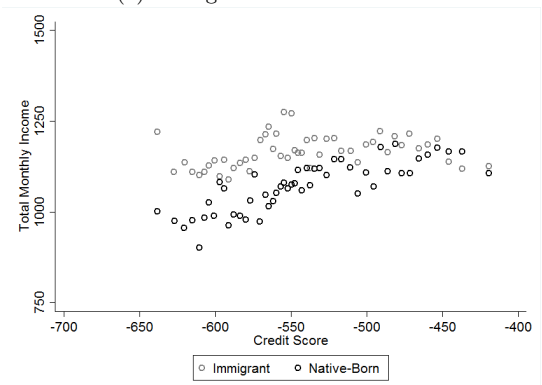
Appendix Figure A7: Relationship Between Observed and Predicted Long-Run Profits



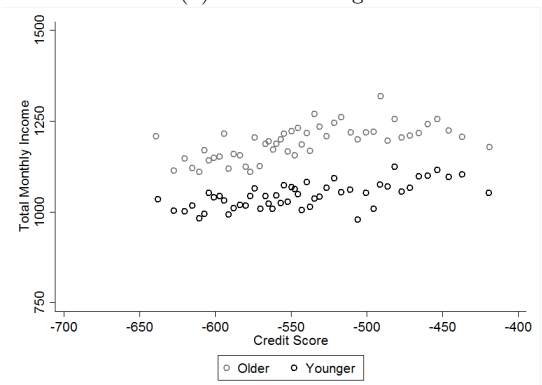
Note: This figure reports the relationship between observed and predicted long-run profits in the test sample. Predicted long-run profits are calculated using the machine learning algorithm described in Section V. The straight dashed line is the 45 degree line. See the text for additional details.

Appendix Figure A8: Joint Distributions of Income and Credit Score

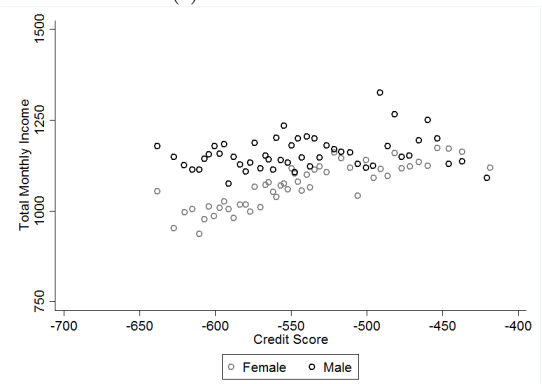
(a) Immigrant vs. Native-Born



(b) Old vs. Young

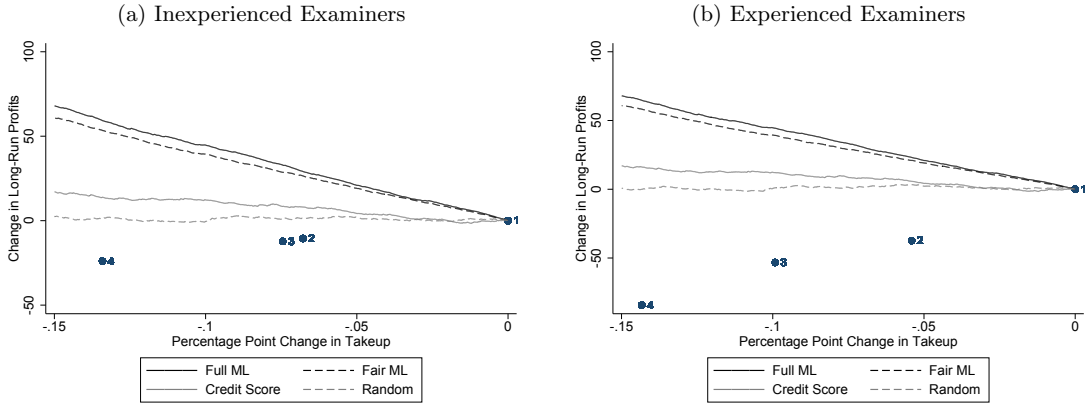


(c) Female vs. Male



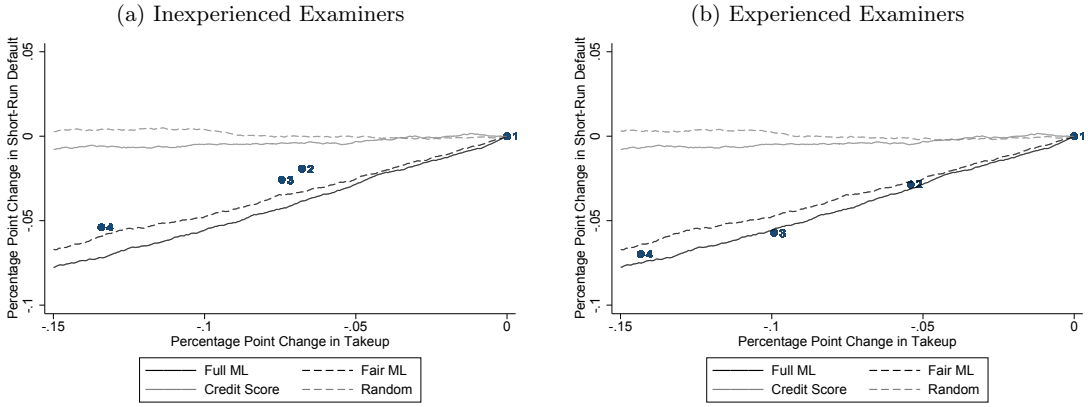
Note: These figures report the relationship between total monthly income and credit score separately by group. The sample excludes applicants with a missing credit score or missing income. See the text for additional details.

Appendix Figure A9: Comparing Additional Profits by Ranking Method and Examiner Tenure



Note: These figures examine the performance of different data-based decision rules versus the actual decisions made by stricter loan examiners. The rightmost point in the graph represents the loan outcomes and loan take-up rate of the most lenient bin of examiners. The additional three points on the graph show loan outcomes and take-up rates for the actual decisions made by the second through fourth most lenient bins of examiners. Each line shows the loan outcome and take-up trade-off that comes from denying additional applicants within the most lenient bin of examinersâĂŹ approval set using different data-based decision rules. The solid black line shows the trade-off when using the machine learning algorithm described in Section V trained using all available variables; the dashed black line for the same machine learning algorithm omitting nationality, gender, and age; the solid gray line for the credit score used to screen applicants; and the dashed gray line for randomly rejecting applicants. Panel A presents these results for inexperienced examiners only, and Panel B presents these results for experienced examiners only.

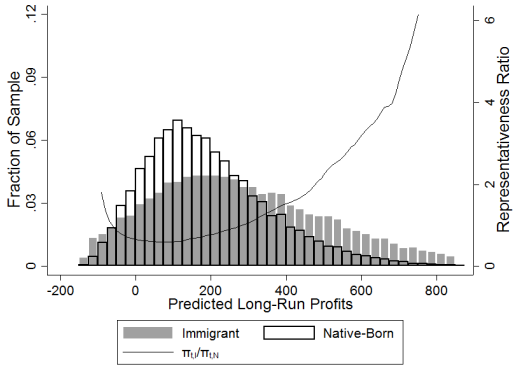
Appendix Figure A10: Comparing Additional Defaults by Ranking Method and Examiner Tenure



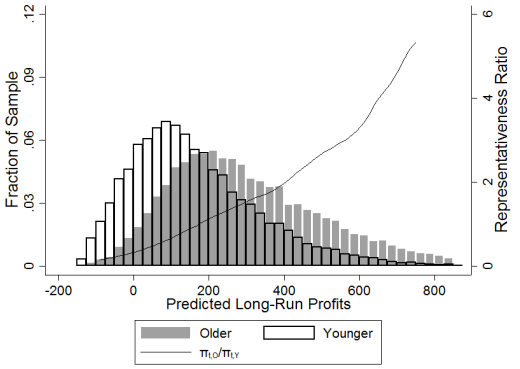
Note: These figures examine the performance of different data-based decision rules versus the actual decisions made by stricter loan examiners. The rightmost point in the graph represents the loan outcomes and loan take-up rate of the most lenient bin of examiners. The additional three points on the graph show loan outcomes and take-up rates for the actual decisions made by the second through fourth most lenient bins of examiners. Each line shows the loan outcome and take-up trade-off that comes from denying additional applicants within the most lenient bin of examinersâĂŹ approval set using different data-based decision rules. The solid black line shows the trade-off when using the machine learning algorithm described in Section V trained using all available variables; the dashed black line for the same machine learning algorithm omitting nationality, gender, and age; the solid gray line for the credit score used to screen applicants; and the dashed gray line for randomly rejecting applicants. Panel A presents these results for inexperienced examiners only, and Panel B presents these results for experienced examiners only.

Appendix Figure A11: Predicted Long-Run Profit Distributions by Group

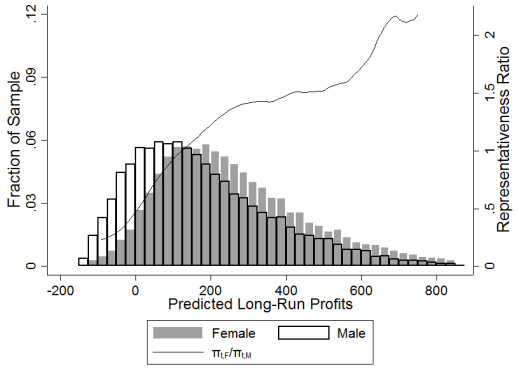
(a) Immigrant vs. Native-Born



(b) Old vs. Young



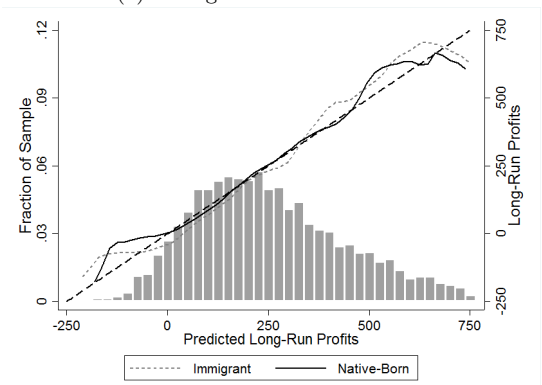
(c) Female vs. Male



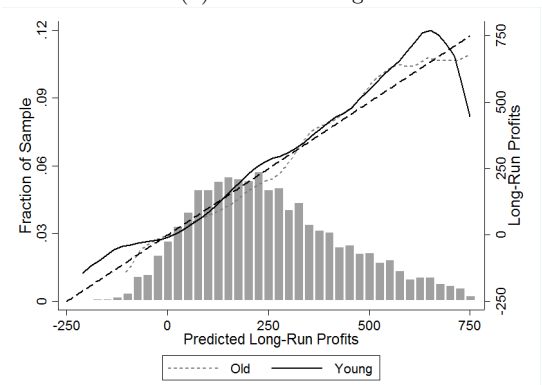
Note: These figures report the distribution of predicted long-run profits separately by group. Predicted long-run profits are calculated using the machine learning algorithm described in Section V. The solid lines report the representativeness ratio for target group versus reference group applicants, or the predicted long-run profits for the target group divided by the predicted long-run profits for the reference group. See the text for additional details.

Appendix Figure A12: Testing for Bias in Machine Learning Predictions

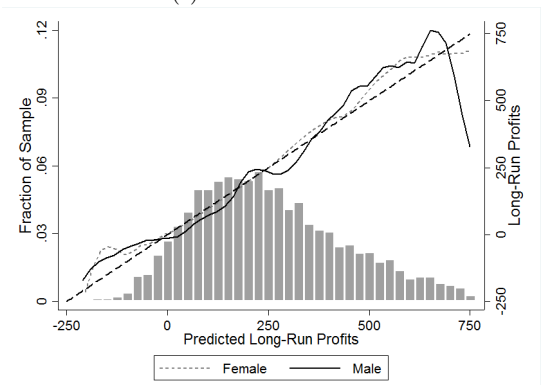
(a) Immigrant vs. Native-Born



(b) Old vs. Young



(c) Female vs. Male



Note: These figures report the relationship between observed and predicted long-run profits by group. Predicted long-run profits are calculated using the machine learning algorithm described in Section V. The straight dashed line is the 45 degree line. See the text for additional details.

Appendix B: Taste-Based and Inaccurate Stereotypes Models

This appendix presents models of taste-based and stereotypes-based bias to complement the incentive-based model of bias presented in the main text.

A. Taste-Based Bias

Let i denote loan applicants and $\mathbf{V}_i$ denote all applicant characteristics, excluding group identity g_i, such as ethnicity or gender. Loan examiners, indexed by e, form an expectation of the long-run profits of lending to applicant i conditional on observable characteristics $\mathbf{V}_i$ and group g_i, $\mathbb{E}[\alpha_i|\mathbf{V}_i, g_i]$.

The perceived cost of lending to applicant i assigned to examiner e is denoted by $t_g^e(\mathbf{V}_i)$, which is a function of observable applicant characteristics $\mathbf{V}_i$. The perceived cost of lending $t_g^e(\mathbf{V}_i)$ includes both the firm's opportunity cost of making a loan and the personal benefits to examiner e from any direct utility or disutility from being known as either a lenient or tough loan examiner, respectively. Importantly, we allow the perceived cost of lending $t_g^e(\mathbf{V}_i)$ to vary by group $g \in T, R$ to allow for examiner preferences to differ for applicants from the target group (e.g., minority applicants) and the reference group (e.g., non-minority applicants), respectively. We do not, however, allow the lender's true opportunity costs of lending to vary by group.

Following Becker (1957, 1993), we define loan examiner e as biased against the target group if $t_T^e(\mathbf{V}_i) > t_R^e(\mathbf{V}_i)$. Thus, biased loan examiners reject target group applicants that they would otherwise approve because these examiners perceive a higher cost of lending to applicants from the target group compared to observably identical applicants from the reference group.

For simplicity, we assume that loan examiners are risk neutral and maximize the perceived net benefit of approving a loan. We also assume that the loan examiner's sole task is to decide whether to approve or reject a loan application given that, in practice, this is the only decision margin in our setting.

Under these assumptions, the model implies that loan examiner e will lend to applicant i if and only if the expected profit is weakly greater than the perceived cost of the loan:

$$\mathbb{E}[\alpha_i|\mathbf{V}_i, g_i = g] \geq t_g^e(\mathbf{V}_i) \tag{10}$$

Given this decision rule, the marginal applicant for examiner e and group g is the applicant i for whom the expected profit is exactly equal to the perceived cost, i.e., $\mathbb{E}[\alpha_i^e|\mathbf{V}_i, g_i = g] = t_g^e(\mathbf{V}_i)$. We simplify our notation moving forward by letting this expected profit for the marginal applicant for examiner e and group g be denoted by α_g^e.

Based on the above framework, the model yields the standard outcome-based test for bias from Becker (1957, 1993).

Outcome Test 1: Taste-Based Bias. If examiner e is biased against applicants from the target group, then the expected profitability for the marginal target group applicant is higher than the

expected profitability for the marginal reference group applicant: $\alpha_T^e > \alpha_R^e$.

Outcome Test 1 predicts that marginal target and marginal reference group loan applicants should have the same profitability if examiners are unbiased, but marginal target group applicants should yield higher profits if examiners are biased against applicants from the target group. The correct procedure to test whether loan decisions are biased is therefore to determine whether loans to marginal target group applicants are more profitable than loans to marginal reference group applicants.

B. Inaccurate Group Stereotypes

In the taste-based model of bias outlined in the main text, we assume that examiners agree on the (true) expected net present profit of lending to applicant i, $\mathbb{E}[\alpha_i|\mathbf{V}_i, g_i]$, but not the perceived cost of lending to the applicant, $t_g^e(\mathbf{V}_i)$. An alternative approach is to assume that examiners disagree on their (potentially inaccurate) predictions of the expected profit, as would be the case if examiners systematically underestimate the profitability of target group applicants relative to reference group applicants in the spirit of Bordalo et al. (2016) and Arnold, Dobbie and Yang (2018). We show that a model motivated by these kinds of biased prediction errors can generate the same predictions as a model of taste-based bias.

Let i again denote applicants and $\mathbf{V}_i$ denote all applicant characteristics considered by the loan examiner, excluding group identity g_i. The perceived cost of lending to applicant i assigned to examiner e is now defined as $t^e(\mathbf{V}_i)$, where we explicitly assume that $t^e(\mathbf{V}_i)$ is independent of the group identity of the applicant.

The perceived profitability of lending to applicant i conditional on observable characteristics $\mathbf{V}_i$, $\mathbb{E}^e[\alpha_i|\mathbf{V}_i, g_i]$, is now allowed to vary across examiners. We can write the perceived profitability as:

$$\mathbb{E}^e[\alpha_i|\mathbf{V}_i, g_i] = \mathbb{E}[\alpha_i|\mathbf{V}_i, g_i] + \tau_g^e(\mathbf{V}_i) \tag{11}$$

where $\tau_g^e(\mathbf{V}_i)$ is a prediction error that is allowed to vary by examiner e and group identity g_i.

Following Arnold, Dobbie and Yang (2018), we define examiner e as making biased prediction errors against target group applicants if $\tau_T^e(\mathbf{V}_i) < \tau_R^e(\mathbf{V}_i)$. Thus, biased loan examiners reject target group applicants that they would otherwise approve because these examiners systematically underestimate the true profitability of lending to target group applicants compared to reference group applicants.

Following the taste-based model, loan examiner e will lend to applicant i if and only if the perceived expected profit is weakly greater than the cost of the loan:

$$\mathbb{E}^e[\alpha_i|\mathbf{V}_i, g_i = g] = \mathbb{E}[\alpha_i|\mathbf{V}_i, g_i = g] + \tau_g^e(\mathbf{V}_i) \geq t^e(\mathbf{V}_i) \tag{12}$$

The prediction error model can be made equivalent to the taste-based model of bias outlined above if we relabel $t^e(\mathbf{V}_i) - \tau_g^e(\mathbf{V}_i) = t_g^e(\mathbf{V}_i)$. As a result, we can generate identical empirical predictions

using the prediction error and taste-based models.

$$\mathbb{E}^e[\alpha_i|\mathbf{V}_i, g_i = g] = \mathbb{E}[\alpha_i|\mathbf{V}_i, g_i = g] + \tau_g^e(\mathbf{V}_i) \geq t^e(\mathbf{V}_i) \tag{13}$$

The prediction error model can be made equivalent to the taste-based model of bias outlined above if we relabel $t^e(\mathbf{V}_i) - \tau_g^e(\mathbf{V}_i) = t_g^e(\mathbf{V}_i)$. As a result, we can generate identical empirical predictions using the prediction error and taste-based models.

Following this logic, our model of biased prediction errors yields a similar outcome-based test for bias.

Outcome Test 2: Inaccurate Stereotypes. If examiner e systematically underestimates the true expected profitability of lending to target group applicants relative to reference group applicants, then the expected profitability for the marginal target group applicant is higher than the expected profitability for the marginal reference group applicant: $\alpha_T^e > \alpha_R^e$.

Parallel to Outcome Test 1, Outcome Test 2 predicts that marginal target group and marginal reference group applicants should have the same profitability if loan examiners do not systematically make prediction errors that vary with group identity, but marginal target group applicants should yield higher profits if examiners systematically underestimate the true expected profitability of lending to target group applicants relative to reference group applicants. The correct procedure to test whether loan decisions are biased is therefore, once again, to determine whether loans to marginal target group applicants are more profitable than loans to marginal reference group applicants.

Appendix C: Data Appendix

Examiner Leniency: We calculate examiner leniency as the leave-out mean residualized take-up decisions of loan examiners within a physical branch location. We use the residual take-up decision after removing store-by-month-by-nationality fixed effects. In our main results, we define loan take-up based on whether the first-time applicant took up their loan with the Lender.

Loan Approved: An indicator for the loan examiner approving the loan application (versus rejecting the loan application).

Loan Take-Up: An indicator for the loan applicant taking out a loan with the Lender (versus not taking out a loan). Loan Take-Up is set to zero if the loan application is rejected or if the application is approved but the applicant decides not to take out a loan.

Loan Top-Up: An indicator for the loan applicant closing the initial loan and taking out a new loan to cover the remaining balance on the initial loan. Loan Top-Up is set to zero for applicants who never take out an initial loan.

Long-Run Profits: We calculate profits as the sum of all payments made from the applicant to the Lender over the course of their relationship, minus the sum of all disbursements made from the Lender to the applicant in pounds. Long-Run Profits are set to zero for applicants who never take out a loan.

Short-Run Default: An indicator for the applicant defaulting on his or her first loan with the Lender. Short-Run Default is set to zero for applicants who never take out a loan.

Immigrant Applicant: An indicator for whether the applicant is an immigrant (versus native-born).

Female Applicant: An indicator for whether the applicant is female (versus male).

Applicant Age: The applicant's age in years. We drop applicants who are younger than 18 years old or older than 75 years old.

Old Applicant: An indicator for whether the applicant is at least 32 years old, the median sample age in the sample (versus less than 32 years old).

Months at Current Residence: The number of months the applicant has spent at their current residence as of the time of their application.

One Year at Residence: An indicator for the applicant having spent at least 12 months at their current residence as of the time of their application (versus less than 12 months at their current residence).

Married Applicant: An indicator for the applicant being married at the time of his or her first loan application (versus unmarried).

Number of Dependents: The applicant's number of dependents at the time of his or her first loan application. The number of dependents variable is winsorized at the 99th percentile of the distribution.

Credit Score: The applicant's credit score from a nationwide credit bureau at the time of his or her first loan application. The credit score variable is set to 0 for the approximately 2.6 percent of the sample with a missing credit score. In all regressions, we include an indicator for whether an applicant is missing the credit score variable or not.

Has Bank Account: An indicator for the applicant having a bank account at the time of his or her first loan application (versus no bank account).

Has Other Loan Payments: An indicator for the applicant having other loan payments at the time of his or her first loan application (versus no other loan payments).

Other Loan Payments: The value of an applicant's other loan payments in pounds at the time of his or her first loan application.

Number Open Lines of Credit: The applicant's number of open lines of credit at the time of his or her first loan application.

Total Credit Outstanding: The value of the applicant's total outstanding credit in pounds at the time of his or her first loan application.

Credit Arrears from Other Lenders: The value in pounds of overdue credit the applicant has to other lenders at the time of his or her first loan application.

Customer was Referred: An indicator for whether the applicant was referred to the Lender.

Loan Amount Requested: The applicant's requested loan amount in pounds. The applicant may take out less than the requested amount.

Loan for Emergency: An indicator for the self-reported reason for the loan being for an emergency.

Loan for Large One-Time Expense: An indicator for the self-reported reason for the loan being large non-recurring expense.

Loan for Overdraft Avoidance: An indicator for the self-reported reason for the loan being to avoid overdraft penalties.

Loan for Shopping or Holiday: An indicator for the self-reported reason for the loan being shopping or holiday expenses.

Loan Amount Net of Fees: The loan amount initially taken out minus all associated Lender fees in pounds. This variable is missing for applicants who did not take out a loan.

Loan APR: The annualized nominal interest rate of the loan. This variable is missing for applicants who did not take out a loan.

Loan Duration: Length of the loan in months if the applicant follows the set payment schedule. This variable is missing for applicants who did not take out a loan.

Male Examiner: An indicator for the examiner being male (versus female). We are missing examiner gender data for two examiners.

Experienced Examiner: An indicator for the examiner being employed by the Lender at the start of the sample period.