Generous to Men or Harsh to Women?

Experimentally Unpacking Gender Bias in Lending

Youngjoo Jung*

March 2026

Abstract

I study gender bias in small business lending by presenting 720 Egyptian loan officers past loan applications with randomized names. The same application is approved 6.2% more often if it has a male name. But is it due to generosity toward men or harshness toward women? I show the gender gap is driven by low-quality loans to men with repayment problems or low credit scores, suggesting generosity toward men. I then test strategies to reduce bias. I find that performance pay tied to repayment outcomes eliminates this gap by reducing approvals of low-quality male applicants, whereas diversity training is ineffective.

Keywords: Gender, Discrimination, Loan officer, Entrepreneurship, Performance pay

JEL Classification: D9, G5, O1

*Youngjoo Jung, yj27@illinois.edu, University of Illinois at Urbana-Champaign. Approval from University of Illinois Human Subjects Committee (IRB 24-0796) and American University in Cairo Human Subjects Committee (IRB 2023-2024-235). This project was registered on the AEA RCT Registry as AEARCTR-0015025 (Jung, 2025). I greatly appreciate my advisor, Adam Osman, for his sincere support. I am grateful for Richard Akresh, Julia Fonseca, Lena Song, Marieke Kleemans, Dan Bernhardt, John List, Robert Metcalfe, Marianne Bertrand, Michelle Brock, Ralph De Hass, Kareem Haggag, Muhammad Meki, Alex Chan, Ben Roth, Natalia Rigol, Reshma Hussam, Joshua Shea, Hoyoung Yoo, Bennedek Kurdi, Maria Loumioti, Leora Klapper, Jules Gazeaud, Carolyn Pelnik, Carolyn Tsao, Diego Ubfal, Sophie Shive, Silvia Prina, Sanghee Mun, Chan Kim, Yeonggyu Yun, Abigail Stocker, William Jeziorski, Sullam Jeoung, and participants at UIUC AMRL, job market seminar, AFE 2025, FMA Doctoral Consortium, EGSC 2025, University of Chicago experimental economics working group, and Accelerating Growth for Women Entrepreneurs Research Conference 2025 for their superb advice. I would also like to thank the Weiss Fund for their funding. For project management and field support, I am grateful to Family Development Foundation and J-PAL MENA staff, especially Mahmoud Hussein and Ahmed Gawish. All errors are those of the author.

1 Introduction

Loan officers act as gatekeepers to capital (Rigol and Roth, 2021), and their biases can contribute to capital misallocation and productivity loss (Hsieh and Klenow 2009; Moll 2014; Hsieh et al. 2019). While the persistent observational gender gap in entrepreneurial lending (Demirguc-Kunt et al., 2018; Ubfal, 2024) suggests bias, recent experimental work has found limited evidence of it at the extensive margin (Brock and De Haas, 2023; Ayalew et al., 2023). Beyond identifying the existence of bias, it remains unclear whether it stems from excessive generosity toward men—approving unqualified men for low-quality loans—or excessive harshness toward women—rejecting qualified women for high-quality loans—or both.

Disentangling generosity from harshness is empirically challenging. It requires evaluating whether decisions are right or wrong against the ground truth of applicant quality. Yet, the distinction is crucial. While both reduce lenders' profits, excessive generosity undermines financial stability, whereas excessive harshness restricts qualified borrowers' access to finance and hurts the economy by keeping qualified women's businesses smaller than optimal. The optimality of debiasing policies also hinges on the nature of bias: policies that induce lenders to raise standards for men are different from those that encourage more lending to women.

To unpack the nature of loan officer bias and find ways to reduce it, I implement a framed field experiment (Harrison and List, 2004) with 720 microfinance loan officers in Egypt. I provide examples of previous small business loan applications and ask officers if they would approve or deny the applicant with randomized names.[1] To test potential debiasing solutions, officers are randomly assigned to a control or one of three treatment groups: (i) higher performance pay tied to repayment outcomes, targeting extrinsic motivation; (ii) diversity training using the Gender–Career Implicit Association Test (IAT) feedback,[2] targeting intrinsic motivation; and (iii) a combined intervention that integrates both.

[1] In this experiment, officers receive performance pay based on ex-post loan repayment outcomes: a base payment, bonuses for approving repaid loans or rejecting defaulted loans, and penalties for approving defaulted loans or rejecting repaid loans.

[2] The IAT measures implicit bias by recording how quickly individuals associate men with careers and women with family relative to the reverse associations. The feedback presents the score and its interpretation.

I find strong evidence of gender bias. The same application is 6.2% more likely to be approved when randomly assigned a male name. The nature of this gender gap stems mainly from excessive generosity toward men: the gender gap is concentrated among low-quality (negative NPV) loans that should not have been extended. This pattern holds across three measures of loan quality: (i) for ex-post defaulted loans, the gender gap is 14.7%, while it is negligible for ex-post repaid loans; (ii) for loans with a history of late repayment, the gap is 21.6%, vs. a marginal gap for those who paid on time; and (iii) the gap is 16.8% for low credit score loans, but it is negligible for high credit score loans. I also find that this bias is concentrated among inexperienced or less-educated loan officers, consistent with overly optimistic priors for men, suggesting statistical discrimination (Phelps, 1972; Aigner and Cain, 1977; Bohren et al., 2023).

Turning to the debiasing interventions, higher performance pay eliminates the gender gap by reducing over-approvals of low-quality male applicants. However, diversity training via IAT feedback—which informs officers of their tendency to associate men with careers and women with family—fails not only to reduce the gender gap but also to reduce approvals of low-quality male applicants that drive the gap through excessive generosity toward men. Moreover, in the combined intervention, adding IAT feedback to higher performance pay yields no additional benefit.

These results indicate that higher performance pay is an effective debiasing policy when the gender gap is driven by excessive generosity toward men: it both narrows the gender gap and reduces the over-approval of low-quality male applicants, thereby targeting the source of the bias. In contrast, IAT feedback remains ineffective. A back-of-the-envelope calculation suggests that generosity toward men costs the bank approximately EGP 3.6 billion (USD 72.9 million) per year, and that increasing performance pay could raise annual net profits by roughly EGP 3.2 billion (USD 64.3 million).

Using a Bayesian theoretical framework, I show that increasing performance pay tied to repayment outcomes, a scalable debiasing tool for lenders, can reduce the gender gap in loan

approval. Under taste-based bias, higher performance pay lowers the relative importance of taste-based utility. Under statistical discrimination, it increases screening effort, reducing officers' reliance on optimistic priors about male borrowers. Empirically, doubling performance pay increases screening time by only 7 seconds (11%) per loan yet fully eliminates the gender gap. The small increase in screening time is unlikely to generate enough new information to close the gap, suggesting that taste-based bias (Becker, 1957) plays a role.

I contribute to the literature on biased decision-making by, to the best of my knowledge, being the first to experimentally separate generosity toward an in-group from harshness toward an out-group. While theoretical decompositions of bias into generosity and harshness were proposed early in psychology (Glick and Fiske, 1996), the economics literature has largely focused on documenting relative gaps[3] rather than identifying whether they arise from generosity toward the in-group or harshness toward the out-group. Building on insights from recent empirical studies suggesting that bias may operate from both sides (Sarsons, 2022; Egan et al., 2022; Hebert et al., 2024), I provide the experimental evidence in the lending context by evaluating loan officers' decisions against verifiable benchmarks of borrower quality. I show that the observed gender gap is concentrated among low-quality loans but not high-quality loans, suggesting that the gender gap is driven by excessive generosity toward unqualified men rather than excessive harshness toward qualified women.[4]

[3]The broader literature on biased decision-making primarily documents relative gaps rather than distinguishing their underlying sources. Evidence spans multiple economic settings, including labor markets (Altonji and Blank, 1999; Bertrand and Mullainathan, 2004; De Mel et al., 2009; Goldin, 2014; Blau and Kahn, 2017; Glover et al., 2017; Baert, 2018; Neumark, 2018; Agan and Starr, 2018; Mocanu, 2022; Egan et al., 2022; Osman et al., 2023), housing (Hanson and Hawley, 2011; Hanson et al., 2016; Flage, 2018; Bao, 2024), teaching (Mengel et al., 2019; Arnold et al., 2020), and judicial decisions (Arnold et al., 2018). A similar pattern appears in the credit market literature, including gender (Carter et al., 2007; Muravyev et al., 2009; Mascia and Rossi, 2017; Beck et al., 2018; Alibhai et al., 2019; Hebert, 2023; Brock and De Haas, 2023; Ayalew et al., 2023; Hebert et al., 2024; Agarwal et al., 2023; Montoya et al., 2024; Bartős et al., 2024), race (Berkovec et al., 1998; Cavalluzzo and Cavalluzzo, 1998; Blanchflower et al., 2003; Ross et al., 2008; Sarsons, 2022; Fuster et al., 2022; Butler et al., 2023), cultural proximity (Fisman et al., 2017), social networks (Haselmann et al., 2018), and immigrant or age status (Dobbie et al., 2021).

[4]This generosity–harshness framework differs from explanations based on information frictions under statistical discrimination, where bias declines as evaluators receive richer information. Hanson and Hawley (2011), for example, study a setting in which landlords initially have limited information about prospective tenants. When additional high-quality signals—such as well-written email inquiries—are provided, landlords rely less on demographic proxies, reducing the racial gap. In contrast, loan officers in my setting already observe the standard information used to screen loans, including credit scores and repayment histories,

I further advance this literature by demonstrating that debiasing interventions must hinge on this distinction. When bias takes the form of in-group generosity, effective solutions must increase the cost of approving low-quality in-group applicants rather than simply encouraging more approvals of out-group applicants. In my setting, the gender gap arises from the over-approval of low-quality male applicants rather than the under-approval of high-quality female applicants. I show that increasing performance pay tied to repayment outcomes eliminates the gender gap by reducing approvals of low-quality male applicants. By contrast, diversity training via Implicit Association Test feedback is ineffective at both reducing the gender gap and decreasing approvals of low-quality male applicants.

I also contribute to the literature on the gender gap in lending by providing the first experimental evidence of gender bias in small business lending at the extensive margin, documenting a gender gap in loan approvals. While previous observational studies report gender bias in lending (Carter et al., 2007; Muravyev et al., 2009; Stefani and Vacca, 2013; Beck et al., 2018), experimental evidence remains scarce and mixed. Prior experimental studies document gender bias in officers' evaluations of business proposals in Uganda (Bartŏs et al., 2024), hypothetical loan size in Turkey (Alibhai et al., 2019), and consumer loan approvals in Chile (Montoya et al., 2024), but none in business loan approvals. This is partly because they examine settings where gender bias is less likely to emerge. For example, Brock and De Haas (2023) found gender bias in collateral requirements in Turkey's banking sector but no bias in loan approval decisions—likely because their sample focused on formal bank lending, which tends to involve high-quality applications, making it harder to detect bias stemming from generosity toward men in low-quality applications. Similarly, Ayalew et al. (2023), using data from a national business competition in Ethiopia, found no gender bias in loan approval, possibly because the applicants were selective, high-performing entrepreneurs and the evaluators were highly experienced officers. In contrast, my experiment deliberately excludes applications with exceptionally high or low credit scores. By focusing on loans

leaving less scope for bias driven by missing information. Furthermore, the observed gender gap is not purely driven by statistical discrimination, providing suggestive evidence of taste-based bias.

where officer discretion is likely greater and gender bias more likely to emerge (Costello et al., 2020; Agarwal et al., 2024; FDIC, 2024), I provide the first experimental evidence of gender bias in small-business loan approval.

Lastly, I contribute to the literature on loan officer incentives by showing that repayment-based pay can serve as a debiasing tool. Prior work has studied loan officer compensation primarily as a general performance incentive (Hertzberg et al., 2010; Cole et al., 2015; Agarwal and Ben-David, 2018), noting that many microfinance institutions in low- and middle-income economies still rely on fixed pay or volume-based commissions. By demonstrating that repayment-based performance pay can close the gender gap in a setting where performance can be objectively measured[5] by repayment outcomes, I provide a scalable debiasing solution for lending institutions.

The paper is structured as follows. Section 2 provides the context and conceptual framework. Section 3 introduces the experimental design and data. Section 4 presents the main results. Section 5 discusses heterogeneous effects. Section 6 investigates the sources of gender bias, distinguishing between taste-based bias and statistical discrimination. Concluding remarks are offered in Section 7.

2 Context and Conceptual Framework

The experiment is conducted in Egypt, a suitable setting to examine whether gender bias reflects excessive generosity toward men or excessive harshness toward women, since identifying the presence of bias is a necessary first step. Discrimination against women in Egypt is rampant (Osman et al., 2023) and the impacts of microcredit are shown to be greater for women than men (Crépon et al., 2024). Evidence from the World Values Survey (Haerpfer

[5]Prior literature shows that performance pay may have a complex relationship with bias, depending on the objectivity of the evaluation process (Heywood and O'Halloran, 2005; Heywood and Parent, 2012; Marinescu et al., 2018; Mejia and Parker, 2021). In the lending context, ex-post repayment outcomes are objectively measurable, leaving little room for evaluator discretion.

et al., 2022) further highlights aspects of gender bias in Egypt.[6]

The experiment targets loan officers at microfinance institutions (MFIs) in Egypt. Between November 2024 and May 2025, 720 officers from 29 MFIs participated in the study. Figure 1 maps the five governorates where the experiment took place: Aswan, Asyut, Luxor, Minya, and Qena. Table A1 provides a list of the MFIs and the number of participating officers from each. The largest participating MFI is the Family Development Foundation, based in Aswan and Luxor, and the experiment uses administrative loan application data from this institution. In this sample, the participating officers are gender-balanced, relatively young (average age 34.6), and well-educated (16.2 years of schooling), with an average of 6.3 years of experience as loan officers (median: 4 years). Loan officers approve 67.3% of the loans assigned to them at their workplace. They exhibit moderate implicit gender bias, with a mean Gender-Career IAT score of 0.37—suggesting a medium or above association of men with career and women with family. Officers report an average score of 6.44 on a 0–10 scale of self-reported risk preference. Column {1} of Table 1 reports summary statistics for the remaining loan officer characteristics.

In Egypt, loan officers receive a small base salary supplemented by performance bonuses tied to loan repayment outcomes. They also face penalties in promotion or even the risk of dismissal when repayment performance is weak. Reflecting this structure, officers receive performance pay based on ex-post loan outcomes in the experiment (Cole et al., 2015; Brock and De Haas, 2023): basic pay, bonuses for approving repaid loans or rejecting defaulted loans, and penalties for approving defaulted loans or rejecting repaid loans.

In Appendix D, I formalize a Bayesian updating framework to show how increasing performance pay tied to ex-post repayment outcomes can reduce the gender gap through two mechanisms. First, if taste-based bias favoring men is present, so that loan officers derive utility from lending to men, stronger performance incentives reduce the relative influence

[6]In a 2017 survey comparing Egypt and the United States, 36.8% of Egyptian respondents strongly agreed that "on the whole, men make better business executives than women do," compared with just 2.2% in the United States. Please see the details in Figure A1

of this utility, narrowing the gender gap. Second, if officers hold overly optimistic priors about men's repayment probabilities, higher performance pay induces greater screening effort, generating more accurate signals and reducing reliance on biased prior beliefs. The model predicts that increasing performance pay tied to repayment outcomes increases loan officers' screening effort and reduces the approval of low-quality male applicants.

3 Experimental Design and Data

3.1 Experimental Design

The experiment is introduced as an assessment of officers' ability to screen for defaulted loans to prevent them from realizing its true purpose—measuring gender bias. The field team coordinates with MFIs and schedules in-person visits. Experiments are conducted individually, with one enumerator assigned per officer. When MFIs lack sufficient space, nearby rented facilities are used. All activities are administered using SurveyCTO. Loan application information is provided in printed form to replicate typical screening practices.

The experiment begins with a baseline survey collecting data on demographics, work characteristics, cognitive ability, risk preferences, psychometrics, and social desirability. Officers then complete a Gender-Career Implicit Association Test (IAT) adapted for SurveyCTO from the publicly available version.[7] Next, officers complete lending decision tasks by screening 10 loan applications.[8] The experiment concludes by asking participants explicit gender-attitude questions drawn from the World Values Survey (Haerpfer et al., 2022). Details of the questionnaires are provided in Appendix C. On average, the total duration of the experiment was 39.3 minutes and no participants dropped out during the course of the experiment.

[7]https://implicit.harvard.edu/implicit/. Following standard IAT procedures (Greenwald et al., 1998, 2003; Lane et al., 2007), trials shorter than 300 milliseconds or longer than 10,000 milliseconds are excluded, and the standard seven-block structure is applied.

[8]Following this, officers were re-randomized into a second experimental stage involving AI-assisted lending. This second stage measures the effectiveness of generative AI and potential human-in-the-loop gender bias, which addresses a distinct research question and is analyzed in a separate paper. (Jung, 2026)

Officers receive a base payment of EGP 400 (≈ USD 8). They earn EGP 10 for approving a loan that was ex-post repaid and EGP 15 for rejecting a loan that was ex-post defaulted. Conversely, they receive a penalty of EGP 15 for approving an ex-post defaulted loan and EGP 10 for rejecting an ex-post repaid loan. This asymmetric incentive structure reflects the real-world context, where loan officers face greater penalties for approving bad loans than for rejecting good ones. Table A2 presents the details of the incentive scheme. Given that the median monthly income of participating officers is approximately EGP 6,000 (≈ USD 120), the rewards for the experiment are substantial: depending on their decision accuracy, officers could earn between EGP 280 and 520 in total compensation, receiving EGP 417 on average, the equivalent of more than a full day of work.

3.1.1 Loan Portfolios

I use 50 previous approved loan applications[9] in the experiment, selected based on credit scores and industries. Given the limited evidence in the literature regarding the experimental presence of gender bias in small business loan approvals (Brock and De Haas, 2023; Ayalew et al., 2023), I focus on cases where loan officers have greater discretion and where gender bias is more likely to emerge (Costello et al., 2020; Agarwal et al., 2024; FDIC, 2024). In the Egyptian credit scoring system, there are five tiers: excellent (above 750), very good (701–750), satisfactory (626–700), not satisfactory (520–625), and bad (below 520). Also, there are many small businesses that operate informally and do not have credit scores due to the limited coverage of the credit system. Then, I exclude loans with very good credit scores (above 750) and bad scores (below 520), as these represent clearly high- or low-quality applicants where gender bias is less likely to occur. I instead select applications with credit scores between 520 and 750 or with no credit score. Next, to reduce the extent to which gender provides meaningful prior information, I restrict the sample to industries where

[9]Data on rejected loans were unavailable and therefore not included. With an approval rate of about 70%, roughly 30% of applications are not captured in the analysis. In microfinance lending, loan officers first screen applications for eligibility and reject loans that are clearly low quality, where gender bias is unlikely to arise due to limited discretion (Costello et al., 2020; Agarwal et al., 2024; FDIC, 2024).

each gender represents at least 30% of borrowers, as gender composition may otherwise affect screening standards (Mocanu, 2022). The selected industries include agriculture, supermarkets, grocery stores, clothing shops, fish sale shops, poultry, and livestock.

The ratio of ex-post repaid to defaulted loans was designed to mirror real-world behavior and provide sufficient variation to capture generosity and harshness. Although precise approval rate data from the Family Development Foundation were unavailable, discussions with a manager indicated typical approval rates of around 70%. Consistently, in the baseline, officers approved 67.3% of assigned applications on average in their real-world workplace. Reflecting this, 60% of the 50 loans used in the experiment were ex-post repaid and 40% were defaulted, making the optimal approval rate 60%. Officers are informed that the repayment rate in the experiment is 60% on average, though it may vary across individuals.

To isolate gender as the only varying attribute when officers screen multiple loans, it is crucial to control the composition of loan bundles, as the evaluation of a loan may differ depending on the other loans presented alongside it. To address this, I group the 50 loans into five bundles, each containing 10 loan portfolios with 60% repaid loan rate: 3 ex-post repaid male-name loans, 3 ex-post repaid female-name loans, 2 ex-post defaulted male-name loans, and 2 ex-post defaulted female-name loans. I then create Bundles 6 through 10 by duplicating Bundles 1 through 5, but swapping each applicant's name to suggest the opposite gender. For example, Bundles 1 and 6 contain the same 10 loan applications in the same order, but names are switched (e.g., "Noha Mahmoud" in Bundle 1 becomes "Ahmed Mahmoud" in Bundle 6). This pairing process is repeated for Bundles 2 and 7, 3 and 8, 4 and 9, and 5 and 10. While gender is not explicitly stated, fictitious first names, selected from commonly-used Egyptian names, implicitly indicate the applicant's gender. See Appendix B for the detailed sample selection procedure.

The loan portfolios contain standard financial information used in real-world loan screenings in the partner MFI, including name, age, number of guarantors, monthly income, business registration status, business start year, location, number of employees, industry, loan

size, loan tenure, interest rate, number of previous loans, number of delayed days on the most recent loan, loan size and interest rate of the most recent loan, and credit score. As in real-world screening processes, the officers review these printed portfolios. Figure A2 displays an example of a loan application used in the experiment.

3.1.2 Randomization

Randomization is conducted at the individual level, stratified by the officer's gender. Upon arrival, each officer receives a unique key that contains the loan bundle they will evaluate. Figure 2 outlines the overall randomization process. Officers are randomized into one of four groups. Specifically, Group 1 is the control group (C), where officers make approval or rejection decisions without any intervention. Group 2 is the higher performance pay group (T1), where officers receive double the standard performance pay tied to ex-post loan outcomes while keeping the same base pay. Table A2 summarizes the incentive structure. Specifically, the minimum compensation—if all decisions are incorrect—is EGP 180, while the maximum—if all decisions are correct—is EGP 640. This extrinsic motivation is designed to raise the opportunity cost of biased behavior and encourage greater screening effort, as outlined in the conceptual framework in Section 2. Group 3 is the IAT feedback group (T2), where officers receive feedback on their Gender-Career Implicit Association Test (IAT) results before making loan decisions, similar to diversity training. The feedback format, shown in Figure A3, is based on the version available at `https://implicit.harvard.edu/implicit/`. This diversity training, grounded in intrinsic motivation, aims to mitigate bias by raising awareness of unconscious attitudes. Group 4 is the combined intervention group (T3), integrating both higher performance pay and IAT feedback. Officers in this group first receive feedback on their IAT results, followed by an explanation of the higher performance pay structure, leveraging both intrinsic and extrinsic motivation.

3.2 Data

3.2.1 Balance Table

Table 1 presents the balance of baseline characteristics based on randomization. Overall, there are no significant differences across the randomized groups, with the exception of weekly working hours. The joint F-test yields a p-value of 0.404, indicating that, collectively, the groups are well balanced on observable characteristics. In the balance table, the monthly approval rate refers to the average proportion of approved loans out of total assigned loans, and loan screening time (in hours) measures the average time spent per loan when making approval or rejection decisions in practice. Appendix C provides additional details on other baseline measures.

3.2.2 Ex-post Repaid vs. Defaulted Loans

A loan officer's main task was to distinguish ex-post defaulted loans[10] from ex-post repaid loans based on observable characteristics. Using logit regression, I first assess whether defaulted and repaid loans differ significantly in their financial characteristics by estimating Equation (1).

$$\text{logit}(\Pr(\text{Default}_i = 1)) = \alpha + \sum_{k=1}^{9} \beta_k X_{ik} \tag{1}$$

Here, i indexes 50 loan portfolios and k indexes 9 loan characteristics, excluding categorical variables such as industry or region due to the small loan sample size, with β_k capturing the association between each characteristic and the likelihood of default. The outcome variable $Default_i$ equals 1 if the loan defaulted and 0 if it was repaid.

In Table 2, the joint F-test yields a p-value of 0.001, indicating that defaulted loans differ significantly from repaid loans based on observable financial characteristics. Specifically, ex-post defaulted loans tend to have larger loan amounts and longer tenures, but not higher

[10]Default is defined as loans with 90 or more total days of late repayment.

monthly borrower income. Ex-post defaulted loans have more late repayment days on their most recent loans and fewer employees. While not individually statistically significant, other patterns—such as a higher number of previous loans, younger borrowers, fewer employee guarantors, and a lower likelihood of being registered with the government—also suggest that ex-post defaulted loans are, on average, of lower quality than repaid loans.

3.2.3 Baseline Gender Bias of Loan Officers

I measure loan officers' baseline gender bias using both an Implicit Association Test (IAT) and survey-based questions on gender attitudes. First, the IAT score captures the officers' implicit bias on connecting male with career and female with family. Figure A4 presents the distribution of officer's IAT scores, categorized into seven ranges.[11] Based on IAT score range, more than half of officers have IAT score above 0.35, which stands for medium or above association between male-career and female-family.

Second, I measure officers' stated gender attitudes using six survey questions drawn from the 2017 World Values Survey (Haerpfer et al., 2022). Figure A5 presents the distribution of average responses across these questions.[12] These questions were asked at the end of the experiment, after lending decisions, and therefore may not capture the baseline explicit bias. Most officers score below 3 on average, suggesting limited explicit gender bias.

[11]The ranges are coded from -3 to 3, where positive numbers indicate an association of male with career and an association of female with family, and negative numbers indicate the opposite association. Specifically, a value of 3 corresponds to an IAT score above 0.65, indicating a strong association between male-career and female-family. A value of 2 corresponds to an IAT score between 0.35 and 0.65, reflecting a moderate male-career and female-family association. A value of 1 represents an IAT score between 0.15 and 0.35, suggesting a slight male-career and female-family association. A value of 0 reflects an IAT score between -0.15 and 0.15, indicating little or no implicit association. On the negative side, a value of -1 represents an IAT score between -0.35 and -0.15, corresponding to a slight female-career and male-family association. A value of -2 corresponds to an IAT score between -0.65 and -0.35, indicating a moderate female-career and male-family association. Finally, a value of -3 corresponds to an IAT score below -0.65, indicating a strong implicit association of female-career and male-family.

[12]Each question uses a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree), with higher values indicating greater stated gender bias.

4 Results

4.1 The Existence of Gender Bias in Loan Approval

Gender bias in lending decisions is measured by observing whether same loan portfolios are more likely to be approved when assigned male names than with female names in the control group, using Equation (2).

$$Y_{ip} = \alpha + \beta Male_p + \mu_p + \delta_{LO\ gender} + \epsilon_{ip} \tag{2}$$

Here, Y_{ip} is a binary indicator equal to 1 if loan officer i approves portfolio p, and 0 otherwise. $Male_p$ is a binary variable indicating whether the portfolio is assigned a male name. μ_p denotes portfolio fixed effects, providing a within-portfolio estimate of gender gap. $\delta_{LO\ gender}$ is the gender of the loan officer fixed effect. Standard errors are clustered at the loan officer level. In this specification, β represents the approval rate difference between male and female portfolios, with a positive β indicating bias against female applicants.

Column {1} of Table 3 reports the estimates of β using the overall sample. Two key features should be noted. First, on average, the same applications are 4.4 p.p. (6.2%) more likely to be approved with male names than with female names, providing strong evidence of gender bias in loan approvals. Second, loan officers over-approve loans. On average, 75.2% of loans with male names are approved, and 70.8% of those with female names are approved. These approval rates substantially exceed the 60% ex-post repayment rate, indicating over-approval. Conceivably, this over-approval may reflect their usual behavior, as they report approving 67.3% of assigned loans on average in their workplace.

4.2 The Nature of Gender Bias: Generosity Toward Men

The nature of gender bias can manifest itself as excessive generosity toward men, excessive harshness toward women, or a combination of both. Bias is characterized as excessive gen-

erosity toward men when the gender gap is more pronounced in low-quality loans, and as excessive harshness toward women when the gap is more concentrated in high-quality loans. To distinguish between high- and low-quality loans and thus assess which applications should have been approved or rejected, I use three measures: (i) ex-post repayment outcomes; (ii) past late repayment history; and (iii) credit scores.

4.2.1 Ex-post Outcome

The ex-post loan outcome—whether repaid or defaulted—offers an indication of quality, as defaulted loans are more likely to have been low-quality. As Table 2 showed, ex-post defaulted loans exhibit lower-quality (i.e., lower NPV) characteristics on average than ex-post repaid loans. Using Equation (2), I conduct separate subgroup analyses for ex-post defaulted and ex-post repaid loans. Column {2} of Table 3 shows that the same loan portfolio is approved 9.6 p.p. (14.5%) more often when assigned a male name than when assigned a female name, suggesting excessive generosity toward men. In contrast, among ex-post repaid loans (Column {3}), there is no longer a statistically significant gender gap in approval rates, indicating no evidence of excessive harshness toward women.

However, ex-post outcomes have limitations due to their stochastic nature: repaid loans are not necessarily ex-ante high-quality, and defaulted loans need not be ex-ante low-quality. To address this limitation, I provide additional evidence of generosity toward men using two alternative, observable measures of quality: past late repayment history and credit scores.

4.2.2 Past Late Repayment History

Out of the 50 loans, 11 have a history of late repayment on their most recent loans, meaning the borrower was at least one day late on their most recent loan repayment. Unlike the ex-post loan outcome—which is not observable to officers during the experiment—late repayment history is observable to officers and serves as a clear signal of default risk. Using Equation (2), I conduct subgroup analyses separately for loans with late repayment histories

and those without them, including first-time loan applicants.

Column {4} of Table 3 shows that, among loans with a history of late repayment, the same portfolio is approved 11.1 p.p. (21.6%) more often when assigned a male name. In contrast, there is no statistically significant gender difference in approval rates for loans without a late repayment history (column {5}). This comprises further evidence that the nature of gender bias in lending is driven by excessive generosity toward men, rather than excessive harshness toward women.

4.2.3 Credit Score

Credit scores, generated by I-Score from the Egyptian Credit Bureau, serve as a key indicator of a borrower's ex-ante risk of default. In this experiment, loan officers were provided a one-page explanation detailing the tiers of credit scores: a score of 751 or above is classified as excellent; 701 to 750 as very good; 626 to 700 as satisfactory; 521 to 625 as unsatisfactory; 520 or below as bad; and "missing" indicates that a credit score does not exist. Using Equation (2), I conduct subgroup analyses based on these credit score tiers. First, I examine 16 loans with credit scores of 700 or below, which fall into the satisfactory or lower categories and are considered low-quality. Greater gender bias in this group would suggest excessive generosity toward men. Second, I analyze 18 loans with credit scores of 701 or above, which are relatively high-quality, where a stronger gender gap in favor of women would indicate excessive harshness toward men. Finally, I analyze 16 loans without available credit scores.

Column {5} of Table 3 shows that, among loans with credit scores of 700 or below, the same portfolio is approved 9.7 p.p. (16.8%) more often when assigned a male name, evidence of generosity toward men. In contrast, Column {6} shows that among loans with credit scores of 701 or above, there is no statistically significant gender gap in approval rates. Column {7} reveals that even among loans without a credit score, there is no statistically significant gender difference in approval.[13]

[13]The lack of gender bias in loans without credit scores is likely to be due to endogeneity in the sample of approved loans. As the experiment includes only loans that were approved in the past, those without

Next, I analyze whether there is a statistically significant monotonic relationship between credit score and loan approval during the experiment. Figure 3 shows that female loan approval rates have a statistically significant monotonic relationship with credit scores. In contrast, the relationship is not statistically significant for male portfolios. Moreover, the difference in slopes between male and female portfolios is statistically significant at the 10% level. The insignificant slope for male portfolios and the steeper, statistically significant slope for female portfolios suggest a pattern of excessive generosity toward men. Specifically, male portfolios receive relatively high approval rates even at lower credit scores, resulting in a flatter approval trend across the credit score range of 680 to 750. This implies that men are too likely to be approved for loans given lower credit scores.

Taken together, the evidence from credit scores supports the previous findings using ex-post loan outcomes and late repayment history. Across all three measures of loan quality, the gender gap in approval rates is concentrated in low-quality loans. This consistent pattern reinforces the interpretation that the nature of gender bias in loan approval is driven by excessive generosity toward men, rather than excessive harshness toward women.

4.2.4 Robustness Check: ML-based Benchmarking of Quality

As a robustness check to benchmark ex-ante repayment probability, I use a Gradient Boosting Machine (GBM) model[14] to predict repayment probabilities for the 50 experimental loans. The model is trained[15] on a dataset of 561 loans from the same MFI, using 9 baseline

a credit score tend to have smaller loan amounts relative to the borrower's income. This pattern suggests that loan officers may have approved smaller loans to reduce risk, or borrowers may have applied for smaller loans to improve their chances of approval. In Table 3, the female approval rate for loans without credit scores is 76%, which is similar to the approval rate for loans with very good credit scores. Moreover, the repayment-to-income ratio for these loans is only 0.287, lower than that of loans in the lower credit score group. This indicates that loans without credit scores were, on average, relatively high-quality loans.

[14]The GBM model is chosen for its superior Area Under the ROC Curve (AUC) performance compared to Random Forest, Support Vector Machines, neural networks, XGBoost, and CatBoost on this dataset.

[15]To address class imbalance, I apply SMOTE and perform cross-validation with 10 folds and 10 repeats. Hyperparameters are tuned over a grid: the number of trees ranges from 300 to 1000 (in increments of 100); tree depth is set at 3, 5, or 7; the learning rate is 0.005; and the minimum number of observations per terminal node is 10.

features.[16] In addition, as the credit scoring system was recently updated, the original application-time credit scores are no longer accessible and thus not included in the training data. Consequently, the ML-based predicted repayment probabilities are trained on slightly less information than what officers originally observed. The most predictive features, in descending order, are: loan tenure, monthly repayment-to-income ratio, days of past late repayment, number of previous loans, and other borrower characteristics.

The model performs well, with an average AUC of 0.88, accuracy of 0.83.[17] For each of the 50 experimental loans, I assign the average predicted repayment probability out of 1,000 runs. As the ML-predicted repayment probabilities may differ from those inferred by loan officers,[18] defining a single approval threshold is challenging. I therefore use two thresholds: $p = 0.50$ as a neutral benchmark, and $p = 0.54$ as the point at which loans are classified into the bottom 40% and top 60%.[19] Based on these thresholds, I define the gender gap below the threshold as reflecting generosity toward men, and the gender gap above the threshold as reflecting harshness toward women.

Table A3 presents the regression results using Equation (2). Column {1} shows that among low-quality loans (predicted repayment probability below 0.5), the gender gap is 6.8 p.p. (10.4%) and statistically significant. In contrast, Column {2}, which includes high-quality loans (probability 0.5 or above), shows a smaller and statistically insignificant gender gap of 3.1 p.p. (4.2%). Similarly, using a threshold of 0.54, Column {3} shows that the gender gap among low-quality loans is 6.1 p.p. (9.1%) and statistically significant, while Column (4) shows a gap of only 3.3 p.p. (4.5%) among high-quality loans, which is not

[16] Monthly repayment-to-income ratio, loan tenure (months), days of past late repayment, number of previous loans, age of borrower, registered business (binary), first-time loan (binary), number of employees, number of employee guarantors. Industry and region are excluded due to the limited sample size.

[17] Using the 0.6 probability threshold, the model achieves an F1 score of 0.85, precision of 0.86, and recall of 0.84—each with a standard deviation of approximately 0.01 across 1,000 iterations.

[18] The ML model is trained using only approved loans, which likely excludes very low-quality applications. As a result, the model may overestimate the probability of repayment. Also, credit score, industry, and geographic variables are excluded from model training due to the limited sample size and data availability.

[19] The ex-post default rate is 40%, and loan officers are informed of this, although the exact rate varies across individuals. Accordingly, the bottom 40% of loans, based on predicted repayment probabilities from the machine-learning model, can be considered low-quality loans.

statistically significant. Although I cannot statistically reject the null hypothesis that the gender gap in low-quality loans equals that in high-quality loans, the pattern suggests that the bias is consistent with excessive generosity toward men.

4.3 Debiasing Treatment Effect

4.3.1 Change in Gender Gap in Loan Approval

I analyze the effects of the three debiasing treatments on loan approval relative to the control group using Equation (3).

$$Y_{ip} = \alpha + \sum_{k=1}^{3} \beta_k T_{ik} Male_p + \sum_{k=1}^{3} \gamma_k T_{ik} + \delta Male_p + \mu_p + \delta_{\text{LO gender}} + \epsilon_{ip} \tag{3}$$

Here, Y_{ip} is a binary indicator equal to 1 if loan officer i approves portfolio p, and 0 otherwise. T_{ik} is a binary variable that equals 1 if loan officer i belongs to treatment group k, and 0 for all other groups. The k can take the values 1, 2, or 3, corresponding to the three treatment groups. All other definitions remain the same as in Equation (2). The interaction coefficient β_k between the treatment dummy T_{ik} and the male applicant dummy $Male_p$ captures the change in the gender gap in loan approval in treatment group k, relative to the gender gap in the control group, captured by δ.

Column 1 of Table 5 presents the estimated coefficients β_k from Equation (3). The results show that higher performance pay eliminates the gender gap by reducing the male–female approval difference by 6.6 p.p. However, diversity training via IAT feedback is not effective in reducing the gender gap. Furthermore, the combined intervention, which adds IAT feedback to higher performance pay, also closes the gender gap but does not differ from higher performance pay alone.

Columns {2} and {3} present subgroup analyses of Column {1} by the Gender–Career IAT score. Column {2} includes officers biased against women (IAT score ≥ 0.35[20]), and

[20]Officers with scores of at least 0.35 are classified as having moderate or stronger implicit associations linking men with careers and women with family roles, following prior studies (Greenwald et al., 1998, 2003;

Column (3) includes the rest. The debiasing effect of IAT feedback should be analyzed primarily among officers whose IAT scores indicate bias against women. The feedback is expected to have a stronger impact among these officers, as it explicitly highlights their implicit association between men and careers. For officers without such bias, the feedback simply confirms neutrality ("you are not associating men with career and women with family") and is therefore unlikely to induce behavioral change.

However, IAT feedback reduces the gender gap by only 2.6 p.p. among officers biased against women, compared with 4.0 p.p. among others. This suggests that IAT feedback is ineffective in reducing the gender gap, possibly because officers who are already aware of their bias are unlikely to change their behavior merely through diversity training.

4.3.2 Change in Screening Performance and Screening Time

Next, I analyze whether the debiasing treatments improve officers' screening performance by increasing approvals of repaid loans and rejections of defaulted loans, and whether they affect screening time using Equation (4). The Bayesian framework in Section 2 predicts that higher performance pay will increase loan officers' screening effort and reduce approvals of low-quality male applicants.

$$Y_{ip} = \alpha + \sum_{k=1}^{3} \gamma_k T_{ik} + \mu_p + \delta_{LO\ gender} + \epsilon_{ip} \tag{4}$$

Here, Y_{ip} is a binary accuracy indicator for male-name and female-name loans in Columns {1} and {2} in Table 5, equal to 1 if the officer approves an ex-post repaid loan or rejects an ex-post defaulted loan, and 0 otherwise. In Column {3}, Y_{ip} represents the average screening time per loan.[21] All other variable definitions remain the same as in Equation (3).

Columns {1} and {2} of Table 5 present the change in screening accuracy in each treatment arm relative to the control group for male-name and female-name loans, respectively.

Lane et al., 2007).

[21]The increase in screening time reflects the overall change in officer effort across all loan applications. Screening time data broken down by the applicant's gender was not collected during the experiment.

Higher performance pay and the combined intervention increase the screening accuracy for male-name loans by 3.6 p.p. (6.6%) and 4.2 p.p. (7.7%), respectively, while the IAT feedback alone has no statistically significant effect. In contrast, there is no statistically significant change in the screening performance for female-name loans across any treatment arm.[22] Column {3} reports the change in average screening time per loan across treatments. Officers with higher performance pay increase their screening time per loan by 7 seconds (11.3%), but there is no statistically significant change with IAT feedback or the combined intervention.

Overall, combining IAT feedback with higher performance pay yields no additional benefit over higher performance pay alone. IAT feedback is ineffective on its own and provides no complementary value when paired with higher performance pay.

4.3.3 Debiasing Treatment and Over-Approval of Low-Quality Men

Beyond simply closing the gender gap, the optimality of debiasing policies hinges on the nature of the bias: policies that induce lenders to raise screening standards for men differ from those that encourage more lending to women. As the gender gap in this context stems from excessive generosity toward men, an optimal debiasing policy should target the over-approval of low-quality male applicants. To assess whether the experimental interventions achieve this, I examine how approval rates for low-quality male and female applicants change under debiasing treatments relative to the control group. I evaluate this across three benchmarks of loan quality: ex-post repayment outcomes, past late-repayment history, and credit scores using Equation (4), where the outcome variable is a binary indicator for loan approval.

Table 6 reports the estimates of γ_k, which capture changes in approval rates relative to the control group for each treatment arm. As these loans are classified as low-quality—loans that ex-post defaulted, loans with late repayment histories, or loans with low credit scores—the coefficient γ_k should be negative if the treatment improves screening accuracy.

[22] Screening performance measured using the ML model is presented in Table A4, showing no statistically significant change across treatments. This may be because the ML-based repayment probability differs from the model used by loan officers, as certain information—such as credit scores—is missing.

Columns {1}, {3}, and {5}, display the changes in approval rates for low-quality male applicants, which must decline if an intervention successfully targets the over-approval of low-quality men. Conversely, Columns {2}, {4}, and {6} report the corresponding results for low-quality female applicants, where approval rates should not increase unless the interventions inadvertently encourage leniency toward low-quality women.

Across Columns {1}, {3}, and {5}, higher performance pay reduces approvals of low-quality male loans, particularly for ex-post defaulted loans in Column {1} and loans with low credit scores in Column {5}. However, IAT feedback fails to reduce approvals of low-quality male applicants. Moreover, adding IAT feedback to higher performance pay yields no additional benefit in the combined intervention and may even produce slight backfire effects.

For female applicants, Columns {2}, {4}, and {6} show a suggestive pattern of higher approval rates for low-quality loans across all treatment arms, although the estimates are mostly statistically insignificant.

Overall, these results indicate that higher performance pay is an effective debiasing policy when the gender gap is driven by excessive generosity toward men.[23] It not only narrows the gender gap but also reduces the over-approval of low-quality male applicants, thereby targeting the source of the bias. In contrast, diversity training via IAT feedback appears ineffective on all dimensions.

[23] Table A5 presents the analysis of screening performance using the ML model, where loans with predicted repayment probability below 0.5 are classified as low quality. Higher performance pay reduces approvals of low-quality male applicants by 3.7 p.p., while IAT feedback reduces them by 1.2 p.p., though neither effect is statistically significant.

5 Heterogeneous Effects

5.1 Gender Gap in Approval

Recall that loan officers are 4.4 p.p. (6.2%) more likely to approve the same portfolio when it is assigned a male name. In this section, I examine the heterogeneity of this gender bias by estimating Equation (2) and exploring variation separately by each of six loan officer characteristics: gender, IAT score, explicit gender bias, years of experience, education level, and risk-attitude.

Table A6 presents the results. In the top panel, Columns {1} and {2} show that both male and female loan officers exhibit a similar degree of gender bias in approval decisions, indicating that the officer's gender does not significantly affect the extent of bias. In Columns {3} and {4}, I divide officers based on their IAT scores. Officers classified as having implicit gender bias exhibit a 6.9% (4.8 p.p.) gender gap in loan approvals, compared to 4.2% (3.0 p.p.) among other officers. However, the difference between the two groups is not statistically significant. Columns (5) and (6) show no heterogeneity by explicit gender bias, measured using the statement, "On the whole, men make better business executives than women do." The gender gap in loan approvals during the experiment does not differ between officers who agree with the statement and those who disagree.

The bottom panel reports heterogeneity by loan officers' years of experience, education level, and risk attitude. Columns {1} and {2} show that officers with below-median experience exhibit a 7.3 p.p. gender gap, while those with above-median experience show a 1.8 p.p. gap. Although the difference is not statistically significant, the pattern suggests that bias may be stronger among less experienced officers, consistent with prior evidence (Alibhai et al., 2019; Brock and De Haas, 2023). Columns {3} and {4} show that officers with below-college education exhibit an 8.3 p.p. gender gap, whereas those with a college degree or higher show a 2.5 p.p. gap. While this difference is also not statistically significant, it suggests that bias may be more pronounced among officers with lower levels of education.

Lastly, Columns {5} and {6} show no heterogeneity by loan officers' risk tolerance.[24]

5.2 Debiasing Treatment Effects

Using the same officer characteristics described above, I examine heterogeneity in the effects of the three debiasing treatments on the gender gap in approvals by estimating Equation 3.

In the top panel of Table A7, heterogeneity by loan officer gender, IAT score, and explicit bias is presented. None of the differences are statistically significant. Although suggestive, the IAT feedback effect appears larger for female loan officers and smaller for officers with higher IAT scores who were informed that they associate men with career and women with family more than necessary. This pattern may indicate that officers with stronger implicit bias are already aware of their attitudes, so the IAT feedback provides limited additional information or behavioral adjustment. Columns {5} and {6} report heterogeneity by explicit bias, but this measure has limitations because it was elicited after treatment assignment, raising concerns about potential experimenter demand effects.

The bottom panel presents heterogeneity by loan officers' years of experience, education level, and risk attitude. Columns {1} to {4} show that the debiasing effect of higher performance pay is stronger among officers with less experience or lower education levels—groups that also exhibit larger baseline gender gaps. Columns {5} and {6} report heterogeneity by risk tolerance, and none of the differences are statistically significant.[25]

[24]The risk tolerance measure has limitations. It is based on the following question: "Imagine you have EGP 10,000 to invest. There is a 50% chance to double the amount of money you choose to 'invest'. It is equally possible that you could lose half the investment. How much of the EGP 10,000 do you invest?" Because the gamble has a positive expected return, higher investment reflects lower risk aversion. However, in the control group, 131 out of 180 loan officers choose to invest exactly half (EGP 5,000), resulting in limited variation.

[25]The risk tolerance measure has limitations. Of the 720 loan officers, 368 choose to invest exactly half (EGP 5,000) of the EGP 10,000 endowment, resulting in limited variation.

5.3 Treatment Effects on Low-Quality Male Approvals

Lastly, I examine heterogeneity in the effectiveness of the three debiasing treatments in reducing approvals of ex-post defaulted male applicants by estimating Equation 4.

Table A8 presents the results. There is no statistically significant heterogeneity in the approval of ex-post defaulted male applicants by loan officer gender, IAT score, explicit gender bias, experience, education level, or risk tolerance. However, the bottom panel of Columns (1) and (2) suggests that inexperienced officers are suggestively more likely to decrease approvals of ex-post defaulted male applicants compared to experienced officers.

6 The Source of Gender Bias

The observed gender gap in lending suggests the joint presence of loan officers' taste-based bias (Becker, 1957) and statistical discrimination (Phelps, 1972; Aigner and Cain, 1977; Bohren et al., 2023), such as overly optimistic priors about male borrowers.

Evidence consistent with statistical discrimination comes from heterogeneity across loan officers. The gender gap is concentrated among less experienced and less educated officers, who are more likely to hold overly optimistic priors about male borrowers.

At the same time, there is suggestive evidence that taste-based bias also plays a role. Doubling performance pay increases screening time by only 7.4 seconds (11.3%) per loan yet fully eliminates the gender gap. If the gap were driven purely by statistical discrimination, the small increase in screening induced by higher performance pay would be unlikely to generate sufficiently informative signals to close the gap. This pattern suggests that taste-based bias[26] operates alongside statistical discrimination.

[26]Moreover, in an endline survey question commonly interpreted as reflecting statistical discrimination—"On the whole, men make better business executives than women do"—the experimental gender gap in loan approval does not differ between officers who agree with the statement and those who do not. Among the 180 loan officers in the control group, 107 report that they disagree or strongly disagree with the statement, yet they approve male-named loans 4.2 p.p. more often. Similarly, among the 73 officers who agree or strongly agree with the statement, consistent with statistical discrimination, they approve male-named loans 4.5 p.p. more often. The results are presented in Columns (5) and (6) of Table A6. One caveat is that loan officers may misreport their views in response to this question.

7 Policy Implications and Conclusion

In summary, the findings highlight the existence of gender bias in small business lending, with the bias driven by excessive generosity toward men rather than excessive harshness toward women. The optimal debiasing policy therefore hinges on reducing the over-approval of low-quality loans driven by this generosity toward men. I show that increasing performance pay tied to repayment outcomes closes the gender gap by reducing approvals of low-quality male applicants. In contrast, diversity training via IAT feedback is ineffective when the gender gap arises from generosity toward men. These results suggest that identifying the source of the gender gap and designing debiasing policies accordingly has important implications for improving efficiency while reducing gender disparities.

According to the Central Bank of Egypt[27] as of end-June 2023, the total outstanding loan portfolio of the banking sector stood at EGP 4,798.9 billion, with a non-performing loan (NPL) ratio of 3.3%.[28] A back-of-the-envelope calculation[29] suggests that 6.9% of NPLs are attributable to generosity toward men. The resulting financial loss is approximately EGP 3.6 billion (USD 72.9 million) per year. In other words, closing the gender bias could potentially increase bank profits by up to this amount—resources that could be used to expand credit access. On the other hand, the estimated increase in labor costs from higher performance pay is about EGP 429.1 million annually.[30]

[27] Central Bank of Egypt, *Annual Report 2022/2023*, https://www.cbe.org.eg/en/economic-research/economic-reports/annual-report

[28] The current NPL rate is relatively low compared to historical peaks—10.9% in 2011 and 26.5% in 2005 (Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org/series/DDSI02EGA156NWDB)

[29] Out of ex-post defaulted loans, 66.1% of female-named loans were approved compared to 75.8% of male-named loans. Without generosity toward men, male-named loans would also have been approved at 66.1%, the same rate as female-named loans. In that case, ex-post defaulted loans overall would have been approved at 66.1%, rather than the observed 71.0% during the experiment. This implies that generosity toward men accounts for $\frac{71.0-66.1}{71.0} = 6.9\%$. Applying this margin to the stock of loans (EGP 4,798.9 billion $\times$ 0.033 $\times$ 0.069) yields an estimated EGP 10.9 billion. Since mortgage loans—typically with longer maturities—constitute only 9.8% of retail loans (Central Bank of Egypt, Financial Stability Report, March 2024, https://www.cbe.org.eg/en/financial-stability/macroprudential/financial-stability-reports), assuming an average loan tenure of three years implies an annualized loss of roughly EGP 3.6 billion (USD 72.9 million) with an exchange rate of 1 USD = 50 EGP.

[30] The higher performance pay raised participation rewards by 2.1% in the experiment. The average annual income of workers in the financial intermediation sector is estimated at EGP 149,928 (https://censusinfo.capmas.gov.eg/metadata-ar-v4.2/index.php/catalog/1920/download/6926, Central Agency for Public

Taken together, these findings imply that increasing performance pay could raise annual net profits by roughly EGP 3.2 billion (USD 64.3 million). The negative consequences of generosity toward men may be particularly pronounced during banking crises, when higher default rates amplify systemic risk and exacerbate economic shocks, especially in low-income countries. Beyond reducing immediate bank losses, the benefits of addressing generosity bias are likely to extend further, as doing so can mitigate capital misallocation and enhance overall productivity. Hence, the total gains from reducing generosity toward men in lending likely exceed the estimated EGP 3.2 billion (USD 64.3 million). Expanding performance pay tied to repayment outcomes (e.g., monthly) is a scalable solution that can be implemented in lending institutions worldwide.

This experiment, however, has several limitations. First, although the officers' decisions were incentivized, they did not have real-world consequences for actual clients in the hypothetical setting, limiting how closely the observed behavior reflects true decision-making. Second, the experiment is conducted in Egypt, where social norms are strong, and draws only on previously approved loans with intermediate or missing credit scores and a relatively small number of loan applications, which may limit the external validity of the findings.

Despite these limitations, the findings have important implications for MFIs and governments. To curb the over-approval of low-quality male applicants, expanding performance-based incentives needs to be considered, particularly for less experienced or educated loan officers, who are more responsive to the higher performance pay. On the other hand, diversity training alone is unlikely to be effective, especially when reducing approvals of male applicants is required rather than increasing approvals of female applicants.

This study also opens several avenues for future research. First, beyond the supply-side aspects of the gender gap in lending, demand-side factors must also be considered, such as

Mobilization and Statistics of Egypt, *Annual Bulletin of Employment, Wages and Working Hours Statistics*, 2023), and the total number of bank employees in 2022 was 136,273 (https://www.cbe.org.eg/-/media/project/cbe/listing/monthly-statistical-bulletin/bulletin/feb/monthly-statistical-bulletin-311.pdf, Central Bank of Egypt, *Monthly Statistical Bulletin* (February 2023)). The implied increase in labor costs is $136{,}273 \times 149{,}928 \times 0.021 \approx$ EGP 429.1 million.

difficulties in the loan application process or behavioral patterns that may contribute to the observed gap. Second, a follow-up study could use ex-ante clearly high- and low-quality loan applications—including previously rejected ones—identified by a machine learning model. It would also be valuable to examine whether loan officers with higher performance pay continue to exhibit debiased behavior over time and what happens when their performance pay returns to the original level.

References

Agan, A. and S. Starr (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics 133*(1), 191–235.

Agarwal, S. and I. Ben-David (2018). Loan prospecting and the loss of soft information. *Journal of Financial Economics 129*(3), 608–628.

Agarwal, S., P. Ghosh, A. Mukherjee, and S. L. Naaraayanan (2023). Rural roads, small business creation, and the gender-gap in entrepreneurship. Technical report, Working Paper.

Agarwal, S., B. C. Morais, A. Seru, and K. Shue (2024). Noisy experts? discretion in regulation. Technical report, National Bureau of Economic Research.

Aigner, D. J. and G. G. Cain (1977). Statistical theories of discrimination in labor markets. *Ilr Review 30*(2), 175–187.

Alibhai, S., A. A. Donald, M. Goldstein, A. A. Oguz, A. Pankov, and F. Strobbe (2019). Gender bias in sme lending: Experimental evidence from turkey. *World bank policy research working paper* (9100).

Altonji, J. G. and R. M. Blank (1999). Race and gender in the labor market. *Handbook of labor economics 3*, 3143–3259.

Arnold, D., W. Dobbie, and P. Hull (2020). Measuring racial discrimination in bail decisions. Technical report, National Bureau of Economic Research.

Arnold, D., W. Dobbie, and C. S. Yang (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics 133*(4), 1885–1932.

Ayalew, S., S. Manian, and K. Sheth (2023). Discrimination and access to capital: Experimental evidence from ethiopia.

Baert, S. (2018). *Hiring discrimination: An overview of (almost) all correspondence experiments since 2005*. Springer.

Bao, H. X. (2024). Is hiding my first name enough? using behavioural interventions to mitigate racial and gender discrimination in the rental housing market. *Urban Studies 61*(11), 2156–2175.

Bartŏs, V., S. Castro, K. Czura, and T. Opitz (2024). Gendered access to finance: The roles of team formation, idea quality, and implementation constraints in business evaluations. Technical report, CESifo Working Paper.

Beck, T., P. Behr, and A. Madestam (2018). Sex and credit: Is there a gender bias in lending? *Journal of Banking and Finance 87*.

Becker, G. S. (1957). The economics of discrimination (chicago: University of chicago).

Berkovec, J. A., G. B. Canner, S. A. Gabriel, and T. H. Hannan (1998). Discrimination, competition, and loan performance in fha mortgage lending. *Review of Economics and Statistics 80*(2), 241–250.

Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review 94*(4), 991–1013.

Blanchflower, D. G., P. B. Levine, and D. J. Zimmerman (2003). Discrimination in the small-business credit market. *Review of Economics and Statistics 85*(4), 930–943.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of economic literature 55*(3), 789–865.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2023). Inaccurate statistical discrimination: An identification problem. *Review of Economics and Statistics*, 1–45.

Brock, J. M. and R. De Haas (2023). Discriminatory lending: Evidence from bankers in the lab. *American Economic Journal: Applied Economics 15*(2), 31–68.

Bryan, G., D. Karlan, and A. Osman (2024). Big loans to small businesses: Predicting winners and losers in an entrepreneurial lending experiment. *American Economic Review 114*(9), 2825–2860.

Butler, A. W., E. J. Mayer, and J. P. Weston (2023). Racial disparities in the auto loan market. *The Review of Financial Studies 36*(1), 1–41.

Carter, S., E. Shaw, W. Lam, and F. Wilson (2007). Gender, entrepreneurship, and bank lending: The criteria and processes used by bank loan officers in assessing applications. *Entrepreneurship theory and practice 31*(3), 427–444.

Cavalluzzo, K. S. and L. C. Cavalluzzo (1998). Market structure and discrimination: The case of small businesses. *Journal of Money, Credit and Banking*, 771–792.

Cheraghali, H. and P. Molnár (2024). Sme default prediction: A systematic methodology-focused review. *Journal of Small Business Management 62*(6), 2847–2905.

Cole, S., M. Kanz, and L. Klapper (2015). Incentivizing calculated risk-taking: Evidence from an experiment with commercial bank loan officers. *The Journal of Finance 70*(2), 537–575.

Costello, A. M., A. K. Down, and M. N. Mehta (2020). Machine+ man: A field experiment on the role of discretion in augmenting ai-based lending models. *Journal of Accounting and Economics 70*(2-3), 101360.

Crépon, B., M. El Komi, and A. Osman (2024). Is it who you are or what you get? comparing the impacts of loans and grants for microenterprise development. *American Economic Journal: Applied Economics 16*(1), 286–313.

De Mel, S., D. McKenzie, and C. Woodruff (2009). Are women more credit constrained? experimental evidence on gender and microenterprise returns. *American Economic Journal: Applied Economics 1*(3), 1–32.

Demirguc-Kunt, A., L. Klapper, D. Singer, S. Ansar, and J. Hess (2018). *The Global Findex Database 2017: Measuring financial inclusion and the fintech revolution*. World Bank Publications.

Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2021). Measuring bias in consumer lending. *The Review of Economic Studies 88*(6), 2799–2832.

Egan, M., G. Matvos, and A. Seru (2022). When harry fired sally: The double standard in punishing misconduct. *Journal of Political Economy 130*(5), 1184–1248.

FDIC (2024). Consumer compliance examination manual. Accessed June 4, 2025.

Fisman, R., D. Paravisini, and V. Vig (2017). Cultural proximity and loan outcomes. *American Economic Review 107*(2), 457–492.

Flage, A. (2018). Ethnic and gender discrimination in the rental housing market: Evidence from a meta-analysis of correspondence tests, 2006–2017. *Journal of Housing Economics 41*, 251–273.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2022). Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance 77*(1), 5–47.

Glick, P. and S. T. Fiske (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology 70*(3), 491–512.

Glover, D., A. Pallais, and W. Pariente (2017). Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *The Quarterly Journal of Economics 132*(3), 1219–1260.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American economic review 104*(4), 1091–1119.

Greenwald, A. G., D. E. McGhee, and J. L. Schwartz (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology 74*(6), 1464.

Greenwald, A. G., B. A. Nosek, and M. R. Banaji (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology 85*(2), 197.

Haerpfer, C., R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, B. Puranen, et al. (2022). World values survey: Round seven-country-pooled datafile version 5.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat 12*(10), 8.

Hanson, A. and Z. Hawley (2011). Do landlords discriminate in the rental housing market? evidence from an internet field experiment in us cities. *Journal of urban Economics 70*(2-3), 99–114.

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

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic literature 42*(4), 1009–1055.

Haselmann, R., D. Schoenherr, and V. Vig (2018). Rent seeking in elite networks. *Journal of Political Economy 126*(4), 1638–1690.

Hays, R., T. Hayashi, and A. Stewart (1989a). Socially desirable response set five-item survey (sdrs-5). *Educ Psychol Measur 49*, 629–39.

Hays, R. D., T. Hayashi, and A. L. Stewart (1989b). A five-item measure of socially desirable response set. *Educational and psychological measurement 49*(3), 629–636.

Hebert, C. (2023). Gender stereotypes and entrepreneur financing. In *10th Miami behavioral finance conference*.

Hebert, C., H. Tookes, and E. Yimfor (2024). Financing the next vc-backed startup: The role of gender. *Available at SSRN*.

Hertzberg, A., J. M. Liberti, and D. Paravisini (2010). Information and incentives inside the firm: Evidence from loan officer rotation. *The Journal of Finance 65*(3), 795–828.

Heywood, J. S. and P. L. O'Halloran (2005). Racial earnings differentials and performance pay. *Journal of Human Resources 40*(2), 435–452.

Heywood, J. S. and D. Parent (2012). Performance pay and the white-black wage gap. *Journal of Labor Economics 30*(2), 249–290.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019). The allocation of talent and us economic growth. *Econometrica 87*(5), 1439–1474.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics 124*(4), 1403–1448.

Jung, Y. (2025). Addressing gender bias in small business lending. AEA RCT Registry.

Jung, Y. (2026). Human-in-the-loop gender bias in ai–assisted lending. Working paper.

Lane, K. A., M. R. Banaji, B. A. Nosek, and A. G. Greenwald (2007). Understanding and using the implicit association test: Iv: What we know (so far) about the method.

Marinescu, I., N. Klein, A. Chamberlain, and M. Smart (2018). Incentives can reduce bias in online reviews. Technical report, National Bureau of Economic Research.

Mascia, D. V. and S. P. Rossi (2017). Is there a gender effect on the cost of bank financing? *Journal of Financial Stability 31*, 136–153.

Mejia, J. and C. Parker (2021). When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Science 67*(1), 166–184.

Mengel, F., J. Sauermann, and U. Zölitz (2019). Gender bias in teaching evaluations. *Journal of the European economic association 17*(2), 535–566.

Mocanu, T. (2022). Designing gender equity: Evidence from hiring practices and committees. Technical report, Working paper.

Moll, B. (2014). Productivity losses from financial frictions: Can self-financing undo capital misallocation? *American Economic Review 104*(10), 3186–3221.

Montoya, A. M., E. Parrado, A. Solis, and R. Undurraga (2024). Bad taste: Gender discrimination in consumer lending.

Muravyev, A., O. Talavera, and D. Schäfer (2009). Entrepreneurs' gender and financial constraints: Evidence from international data. *Journal of comparative economics 37*(2), 270–286.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature 56*(3), 799–866.

Osman, A., J. Speer, and A. Weaver (2023). Discrimination against women in hiring.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review 62*(4), 659–661.

Rigol, N. and B. N. Roth (2021). Loan officers impede graduation from microfinance: Strategic disclosure in a large microfinance institution. Technical report, National Bureau of Economic Research.

Ross, S. L., M. A. Turner, E. Godfrey, and R. R. Smith (2008). Mortgage lending in chicago and los angeles: A paired testing study of the pre-application process. *Journal of Urban Economics 63*(3), 902–919.

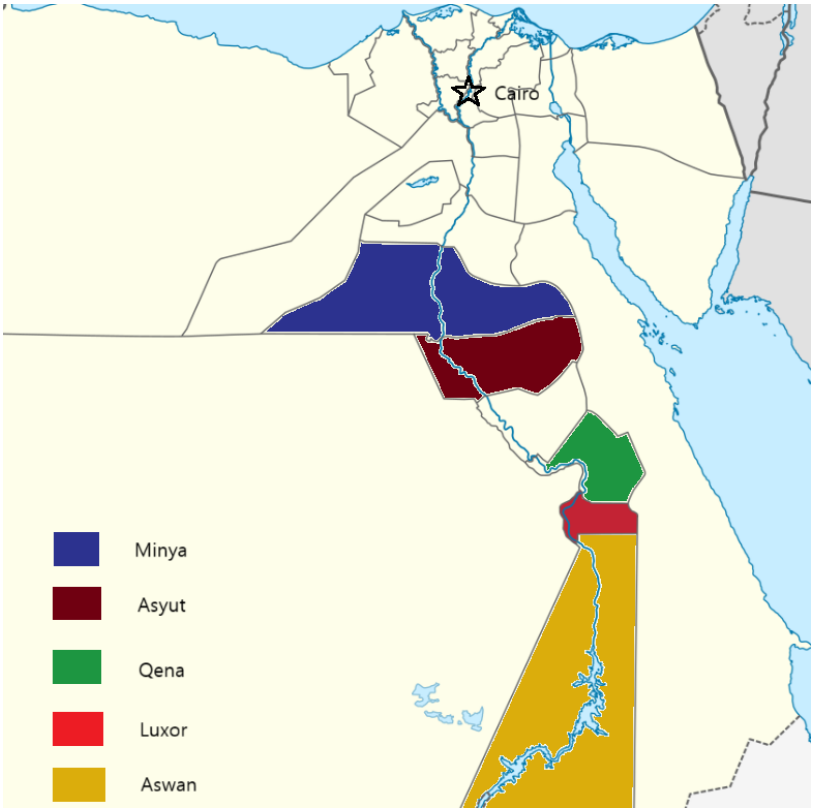
Sarsons, H. (2022). Interpreting signals in the labor market: evidence from medical referrals. *Job market paper*, 141–145.

Stefani, M. L. and V. P. Vacca (2013). Credit access for female firms: Evidence from a survey on european smes. *Bank of Italy Occasional Paper* (176).

Ubfal, D. (2024). What works in supporting women-led businesses? *IZA Discussion Papers*.

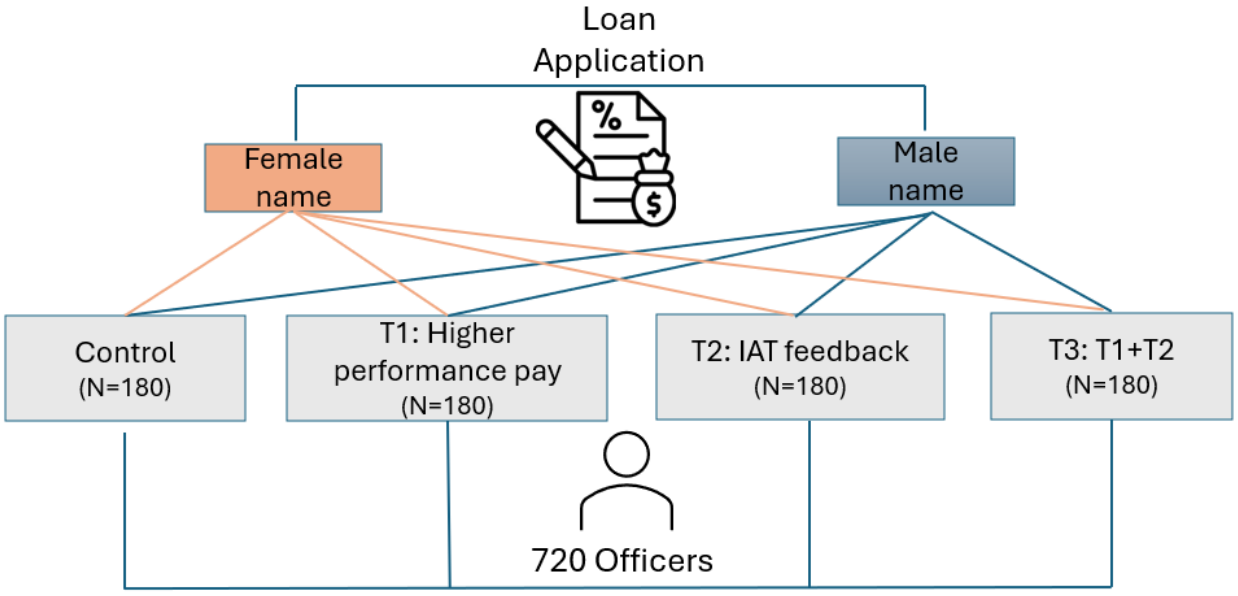
Figures

Figure 1. Geographical Distribution of Loan Officers by Governorates



Notes: This figure shows geographical distribution of the five governorates where the experiment takes place—Aswan, Qena, Luxor, Asyut, and Minya. Egypt's capital, Cairo, is marked with a star. Source: Wikipedia: Luxor Governorate; map recolored by author.

Figure 2. Randomization Design



Notes: The randomization is conducted at the individual level stratified by loan officers' gender. T3 refers to the combined treatment group, where officers receive both higher performance pay (T1) and IAT feedback (T2) together.

Figure 3. Credit Score and Approval Rate by Gender of Applicants



Notes: This figure shows the linear relationship between credit score and approval rate by gender of applicants. It uses 31 out of 50 loans with available credit scores ranging from 685 to 750. Of the remaining 19 loans, 16 are missing credit scores, and 3 have low scores of 536, 537, and 622, respectively. These three loans are excluded from the figure because their credit scores are outliers compared to the typical range of 685 to 750, which could distort the overall trend. A credit score between 626 and 700 is considered satisfactory, while scores between 701 and 750 are classified as very good. The shaded area around the fitted line represents the 90% confidence interval. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Tables

Table 1: Baseline Balance Check

	(1) All	(2) Control	(3) Higher Performance Pay (Treatment1)	(4) IAT Feedback (Treatment2)	(5) Combined Intervention (Treatment3)	(6) Joint F-test p-value
Female (=1)	0.50	0.50	0.50	0.50	0.50	1.00
	{0.50}	{0.50}	{0.50}	{0.50}	{0.50}	
Age	34.62	34.25	35.28	35.02	33.94	0.34
	{8.22}	{7.83}	{8.79}	{8.57}	{7.62}	
Married (=1)	0.66	0.70	0.66	0.66	0.63	0.52
	{0.47}	{0.46}	{0.47}	{0.47}	{0.48}	
Year of Education	16.20	16.34	16.01	16.14	16.30	0.31
	{1.83}	{1.75}	{1.90}	{1.88}	{1.77}	
Total Monthly Income (EGP)	6048	6200	6037	5962	5995	0.93
	{3843}	{3827}	{3936}	{3584}	{4038}	
Senior or above Rank LO Binary	0.29	0.29	0.28	0.34	0.26	0.41
	{0.46}	{0.46}	{0.45}	{0.47}	{0.44}	
Number of Experience (Year)	6.28	5.82	6.69	6.44	6.15	0.47
	{5.72}	{5.45}	{6.20}	{5.90}	{5.28}	
Household Size	3.70	3.62	3.68	3.70	3.80	0.80
	{1.68}	{1.68}	{1.67}	{1.67}	{1.72}	
Monthy Approval Rate	67.33	67.09	66.78	67.70	67.74	0.94
	{16.90}	{16.05}	{16.79}	{17.23}	{17.62}	
Loan Screen Time (hour)	10.51	11.05	9.56	12.03	9.41	0.31
	{15.21}	{15.34}	{14.30}	{16.68}	{14.35}	
LO Skill Binary (1=Soft, 0=Hard)	0.69	0.70	0.73	0.70	0.65	0.46
	{0.46}	{0.46}	{0.45}	{0.46}	{0.48}	
Weekly Working Hours	40.93	42.57	39.81	41.76	39.60	0.02
	{11.12}	{11.87}	{10.97}	{11.10}	{10.28}	
Accuracy rate in Cognitive Test	0.30	0.29	0.30	0.29	0.31	0.59
	{0.23}	{0.24}	{0.23}	{0.23}	{0.23}	
Digit Span Recall	3.98	4.00	4.07	3.93	3.92	0.59
	{1.24}	{1.27}	{1.18}	{1.41}	{1.09}	
Psychometric Test	3.65	3.63	3.66	3.64	3.65	0.89
	{0.45}	{0.43}	{0.45}	{0.42}	{0.49}	
Social Desirabilty Bias (Out of 100)	23.22	24.33	24.78	21.89	21.89	0.45
	{22.76}	{23.47}	{22.90}	{22.04}	{22.64}	
IAT Experience Before (=1)	0.14	0.16	0.11	0.16	0.14	0.37
	{0.35}	{0.36}	{0.31}	{0.37}	{0.35}	
IAT Score	0.37	0.40	0.41	0.32	0.33	0.26
	{0.53}	{0.49}	{0.55}	{0.54}	{0.54}	
Explicit Gender Bias (Business)	2.82	2.80	2.93	2.72	2.82	0.47
	{1.26}	{1.24}	{1.26}	{1.28}	{1.27}	
Explicit Gender Bias (All)	2.98	2.98	3.05	2.90	2.98	0.30
	{0.73}	{0.70}	{0.72}	{0.71}	{0.78}	
Risk Preference (Out of 10)	6.44	6.49	6.73	6.11	6.42	0.25
	{2.95}	{2.97}	{2.88}	{3.08}	{2.86}	
Investment Choice on Fair 50/50 Chance (Out of 100)	52.30	51.53	54.15	50.92	52.61	0.61
	{24.21}	{24.75}	{25.61}	{22.61}	{23.86}	
Gender Education Experience Binary	0.44	0.41	0.46	0.46	0.43	0.78
	{0.50}	{0.49}	{0.50}	{0.50}	{0.50}	
Existence of Second Job Binary	0.20	0.19	0.22	0.17	0.24	0.34
	{0.40}	{0.39}	{0.41}	{0.37}	{0.43}	
Joint F-Test p-value						0.404
Number of Observations	720	180	180	180	180	720

Notes: This table shows balance check of baseline variables by treatment arms. Means are presented with standard deviations shown in brackets below. Due to a technical error, IAT scores were not collected for 11 loan officers—specifically, those in the control group and the higher performance pay group (T1), both of which did not receive IAT feedback as part of their treatment. For these officers, missing IAT scores were imputed using the median value. The monetary variables are winsorized by top and bottom 1%. Appendix C presents the details of the baseline characteristic variables.

Table 2: Logit Regression Results (Ex-post Defaulted vs. Repaid Loans)

	Ex-post Defaulted Loans
log (Loan Size)	3.31*
	(1.88)
Loan Tenure (Months)	0.31***
	(0.07)
log (Monthly Income)	0.71
	(3.85)
Late Repayment History Binary	1.71*
	(0.99)
Number of Previous Loans	0.14
	(0.26)
Number of Employees	-2.98**
	(1.49)
Registered Business Binary	-2.52
	(2.66)
Age	-0.05
	(0.07)
Number of Employee Guarantors	-3.09
	(2.26)
Joint F-test Pvalue	0.001
Number of Loans	50

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table presents the coefficients from a logistic regression that compares ex-post defaulted loans to ex-post repaid loans. Standard errors are reported in parentheses. Among the 50 loans analyzed, 30 were repaid and 20 defaulted. The regression includes 9 continuous loan characteristics, excluding categorical variables such as industry and region.

Table 3: Approval Rate Gap by Gender of Loan Applications

Outcome: Approval Binary	Total	Ex-post Outcomes		Late Repayment History		Credit Score		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Defaulted (Generous to Men)	Repaid (Harsh to Women)	Yes (Generous to Men)	No (Harsh to Women)	700 or below (Generous to Men)	701 or Above (Harsh to Women)	Not Available
Male Names	0.044**	0.097***	0.009	0.111**	0.026	0.097***	0.006	0.035
	(0.019)	(0.032)	(0.022)	(0.047)	(0.021)	(0.037)	(0.031)	(0.034)
Female Name Approval Mean	0.708	0.661	0.739	0.515	0.762	0.576	0.778	0.760
T-test P-value (Generosity vs. Harshness)			0.022		0.094		0.058	.
Number of Portfolios	1800	720	1080	396	1404	576	648	576
Number of Loan Officers	180	180	180	180	180	180	180	180
Portfolio FE	✓	✓	✓	✓	✓	✓	✓	✓
Loan Officer Gender FE	✓	✓	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports the gender gap in loan approvals by comparing approval rates for identical applications that differ only in having a male versus a female applicant name. Ex-post defaulted loans are those with repayment delays of 90 days or more. Late-repayment history is defined as having at least one day of delayed repayment in the most recent loans. A credit score of 701 or above is considered very good, while a score of 700 or below is classified as satisfactory. Loan officers received explanations about these credit-score ranges prior to the task. The outcome is a binary variable equal to 1 if the loan officer approves a loan, and 0 otherwise. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table 4. Debiasing Treatment Effects in Loan Approval

Outcome: Approval Binary	(1) All Officers	(2) Officers Biased Against Women (IAT Score $\geq$0.35)	(3) Others (IAT Score $<$0.35)
Higher Performance Pay (T1) x Male Names	-0.066**	-0.080**	-0.019
	(0.027)	(0.035)	(0.043)
IAT Feedback (T2) x Male Names	-0.039	-0.026	-0.040
	(0.026)	(0.039)	(0.037)
Combined (T3) x Male Names	-0.048*	-0.064*	-0.017
	(0.027)	(0.037)	(0.040)
Male Names	0.044**	0.047*	0.028
	(0.019)	(0.026)	(0.030)
T-test p-value (T2 vs. T1)	0.478	0.306	0.707
T-test p-value (T3 vs. T1)	0.640	0.764	0.970
Number of Portfolios	7200	3790	3300
Number of Loan Officers	720	379	330
Portfolio FE	✓	✓	✓
Loan Officer Gender FE	✓	✓	✓
Cohort FE	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table presents the change in the gender gap in loan approval rates for each treatment arm relative to the control group. The combined group refers to the treatment group 3, where officers receive both the higher performance pay and IAT feedback together. Column (1) reports results for all loan officers. Column (2) shows results for officers with a Gender–Career IAT score of 0.35 or higher, indicating a medium or stronger association between men and careers and women and family. Column (3) reports results for officers with an IAT score below 0.35, indicating a weaker, no, or opposite association. Due to a technical issue, IAT scores were not recorded for 11 officers in the control and higher-performance-pay groups; these observations are excluded from Columns (2) and (3). Standard errors, clustered at the loan officer level, are reported in parentheses.

Table 5. Debiasing Treatment Effects in Loan Screening Performance and Time

Outcome: Accuracy Binary & Screening Time	(1) Accuracy (Male-name Loans)	(2) Accuracy (Female-name Loans)	(3) Average Screening Time per Loan (minutes)
Higher Performance Pay (T1)	0.036*	-0.021	0.124**
	(0.020)	(0.019)	(0.052)
IAT Feedback (T2)	0.021	-0.004	-0.033
	(0.019)	(0.018)	(0.045)
Combined (T3)	0.042**	-0.001	0.006
	(0.019)	(0.018)	(0.044)
Control Group Mean	0.546	0.579	1.098
T-test p-value (T2 vs. T1)	0.601	0.523	0.023
T-test p-value (T3 vs. T1)	0.809	0.445	0.083
Number of Portfolios	3600	3600	.
Number of Loan Officers	720	720	704
Portfolio FE	✓	✓	
Loan Officer Gender FE	✓	✓	✓
Cohort FE	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports changes in loan screening performance (accuracy) and screening time for each treatment arm relative to the control group. The combined group refers to the treatment group 3, where officers receive both the higher performance pay and IAT feedback together. Screening performance is a binary variable equal to 1 if an ex-post repaid loan is approved or an ex-post defaulted loan is rejected. It is equal to 0 if an ex-post defaulted loan is approved or an ex-post repaid loan is rejected. Screening time is measured as the total duration spent reviewing all 10 loan applications; therefore, it is not possible to calculate screening time separately for male and female applicants. For the first 16 loan officers in the experiment, screening time was not recorded due to a technical error, and their observations are excluded from the Column (3) analysis. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table 6. Debiasing Treatment Effects in Approval of Low-quality Loans by Applicant Gender

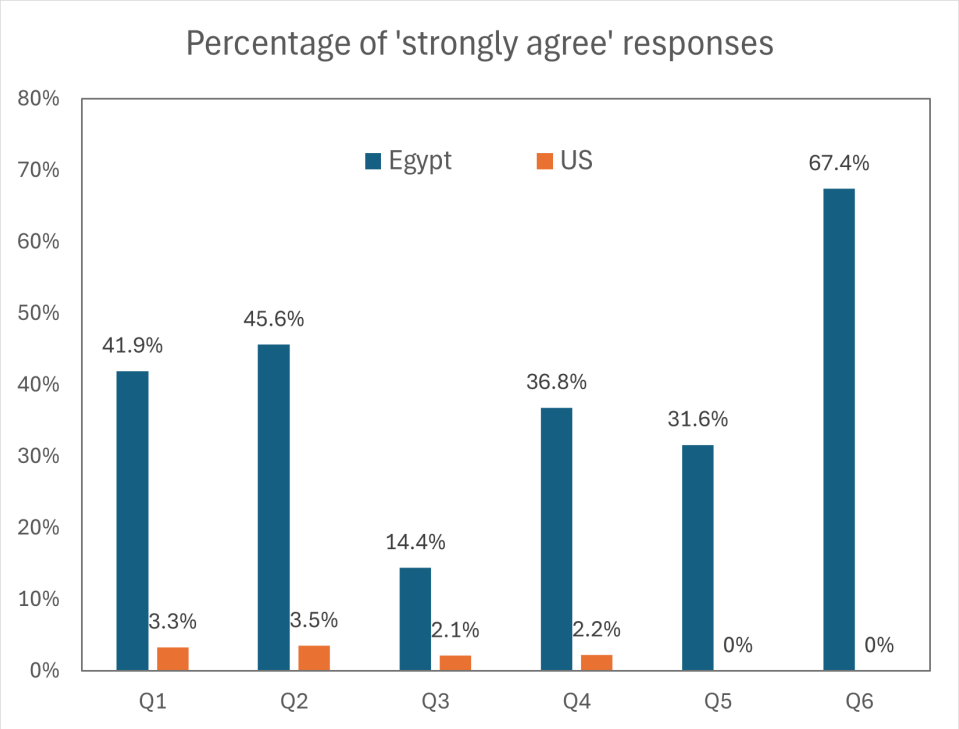
Outcome: Approval Binary	Ex-post Defaulted		With Late Repayment Histories		Low Credit Score	
	(1) Male	(2) Female	(3) Male	(4) Female	(5) Male	(6) Female
Higher Performance Pay (T1)	-0.097***	0.056*	-0.066	0.071	-0.090**	0.047
	(0.032)	(0.032)	(0.047)	(0.047)	(0.043)	(0.045)
IAT Feedback (T2)	-0.031	0.050	-0.061	0.056	0.004	0.060
	(0.031)	(0.032)	(0.047)	(0.045)	(0.041)	(0.042)
Combined (T3)	-0.061*	0.053	-0.051	0.081*	-0.043	0.090**
	(0.034)	(0.032)	(0.048)	(0.044)	(0.042)	(0.043)
Control Group Approval Mean	0.758	0.661	0.626	0.515	0.748	0.624
Number of Portfolios	1440	1440	792	792	936	936
Number of Loan Officers	720	720	720	720	720	720

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports changes in the approval of low-quality loans by applicant gender across three benchmarks of loan quality. The combined group refers to the treatment group 3, where officers receive both the higher performance pay and IAT feedback together. Loans with late repayment histories are defined as those with at least one day of late repayment on the most recent loan. Low credit score loans are defined as those with a credit score of 700 or below, excluding loans without a recorded credit score. Standard errors, clustered at the loan officer level, are reported in parentheses.

8 Online Appendix Figures and Tables

Figure A1. Response to 6 gender bias questions in the 2017 World Value Survey



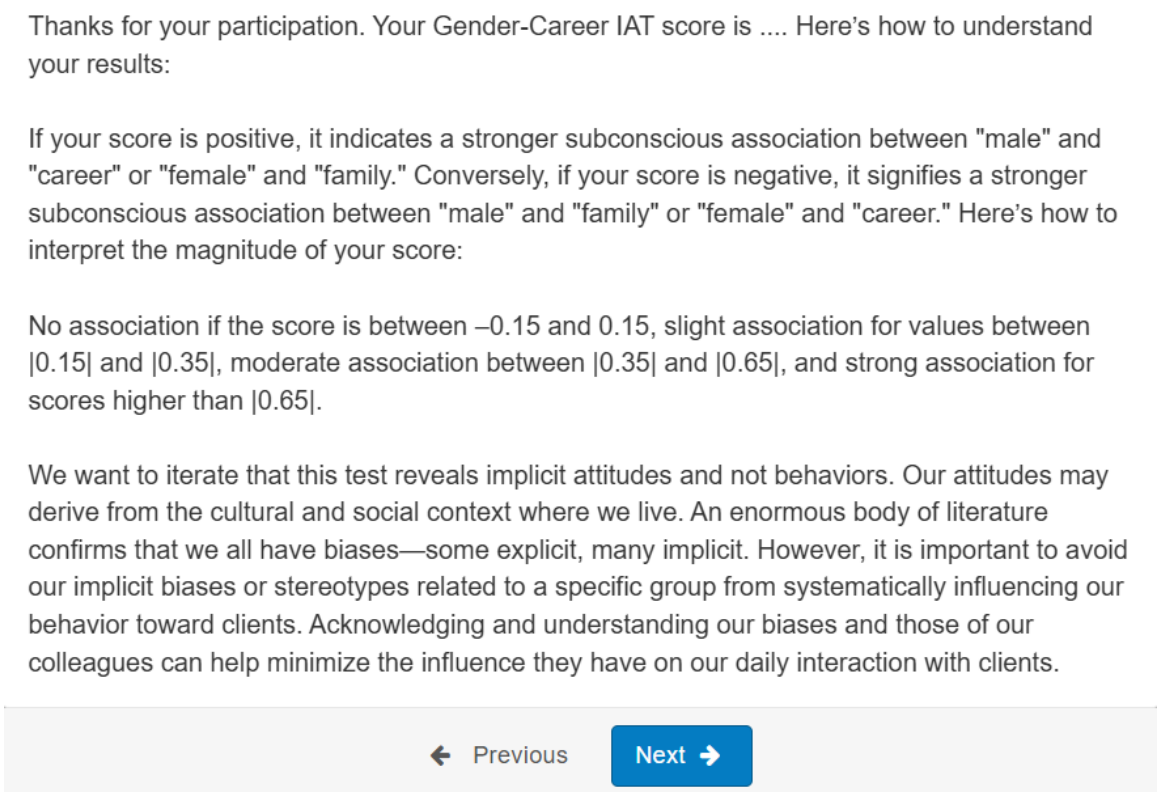
Notes: This figure compares the percentage of "Strongly Agree" responses between the United States and Egypt in the 2017 World Value Survey. The questions are: (1) When a mother works for pay, the children suffer. (2) On the whole, men make better political leaders than women do. (3) A university education is more important for a boy than for a girl. (4) On the whole, men make better business executives than women do. (5) If a woman earns more money than her husband, it's almost certain to cause problems. (6) When jobs are scarce, men should have more right to a job than women.

Figure A2. Illustrative Example of a Loan Application in the Experiment

Profile	Value	Profile	Value	Profile	Value
Client Name	Ahmed Hassan	**Age**	38	**Number of Employee Guarantor**	1
Number of Non-Employee Guarantor	1	**Monthly Income**	15,000	**Registered Business**	1
Business Start Year	2012	**Location**	Edfu, Aswan	**Number of Employees**	1
Industry	Clothing	**Loan Size**	100,000	**Loan Term (Months)**	29
Interest Rate	22%	**Number of Previous Loans**	5	**Days Late on Recent Loan**	51
Loan Size on Most Recent Loan	100,000	**Interest Rate on Recent Loan**	22%	**Credit Score**	704

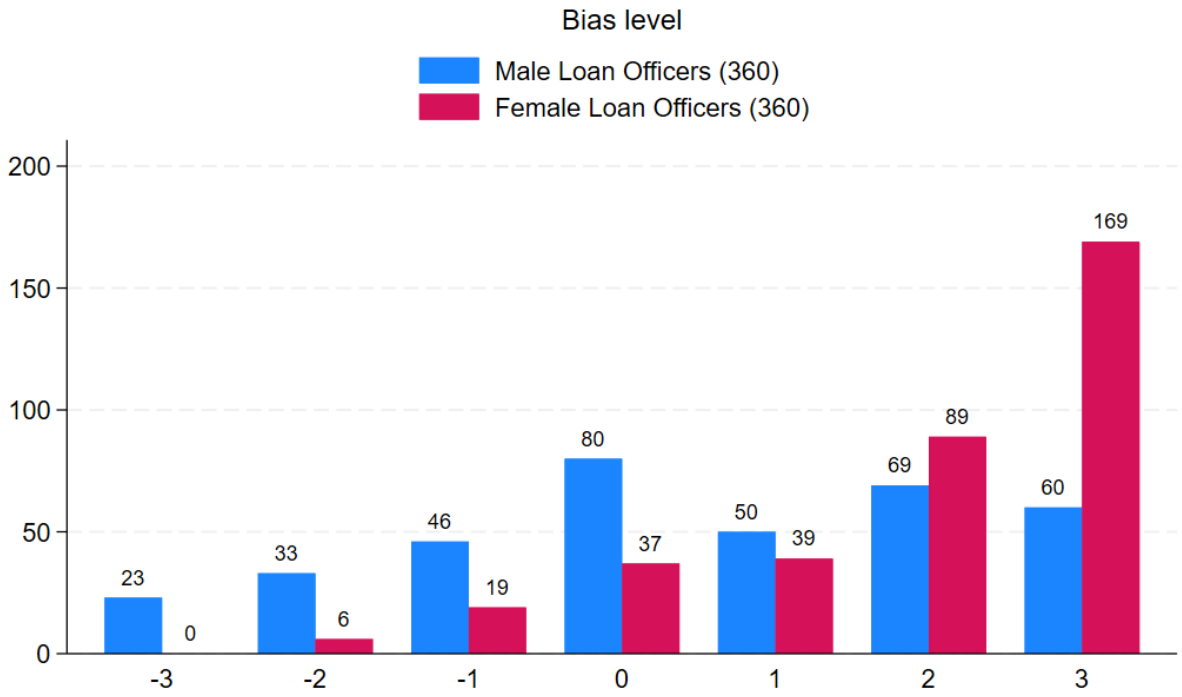
Notes: This figure presents a pseudo-example of one portfolio used in the experiment. All information has been replaced with fictitious data and does not contain any personal details.

Figure A3. IAT Feedback Treatment



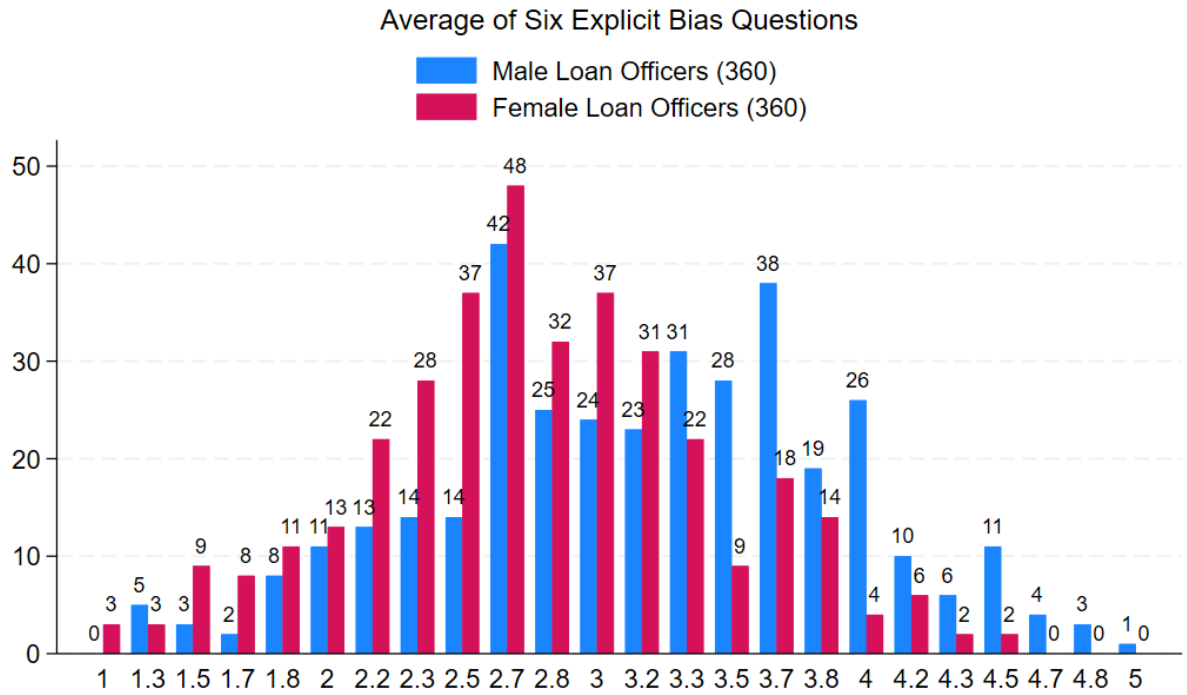
Notes: This figure shows an example of the IAT feedback provided to the IAT Feedback group (T2) and the Combined group (T3). The format and feedback are adapted from the widely used, publicly available version at `https://implicit.harvard.edu/implicit/`. In contrast, the Control group (C) and the Higher Performance Pay group (T1) received only a simple message after completing the IAT test: Thanks for your participation.

Figure A4. The range of Gender-Career IAT score of loan officers



Notes: This figure shows the Gender-Career IAT score of 360 male loan officers and 360 female loan officers. Gender-Career IAT score labeling: -3: Strong bias against men (IAT <= -0.65), -2: Moderate bias against men (-0.65 < IAT <= -0.35), -1: Weak bias against men (-0.35 < IAT <= -0.15), 0: No gender bias (-0.15 < IAT < 0.15), 1: Weak bias against women (0.15 <= IAT < 0.35), 2: Moderate bias against women (0.35 <= IAT < 0.65), 3: Strong bias against women (IAT > 0.65).

Figure A5. The average responses to explicit gender bias questions



Notes: This figure compares the average responses on a 5-point Likert scale for six explicit gender bias questions derived from the 2017 World Values Survey. The Likert scale ranges from 1 (strongly disagree) to 5 (strongly agree), with 3 representing a neutral response. The questions are: (1) When a mother works for pay, the children suffer. (2) On the whole, men make better political leaders than women do. (3) A university education is more important for a boy than for a girl. (4) On the whole, men make better business executives than women do. (5) If a woman earns more money than her husband, it's almost certain to cause problems. (6) When jobs are scarce, men should have more right to a job than women.

Table A1. Summary of Participating Microfinance Institutions (MFIs)

MFI Number	Governorate	Number of Loan Officers		
		Total	Female	Male
1	Luxor and Aswan	185	88	97
2	Luxor and Qena	152	79	73
3	Asyut	100	68	32
4	Minya	51	1	50
5	Aswan	24	21	3
6	Minya	21	5	16
7	Minya	19	14	5
8	Minya	17	4	13
9	Minya	16	5	11
10	Minya	13	8	5
11	Minya	11	8	3
12	Qena	11	5	6
13	Minya	10	3	7
14	Minya	9	5	4
15	Minya	9	5	4
16	Minya	9	5	4
17	Minya	9	0	9
18	Minya	8	3	5
19	Minya	6	6	0
20	Minya	6	0	6
21	Aswan	5	5	0
22	Minya	5	2	3
23	Minya	5	5	0
24	Minya	5	5	0
25	Aswan	4	4	0
26	Minya	4	3	1
27	Minya	3	2	1
28	Minya	2	1	1
29	Minya	1	0	1
	Total	**720**	**360**	**360**

Notes: This table shows the number of loan officers participated in the experiment by MFIs with governorate and gender of loan officers. The specific names of MFIs are not disclosed due to the some MFIs did not want to disclose their names. The MFIs are ordered by the number of loan officers.

Table A2. Payoff structure by treatment arms

	Control Group IAT Feedback (Treatment 2)	Higher Performance Pay (Treatment 1) Combined Intervention (Treatment 3)
Basic payment	400 EGP	400 EGP
Approval of an ex-post repaid loan	10 EGP	20 EGP
Rejection of an ex-post defaulted loan	15 EGP	30 EGP
Approval of an ex-post defaulted loan	-15 EGP	-30 EGP
Rejection of an ex-post repaid loan	-10 EGP	-20 EGP

Notes: This table compares the incentive and penalty structure of loan screening by treatment arms. The minimum payment (when all answers are wrong) is 280 EGP for the control group and 160 EGP for the higher performance pay group. The maximum payment (when all answers are correct) is 520 EGP and 640 EGP, respectively. As of the end of 2024, 1 USD was exchanged for approximately 50 EGP.

Table A3. Approval Rate Gap by Gender of Portfolio: ML Outcomes

Outcome: Approval Binary	Predicted Probability of Repayment by ML (0.5 thereshold)		Predicted Probability of Repayment by ML (Bottom 40% vs. Top 60%)	
	(1) Optimal Decision: Reject (<0.50) Generous to Men	(2) Optimal Decision: Approve (>=0.50) Harsh to Women	(3) Optimal Decision: Reject (Bottom 40%) Generous to Men	(4) Optimal Decision: Approve (Top 60%) Harsh to Women
Male Names	0.068**	0.031	0.061**	0.033
	(0.031)	(0.023)	(0.030)	(0.024)
Female Names Approval Mean	0.654	0.738	0.667	0.735
T-test p-value		0.342		0.232
Number of Portfolios	648	1152	720	1080
Number of Loan Officers	180	180	180	180
Portfolio FE	✓	✓	✓	✓
Loan Officer Gender FE	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Each loan has a predicted repayment probability generated by the GBM model. Columns 1 and 2 classify loans using a threshold of 0.5. Loans with predicted probabilities below 0.5 are considered high-risk and optimally rejected, while those with probabilities above 0.5 are considered low-risk and optimally approved. Columns 3 and 4 apply the same classification process but use a higher threshold of 0.54 at which loans are classified into the bottom 40% and top 60%. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table A4. Debiasing Treatment Effects in Loan Screening Performance by ML Outcomes

Outcome: Accuracy Binary	(1) Accuracy (Male-name Loans)	(2) Accuracy (Female-name Loans)
Higher Performance Pay (T1)	-0.016	-0.017
	(0.018)	(0.020)
IAT Feedback (T2)	0.006	-0.002
	(0.018)	(0.019)
Combined (T3)	0.007	-0.001
	(0.018)	(0.019)
Control Group Accuracy Mean	0.592	0.597
Number of Portfolios	7200	7200
Number of Loan Officers	720	720

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Each loan has a predicted repayment probability generated by the GBM model. The outcome is a binary screening performance (accuracy) variable that equals 1 if the officer approves loans with predicted repayment probabilities above the threshold or rejects loans below the threshold, and 0 if the officer approves loans below the threshold or rejects loans above the threshold. A threshold of 0.50 is used to measure accuracy. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table A5. Debiasing Treatment Effects in Approval of Low-quality Loans (ML)

Outcome: Approval Binary	ML-Predicted Repayment Probability (< 0.5)	
	(1) Male	(2) Female
Higher Performance Pay (T1)	-0.037	0.056*
	(0.032)	(0.033)
IAT Feedback (T2)	-0.012	0.052
	(0.031)	(0.032)
Combined (T3)	-0.019	0.059*
	(0.033)	(0.034)
Control Group Approval Mean	0.722	0.654
Number of Portfolios	1296	1296
Number of Loan Officers	720	720

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports changes in the approval of low-quality loans by applicant gender by using a predicted repayment probability generated by the GBM model. The combined group refers to the treatment group 3, where officers receive both the higher performance pay and the IAT feedback together. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table A6. Heterogeneity by Officers' Characteristics: Gender Gap in Loan Approval

Loan Officers' Characteristics

Outcome: Approval Binary	Gender		IAT Score		Explicit Bias	
	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Biased against Women (≥0.35)	Others (<0.35)	Biased against Women	Others
Male Names	0.044	0.044	0.048*	0.030	0.045	0.042*
	(0.028)	(0.027)	(0.027)	(0.032)	(0.036)	(0.023)
Female Loan Approval Rate Mean	0.73	0.69	0.70	0.72	0.72	0.70
Number of Portfolios	900	900	980	760	730	1070
Number of Loan Officers	90	90	98	76	73	107
T-test p-values		1.000		0.674		0.929

	Experience		Education		Risk Tolerance	
	Above Median	Below Median	College or Above	Others	Above Median	Below Median
Male Names	0.018	0.073**	0.025	0.083***	0.054	0.035
	(0.026)	(0.028)	(0.025)	(0.031)	(0.033)	(0.024)
Female Loan Approval Rate Mean	0.73	0.68	0.72	0.68	0.72	0.70
Number of Portfolios	940	860	1160	640	480	1320
Number of Loan Officers	94	86	116	64	48	132
T-test p-values		0.145		0.141		0.648

Notes: This table presents the heterogeneity in the gender gap in loan approval based on loan officers' characteristics. The IAT score refers to the Gender–Career Implicit Association Test (IAT) score. Explicit bias denotes responses to the statement "On the whole, men make better business executives than women do." Loan officers who answered "strongly agree" or "agree" are classified as biased against women, while those who answered "strongly disagree" or "disagree" are classified as others. A neutral option was not provided (four response options only). Experience denotes the number of years the loan officer has worked in that role. The risk tolerance is measured using the following question: "Imagine you have EGP 10,000 to invest. There is a 50% chance to double the amount of money you choose to 'invest'. It is equally possible that you could lose half the investment. How much of the EGP 10,000 do you invest?" The median investment amount is EGP 5,000 (131 loan officers chose EGP 5,000). Above-median is defined as investing EGP 5,000 or more, and below-median as investing less than EGP 5,000. Results are robust to classifying EGP 5,000 in the below-median category. Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table A7. Heterogeneity: Treatment Effects on the Approval of Ex-post Defaulted Male Applicants

Loan Officers' Characteristics

Outcome: Approval Binary	Gender		IAT Score		Explicit Bias	
	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Biased against Women (≥0.35)	Others (<0.35)	Biased against Women	Others
Higher Performance Pay (T1) x Male Names	-0.060	-0.071*	-0.078**	-0.022	-0.067	-0.062*
	(0.038)	(0.038)	(0.035)	(0.043)	(0.045)	(0.033)
IAT Feedback (T2) x Male Names	-0.029	-0.049	-0.026	-0.040	-0.061	-0.020
	(0.036)	(0.038)	(0.039)	(0.037)	(0.041)	(0.034)
Combined (T3) x Male Names	-0.033	-0.062	-0.064*	-0.017	-0.031	-0.056
	(0.037)	(0.039)	(0.038)	(0.040)	(0.044)	(0.034)
Male Names	0.044	0.044	0.047*	0.028	0.045	0.041*
	(0.028)	(0.028)	(0.026)	(0.030)	(0.034)	(0.023)
T-test p-values (T1)		0.836		0.313		0.936
T-test p-values (T2)		0.705		0.794		0.446
T-test p-values (T3)		0.594		0.393		0.648
Number of Portfolios	3600	3600	3760	3330	3010	4190
Number of Loan Officers	360	360	376	333	301	419

	Experience		Education		Risk Tolerance	
	Above Median	Below Median	College or Above	Others	Above Median	Below Median
Higher Performance Pay (T1) x Male Names	-0.018	-0.119***	-0.033	-0.118***	-0.091**	-0.058*
	(0.035)	(0.040)	(0.035)	(0.041)	(0.045)	(0.033)
IAT Feedback (T2) x Male Names	-0.024	-0.044	-0.034	-0.061	-0.068	-0.027
	(0.036)	(0.038)	(0.033)	(0.042)	(0.045)	(0.032)
Combined (T3) x Male Names	-0.025	-0.061	-0.037	-0.065	-0.068	-0.043
	(0.035)	(0.042)	(0.034)	(0.044)	(0.047)	(0.033)
Male Names	0.018	0.068**	0.026	0.080***	0.072**	0.034
	(0.026)	(0.027)	(0.025)	(0.030)	(0.033)	(0.024)
T-test p-values (T1)		0.061		0.116		0.561
T-test p-values (T2)		0.697		0.611		0.463
T-test p-values (T3)		0.502		0.626		0.660
Number of Portfolios	4070	3130	4420	2780	2000	5200
Number of Loan Officers	407	313	442	278	200	520

Notes: This table presents the heterogeneity in changes in loan approval of ex-post defaulted male applicants across treatment arms, relative to the control group, based on loan officers' characteristics. The IAT score refers to the Gender–Career Implicit Association Test (IAT) score. Explicit bias denotes responses to the statement "On the whole, men make better business executives than women do." Loan officers who answered "strongly agree" or "agree" are classified as biased against women, while those who answered "strongly disagree" or "disagree" are classified as others. A neutral option was not provided. Experience denotes the number of years the loan officer has worked in that role. The risk tolerance is measured using the following question: "Imagine you have EGP 10,000 to invest. There is a 50% chance to double the amount of money you choose to 'invest'. It is equally possible that you could lose half the investment. How much of the EGP 10,000 do you invest?" The median investment amount is EGP 5,000 (368 loan officers chose EGP 5,000). Above-median is defined as investing EGP 5,000 or more, and below-median as investing less than EGP 5,000. Results are robust to classifying EGP 5,000 in the below-median category. Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table A8. Heterogeneity: Debiasing Treatment Effects in Approval of Ex-post Defaulted Male Applicant Loan Approval

Loan Officers' Characteristics

	Gender		IAT Score		Explicit Bias	
	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Biased against Women (≥0.35)	Others (<0.35)	Biased against Women	Others
Higher Performance Pay (T1)	-0.094**	-0.100**	-0.120***	-0.065	-0.131***	-0.084*
	(0.042)	(0.048)	(0.043)	(0.050)	(0.047)	(0.043)
IAT Feedback (T2)	-0.067	0.006	-0.041	-0.007	-0.069	-0.009
	(0.042)	(0.043)	(0.042)	(0.046)	(0.046)	(0.041)
Combined (T3)	-0.100**	-0.022	-0.047	-0.065	-0.035	-0.079*
	(0.046)	(0.049)	(0.049)	(0.048)	(0.047)	(0.045)
Control Group Approval Mean	0.718	0.704	0.707	0.712	0.731	0.697
T-test p-values (T1)		0.931		0.410		0.465
T-test p-values (T2)		0.233		0.585		0.329
T-test p-values (T3)		0.248		0.795		0.499
Number of Portfolios	720	720	752	666	602	838
Number of Loan Officers	360	360	376	333	301	419

	Experience		Education		Risk Tolerance	
	Above Median	Below Median	College or Above	Others	Above Median	Below Median
Higher Performance Pay (T1)	-0.057	-0.128**	-0.073*	-0.129**	-0.142**	-0.072*
	(0.044)	(0.051)	(0.041)	(0.051)	(0.063)	(0.039)
IAT Feedback (T2)	-0.011	-0.050	-0.020	-0.056	-0.062	-0.009
	(0.043)	(0.045)	(0.041)	(0.045)	(0.065)	(0.035)
Combined (T3)	-0.026	-0.096*	-0.037	-0.088*	-0.076	-0.056
	(0.046)	(0.050)	(0.043)	(0.053)	(0.065)	(0.039)
Control Group Approval Mean	0.716	0.704	0.714	0.707	0.713	0.711
T-test p-values (T1)		0.291		0.400		0.345
T-test p-values (T2)		0.530		0.546		0.466
T-test p-values (T3)		0.299		0.457		0.797
Number of Portfolios	814	626	884	556	400	1040
Number of Loan Officers	407	313	442	278	200	520

Notes: This table presents the heterogeneity in changes in loan approval rates across treatment arms, relative to the control group, based on loan officers' characteristics. The IAT score refers to the Gender–Career Implicit Association Test (IAT) score. Explicit bias denotes responses to the statement "On the whole, men make better business executives than women do." Loan officers who answered "strongly agree" or "agree" are classified as biased against women, while those who answered "strongly disagree" or "disagree" are classified as others. A neutral option was not provided. Experience denotes the number of years the loan officer has worked in that role. The risk tolerance is measured using the following question: "Imagine you have EGP 10,000 to invest. There is a 50% chance to double the amount of money you choose to 'invest'. It is equally possible that you could lose half the investment. How much of the EGP 10,000 do you invest?" The median investment amount is EGP 5,000 (368 loan officers chose EGP 5,000). Above-median is defined as investing EGP 5,000 or more, and below-median as investing less than EGP 5,000. Results are robust to classifying EGP 5,000 in the below-median category. Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Appendix B: Selection of Experimental Loans

B.1. Procedure for Selecting the 50 Loans Used in the Experiment

The 50 loans used in this experiment were selected from a pool of 200 previously approved applications provided by the Family Development Foundation. From this initial set, 49 loans were excluded because their tenures had not yet concluded, leaving 151 eligible applications. Based on the power analysis, the maximum number of loans that could be used to achieve sufficient statistical power was 50. To arrive at the final sample of 50 loans, a multi-stage filtering procedure was implemented. This procedure was designed to mitigate potential statistical discrimination and to exclude clearly very high- or very low-quality loans, where gender bias is less likely to appear, thereby increasing statistical power.

The first filter addressed industry-level gender representation. To reduce the potential impact of statistical bias, the sample was restricted to industries where each gender constituted at least 30% of borrowers. This step ensured that an applicant's gender was not correlated with industry-specific risk. The qualifying industries included agriculture, supermarkets, grocery stores, clothing shops, fish sale shops, poultry, and livestock businesses.

The second filter focused on identifying borderline applications where loan officers have greater discretion and gender bias is more likely to emerge (Costello et al., 2020; Agarwal et al., 2024; FDIC, 2024). In the Egyptian credit scoring system, scores are categorized into five tiers: Excellent (above 750), Very Good (701–750), Satisfactory (626–700), Not Satisfactory (520–625), and Bad (below 520). Applications with clear-cut approval or rejection signals were excluded, specifically those with "Excellent" scores (above 750) or "Bad" scores (below 520). The selection was narrowed to loans with scores between 520 and 750, as well as those with no credit score, which is common for entrepreneurs in the informal sector. After applying both the industry and credit score filters, 68 candidate loans remained.

From these 68 candidates, the final 50 loans were selected using a machine learning (ML) methodology designed to identify loans with the most predictable outcomes based on

application data. The objective was not to develop a high-accuracy prediction model but rather to filter out loans whose ex-post outcomes appeared random or noisy. To reduce the over-fitting issue from a small sample size, an ensemble of seven ML algorithms was employed: Lasso, Random Forest, Support Vector Machines (SVM), Neural Networks, Gradient Boosting, AdaBoost, and HistGradientBoost. These models were selected for their predictive performance in related literature (Cheraghali and Molnár, 2024) and their widespread use in economics and applied ML. The selection process involved 100 independent iterations. In each iteration, the full 151-loan dataset was randomly split into a 70% training set and a 30% test set. For each loan and each of the seven models, a prediction was deemed accurate if its predicted probability of repayment (threshold > 0.5) matched its ex-post outcome (repaid or defaulted). A final "predictability benchmark" score was then calculated for each of the 68 candidate loans by averaging the accuracy scores from its four best-performing models across all iterations in which it appeared in a test set.

Finally, the 50-loan sample was constructed by stratifying based on gender and loan outcome, using the predictability benchmark to guide selection. The target sample composition was 30 repaid loans (15 male, 15 female) and 20 defaulted loans (10 male, 10 female). The pool of 68 candidates contained 17 male-repaid, 28 female-repaid, 11 male-defaulted, and 12 female-defaulted loans. To meet the target strata, loans with the lowest predictability scores within each category were removed. This entailed excluding the 2 lowest-scoring male-repaid, 13 lowest-scoring female-repaid, 1 lowest-scoring male-defaulted, and 2 lowest-scoring female-defaulted loans.

This ML-based selection method has limitations, including a non-standard cross-validation design and a relatively low average Area Under the Curve (AUC) of approximately 0.60 across iterations, reflecting the small sample size. However, a post-hoc validation exercise suggests the method's utility. Using an expanded dataset of 500 additional loans after the experiment, the selection procedure was repeated 1,000 times with a Gradient Boosting Machine (GBM) model introduced in Section 4.2.4 with AUC 0.88. The model is trained on 543

loans whose data became available later and tested on the original 68 loans. This procedure re-selected an average of 44 of the original 50 loans, yielding an 88% overlap. This result indicates that, despite its limitations, the selection process likely performed better than random sampling by successfully filtering out loans with highly unpredictable outcomes.

B.2. Procedure for Grouping the 50 Loans into Bundles

For the experiment, the 50 selected loans were organized into five bundles, each containing 10 loans. A simpler approach, such as randomly assigning 10 individual loans to each officer from the total pool of 50, was intentionally avoided. Under such a design, two officers evaluating the same loan application (where only the applicant's gender was experimentally altered) would make decisions within the context of two different sets of nine other loans. The composition of an evaluation batch can influence a loan officer's judgment. For example, an officer's assessment of a specific borderline application might become stricter if it is evaluated alongside exceptionally strong applications, or more lenient if it is surrounded by particularly weak ones. These spillover effects could introduce systematic noise, making it difficult to isolate the causal impact of applicant gender. By creating fixed bundles, this procedure ensures that when different loan officers evaluate a given application, they do so within an identical context of the other nine loans in the bundle.

To balance the overall difficulty level across the five bundles, I used the predictability benchmark derived from the machine learning (ML) model described in Appendix B.1. This benchmark captures how well each loan's repayment outcome could be predicted from its application data. Within each subgroup defined by applicant gender and ex-post repayment status (i.e., male-repaid, female-repaid, male-defaulted, and female-defaulted), I ordered the loans by their predictability scores. I assigned the loans to bundles using a symmetric "zigzag" pattern to ensure an even distribution of predictability across the five bundles. For instance, among the 15 ex-post male-name repaid loans, I assigned the most predictable loan to Bundle 1, the second to Bundle 2, the third to Bundle 3, and so on until Bundle 5. Then,

I reversed the direction: the sixth most predictable loan went into Bundle 5, the seventh into Bundle 4, and so on back to Bundle 1. This zigzag pattern continued until all 15 loans were assigned. I applied the same procedure separately to the 15 ex-post female repaid loans, 10 male defaulted loans, and 10 female defaulted loans.

Finally, I created five additional "clone" bundles (Bundles 6 to 10) that are exact clones of Bundles 1 to 5 in terms of loan content and order, but with first names switched to indicate the opposite gender. For example, if the first three loans in Bundle 1 featured applicants with genders female, male, and female, then the corresponding loans in Bundle 6 would have the same loan details and ordering, but with the genders switched to male, female, and male, respectively. To implement the gender manipulation, I removed the gender variable from the loan application and replaced the applicant's real first name with a common male or female name to signal gender. The replacement names were selected from the top 10 most frequently appearing male and female first names in the loan portfolios provided by the Family Development Foundation. The male names included Muhammad, Ahmed, Abdul, Mostafa, Mahmoud, Hussein, Ali, Hassan, Jaber, and Ashraf; the female names included Fatima, Karima, Iman, Shadia, Soheir, Hamida, Hana, Heba, Marwa, and Noha. This allows for a clean experimental comparison: by comparing loan officers assigned to Bundle 1 with those assigned to Bundle 6, I can examine how the same portfolio of loan applications is evaluated differently when only the applicant's gender is altered. In the Egyptian context, these names unambiguously signal male or female identity and do not carry gender-neutral interpretations.

Appendix C: Survey Questions

This section explains the baseline survey questions presented in Table 1. The "Loan Officers Skill Binary", "Cognitive Test", "Digit Span Recall", "Psychometric", and "Risk Attitude" questions are adopted from Bryan et al. (2024). The Social Desirability Test questions are from Hays et al. (1989a,b), and the Explicit Bias Questions are taken from the World Values Survey (Haerpfer et al., 2022). The format and instructions for the Gender-Career Implicit Association Test (IAT) are based on Greenwald et al. (1998, 2003); Lane et al. (2007), and `https://implicit.harvard.edu/implicit/`.

Loan Officers Skill Binary

There are two types of loan officers. Which characteristics do you think are similar to you?

- **PERSON A:** Person A has good communication skills; he is patient and a good listener and also has the ability to convince people, motivate them and give them valuable feedback.
- **PERSON B:** Person B is firm in his dealings and knows how to say no when required. He has good analytical skills along with excellent supervisory and management skills.

Cognitive Test

Now I'm going to ask you a series of intelligence and general knowledge questions.

- Q1. A bat and a ball cost $11 in total. The bat costs 1 dollars more than the ball. How much does the ball cost?
- Q2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

- Q3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake from the beginning?

Digit Span Recall

Now we're going to play a game. A series of digits will appear for a few seconds. Try to remember the digits and repeat them back in order. After each successful turn an extra digit will be added.

Social Desirability Test

Now I will read some phrases to you, and I want you to rate your level of agreement for each one. Please select one of the following options for each statement: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree.

- I am always courteous even to people who are disagreeable.
- There have been occasions when I took advantage of someone.
- I sometimes try to get even rather than a forgive and forget.
- I sometimes feel resentful when I don't get my way.
- No matter who I'm talking to, I'm always a good listener.

Psychometric Questions

Now I will read some phrases to you, and I want you to rate your level of agreement for each one. Please select one of the following options for each statement: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree.

- I tend to act first and worry about the consequences later.

- I can think of several solutions to any problem.
- When I make decisions I usually go with my first gut feeling.
- I always get things done ahead of time.
- I would work seven days a week if I could.
- I spend a lot of time planning for my future.
- I prefer to have a flexible schedule - I don't like being tied down.

Explicit Bias Questions

Now I will read some phrases to you, and I want you to rate your level of agreement for each one. Please select one of the following options for each statement: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree.

- When a mother works for pay, the children suffer.
- On the whole, men make better political leaders than women do.
- A university education is more important for a boy than for a girl.
- On the whole, men make better business executives than women do.
- If a woman earns more money than her husband, it's almost certain to cause problems.
- When jobs are scarce, men should have more right to a job than women.

Risk Attitude

- R1: How do you see yourself? Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please select one number on the scale, where 0 equals "unwilling to take risks" and 10 equals: "fully prepared to take risks".

- R2: Imagine you have EGP 10,000 to invest. You may choose to "invest" all or a portion of your money. There is a 50% chance to double the amount of money you choose to "invest". It is equally possible that you could lose half the investment. For example, if you choose to invest EGP 1,000, there is a 50-50 chance of getting EGP 2,000 or EGP 500. How much of the EGP 10,000 do you invest?

IAT Introduction

You will be presented with a series of nouns that are typically associated with Career or Family. You must state if that word is most related to career or family. To select "Career", press E on your keyboard. To select "Family", press I on your keyboard. If you select the incorrect answer, we will let you know, and you should then select the correct answer.

Answer as quickly as possible. When you are ready, go to the next field. There, you will either press the spacebar or tap the box at the bottom to start.

IAT Feedback: Control and Higher Performance Pay Groups

Thanks for your participation. Let's go to the next page.

IAT Feedback: IAT Feedback and Combined Intervention Groups

Thanks for your participation. Your Gender-Career IAT score is `{IAT_score_final}`. Here's how to understand your results:

If your score is positive, it indicates a stronger subconscious association between "male" and "career" or "female" and "family." Conversely, if your score is negative, it signifies a stronger subconscious association between "male" and "family" or "female" and "career." Here's how to interpret the magnitude of your score:

- No association if the score is between –0.15 and 0.15,
- Slight association for values between |0.15| and |0.35|,

- Moderate association between $|0.35|$ and $|0.65|$,
- Strong association for scores higher than $|0.65|$.

We want to iterate that this test reveals implicit attitudes and not behaviors. Our attitudes may derive from the cultural and social context where we live. An enormous body of literature confirms that we all have biases—some explicit, many implicit. However, it is important to avoid our implicit biases or stereotypes related to a specific group from systematically influencing our behavior toward clients. Acknowledging and understanding our biases and those of our colleagues can help minimize the influence they have on our daily interaction with clients.

Appendix D: Theoretical Framework

D.1. Model Primitives

D.1.1. Loan Officer

Risk-neutral[31] loan officers receive performance pay based on the accuracy of their approval decision by the loan's ex-post outcome. To capture the fact that the stakes from defaults are typically greater than those from repayments, I assign two stake parameters with $\ell_1 > \ell_2 > 0$ where $\pi \geq 1$ represents the strength of performance pay.

- Approve: $+\pi\ell_2$ if repaid, $-\pi\ell_1$ if defaulted.
- Reject: $+\pi\ell_1$ if defaulted, $-\pi\ell_2$ if repaid.

The loan officer derives taste-based utility δ_M, a non-negative parameter, when approving loans to male applicants; otherwise, taste-based utility is zero.

D.1.2. Loan Applicant (j)

Each loan applicant's project has a true, unobservable revenue, r. The loan amount is normalized to 1, so a loan is repaid if $r \geq 1$ and defaulted if $r < 1$. The officer's prior belief about revenue r for an applicant follows a normal distribution with a gender-specific mean m_{G_j} and a common variance σ_r^2.

$$r \sim N(m_{G_j}, \sigma_r^2)$$

Officers may hold more favorable prior beliefs toward men, and loan applications are on average profitable (i.e., $m > 1$). Therefore, I assume $m_M \geq m_F > 1$. Moreover, to reflect the high uncertainty in microfinance lending (e.g., informal and micro-business contexts), I assume the prior variance σ_r^2 is sufficiently large.

[31] To simplify the model, I assume risk neutrality. In a self-reported risk-attitude measure (on a 0–10 scale, where higher values indicate greater risk tolerance), loan officers report an average of 6.44, suggesting behavior close to risk neutrality. If loan officers are risk-averse, increasing performance pay tied to repayment outcomes would likely reduce approvals of unqualified male applicants even more than under risk neutrality.

D.2. Information Structure and Decision Rule

The officer cannot observe r. Instead, they choose an effort e_j and receive a noisy signal, $s = r + \epsilon$. The error term ϵ follows a normal distribution with mean zero and variance $\sigma_\epsilon^2(e_j)$, which decreases with effort.

$$\epsilon \sim N(0, \sigma_\epsilon^2(e_j)), \quad \text{where} \quad \sigma_\epsilon^2(e_j) = \frac{1}{e_j}$$

With the signal s, the officer uses Bayesian updating to form a posterior belief about the loan's revenue:

$$r \mid s, e_j, G_j \sim \mathcal{N}(\hat{m}_{G_j}, \hat{\sigma}^2), \quad \text{where } \hat{m}_{G_j} = \frac{m_{G_j} + e_j \sigma_r^2 s}{1 + e_j \sigma_r^2} \text{ and } \hat{\sigma}^2 = \frac{\sigma_r^2}{1 + e_j \sigma_r^2}$$

The officer's belief about the loan being repaid, $p(s, e_j, G_j)$, is the posterior probability that loan revenue $r \geq 1$, expressed using the cumulative distribution function of the standard normal distribution Φ.

$$p(s, e_j, G_j) = P(r \geq 1 \mid s, e_j, G_j) = \Phi\left(\frac{(m_{G_j} - 1) + e_j \sigma_r^2 (s - 1)}{\sqrt{\sigma_r^2 (1 + e_j \sigma_r^2)}}\right)$$

The officer approves a loan if the expected utility of approval exceeds that of rejection. Since $p(s, e_j, G_j)$ is monotonically increasing in the signal s, this is equivalent to approving the loan when $s \geq s^*_{G_j}$.

$$s^*_{G_j}(e_j, \pi, d) = 1 + \frac{\sqrt{\sigma_r^2 (1 + e_j \sigma_r^2)} \cdot \Phi^{-1}(k^*_{G_j}) - (m_{G_j} - 1)}{e_j \sigma_r^2}$$

where $k^*(G_j = M) = \frac{2\pi\ell_1 - \delta_M}{2\pi(\ell_1 + \ell_2)} \leq k^*(G_j = F) = \frac{2\pi\ell_1}{2\pi(\ell_1 + \ell_2)}$. The full derivations are provided in Appendix E.

D.3. The Officer's Optimal Effort Choice

The officer chooses an optimal effort level e_j to maximize expected utility subject to the convex cost ce^2.

$$\max_{e_j \geq 0} U(e_j; G_j, \pi, d) = \underbrace{\int_{-\infty}^{s_G^*(e,\pi,d)} V_R(s, e, G, \pi)\, g(s \mid e, G)\, ds}_{\text{reject region}} + \underbrace{\int_{s_G^*(e,\pi,d)}^{\infty} V_A(s, e, G, \pi, d)\, g(s \mid e, G)\, ds}_{\text{approve region}} - ce^2.$$

where $V_R(s, e, G, \pi) = \pi\Big[\ell_1 - (\ell_1 + \ell_2)p(s, e, G)\Big]$, $V_A(s, e, G, \pi, d) = \pi\Big[-\ell_1 + (\ell_1 + \ell_2)\, p(s, e, G)\Big] + d(G)$,

and $g(s \mid e, G)$ denotes the probability density function of the signal.

First-Order Condition can be written as three terms: (i) higher effort improves decision accuracy, increasing monetary benefits (A term); (ii) higher effort shifts the approval threshold, changing the relative weight of taste-based bias (B term); and (iii) the marginal cost of effort (see Appendix E for full derivations).

$$\frac{dU}{de} = \underbrace{\pi\, A\Big(e; k_G^*(\pi, d(G))\Big)}_{\text{(i)direct monetary return}} + \underbrace{d(G)\, B\Big(e; k_G^*(\pi, d(G))\Big)}_{\text{(ii) taste-based bias return}} - \underbrace{2ce}_{\text{(iii) marginal cost}}$$

D.4. Testable Prediction

The model formalizes how taste-based bias or priors about repayment likelihood can shape generosity toward men and harshness toward women in approval thresholds. It also shows how higher performance pay affects both screening effort and approval thresholds, yielding testable predictions. Appendix E presents the proofs.

Proposition 1. *For a sufficiently large increase in performance pay (π), the officer's optimal screening effort (e^*) rises, i.e., $\frac{de^*}{d\pi} > 0$.*

Proposition 2. *If either δ_M or m_M is sufficiently large, higher performance pay raises the*

approval threshold for male applicants (s_M^*), i.e., $\frac{ds_M^*}{d\pi} > 0$.

Testable Prediction:

Higher performance pay is expected to increase loan screening time and reduce the approval of low-quality male applicants.

Appendix E: Theoretical Model Derivations

This appendix provides the detailed derivations for the theoretical model presented in Appendix D.

E.1. Bayesian Updating of Posterior Beliefs

Before exerting effort—i.e., prior to receiving any signal—the loan officer holds a prior belief about the revenue r of an applicant, modeled as:

$$r \sim \mathcal{N}(m_{G_j}, \sigma_r^2)$$

where m_{G_j} is the prior mean revenue for an applicant of gender G_j, and σ_r^2 is the prior variance.

After exerting effort e_j, the officer observes a noisy signal:

$$s = r + \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}\left(0, \frac{1}{e_j}\right)$$

Using Bayes' rule, the officer updates their belief about the revenue r. The posterior distribution is:

$$r \mid s, e_j, G_j \sim \mathcal{N}(\hat{m}_{G_j}, \hat{\sigma}^2)$$

with posterior mean and variance given by:

$$\hat{m}_{G_j} = \frac{m_{G_j} + e_j \sigma_r^2 s}{1 + e_j \sigma_r^2}, \qquad \hat{\sigma}^2 = \frac{\sigma_r^2}{1 + e_j \sigma_r^2}$$

These expressions reflect precision-weighted updating, where the prior and signal are combined based on the inverse of their variances (i.e., their precision).

- **Posterior Mean ($\hat{m}$):** A weighted average of the prior mean and the observed signal,

with weights proportional to their precision:

$$\hat{m}(s, e_j, G_j) = \frac{\left(\frac{1}{\sigma_r^2}\right) m_{G_j} + e_j s}{\frac{1}{\sigma_r^2} + e_j} = \frac{m_{G_j} + e_j \sigma_r^2 s}{1 + e_j \sigma_r^2}$$

- **Posterior Variance ($\hat{\sigma}^2$):** The inverse of the total precision:

$$\hat{\sigma}^2(e_j) = \frac{1}{\frac{1}{\sigma_r^2} + e_j} = \frac{\sigma_r^2}{1 + e_j \sigma_r^2}$$

E.2. Derivation of Decision Thresholds

Derivation of the Approval Probability Threshold (k_G^*)

The officer approves an application if the expected utility of approving, $E[U(\text{Approve})]$, exceeds that of rejecting, $E[U(\text{Reject})]$. The threshold k_G^* is the posterior probability of repayment, $p(s, e_j, G_j)$, where the officer is exactly indifferent. I find this by setting the expected utilities equal:

$$E[U(\text{Approve})] = E[U(\text{Reject})]$$

Let $p \equiv p(s, G_j)$. The expected utilities for each action are:

$$E[U(\text{Approve})] = p \cdot (+\pi \ell_2) + (1 - p) \cdot (-\pi \ell_1) + d_j(G_j)$$
$$E[U(\text{Reject})] = p \cdot (-\pi \ell_2) + (1 - p) \cdot (+\pi \ell_1)$$

Setting these equal and solving for p yields the threshold k_G^*:

$$\pi p \ell_2 - \pi(1 - p)\ell_1 + d_j(G_j) = -\pi p \ell_2 + \pi(1 - p)\ell_1$$
$$k_G^* = \frac{2\pi \ell_1 - d_j(G_j)}{2\pi(\ell_1 + \ell_2)}$$

Substituting $d_j(M) = \delta_M$ and $d_j(F) = 0$, I obtain:

$$k_M^* = \frac{2\pi\ell_1 - \delta_M}{2\pi(\ell_1 + \ell_2)} \quad \leq \quad k_F^* = \frac{2\pi\ell_1}{2\pi(\ell_1 + \ell_2)}$$

where $0 \leq \delta_M \leq 2\pi\ell_1$.

Derivation of the Signal Threshold (s_G^*)

The officer's belief about the loan being repaid is captured by $p(s, e, g)$, defined as the posterior the loan's revenue $r \geq 1$, given the observed signal s_j by effort e_j and the applicant's gender G_j.

$$p(s_j, e_j, G_j) = P(r \geq 1 \mid s_j, e_j, G_j) = 1 - P(r < 1 \mid s_j, e_j, G_j)$$

Since $r \mid s, e_j, G_j \sim \mathcal{N}(\hat{m}_{G_j}, \hat{\sigma}^2)$, this probability can be expressed using the standard normal CDF $\Phi(\cdot)$:

$$p(s, e_j, G_j) = 1 - \Phi\left(\frac{1 - \hat{m}(s, e_j, G_j)}{\hat{\sigma}(e_j)}\right) = \Phi\left(\frac{\hat{m}(s, e_j, G_j) - 1}{\hat{\sigma}(e_j)}\right) = \Phi\left(\frac{(m_{G_j} - 1) + e_j\sigma_r^2(s-1)}{\sqrt{\sigma_r^2(1 + e_j\sigma_r^2)}}\right)$$

The signal threshold, s_G^*, is the signal value that yields a posterior probability equal to k_G^*. I solve for s_G^* by setting $p(s_G^*) = k_G^*$:

$$\begin{aligned}
\Phi\left(\frac{(m_{G_j} - 1) + e_j\sigma_r^2(s_G^* - 1)}{\sqrt{\sigma_r^2(1 + e_j\sigma_r^2)}}\right) &= k_G^* \\
\frac{(m_{G_j} - 1) + e_j\sigma_r^2(s_G^* - 1)}{\sqrt{\sigma_r^2(1 + e_j\sigma_r^2)}} &= \Phi^{-1}(k_G^*) \\
s_G^* &= 1 + \frac{\sqrt{\sigma_r^2(1 + e_j\sigma_r^2)} \cdot \Phi^{-1}(k_G^*) - (m_{G_j} - 1)}{e_j\sigma_r^2}
\end{aligned}$$

Then, the difference in approval thresholds between male and female applicants is:

$$s_M^* - s_F^* = \frac{\sqrt{\sigma_r^2(1 + e_M^*\sigma_r^2)} \cdot \Phi^{-1}(k_M^*) - (m_M - 1)}{e_M^*\sigma_r^2} - \frac{\sqrt{\sigma_r^2(1 + e_F^*\sigma_r^2)} \cdot \Phi^{-1}(k_F^*) - (m_F - 1)}{e_F^*\sigma_r^2}$$

$$= \underbrace{\left(\frac{\sqrt{(1+e_M^*\sigma_r^2)}\cdot\Phi^{-1}(k_M^*)}{e_M^*\sigma_r} - \frac{\sqrt{(1+e_F^*\sigma_r^2)}\cdot\Phi^{-1}(k_F^*)}{e_F^*\sigma_r} \right)}_{\text{Taste-Based Bias}} + \underbrace{\left(\frac{1-m_M}{e_M^*\sigma_r^2} - \frac{1-m_F}{e_F^*\sigma_r^2} \right)}_{\text{Prior Belief}}$$

E.3. Derivation of the First-Order Condition

The officer chooses effort e to maximize expected utility (with convex cost ce^2):

$$U(e;G,\pi,d) = \int_{-\infty}^{s_G^*(e,\pi,d)} V_R(s,e,G,\pi)\, g(s \mid e,G)\, ds + \int_{s_G^*(e,\pi,d)}^{\infty} V_A(s,e,G,\pi,d)\, g(s \mid e,G)\, ds - ce^2,$$

where $V_R(s,e,G,\pi) = \pi\Big[\ell_1 - (\ell_1+\ell_2)p(s,e,G)\Big]$, $\quad V_A(s,e,G,\pi,d) = \pi\Big[-\ell_1 + (\ell_1+\ell_2)p(s,e,G)\Big] + d(G)$, and $g(\cdot \mid e,G)$ is the pdf of the signal.

Differentiating with respect to e using Leibniz' rule yields

$$\frac{dU}{de} = \int_{-\infty}^{s^*} \partial_e\Big(V_R g\Big)\, ds + \int_{s^*}^{\infty} \partial_e\Big(V_A g\Big)\, ds - 2ce,$$

since boundary terms vanish by the cutoff condition $V_A(s^*) = V_R(s^*)$. Applying the product rule $\partial_e(Vg) = V_e g + V g_e$, and using $V_{R,e} = -\pi(\ell_1+\ell_2)p_e$, $V_{A,e} = \pi(\ell_1+\ell_2)p_e$, gives

$$\frac{dU}{de} = \int_{s^*}^{\infty} \Big(\pi(\ell_1+\ell_2)p_e\, g + V_A g_e\Big)\, ds + \int_{-\infty}^{s^*} \Big(-\pi(\ell_1+\ell_2)p_e\, g + V_R g_e\Big)\, ds - 2ce.$$

Grouping terms, the first-order condition is

$$\boxed{\frac{dU}{de} = \pi\, A\Big(e; k_G^*(\pi, d(G))\Big) + d(G)\, B\Big(e; k_G^*(\pi, d(G))\Big) - 2ce.}$$

where $A(e;k) = \displaystyle\int_{s^*}^{\infty} \Big((\ell_1+\ell_2)p_e\, g + [-\ell_1 + (\ell_1+\ell_2)p]\, g_e\Big)\, ds + \int_{-\infty}^{s^*} \Big((\ell_1+\ell_2)p_e\, g + [\ell_1 - (\ell_1+\ell_2)p]\, g_e\Big)\, ds,$

$$B(e;k) = \int_{s^*}^{\infty} g_e(s \mid e,G)\, ds, \qquad g_e(s \mid e,G) = \frac{\partial g(s \mid e,G)}{\partial e} \quad \text{and} \quad p_e(s \mid e,G) = \frac{\partial p(s \mid e,G)}{\partial e}.$$

Here, $A(e;k)$ is the monetary channel (stakes $\times$ informativeness), strictly positive under the standard regularity assumption that higher effort increases signal precision, while $B(e;k)$ captures how effort shifts the approval region above s^*, scaling $d(G)$. The sign of B_k is not definitive, as it depends on how effort shifts the signal distribution around the threshold s^*.

E.4. Proof of Proposition 1

Proposition 1. *For a sufficiently large increase in performance pay (π), the officer's optimal screening effort (e^*) rises, i.e., $\frac{de^*}{d\pi} > 0$.*

To study how performance pay π affects optimal effort, define the first-order condition

$$F(e,\pi;d) \;\equiv\; \pi A\Big(e;k_G^*(\pi,d)\Big) \;+\; d(G)\,B\Big(e;k_G^*(\pi,d)\Big) \;-\; 2ce,$$

where $k_G^*(\pi,d) = \frac{2\pi\ell_1 - d(G)}{2\pi(\ell_1+\ell_2)}$ and $\frac{\partial k_G^*}{\partial \pi} = \frac{d(G)}{2\pi^2(\ell_1+\ell_2)} \geq 0$.

The optimal effort e_G^* is characterized by $F(e_G^*,\pi;d) = 0$. By the implicit function theorem,

$$\frac{de_G^*}{d\pi} \;=\; -\frac{F_\pi}{F_e}\Big|_{e=e_G^*}.$$

If the second-order condition $F_e < 0$ holds, the sign of $\frac{de_G^*}{d\pi}$ is determined by the sign of F_π:

$$F_\pi = A\Big(e_G^*;k_G^*\Big) + \frac{d(G)}{2\pi(\ell_1+\ell_2)}\,A_k\Big(e_G^*;k_G^*\Big) + \frac{d(G)^2}{2\pi^2(\ell_1+\ell_2)}\,B_k\Big(e_G^*;k_G^*\Big),$$

where subscripts denote partial derivatives with respect to k.

- $A(e;k) > 0$ is the marginal monetary return to effort (via repayment payoffs).
- A_k measures how the monetary return changes as the cutoff shifts; its sign is ambiguous because a tighter cutoff both reduces the mass affected by effort and increases the

sensitivity of marginal loans.

- $B_k \leq 0$ under standard approval monotonicity, since tightening the cutoff reduces approval probability.

As $\pi \to \infty$, $k_G^*(\pi, d)$ converges to $\bar{k} = \ell_1/(\ell_1 + \ell_2)$, and $A\Big(e_G^*; k_G^*\Big)$ converges to a positive finite limit, while A_k and B_k remain bounded. Meanwhile, both $\frac{d(G)}{2\pi(\ell_1+\ell_2)}$ and $\frac{d(G)^2}{2\pi^2(\ell_1+\ell_2)}$ converge to zero. Therefore,

$$\lim_{\pi\to\infty} F_\pi = A\Big(e_G^*; k_G^*\Big) \;>\; 0.$$

It follows that for sufficiently large π,

$$\frac{de_G^*}{d\pi} \;=\; -\frac{F_\pi}{F_e} \;>\; 0.$$

E.5. Proof of Proposition 2

Proposition 2. *If either δ_M or m_M is sufficiently large, higher performance pay raises the approval threshold for male applicants (s_M^*), i.e., $\frac{ds_M^*}{d\pi} > 0$.*

The male approval threshold is:

$$s_M^* = 1 + \underbrace{\frac{\sqrt{\sigma_r^2(1+e_M^*\sigma_r^2)}\cdot\Phi^{-1}(k_M^*)}{e_M^*\sigma_r^2}}_{\text{Taste-Based Bias Effect}} - \underbrace{\frac{(m_M-1)}{e_M^*\sigma_r^2}}_{\text{Prior Belief Effect}}, \quad k_M^* = k_M^*(\pi,\delta_M) = \frac{2\pi\ell_1 - \delta_M}{2\pi(\ell_1+\ell_2)}.$$

Let $a = \sigma_r^2$, $e = e_M^*(\pi)$, $k = k_M^*(\pi, \delta_M)$, $z = \Phi^{-1}(k)$, and

$$f(e) = \frac{\sqrt{1+ea}}{e\sqrt{a}}, \qquad f'(e) = -\frac{ea+2}{2e^2\sqrt{a}\sqrt{1+ea}}, \qquad \frac{d}{dk}\Phi^{-1}(k) = \frac{1}{\phi(\Phi^{-1}(k))}.$$

By the chain rule,

$$\frac{ds_M^*}{d\pi} = \Big[z\, f'(e) + \frac{m_M - 1}{e^2 a}\Big]\frac{de_M^*}{d\pi} \;+\; \frac{f(e)}{\phi(z)}\frac{\partial k_M^*}{\partial \pi}, \qquad \frac{\partial k_M^*}{\partial \pi} = \frac{\delta_M}{2\pi^2(\ell_1 + \ell_2)}.$$

By Proposition 1, $\dfrac{de_M^*}{d\pi} > 0$, and assume $m_M > 1$. Then, $\frac{ds_M^*}{d\pi} > 0$ under either of the following conditions[32]:

1. **Strong taste based bias favoring men:** If

$$\delta_M \geq \pi(\ell_1 - \ell_2) \;\; \text{with} \;\; \ell_1 > \ell_2,$$

then $k_M^*(\pi, \delta_M) \leq \frac{1}{2}$, so $z = \Phi^{-1}(k_M^*) \leq 0$. In this case,

$$\frac{ds_M^*}{d\pi} = \Big[z\, f'(e_M^*) + \frac{m_M - 1}{e_M^{*2}\sigma_r^2}\Big]\frac{de_M^*}{d\pi} \;+\; \frac{f(e_M^*)}{\phi(z)}\,\frac{\partial k_M^*}{\partial \pi},$$

$$f'(e) = -\frac{e\sigma_r^2 + 2}{2e^2\sigma_r\sqrt{1 + e\sigma_r^2}} < 0, \frac{\partial k_M^*}{\partial \pi} = \frac{\delta_M}{2\pi^2(\ell_1 + \ell_2)} > 0.$$

Since $z \leq 0$ and $f'(e_M^*) < 0$, I have $z\, f'(e_M^*) \geq 0$. With $m_M > 1$, it follows that $\dfrac{m_M - 1}{e_M^{*2}\sigma_r^2} > 0$. By Proposition 1, $\dfrac{de_M^*}{d\pi} > 0$. Moreover, $\dfrac{f(e_M^*)}{\phi(z)} > 0$ and, as $\delta_M > 0$, $\dfrac{\partial k_M^*}{\partial \pi} > 0$. Therefore, $\dfrac{ds_M^*}{d\pi} > 0$.

2. **Strong prior belief in male borrowers' repayment likelihood :** If

$$m_M \;>\; 1 \;+\; z\,\frac{\sigma_r\,(e_M^*\,\sigma_r^2 + 2)}{2\,\sqrt{1 + e_M^*\,\sigma_r^2}},$$

then $z\, f'(e_M^*) + \dfrac{m_M - 1}{e_M^{*2}\sigma_r^2} > 0$, hence $\dfrac{ds_M^*}{d\pi} > 0$ regardless of whether $\delta_M > 0$ or $\delta_M = 0$: $\dfrac{f(e_M^*)}{\phi(z)}\,\dfrac{\partial k_M^*}{\partial \pi}$ is positive if $\delta_M > 0$ and equals zero if $\delta_M = 0$. Therefore, $\dfrac{ds_M^*}{d\pi} > 0$.

[32] A sufficiently large bias is required because with $\ell_1 > \ell_2$, the sign of z depends on the bias magnitude. On the other hand, when $\ell_1 < \ell_2$, z is always negative, so even a small bias ensures $\frac{ds_M^*}{d\pi} > 0$