NBER WORKING PAPER SERIES

SMILES IN PROFILES:
IMPROVING FAIRNESS AND EFFICIENCY USING ESTIMATES OF USER PREFERENCES IN ONLINE MARKETPLACES

Susan Athey
Dean Karlan
Emil Palikot
Yuan Yuan

Working Paper 30633
http://www.nber.org/papers/w30633

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2022, Revised March 2025

We thank Kiva Microfunds for generously sharing data and discussing the research questions. We thank Herman Don-ner and Kristine Koutout at Stanford Graduate School of Business and Allison Koenecke at Cornell University for comments and suggestions. The Golub Capital Social Impact Lab at Stanford Graduate School of Business provided funding for this research. This research has been subject to review and approval by the Research Compliance Office at Stanford University, protocol number IRB-62442 and registered at the AEA RCT registry with the number 0010030. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w30633

Smiles in Profiles: Improving Fairness and Efficiency Using Estimates of User Preferences
in Online Marketplaces
Susan Athey, Dean Karlan, Emil Palikot, and Yuan Yuan
NBER Working Paper No. 30633
November 2022, Revised March 2025
JEL No. D0, D40, J0, J02, O1

ABSTRACT

Online platforms often have conflicting goals: they face tradeoffs between increasing efficiency and reducing disparities, where the latter may relate to objectives such as the longer-term health of the marketplace or the organization's mission. We examine how participants' profile pictures shape this trade-off in the context of a peer-to-peer lending platform. We develop and apply an approach to estimate marketplace participants' preferences for different profile features, distinguishing between (i) "type" (e.g., gender, age) and (ii) "style" (e.g., smiling in the photo). Relative to type, style features are easier to change, and platforms may be more willing to encourage such changes. Our approach starts by using causal inference methods together with computer vision algorithms applied to observational data to identify type and style features of profiles that appear to affect demand for transactions. We further decompose type-based disparities into a component driven by demand for certain types and a component that arises because different types have different distributions of style features; we find that style differences exacerbate type-based disparities. To improve internal validity, we then carry out two randomized survey experiments using generative models to create multiple versions of profile images that differ in one feature at a time. We then evaluate counterfactual platform policies based on the changeable profile features and identify approaches that can ameliorate the disparity-efficiency tension. We identify marketplace feedback effects, where encouraging certain style choices attracts participants who value these choices.

Susan Athey
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
athey@stanford.edu

Dean Karlan
Kellogg Global Hub
Northwestern University
2211 Campus Drive
Evanston, IL 60208
and CEPR
and also NBER
dean.karlan@gmail.com

Emil Palikot
306 Fair Oaks st
San Francisco, CA 94110
emil.palikot@gmail.com

Yuan Yuan
5000 Forbes Avenue
Pittsburgh, PA 15213
yuany3@andrew.cmu.edu

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/10030

1 Introduction

Profile images are a key design feature of many online platforms, shaping user interactions and platform outcomes (Ert et al., 2016). A well-established literature documents how profile images reveal socio-demographic characteristics, often enabling discrimination and leading to disparities across groups in outcomes (Pope and Sydnor, 2011). However, profile images contain abundant information beyond socio-demographic characteristics that impact outcomes (Zhang et al., 2022b). In this paper, we study how seemingly innocuous stylistic choices in profile images can influence both efficiency and across-group disparities for online platforms.

We distinguish between two categories of profile features: fixed or difficult-to-change characteristics (*type*, e.g., gender or age) and features that are easier to alter (*style*, e.g., whether an individual smiles in the photo). The *style* features we consider are features that platform managers might encourage platform participants to change. If *type* and *style* choices are uncorrelated, two distinct sources of disparities might emerge on the platform: some types may transact more frequently, and some styles may be more successfu. However, when *type* and *style* are positively correlated, disparities compound; when negatively correlated, they partially offset one another.

This paper analyzes the *type* and *style* features of online profiles on Kiva, a non-profit microlending platform. On Kiva, individual lenders allocate capital to borrowers by selecting from a curated catalog of borrowing campaigns.[1] In designing this marketplace, Kiva aims to balance efficiency—measured by the volume of transactions—with a notion of fairness, understood as the equitable distribution of capital among borrowers of similar reliability on the platform (Burke et al., 2022).

Fairness can be defined in various ways, depending on a platform's objectives, the groups of interest, and the broader context (Kleinberg et al., 2016; Dwork et al., 2012). A key input for implementing any fairness policy is an empirical understanding of disparities across relevant groups. This paper focuses on disparities by *type*, which captures socio-demographic characteristics—such as *gender* and *age*—that are commonly considered in fairness policies.

We develop a two-step approach to understanding how *type* and *style* profile features contribute to efficiency and type-based disparities, selecting features that are managerially relevant in the sense that the platform can design interventions based on them. In the first step, using observational data from the Kiva platform, we identify features that appear to matter to lenders in that they affect funding out-

[1] Technically, the loan is made to a microcredit institution and earmarked for the specific borrower.

comes conditional on exposure to lenders, and analyze the contribution of *style* features to *type*-based disparities. To extract profile image features, we apply an off-the-shelf machine learning algorithm that detects over one hundred features. By comparing the predictive performance of models trained with and without *style* features, we show that these features collectively predict funding outcomes. We further identify specific stylistic elements with a large and statistically significant impact on funding success, both unconditionally and after controlling for proxies for exposure to lenders as well as other observable borrower characteristics. For example, a *smile* is associated with higher funding, while wearing *sunglasses* or featuring a *body-shot*-image correlates with lower funding outcomes.[2]

We next show that borrowers' *style* choices aggravate disparities between many *types*. First, *style* and *type* tend to be correlated - borrowers' from different socio-demographic groups systematically create different profiles. *Men* are less likely to *smile* and more likely to wear *sunglasses* than *women*, while *young* borrowers are more likely to both *smile* and wear *sunglasses* than older borrowers. Second, we carry out covariate decomposition of *type*-based disparities (Gelbach, 2016) and show that *style* features jointly increase disparities between *men* and *women* or *old* and *young* borrowers.

Estimates of the impact of profile features on outcomes from observational data rely on the assumption of unconfoundedness, that is, the assumption that conditional on other observable features of profiles, the assignment of style is as good as random, with no important omitted variables correlated with both the target feature and funding outcomes. This assumption is not directly testable, and there are reasons to question it in our context. For this reason, we view our estimates from observational data as suggestive but not definitive, and we use our findings to prioritize *type* and *style* features to analyze further in recruited experiments.

The second step of our approach aims to provide internally valid estimates of the magnitude of the impact of profile features on funding decisions. In a sequence of two recruited experiments, we ask subjects to make a series of choices, each time selecting between two profiles. The profiles that we show them are mostly artificial profiles based on real images. For each original picture, we use generative artificial intelligence models to create several variants where each alters one profile feature and holds the rest of the photo fixed.[3]

In the first experiment, we examine two *style* features (*smile* and *body-shot*) and one *type* feature (*gender*), including only hypothetical borrowers with AI generated images. The second experiment introduces financial incentives: subjects evaluate both hypothetical borrowers and real borrowers

[2] A *body-shot* refers to an image where the person's body occupies a large portion of the frame.

[3] The original picture is never used in a profile.

currently active on Kiva. For each participant, we allocate $10 to a selected real borrower, aligning experimental choices with real financial stakes. In this experiment, we analyze the impact of three additional *style* features (*sunglasses*, *glasses*, and *dark hair*) and one additional *type* feature (*age*).

Across a range of empirical specifications, we find that a *smile* has a large and statistically significant positive effect on the probability of being selected by a subject. In contrast, wearing *sunglasses* or *glasses* has a negative and statistically significant impact. We also find suggestive evidence that *body-shot* has a negative effect and *dark hair* a positive effect, though these results are not statistically significant across all specifications. Finally, we find that experimental subjects prefer *women* borrowers over *men*, consistent with patterns observed in the Kiva observational data, while the difference in selection probabilities between *young* and *old* profiles is small and statistically insignificant.

We then explore the mechanisms through which *style* features impact funding outcomes, distinguishing between monetary and non-monetary channels. To investigate monetary mechanisms, we examine whether *style* features correlate with loan repayment probability, as lenders may use these visual cues to infer creditworthiness. Using machine learning models trained on Kiva data, we find that incorporating *style* features does not improve predictions of repayment probability, suggesting that these features do not contain relevant financial information. To explore non-monetary mechanisms, we analyze whether *style* features evoke psychological responses that influence lender preferences. Leveraging the deep learning model of (Peterson et al., 2022), which is trained on human judgments, we estimate psychological traits from borrower images and find strong correlations between *style* features and perceived attributes such as *trustworthiness* and *dominance*. These findings are consistent with the hypothesis that lenders' decisions are shaped more by psychological perceptions of borrower images than by financial considerations.

Style features influence funding outcomes and contribute to disparities between borrowers. However, unlike fixed borrower characteristics, *style* features can be modified, making them a potential lever for platform policies aimed at reducing disparities or improving efficiency. We use the estimates of impact of image features on demand to examine counterfactual platform policies that modify the conditional distribution of *style* features in borrower profiles and adjust the probability of borrowers appearing in lenders' choice sets based on borrower characteristics. To assess these policies, we calibrate a model of lender demand using estimates from our recruited experiments. In our model, lenders are heterogenous with respect to *style* and *type* preference parameters. They, first, decide whether to participate and then choose one of the available borrowers. The choice set they observe depends on the platform policy.

Our findings suggest that policies reducing the correlation between desirable *style* features and borrower *type* reduce disparities and increase efficiency. A policy of *style* recommendation — for example, encouraging borrowers to include a *smile* in their profile while avoiding *body-shots* - reduces disparities, as measured by a lower Gini coefficient and a reduced gender gap. This policy also increases the total number of transactions. In contrast, a policy that increases the visibility of campaigns featuring *smiles* and excluding *body-shot*s—such as ranking them higher in search results—improves efficiency but also increases disparities. This occurs because emphasizing these *style* features disproportionately benefits high-*type* borrowers. Notably, if a platform trains a recommendation system based on funding data and incorporates image features, the system will likely prioritize borrower profiles with *style* attributes that appeal to lenders. Thus, this policy reflects the expected outcome of implementing a recommendation system that accounts for image features.

Platform policies not only influence which borrowers receive funding but also shape the type of lenders who are active on the platform. We show that interventions such as increasing the visibility of borrowers with certain stylistic features or recommending *style* features attract lenders with stronger preferences for those features. Conversely, efforts to promote borrowers from under-performing *types* reduce participation among lenders who prioritize *style*. These shifts in lender composition represent an equilibrium effect—platform policies do not merely alter individual choices but also reshape the overall dynamics of lender engagement.

We examine a specific dimension of *type*-based inequity: the gender gap in favor of campaigns featuring *women* profiles.[4] In our study of observational data from the Kiva platform, we find that campaigns featuring *men* profiles receive 32% less funding per day than those featuring *women* profiles. This result is corroborated by our recruited experiment, where subjects are 31% more likely to select a *women* profile than a *men* profile.[5] The distribution of selected *style* features further amplifies this gap. Among borrowers classified as *women*, 77% have profiles with a *smile*, while 22% have *body-shot* images. In contrast, only 33% of *men* borrowers are depicted with a *smile*, and 26% have a *body-shot* profile image. These differences in *style* feature distribution suggest that stylistic choices contribute to the observed gender gap in funding outcomes.

[4] Throughout, we use *men* and *women* to denote the gender classification assigned by the feature detection algorithm.

[5] Lenders may prefer campaigns with female profiles for several reasons. For example, extensive evidence suggests that female entrepreneurs who receive microfinance funding tend to use the funds effectively (D'Espallier et al., 2011; Aggarwal et al., 2015). Additionally, lenders may seek to counteract gender discrimination in traditional entrepreneurial finance (Alesina et al., 2013).

Our findings demonstrate that in marketplaces where users have preferences for certain profile image features, the correlation between *type* and *style* characteristics can matter for disparities and efficiency. Platforms seeking to balance these objectives need to account for this correlation before implementing policies based on profile images.

The paper is organized as follows: Section 2 presents related literature. Section 3 describes how micro-lending platforms operate and provide institutional details about Kiva. Section 4 presents the observational data and its analysis. Section 5 describes the design of the experiment and its results. Section 6 focuses on counterfactual simulations, and Section 7 concludes.

2 Literature Review

In this paper, we show that mutable *style* features in user-generated content contributes to disparities in outcomes across different *type*s. Numerous papers document disparities in outcomes by race and gender (*type* features) in online platforms. Users with African-American-sounding names face higher cancellation rates on ride-sharing platforms (Ge et al., 2016) and Airbnb (Edelman et al., 2017); drivers with Arabic or African-sounding names earn less on BlaBlaCar (Lambin and Palikot, 2022); Airbnb hosts with distinctively Asian names received fewer guests following the onset of the COVID-19 pandemic (Luca et al., 2024); non-Caucasian online profiles are shown fewer and different types of housing ads (Asplund et al., 2020); black and female NFT avatars are valued less (Yuan et al., 2024); and women earn less on Lyft (Cook et al., 2021) and Airbnb (Davidson and Gleim, 2023). Specifically for platforms related to lending, several papers document disparities in funding outcomes by race, including on Prosper.com (Theseira, 2009; Pope and Sydnor, 2011) and Kickstarter.com (Younkin and Kuppuswamy, 2018). Beauty, which can be thought of as a function of both *type* (e.g., young) and *style* features (e.g., smiling), has also been shown to impact funding outcomes on Prosper.com (Ravina, 2019), as well as on Kiva (Park et al., 2019), the platform used in our study.

Another literature considers disparities by *style* features. For example, Zhang et al. (2022b) show that properties on Airbnb with verified photos have 9% higher occupancy rates than those without verification.[6] Dupas et al. (2024) shows that clothing and image background impact job interview chances. There is also evidence that different *type*s choose different *style* features. For example, women

[6] Zhang et al. (2022b) use Convolutional Neural Networks (CNN), the tool we use in this paper to identify image features, to classify images as high or low quality, finding that including image quality in their linear regression model reduces the estimated treatment effect of verified photos by 41%. They also define 12 image features based on the art and photography literature, which they sort into artistic categories like composition and color, that completely explain the difference between verified and unverified photos.

write more, and more enthusiastically, in their profiles on OkCupid compared to men (Shishido et al., 2016). Livestreamers' smiles and studio color design can affect their audience engagement and sales performance in the online live-streaming marketplace (Lin et al., 2021; Han et al., 2024). In line with our results, Haferkamp et al. (2012) find that women are less likely to choose body-shots than men in their online social media profile.

Other papers evaluate how much of a disparity by type can be explained by other factors. Most closely resembling our own study, Marchenko (2019) evaluates whether features of the property listing *text* on Airbnb, measured using natural language processing (NLP), can explain earnings disparities between minority and white male hosts. She finds that *style* features, specifically the subjectivity and polarity of the text, have at most a marginal impact on estimates after accounting for hard-to-change property features like location, property type, and number of bedrooms, which explain most of the disparity. Other studies find that immutable or difficult-to-change factors account for a portion of *type*-based disparities. For example, a Superhost designation and their rating, as well as guest ratings impact bi-directional racial disparities (i.e., preferences for own-race) (Zhang et al., 2022a); number of guests accommodated, median home value, years of experience, and room type can account for more than half of the gender earning gap on Airbnb (Davidson and Gleim, 2023); reputation closes the gap between white and minority drivers on BlaBlaCar as minority drivers gain experience (Lambin and Palikot, 2022); facial femininity increases the disparity between women and men in whether potential customers online seek information about their tutoring services (Luo and Toubia, 2024). In this paper, we consider how different *type*s making different choices in *style* features in profile images, features characterized by being easy to alter (particularly in comparison with other features listed in this paragraph), can explain disparities by *type*.

Furthermore, our paper uses a model to study how a policy encouraging all users to change mutable *style* features can both improve efficiency and decrease disparities compared to policies aimed directly at either objective. This result relates to the vast literature on improving the efficiency of algorithms, and recommendation systems in particular.[7] More relevantly to our paper, there is also evidence on how algorithms impact disparities by *type*. In one example, Zhang et al. (2021) shows that Airbnb's smart-pricing algorithm decreased the white-black host earnings gap by 71%. A more common theme in the literature is that algorithms may create or exacerbate disparities by *type*.[8] For

[7] See, for example, Lin et al. (2023) for a recent survey on how large language models can improve recommendation system efficiency.

[8] See Williams et al. (2018) for a discussion of how a lack of data can cause algorithms to discriminate in a variety of ways;

example, Lambrecht and Tucker (2019) shows that a gender-neutral ad for job opportunities in science, technology, engineering and math (STEM) fields was shown to fewer women than men by an algorithm optimizing for cost-effectiveness because young women are a more expensive demographic for ads.

Current approaches for addressing algorithmic disparities by *type* generally intervene by changing either the training data, the algorithm, or the testing data post-processing (Mehrabi et al., 2021).[9] Interventions that change training data use a technique such as reweighing or resampling the data to remove the disparity (Kamiran and Calders, 2012). Many approaches have been proposed that intervene at the algorithm level, such as algorithms that integrate fairness constraints (Naghiaei et al., 2022), fairness regularizers (Berk et al., 2017; Wang et al., 2024), and fairness-efficiency trade-offs (Wang et al., 2021). With respect to altering testing data post-processing, these interventions take the output of a model and alter it to reduce disparities through, for example, modifying embeddings that associate the words "receptionist" and "female" (Bolukbasi et al., 2016). Instead of following one of these three approaches, Kleinberg et al. (2018) proposes that interventions should be implemented post-estimation, where a policymaker can take the output of a model that optimizes efficiency and use it in a way that reduces disparities, for example, implementing minority-specific cut-offs for college admissions. In this paper, we propose an intervention that leverages an existing managerial tool for Kiva and other platforms, *style* recommendations for profile images, to implement data-driven policy decisions that address disparities by *type*, disparities that exist given the training data, algorithm, and post-processing.

On Kiva, three papers have studied how to reduce disparities in terms of equitable funding outcomes, generally measured through funding to underfunded projects. Closest to our paper, Burke et al. (2022) interferes outside of typical avenues through adding a "slate," meaning a horizontal list of scrollable loans, of underfunded loans directly on the Kiva website, studying the impact on "adds to basket" (ATBs) to proxy for loans funded. They find that this additional slate of underfunded loans does not impact the ATBs on other slates, while attracting more than twice as many ATBs as the control slate from anonymous users. Two other papers study how algorithmic changes impact underfunded loans using collaborative filtering (Lee et al., 2014), and a combination of classification and the ϵ-greedy algorithm (Hapek, 2021).

Mehrabi et al. (2021) for a detailed description and categorization of different types of biases that can caused or exacerbated by algorithms.

[9] Chen et al. (2023) describe a similar categorization of interventions aimed at debiasing recommendation systems, namely a feedback loop involving data collection, model learning, and user serving.

Another paper that bears similarities to ours is Ludwig and Mullainathan (2024), which uses a combination of machine learning to identify facial features in mug shots and humans on mTurk to label those features to generate testable hypotheses about disparities in judicial outcomes by facial features unrelated to type, like gender and age. They find that features previously studied in the literature, both *type* and other features that may be mutable (e.g., attractiveness, appearance of dominance), can explain about 22% of the predicted variation in judge's detention rates, demonstrating that features identified through deep learning have predictive power beyond these previously studied features. Sisodia et al. (2024) applies a similar methodology to identify impactful features of watches. Our approach is from the opposite direction, testing hypotheses suggested by our observational data analysis.

One of the novel elements of this study is the use of generative AI, in the form of Generative Adversarial Networks (GANs), to create artificial profiles that vary a single feature of an image while holding other features fixed, in particular, *type* features like gender. Other approaches that have been used in the experimental economics literature to signal *type* features like gender and race include: names (e.g., Bertrand and Mullainathan (2004)); images (e.g., Andreoni and Petrie (2008)), and subjects' physical identity in laboratory experiments (e.g., Reuben et al. (2014)). Generative AI tools like GANs provide the experimenter with the opportunity to finely control the content of an image, offering new avenues of research, particularly with respect to experiments aimed at isolating mechanisms that may contribute to disparities by *type* and that also may be observed or altered in images. Other work that uses GANs in an approach similar to the one described in our working paper (Athey et al., 2022) include Luo and Toubia (2024) and Sisodia et al. (2024). Dash et al. (2023) surveys several applications of GANs to images, ranging from astronomy (e.g., to generate realistic simulations of deep space) to marketing (e.g., to generate many variations of a logo) to fashion design (e.g., to generate clothing designs).[10]

3 The Setting, Institutional Background and Data

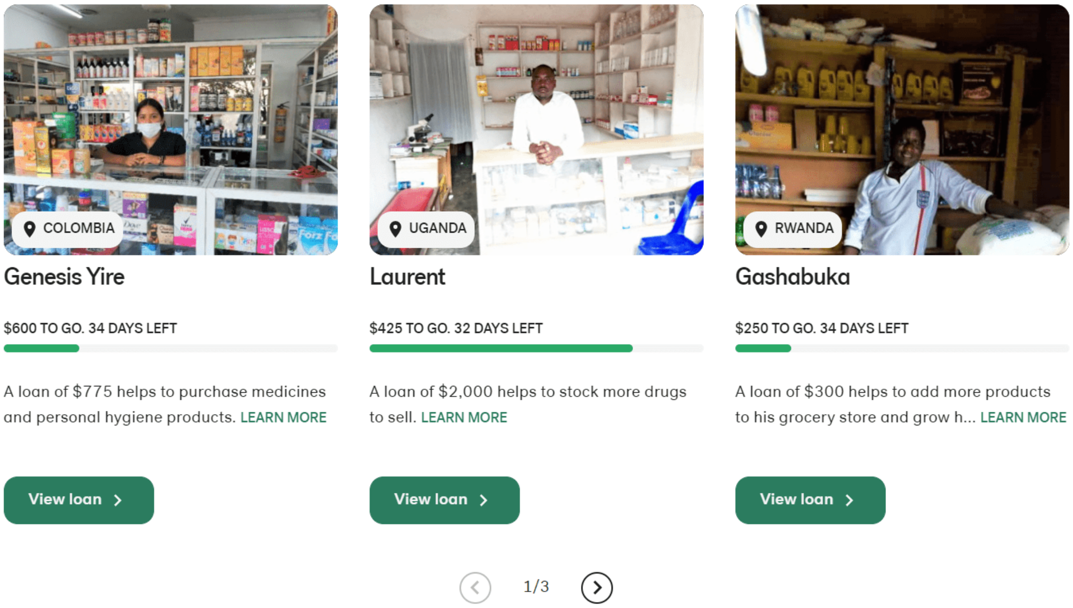
Kiva is one of the most prominent online, non-profit, peer-to-peer microcredit platforms.[11] Serving borrowers in more than 80 countries, Kiva has issued over 1.6 million loans, funded by more than 2 million lenders, totaling $1.7 billion since its founding in 2005. On the Kiva marketplace, borrowers

[10] See also Jabbar et al. (2021) for additional applications of GANs in a wide variety of other domains.

[11] Zidisha, Lend with Care, and Lend a Hand are other major peer-to-peer microfinance platforms sharing many features with Kiva.

have individual profile pages featuring pictures that prospective lenders can browse when selecting borrowing campaigns to invest in. Kiva collaborates with local microcredit agencies to vet, curate, and promote borrowers. The platform aims to enhance efficiency, measured by the total number of transactions, while also reducing disparities, defined as achieving a more equitable distribution of funds across borrowers (Burke et al., 2022).

Figure 1: Kiva category page.



Genesis Yire

$600 TO GO. 34 DAYS LEFT

A loan of $775 helps to purchase medicines and personal hygiene products. LEARN MORE

Laurent

$425 TO GO. 32 DAYS LEFT

A loan of $2,000 helps to stock more drugs to sell. LEARN MORE

Gashabuka

$250 TO GO. 34 DAYS LEFT

A loan of $300 helps to add more products to his grocery store and grow h... LEARN MORE

Note: Screenshot from kiva.org collected on 3/3/2022.

Images play a significant role in the way lenders discover borrowers and help borrowers convey the reason they need a loan. When searching for prospective borrowers, a potential lender typically begins on the category page, as illustrated in Figure 1. By clicking on "View loan," lenders can access more information about the loan's purpose and the geographical location of the borrower. These details are crucial for lenders when deciding whether to invest (Park et al., 2019).

Borrowers submit photographic images that can vary greatly in quality, content, and composition. Some images primarily feature the borrower, while others showcase the borrower's business. Additionally, the facial expressions of borrowers—whether serious or smiling—as well as technical elements like lighting and resolution can differ noticeably. To assist borrowers in presenting themselves effectively in this important application component, Kiva offers several recommendations. These include using high-resolution photos, ensuring a horizontal orientation, and incorporating both the business owner and their business in the background.[12]

3.1 Kiva Data

We construct *Kiva data* by combining three datasets: a publicly available dataset with loan characteristics and lending outcomes, a dataset that captures features in images associated with the borrowing campaigns, obtained using the methodology described in Appendix A, and a dataset on repayments provided by Kiva.[13]

Public Available Data. The publicly available dataset covers borrowing campaigns from April 2006 to May 2020, with over 500,000 observations. Each observation corresponds to a borrowing campaign, the primary unit of analysis. It includes key attributes such as sector, name of activity, country, loan amount, and repayment schedule. We also observe when the campaign was posted, from which we construct weekly and monthly time fixed effects and interaction terms between month and sector and month and country to control for sector- and country-specific fluctuations in loan availability and demand. We observe all borrowing campaigns that were active on Kiva at a point in time. We construct measures of the competition from other borrowers, and lender behavior, we construct measures of market conditions, including the total number of borrowing campaigns available at a given time, the number of concurrently listed borrowers from the same country and sector, and the number of borrowers of the same race and gender. At the time our data was collected, Kiva was displaying borrowers chronologically within categories. We thus include interactions of time-fixed effects and categories to proxy for exposure. Additionally, we construct a measure of lender supply, specifically the number of active lenders per week.

For funding outcomes, we focus primarily on money collected per day, as it captures how lenders allocate capital among competing campaigns. We also examine the number of days it took to raise the capital (campaigns generally stay active until they collect all funds), and the number of lenders that loaned money to the borrower.[14]

Image Data, *Type-Style* Classification, and Extraction. Borrowers on Kiva can upload profile images, which are publicly visible to prospective lenders. We process these images using a Convolutional Neural Network (CNN) to extract structured visual attributes, which we classify into *type* and

[12] See https://www.kivaushub.org/profile-photo.

[13] See here: https://www.kiva.org/build/data-snapshots for the publicly available dataset.

[14] In Appendix G.1, we expand the set of outcome variables and consider a constructed variable, which adjusts for differences across requested loans and funding success across categories.

style features.[15]

We distinguish *type* and *style* features based on the effort required to change them. *Type* features, such as age, race, and facial structure, are intrinsic and biologically determined, making them difficult to alter without significant intervention. Although borrowers might alter their appearance in a photograph, such a misrepresentation might incur psychological costs,[16] and the intermediaries who screen borrowers might object. In contrast, *style* features are extrinsic and more easily modified, varying across images due to individual choices, environmental conditions, or temporary factors. Because *style* features can be adjusted with relatively low effort, they are particularly relevant for platform interventions, as platforms can influence them through guidance on image composition and presentation.[17] In our analysis, we posit that the platform, acting as the decision-maker, possesses the capability to discern which image attributes are managerially pertinent and which are not. Some features may be deemed unsuitable for modification due to the substantial effort required from borrowers or the potential for such recommendations to be perceived as insensitive or politically incorrect.

Examples of features that we classify as *style* include: *No Eyewear*, *Sunglasses*, *Smile*, *Blurry*, *Eyes Open*, *Mouth Wide Open*, *Harsh Lighting*, *Flash*, *Soft Lighting*, *Outdoor*, *Partially Visible Forehead*, *Color Photo*, *Posed Photo*, *Flushed Face*, *Top* (person's face in the top part of the image), *Right* (person's face in the right part of the image), *Bottle* (there is a bottle in the image), *Chair* (there is a chair in the image), *Person* (there is another person in the image), and *Body-shot* (the body of the borrower occupies a substantial part of the image).

The CNN assigns probability scores to approximately 140 features, including object presence (e.g., *cup*, *chair*), technical aspects of the image (e.g., *blurry*, *flash*), facial expressions (e.g., *smiling*, *frowning*), and demographic characteristics such as *race* and *age*. We filter out features that appear in fewer than 0.01% of images and remove highly correlated features (those with a Pearson correlation coefficient above 0.75, such as *smiling* and *frowning*). After these steps, we retain 55 key features for analysis (full list in Appendix A). Throughout the paper, we use *italics* when referring to demographic features predicted using CNN.

[15] In the field of computer vision, it is customary to differentiate between mutable and immutable aspects of facial images. This distinction is often guided by frameworks such as those provided by the Facial Identification Scientific Working Group (FISWG), which offers comprehensive guidelines for facial comparison methodologies `https://www.fiswg.org/fiswg_facial_comparison_overview_and_methodology_guidelines_V1.0_20191025.pdf`

[16] For example, there is robust evidence from the field (Abeler et al., 2014) and lab Kajackaite and Gneezy (2017) that there are costs to lying.

[17] See (`https://www.kivaushub.org/profile-photo`) for Kiva guidelines for profile images highlighting image composition and style, and (`https://www.airbnb.com/resources/hosting-homes/a/how-to-take-a-great-airbnb-profile-photo-581`) for Airbnb's recommendations focused on facial expressions and composition.

CNN-based feature extraction introduces the potential for misclassification, particularly for complex or ambiguous image attributes. To assess the accuracy of our image-based features, we conduct an audit study in which we compare algorithmic predictions with human-annotated labels. The results indicate high correlation between human judgments and CNN predictions but reveal a greater incidence of false negatives than false positives, meaning that the algorithm is more likely to miss features that are actually present rather than falsely detecting them. This tendency suggests that our estimates based on image features may understate their actual impact on funding outcomes. We assess the implications of this potential bias in Appendix B.

Loan Repayment Data and Defaults. The repayment dataset spans 2006–2016, covering approximately 420,000 borrowing campaigns. In this dataset, each observation corresponds to an individual lender's contribution to a borrower. We aggregate this data to construct an indicator for whether a campaign was fully repaid, meaning all lenders who contributed received their funds back; 95% of campaigns are fully repaid. Defaulted loans are categorized as follows: (i) borrower default (75%), where the borrower fails to repay; (ii) microfinance partner default (23%), where the intermediary organization managing the loan defaults; and (iii) joint default (2%), where both the borrower and the microfinance partner default. Since lenders face the same financial consequences regardless of the reason for default, we treat all defaults equivalently in our analysis. In Appendix H, we carry out the analysis across default categories.

Data Merging, Platform Adjustments, and Considerations. We merge the three datasets to create a panel covering 2006–2016, enabling us to track borrowing campaigns, funding dynamics, and repayment outcomes over time. Since Kiva has undergone multiple platform design changes during this period, we include time-fixed effects in all our analyses to account for structural shifts in borrower composition, lender behavior, and macroeconomic conditions.

Several limitations should be noted. For instance, while we observe lender funding decisions, we do not have direct data on how lenders search for or browse campaigns, which limits our ability to account for visibility effects. Additionally, our repayment data extends only through 2016, restricting our analysis of long-term repayment patterns.

Table 1 presents summary statistics for the key variables, and a full list of covariates is provided in Appendix E.

Table 1: Summary statistics of the main variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
cash per day	420,765	104.587	136.378	1	25	116.7	621
days to raise	420,765	13.175	10.947	1	5	20	38
default	420,765	0.050	0.218	0	0	0	1
loan amount	420,765	800.107	993.370	25	275	950	50,000
no. competitors	420,765	0.091	0.173	0.003	0.006	0.075	1.000
share same race and gender	420,765	0.665	0.294	0	0.4	1	1
male	420,765	0.198	0.398	0	0	0	1
smile	420,765	0.498	0.177	0	0	1	1
body-shot	420,765	0.406	0.491	0	0	1	1

Note: Summary statistics of selected variables. Cash per day and days to raise are Winsorized at the top 97th percentile. Cash per day and loan amount are in USD dollars; male and smile take the value of 1 when CNN predicted probability is above 0.5 and zero otherwise. No. competitors is the number of borrowing campaigns from the same sector and country posted concurrently; the value is standardized by the maximum.

4 Funding Outcomes and Profile Features in Kiva Data

This section presents the first step in our two-step approach to understand how *type* and *style* features contribute to platform efficiency and type-based disparities. The end product of this step is a prioritized list of style features for interventions in the second step, recruited experiments, discussed in Section 5.

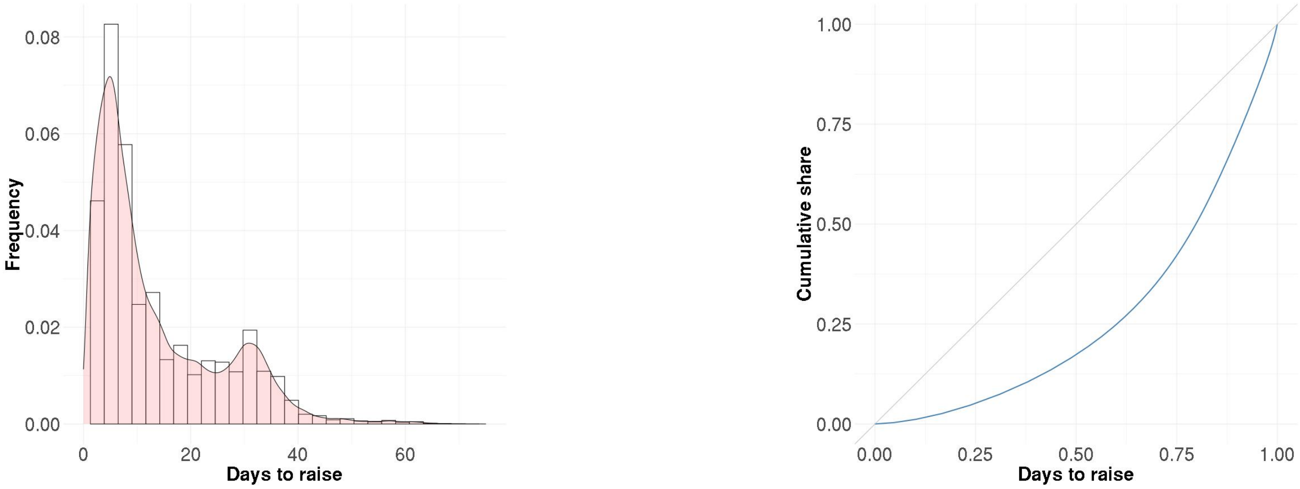
4.1 Disparities in Funding Outcomes

While loans on Kiva stay active for a long time, so that the vast majority of them eventually get funded, there is substantial variation in how long it takes to reach a campaign's funding goal or how much money is collected per day. In Figure 2, we show a histogram of the number of days it takes to collect the entire amount (*days to raise*) and a Lorenz curve documenting inequity in this outcome. If every borrowing campaign took the same number of days to get funded, the blue (actual distribution) and gray (perfect equality) curves would overlap.

From the left panel of Figure 2, we can observe that there is substantial dispersion in how long it takes to collect the entire loan amount. The mean outcome is 14.5 days, but many campaigns are entirely funded almost immediately, while others take over a month to reach their funding goals.

An important driver of how long it takes to raise all the funds is the loan size. Thus, the amount raised per campaign per day is a useful measure of how quickly borrowers raise funds. In Figure 3, we show a histogram of funds in dollars collected per day (*cash per day*). The y-axis shows the share

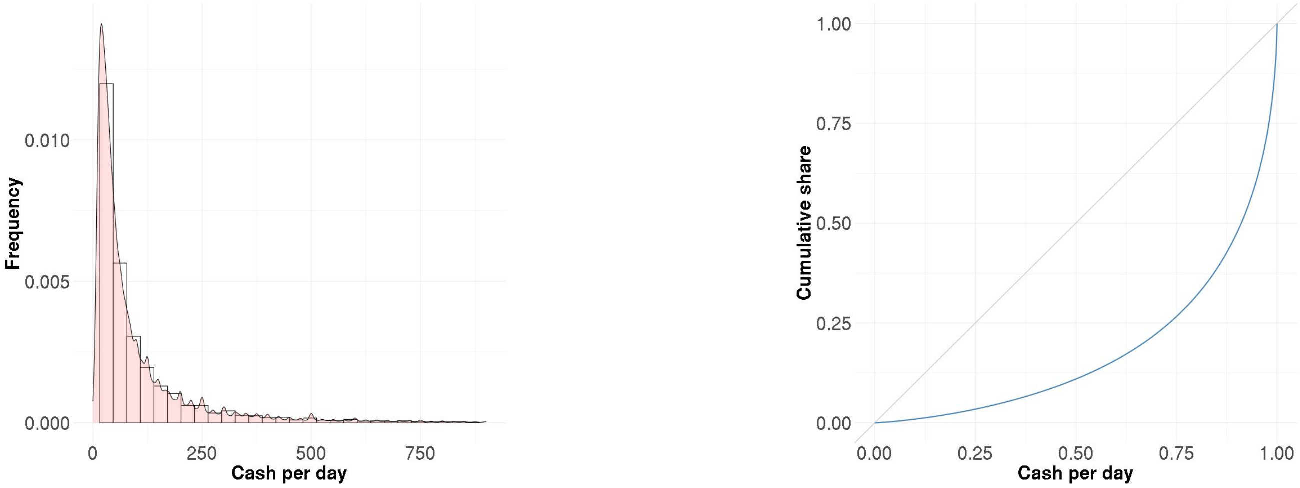
Figure 2: *Days to raise*: histogram and Lorenz curve



Note: Left panel - histogram of days to raise capped at 75 USD. The fitted density curve is shown in pink. Right panel - Lorenz curve of days to raise.

of campaigns attracting the sum of money shown on the x-axis and an associated Lorenz curve. There is substantial variation in *cash per day*. The mean is 118 USD, but many campaigns raise just a few dollars per day. Focusing on the Lorenz curve (right panel), we observe even higher dispersion than in Figure 2.

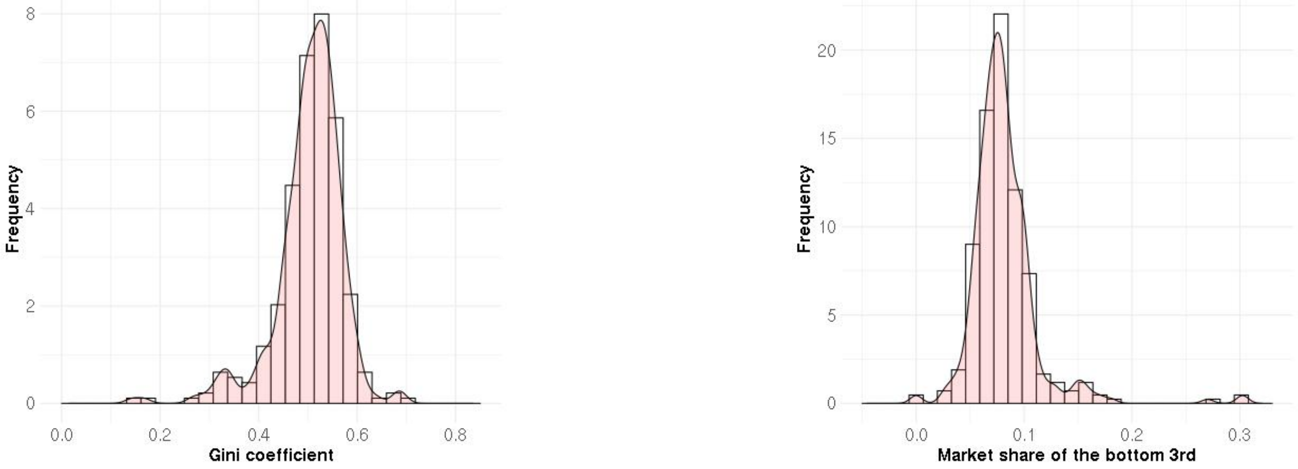
Figure 3: *Cash per day*: histogram and Lorenz curve



Note: Left panel - histogram of cash per day capped at 1250 USD. The fitted density curve is shown in pink. Right panel - Lorenz curve of cash per day.

The evidence presented in Figures 2 and 3 is based on data collected over ten years. Time trends in factors such as differences in the number of available lenders or borrowers may explain some of the variation. In Figure 4, we group campaigns into weekly intervals such that campaigns that were listed online during the same week are in the same group. Thus, a group of borrowing campaigns approximates a choice set available to lenders that were active in that week.[18] We use two measures of disparities, the Gini coefficient and the sum of market shares of the 33% of borrowers with the lowest amount of money collected per day.[19] The Gini coefficient of 0 expresses perfect equality, where all values are the same. The market share of the bottom third amounting to 33% would indicate that the outcomes are equally distributed across tertiles. We can observe that both metrics reveal that outcomes are far from equally distributed.

Figure 4: *Cash per day* distribution within weeks: Gini coefficient and share of the bottom tertile.



Note: Statistics in both panels are computed on a weekly basis. Left panel - Gini coefficients of weekly distributions of cash collected per day. Right panel - weekly sums of cash collected per day by the 33% lowest performing borrowers.

First, consider disparities according to *type*s. *Men* raise $30.2 (± 0.9) less than *women*; *Senior* borrowers raise $57.2 (± 4.7) less than *young* borrowers; *Black* borrowers raise $12.1 (± 1.4) less than non-*black* borrowers. In the following sections, we decompose the disparities between *type*s to characterize the parts of these gaps that are due to differences in *style*.

[18] We say that this only approximates the choice set of lenders because, in some cases, a borrowing campaign could have been posted at the beginning of the week, quickly collect all the funds and disappear from the platform. Thus, a lender active only at the end of the week would not see such a campaign. On average, there are 450 borrowing campaigns active in a given week, which raises, on average, over USD 400,000 in loans.

[19] The Gini coefficient is defined as

$$Gini = \frac{\sum_{j=1}^{n}\sum_{j'=1}^{n}|x_j - x_{j'}|}{2n\bar{x}}$$

where x_j is the outcome for borrower j and $x_{j'}$ for borrower j', n is the number of borrowers available in that week and $\bar{x}$ is the average cash collected per day.

4.2 *Style* Features and Funding Outcomes

Disparities in borrower outcomes can arise due to differences in borrowers' *type*s and their *style*. *Style* features are central to this analysis because a platform can design interventions to modify them. In Table 2, we show that part of the variation in *cash per day* can be explained by *style* features in images.

We train three models to predict *cash per day*. The first is a benchmark model with just an intercept, the second is a model that includes only *style* features, and the third is a full model incorporating all variables in the *Kiva data*. We use gradient boosted machine (GBM) (Friedman (2001)) as the predictive model.[20] The dataset is split 70:30 into train and test sets. Table 2 reports the predictive performance in the test set, measured using mean squared error. We find that including *style* features improves predictive performance compared to the mean model. Furthermore, the full model, which includes all covariates in the *Kiva data*, performs better than the model with only *style* features. Appendix G extends this analysis to an outcome variable that adjusts for systematic differences in *cash per day* and total loan amounts requested across business categories. The results indicate that differences in *style* are predictive of outcomes even within business categories. Appendix G also provides diagnostics for the GBM models.

Table 2: Image Features as Predictors of *Cash per day*.

specification	MSE	SE
Mean	22367	252
Style features	19373	224
Full model	10996	138

Note: Test set performance of a GBM trained using all available covariates (full model), models with only image style *features, and a mean model. Mean squared errors are in the second column. Standard errors of MSE are in the third column.*

Specific style features. Results presented in Table 2 show that *style* features are predictive of funding outcomes. However, if Kiva is to construct platform policies around *style* features, it is important to select individual impactful features. We want to know: "What would happen if a profile was presented with a change in one characteristic and remained unchanged otherwise." In other words, we want to know the average treatment effect (ATE) of a specific feature in an image.

To estimate ATEs we use the Augmented Inverse Propensity Weighing (AIPW) estimator (Robins et al., 1994; Glynn and Quinn, 2010). AIPW is a doubly robust method: it adjusts for covariates in

[20] Appendix F presents a comparison of the accuracy of the GBM to other predictive models.

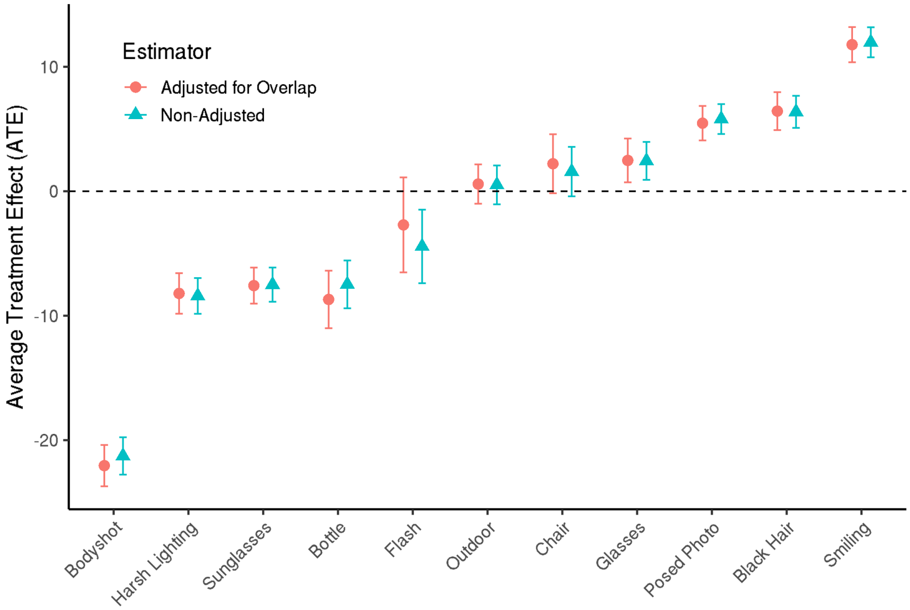
the outcome model and the propensity score. We use the *grf* implementation of the AIPW estimator (Athey et al. (2019)). We consider a rich set of covariates: we control for *(i.)* the requested amount of loan, the week in which the fundraising campaign was posted on Kiva, sector and country of the business, repayment schedule, time fixed effects and interaction terms between the month and sector and the month and country; *(ii.)* all *type* and other *style* features, *(iii.)* measures of the other loans available at the time on Kiva (we observe the entire choice set available to the lender), specifically, the total number of loans available and the number of loans from borrowers of the same race and of the same gender, *(iv.)* the number of lenders active on Kiva in the specific week to account for intertemporal differences in the supply of money. In the estimation of the effect of each feature, we drop other features that are very highly correlated, where it may be difficult to hold the highly correlated features fixed while changing the target feature. Then, we interpret our results as the treatment effect of the relevant feature and other features that covary strongly with it.

Some *style* features are observed only in specific subsets of the data, such as particular sectors or countries. As a result, it is not possible to credibly estimate the impact of these features on borrowers outside these subsets. To address this limitation, we estimate a propensity score function and exclude cases where either the treatment group (borrowers with the feature in the *Kiva* data) or the control group (borrowers without the feature) falls outside the range of common support. Specifically, we drop features for which a large probability mass of either group exceeds 0.9 or falls below 0.1. Appendix I.2 provides density plots illustrating examples of the features removed through this process. Furthermore, we also report estimates adjusted for overlap following the methodology of Li et al. (2018).

Figure 5 shows estimates of average treatment effects on *cash per day* for selected features. We find that several features have negative ATE, like *Body-shot* and *Sunglasses*, while others like *Posed Photo* or *Smiling* have positive effects.

From the perspective of a causal diagram or directed acyclic graph (DAG), certain image-extracted features may, in principle, be *caused* by a given *treatment feature*. In such cases, these features would moderate the effect of the treatment feature and should not be included as controls in an AIPW estimator. For instance, if *Sunglasses* is the treatment feature, then *Bags under eyes* may only be observed when *Sunglasses* = 0, making it inappropriate as a control. Similarly, if *Lighting* is the treatment feature, other features—such as *Age*—may only be detected by the algorithm under favorable lighting conditions. To assess the robustness of our estimates to this concern, we conduct a sensitivity analysis

Figure 5: Estimates of the Average Treatment Effect of Selected Style Features



Note: Estimates of the average treatment effect of selected features on cash collected per day with 95% confidence interval. The propensity and outcome model was estimated using Regression Forest. We transform the treatment variable to a binary variable that takes the value of one when the predicted probability of the feature is above 0.5 and zero otherwise.

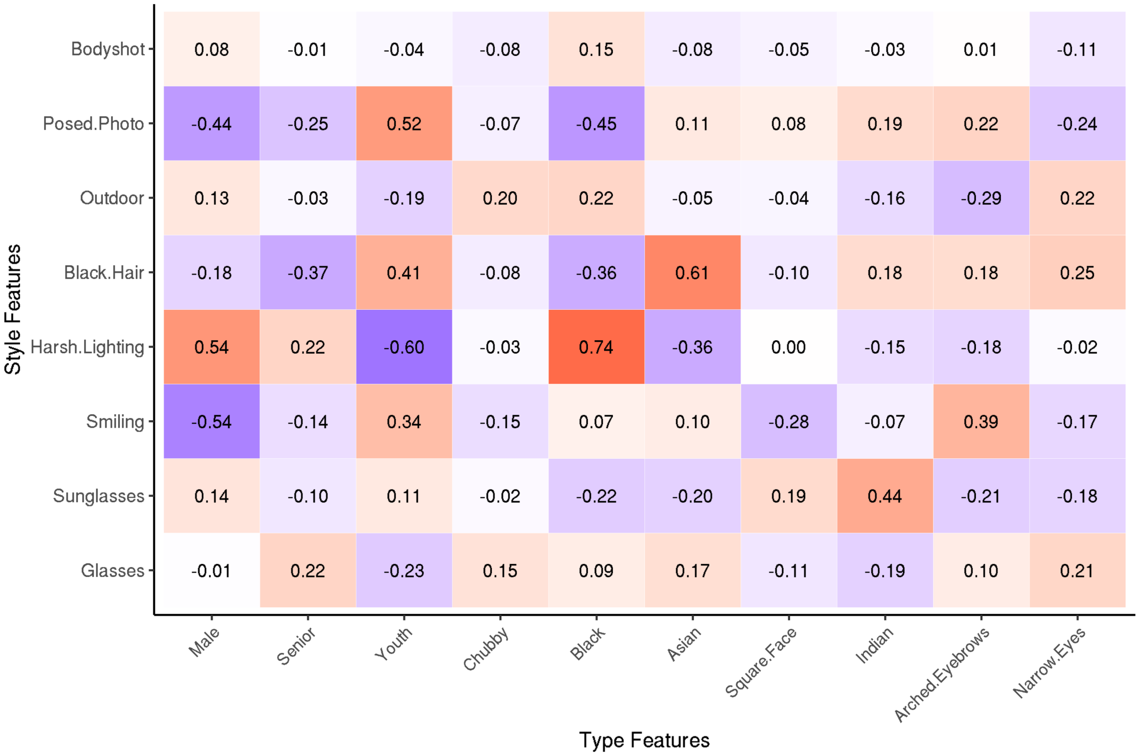
in Appendix J, where we exclude features potentially subject to this type of mechanism. The results indicate that our treatment effect estimates remain stable.

In the estimation of the ATE, we are assuming unconfoundedness. While we control for a rich set of borrower and market characteristics, we may still be missing some variables (both features in images and other variables) that might be correlated with the treatment variable and also influence lenders' decisions. Thus, our estimates should be interpreted as suggestive rather than definitive. The analysis of the observational data constitutes the first step in our approach; we use the results to select features to be tested in the recruited experiment, which has greater internal validity (omitted variable bias eliminated, in principle) at some cost to external validity (given the artificial context of the experiment).

4.3 *Style* Features Impact Disparities Between *Type*s.

Evidence presented in Section 4.2 suggests that borrowers' style choices impact their funding outcomes. These choices can aggravate or mitigate inequities in outcomes across *type*s. When borrowers with *type* features associated with high outcomes choose to include attractive *style* features in their profiles, the disparities due to *type*s will increase further; in contrast, if borrowers with less desir-

Figure 6: Correlation Between Selected *Type* and *Style* Features.



Note: Pearson correlation coefficient between selected type *features in columns and* style *features in rows.*

able *type* features choose attractive profile *style*s, outcomes will be more equitable. In this section, we document the correlation between *type*s and *style*s.

Figure 6 shows a correlation between selected *type* and *style* features. Some of the features are highly correlated; for example, *Smiling* is less common among *Male* and *Senior* and more common for *Youth*. Correlations presented in Figure 6 are unadjusted for other differences across *type*. To argue that the choices of *style* features exacerbate disparities due to *type*s, we need to show that the distribution of *type*s across borrowing campaigns results in disparities in outcomes and that the desirable *style* features are more prevalent amongst borrowing campaigns whose *type*s lead to better funding outcomes. To do that, we carry out a *Gelbach Decomposition* (Gelbach, 2016) of selected *type* variables. This is a method for measuring the extent to which adjusting for a group of variables changes the effect of a selected variable. It tells us what the estimated effect would be if the means of the adjusted variables were the same across the levels of the evaluated variables.

Table 3 presents the results for selected *type* variables. The first column shows the name of the variable. The second column shows the coefficient associated with the selected variable from a linear regression of the variable on *cash per day*. The third column shows the coefficient adjusted for all variables in *Kiva data*. In the final column, we have the adjustment due to *style* features; that is, the impact of *style* features on the observed, unadjusted difference in *cash per day*. For example, if we partial out differences in the distribution of *style* features, profiles with images of *bald type* receive USD 10.21 less than those of not-*bald type*. Thus, differentially distributed *style* features aggravate the disparity between *bald* and not *bald*.

Table 3: Gelbach decomposition of *type*-based disparities in *cash per day*

Feature	Coefficient base	Std. error base	Coefficient full	Std. error full	Delta style
Bald	-71.57	4.86	-22.14	7.31	-10.21
Chubby	16.92	2.02	-7.65	4.17	2.08
Narrow Eyes	-9.14	1.87	3.77	3.66	2.01
Square Face	-87.70	8.06	-27.64	10.46	-52.59
Black	-12.06	1.37	-1.61	3.51	4.61
Senior	-57.18	4.70	-6.02	5.96	-6.02
Attractive Woman	75.32	2.43	10.72	4.13	9.48
Asian	12.39	1.52	0.72	2.74	2.30

Note: Gelbach decomposition of selected type *features (Gelbach, 2016).* Coefficient base *refers to coefficient of a univariate model with the selected* type *feature;* coefficient full *is the coefficient from a model adjusting for all covariates in* Kiva data*;* delta style *is the impact of style features on the disparity between* types*. We use the R implementation by Stigler (2018).*

When the estimated impact of *style* (delta style) has the same sign as the unadjusted difference between *types* (coefficient base), *style* increases the disparity between groups; if the signs are opposite, the gap is decreased. Results presented in Table 3 indicate that *style* choices aggravate disparities due to *bald, square face, senior, attractive woman*, and *Asian types* and mitigate for *chubby, narrow eyes*, and *black types*.

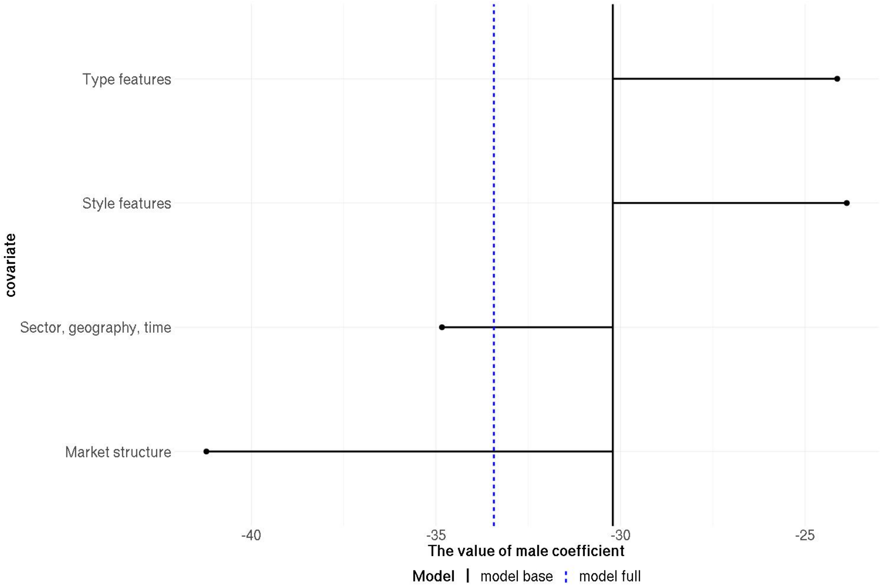
***Gender* gap.** A specific *type* feature that matters in our context is *gender*.[21] We find that campaigns classified as *male* raise on average USD 36 less per day and take 5.8 more days to be funded fully (differences in means).[22]

The results presented in Table 3 show that the differences between *types* can be partially accounted for by differences by type in the distributions of other characteristics, e.g., *style*. To decompose the un-

[21] We use an algorithmic prediction of *male*. Thus, the variable *male* indicates that the feature detection algorithm assigns a probability of at least 0.5 that the person in the image is a male.

[22] In the context of microfinance, the gender gap might be driven by users that aim to correct for discrimination against women in traditional finance. There is a rich literature documenting discrimination against women in traditional entrepreneurial lending. Alesina et al. (2013) shows that women entrepreneurs pay higher rates for access to credit and Brock and De Haas (2021) use a randomized experiment to show that loan officers grant loans to women under less favorable conditions than to men. The phenomenon of over-correcting for discrimination is well documented in experimental psychology (Mendes and Koslov (2013), Nosek et al. (2007)). It is also plausible that Kiva lenders follow broader policy discussions, where the emphasis on developmental policies and aid targeting women is common (Kristof and WuDunn, 2010). Furthermore, Ozer et al. (2023) show that peer-to-peer microlending can effectively advance such social goals.

Figure 7: Gelbach Decomposition of *Male* Coefficient for *cash per day*



Note: The solid line is an estimate of the coefficient associated with male *from a univariate linear regression; the dashed line is the coefficient adjusted for all variables in Kiva data using an OLS estimator. Horizontal lines represent contributions of the variables group to the coefficient associated with* male; type *features include all other* type *features from the image;* sector, geography, time *includes* sector, country, *and* week *fixed effects,* loan amount *and* repayment details, *and* market structure *features including interactions of* month *and* country, month *and* sector, number of lenders in the week, number of competing campaigns, *and* share of campaigns of the same race and gender. *We use the R implementation of Gelbach (2016) by Stigler (2018).*

adjusted difference we perform another Gelbach Decomposition (Gelbach, 2016), where we compare a baseline model with only a *male* indicator variable to a full model that includes all the variables in the *Kiva data* and measure the contribution of each added variable to the change in the coefficient of interest. Figure 7 presents the results.

Figure 7 depicts the differences in the coefficient associated with *male* between a univariate linear regression (solid line) and a full model which includes all variables from the *Kiva data* (dashed line). The length of each horizontal line indicates the contribution from adjusting for the corresponding variable group to the *male* coefficient in the full model; in other words, it is the partial effect of the unequal distribution of features within the group. We observe that changing the distributions of *style* features decreases the gender gap; additionally, we find that *male* campaigns also have an undesirable distribution of other *type* features, but the distribution of *sector, geography, time* and *market structure* decreases the *gender* gap.

We further analyze the prevalence of the desirable and non-desirable features. For example, the frequencies of *body-shot* and *smile* differ substantially between genders: 77% of *female* borrowers *smile* in the image, compared to 33% of *male* borrowers. 26% of *male* borrowers use a *body-shot*, compared to

Table 4: Image Features as Predictors of Repayment Probability.

Specification	MSE	SE
Mean model	0.065	0.001
Style features	0.064	0.001
Full model	0.059	0.001

Note: Test set performance of a gradient boosted machine (GBM) trained using all available covariates (full model), a model using image style features (Style features) and a model with only an intercept (Mean model). Models trained on 70% of data and tested on 30%. Mean squared errors are in the second column. Standard errors of MSE are in the third column.

22% of *female* borrowers (both are statistically significant).

4.4 Potential Mechanisms Linking Profile Features to Outcomes

Lender decisions about funding on Kiva potentially respond to both monetary and non-monetary considerations, and both *type* and *style* features can impact them. A lender's utility function can be expressed as a combination of the expected financial return from the loan and the non-monetary benefits of supporting a borrower they prefer. Borrower images may influence non-monetary benefits by evoking psychological and emotional responses. In this section, we explore both mechanisms.

Monetary benefits. The monetary benefit of selecting a specific borrower can be proxied by the probability of loan repayment, assuming lenders prefer borrowers with a higher likelihood of repayment. Repayment probability varies across loan size, country, and sector of activity, but it could also be influenced by profile image features. To evaluate this, Table 4 compares the predictive performance of three GBM models trained to predict loan repayment. The first model is a baseline model with only an intercept, the second includes *style* features, and the third incorporates all available covariates in the Kiva data. The results show that including *style* features does not improve predictive performance compared to the baseline model. This suggests that *style* features do not contain information about repayment probability.[23] Appendix H examines the predictive power of specific *style* features.

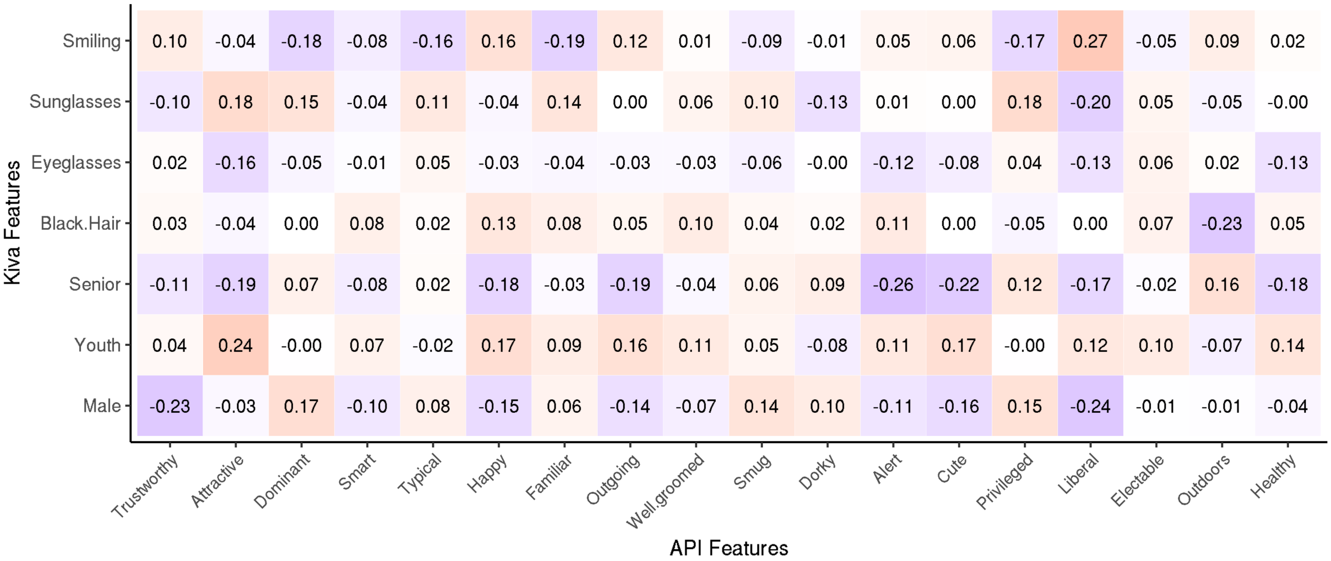
Non-monetary benefits. Borrower images may also evoke psychological traits that influence lenders' preferences. For example, lenders may perceive a borrower who is *Smiling* as more trustworthy or positive. Advances in deep learning have enabled the inference of psychological traits from facial attributes based on human judgments. For instance, Peterson et al. (2022) trained a model on over one

[23] Our findings contrast with studies on the Prosper crowdfunding platform, which show that free text in loan listings can predict default probability even after adjusting for financial and socio-demographic characteristics (Herzenstein et al., 2011; Netzer et al., 2019).

million judgments to infer more than 30 traits. Using this model, we estimate psychological traits from borrower images.[24] While the Peterson et al. (2022) model was trained on a general sample of judges rather than Kiva lenders, it provides a useful approximation of how facial attributes may influence perceptions.

Figure 8 illustrates the correlation between features of Kiva images, estimated using our CNN, and psychological traits inferred from the Peterson et al. (2022) model. The results indicate significant correlations between these attributes. For instance, *Male* borrowers are perceived as less *Trustworthy*, more *Dominant*, and less *Smart*.

Figure 8: Correlation Between Image Features and Psychological Traits



Note: Pearson correlation coefficients between selected profile image features and psychological traits inferred using the Peterson et al. (2022) model.

We find no evidence that *style* features predict repayment probability, suggesting that rational lenders should not use them as signals for loan performance.[25] However, *style* features are strongly associated with psychological traits, supporting the hypothesis that lenders choose borrowers whose images align with their preferences due to the psychological responses these images evoke.

[24] We thank the authors for providing API access to their model.

[25] The literature highlights the issue of inaccurate statistical discrimination, where decisions are based on signals that are not predictive of outcomes (Bohren et al., 2023). It is possible that Kiva lenders misinterpret signals from *style* features as informative of repayment probability.

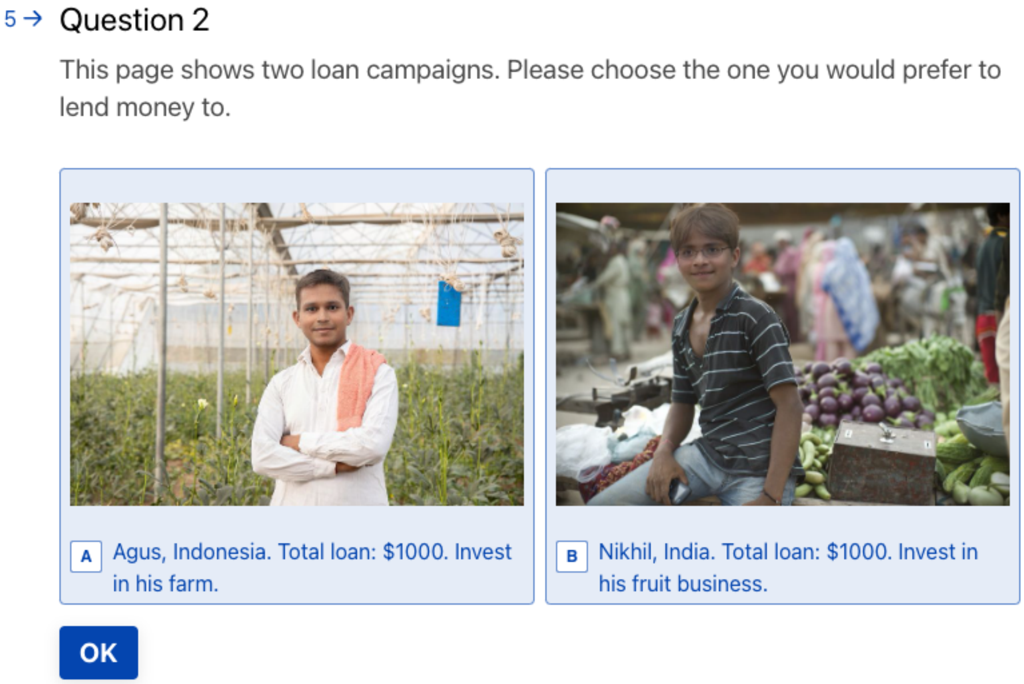
5 Recruited Experiments

Our analysis of the *Kiva Data* identifies a set of candidate *style* features for intervention based on their potential to explain disparities in funding outcomes across *type* features. In the second step of our approach, we conduct two recruited experiments designed to causally estimate the effects of a subset of profile image features on the probability of being selected by potential funders. These experiments also validate which *style* features identified in data from the Kiva platform are good candidates for policy intervention by providing internally valid treatment effect estimates.

5.1 Experiment Design

Both experiments employ a conjoint methodology, where the experimental stimuli have variations in selected image features.[26] Recruited participants are presented with pairs of simulated fundraising campaigns and asked to select the profile they prefer. Figure 9 illustrates an example of a choice instance shown to participants during the experiment.

Figure 9: Example of a Choice Instance



5 → Question 2

This page shows two loan campaigns. Please choose the one you would prefer to lend money to.

A Agus, Indonesia. Total loan: $1000. Invest in his farm.

B Nikhil, India. Total loan: $1000. Invest in his fruit business.

Note: This figure presents an example of a choice instance from the recruited experiment. Subjects were instructed to select between two fundraising campaigns based on the provided profiles.

The profile pictures of hypothetical borrowers displayed to participants are generated using fine-

[26] For a comprehensive review of conjoint design and its underlying assumptions, see Hainmueller et al. (2014).

tuned GANs. The resulting images closely resemble actual Kiva borrower profiles while systematically varying specific features of interest.[27] This approach ensures that paired images are identical except for the feature being analyzed, allowing us to isolate its causal effect on selection probability. The treatment effect for each profile is defined as the difference in the probability of being chosen between the image variant with the feature and the variant without it. By aggregating these differences across all profiles, we estimate the average treatment effect.

Both experiments were conducted on Prolific.com, where we recruited 410 participants in the first experiment and 436 for the second one.[28] All subjects were fluent in English and had donated money to a charitable cause within the past year. Each session lasted approximately five minutes. Attention checks were included to ensure data quality, and participants generally performed well on these checks (details in Appendix L). Appendix P provides covariate balance checks in the two experiments as well as mean outcomes across base Shutterstock images.

Although the experimental design ensures internal validity of causal estimates, several limitations remain. First, the images used in the experiment differ from those typically found on Kiva. To assess the similarity between GAN-generated and actual Kiva images, we compared the distribution of psychological trait scores for both sets of images and found substantial overlap in distribution with similar means and standard deviations (Appendix D). Second, our participant pool may not fully represent Kiva lenders, so the preferences observed in the experiment may not perfectly mirror those of actual users.

5.2 Experiment Implementation

Experiment 1. In the first experiment, we estimate the impact of *Smile* and *Bodyshot*, two stylistic features that demonstrate high treatment effect estimates in the Kiva data. These features are also correlated with several *type* characteristics, most notably gender: *Male* profiles are less likely to *Smile* and more likely to feature *Bodyshots*. To explore these relationships experimentally, we exogenously vary the gender of the borrower in the generated images.

Subjects began the experiment by receiving a brief introduction to the concept of microloans. They were then presented with six pairs of borrower profiles and asked to select their preferred option from

[27] To address privacy and ethical concerns, we do not alter or use images of actual Kiva borrowers, as their consent for such modifications cannot be obtained. Instead, we purchase images from *Shutterstock.com*, selecting those that closely match the characteristics of Kiva profiles. These images are then used to train a deep learning model, which enables controlled modification of selected features.

[28] The slight differences in the number of subjects are due to some recruited subjects not consenting to participate in the study or dropping out before completing it.

each pair. We kept the number of questions per subject relatively low to mitigate the risk of subject being fatigued and responding to the later questions with less deliberation; in doing so we follow the recommendation of Hainmueller et al. (2014).

To generate the experimental stimuli, we first created a pool of images. Starting with 20 original Shutterstock images resembling Kiva borrower profiles, we generated artificial versions by varying the *Smile*, *Bodyshot*, and *Male* features. This resulted in eight variants of each base image. All images used in the experiment were artificial, generated using a pre-trained GAN model. Thus, participants were always choosing between two fabricated profiles. We use the term "profile" to refer to all eight variants derived from the same original image.

Experimental protocols were created by pairing images systematically. For each protocol, we randomly drew one image without replacement and paired it with another image that differed in at least one feature and was not derived from the same base image. This process was repeated to create six pairs per protocol.[29] Our goal was to isolate the effect of borrower profile image features on funding choices, and so we did not vary any other aspect of borrower profiles beyond their style and type features in profile images. In total, we generated 15 protocols, which were randomly assigned to participants.

Experiment 2. The second experiment builds on the design of the first, introducing a key modification: the inclusion of financial incentives to ensure real-world consequences for subjects' decisions. This alignment of incentives encourages more deliberate consideration of profile choices. Subjects may otherwise have limited motivation to pay close attention. Each experimental protocol included a pair of profiles featuring actual Kiva borrowers.[30] Subjects were informed that some profiles represented real borrowers currently active on Kiva.org and that we would loan $10 to one borrower selected by the subject from this subset. This design ensured that participants recognized the potential impact of their decisions.

In this experiment, we expanded the set of features under analysis; we examined the effects of *Sunglasses*, *Glasses*, *Dark Hair*, and *Age*. Among these, *Glasses*, *Sunglasses*, and *Dark Hair* are *style* choices, while *Age* is a *type* feature. We selected these features because our analysis of the *Kiva data* established that they have a substantial impact on funding outcomes.

[29] We intentionally did not present the same base image with different style variations to the same respondent since we believed that repeatedly showing the same profile with different attributes could reduce the realism of their task.

[30] Kiva borrower profiles were obtained using the Kiva developer API: https://www.kiva.org/blog/introducing-the-kiva-api.

We generated artificial profiles by modifying Shutterstock images using a pre-trained GAN algorithm to resemble Kiva borrower profiles. Each base image was altered to create multiple variants, systematically varying the features of interest. For each protocol, we included nine (as opposed to six in Experiment 1) pairs of profiles: eight pairs consisted of GAN-generated profiles differing in one or more features, and one pair featured actual Kiva borrowers (Kiva borrowers were randomly drawn from the pool available in Kiva API). Subjects were randomly assigned to protocols and asked to select their preferred profile from each pair; they were not informed which pair were actual Kiva borrowers.

5.3 Generated Images Used in the Experiment

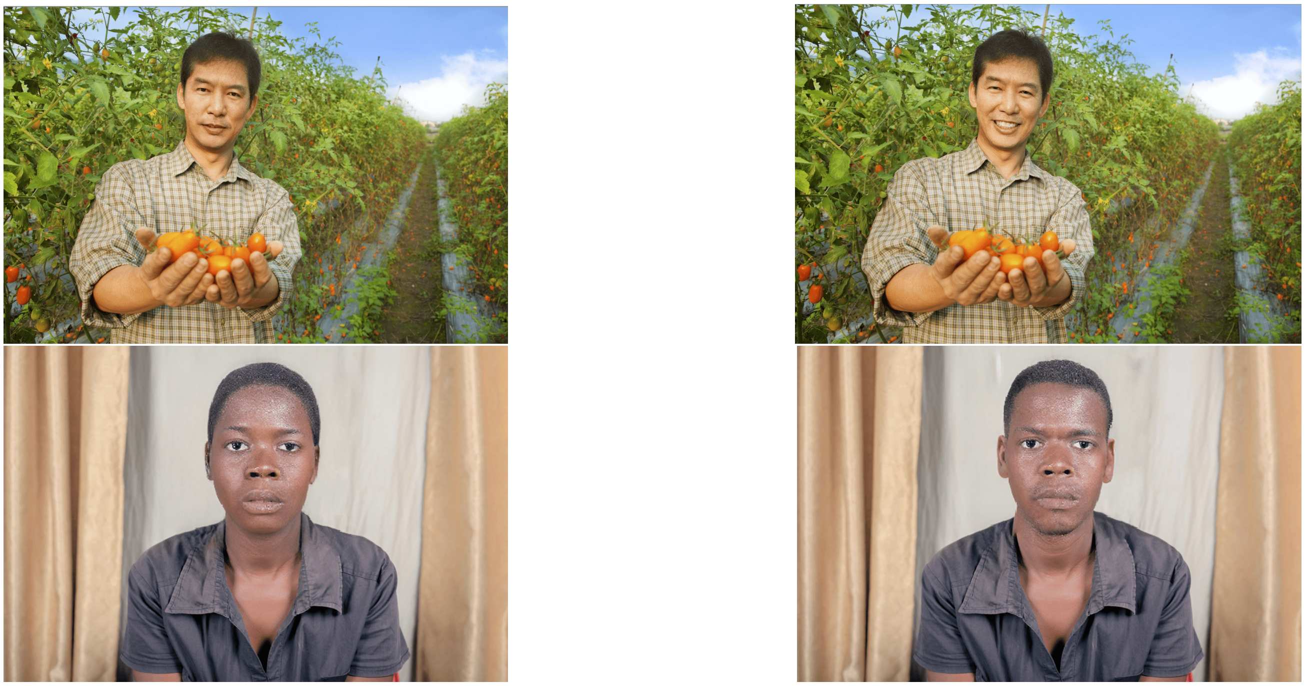
To generate images that differ in our specified features, we used Generative Adversarial Networks (GANs), a deep-learning-based approach to generative modeling introduced by Goodfellow et al. (2014). GANs generate data that closely resemble the original distribution and have been applied in the social sciences to create realistic images (Ludwig and Mullainathan, 2024) and synthetic datasets (Athey et al., 2021). Shen et al. (2021) shows that humans cannot distinguish synthetic images generated by GANs from real images.

We specifically adopt the Style-GAN model developed by Karras et al. (2019) to generate images modified in a specified feature while preserving all other attributes. We first encode images into the latent space using a pre-trained GAN to obtain their latent representations (image embeddings). Next, we compute the direction vector corresponding to our feature of interest (e.g., mapping *smile* to *non-smile*). We then adjust the latent representation of each image along this direction vector and regenerate the images using the GAN. The resulting outputs remain unchanged in every other aspect except for the targeted feature. We use this method to generate images that vary in *Gender*, *Smile*, *Age*, *Hair Color*, and *Glasses*.

Finally, to ensure the altered images appear realistic, we apply deblurring, inpainting, and auto-blending. See Appendix C for further details and Figure 10 for examples of GAN-generated images. Other features, such as *Body Shot* and *Sunglasses*, are produced using image detection and Photoshop, starting from the GAN-generated images.

In Appendix D, we compare the estimated distributions of emotional traits in Kiva Data and the GAN-generated images. The results indicate that these distributions are closely aligned, suggesting that the images used in the experiment are likely to evoke emotional responses similar to those elicited by the real Kiva borrower images.

Figure 10: Variation in *Smile* and *Male*



Note: Examples of pair of images generated using GANs. The image on the top has variation in Smile *and the image at the bottom in* Male.

5.4 Experiment results

Subjects in the experiment chose between two profiles in each choice instance. The outcome is whether the profile is chosen or not, and its mean is 0.5. Suppose a lender's utility has a systematic component that depends on *male*, *smile*, and *body-shot*, profile fixed effect μ_j and an idiosyncratic random component ϵ_{ij}. The utility of subject i choosing option j is written:

$$u_{ij} = \alpha \cdot male_j + \beta \cdot smile_j + \gamma \cdot bodyshot_j + \mu_j + \epsilon_{ij}. \tag{1}$$

Assuming that ϵ is distributed following a type I extreme value distribution, the probability of subject i choosing option j is written:

$$u_{ij} = \frac{\exp(\alpha \cdot male_j + \beta \cdot smile_j + \gamma \cdot bodyshot_j + \mu_j)}{\sum_{k=j,j'} \exp(\alpha \cdot male_k + \beta \cdot smile_k + \gamma \cdot bodyshot_k + \mu_j)}. \tag{2}$$

We are interested in the estimates of parameters α, β, and γ. We obtain them by estimating a logistic regression model that maximizes the conditional likelihood.

Table 5 presents the results for Experiments 1 and 2. For each experiment, column (1) shows the baseline specification using conditional logit. Column (2) extends the model by including subject-specific covariates, and column (3) applies a logit model with fixed effects. All regressions include borrowing campaign fixed effects. The results for Experiment 1 include variables related to *Gender*, *Smile*, and *Bodyshot*, while Experiment 2 focuses on *Age*, *Sunglasses*, *Glasses*, and *Dark Hair*. Subject-specific characteristics and fixed effects are incorporated in columns (2) and (3) for both experiments, allowing for a detailed analysis of the interaction between profile features and subject-level adjustments.

Table 5: Average Treatment Effects Estimates from Experiments 1 and 2

	Experiment 1			*Experiment 2*		
	Conditional Logit	+ Covariates	+ Fixed Effects	Conditional Logit	+ Covariates	+ Fixed Effects
	(1)	(2)	(3)	(4)	(5)	(6)
Male	-0.385*** (0.079)	-0.299*** (0.079)	-0.146** (0.066)			
Smile	0.298*** (0.074)	0.326** (0.078)	0.160** (0.072)			
Bodyshot	-0.191** (0.079)	-0.118 (0.086)	-0.050 (0.075)			
Age				-0.023 (0.057)	-0.026 (0.058)	-0.020 (0.058)
Sunglasses				-0.475*** (0.073)	-0.466*** (0.074)	-0.452*** (0.074)
Glasses				-0.254*** (0.075)	-0.247*** (0.076)	-0.282*** (0.076)
Dark Hair				0.066 (0.057)	0.062 (0.058)	0.098* (0.058)
Image FE	x	x	x	x	x	x
Subject's Characteristics		x	x		x	x
Restricted Sample			x			x
Observations	4,920	4,428	4,920	6,119	5,919	6,119

Note: Estimates of logistic regression models. Specifically, conditional logit models (Columns 1, 4), conditional logit with covariates (Columns 2, 5), and logit with subject fixed effects (Columns 3, 6). Borrower profile fixed effects are included in all regressions. Experiment 1 variables: Male, Smile, Bodyshot. Experiment 2 variables: Age, Sunglasses, Glasses, and Dark Hair. *$p<0.1$; **$p<0.05$; ***$p<0.01$.

In *Experiment 1*, the coefficients for *Male* and *Smile* are consistently statistically significant across all specifications, indicating that these attributes strongly influence the likelihood of selection. Specifically, males are significantly less likely to be chosen, while individuals displaying a smile are more likely to be selected. *Bodyshot*, however, shows less consistent significance, with its effect diminishing in the presence of subject-specific covariates and profile fixed effects. In *Experiment 2*, both *Sunglasses* and *Glasses* are consistently significant, with negative coefficients suggesting that these attributes reduce the likelihood of selection. *Dark Hair* and *Age* are not statistically distinguishable from zero.

The experimental estimates of the impact of *male* are very similar to those based on observational data; in contrast, we estimate a considerably higher impact of *smile* on outcomes in the experimental compared to the observational data. In Appendix B, we document that the feature prediction model has a substantial rate of false negatives in the task of detecting *smile*, and we show that might account for the discrepancy, as false negatives introduce a negative bias in the observational estimate of the impact of *smile* on outcomes. We also find similar estimates across the two methods for *Sunglasses*. We find that *Age* has a statistically insignificant estimate, and *Glasses* has the opposite effect in *Kiva Data* compared to the recruited experiment.

In summary, evidence from the recruited experiment supports several findings from the observational data. While neither analysis is definitive, consistency between estimates from *Kiva Data* and the experiment suggests that several of the *style* features are strong candidates for policy interventions.

6 Efficiency-Disparity tradeoff: counterfactual simulations

There are many different platform policies that could exploit our finding that certain *style* features impact funding outcomes, and type-based disparities in particular. In this section, we propose several such policies, simulate counterfactual outcomes, and evaluate their impact on disparities and efficiency. To do that, we consider a simplified model of interactions on Kiva characterized by the parameters from the recruited experiment.

Although our approach is stylized, our findings provide insight into which types of policies are likely to be effective. In practice, our method can help prioritize policies for randomized experiments. The policies we study, policies based on *style* features, are particularly relevant from a managerial perspective, as platforms like Kiva often provide guidelines for *style* choices, given that these features are easily modifiable.

6.1 A model of a micro-lending platform

Pool of borrowers. The pool of available borrowers is a set of borrowing campaigns from which Kiva selects a subset to display to lenders. The pool of borrowing campaigns can be summarized by a vector of profiles $\mathbf{x}$, where each element is a profile $x_i = (type_i, style_i, \eta_i)$, which describes features of the borrower i; the first element corresponds to the borrower's *type* and we consider two dimensions *male* or *female* and *young* and *old*. The *style* features encompass *smile*, *body-shot*, *sunglasses*, and *dark hair*, and η_i is a fixed effect which summarizes all other characteristics of the borrower.

The pool of borrowers is exogenously determined and the joint distribution of borrowers' characteristics is denoted as G. The expected pool of borrowers is denoted as $\tilde{x}$.

Policy and markets. A market is a set of borrowers shown to a lender. Platform policy $\mathbb{H}$ transforms the joint distribution of borrowers' characteristics from G to H. Specifically, the policy defines $\mathbf{E}_H\left[style|type, \eta\right]$ the conditional probability of *style* features in the pool of borrowers. Additionally, a policy applies the probability of being shown to lenders $h : (\eta, type, style) \to [0, 1]$ to the pool of borrowers. Thus, a policy can be summarized as $\mathbb{H} = \{\mathbf{E}_H [style|type, \eta], \mathbf{h}\}$. The expected pool of borrowers shown to lenders under the policy $\mathbb{H}$ is denoted as $\tilde{x}^H$.

The policies that we consider have two elements. First, they can impact the distribution of *style* features in the pool of borrowers. Examples of this include advice on profile creation, a protocol that requires borrowers to upload several images and selects the most compliant one, or behavioral interventions that nudge borrowers to create compliant profiles. Second, a policy can modify the probabilities with which borrowers in the pool appear in the market as a function of image features and other borrowers' characteristics.

Lenders. Lenders, indexed by j, are heterogeneous with respect to their preference parameters $\boldsymbol{\alpha}_j$ (for *type*) and $\boldsymbol{\beta}_j$ (for *style*).[31] Preference parameters are random variables that are realized before a lender sees any profiles. Each parameter is drawn from a normal distribution centered at the corresponding average treatment effect estimate from the recruited experiment, and the standard deviation is equal to the standard error of the estimate. After the parameters are determined, the lender decides whether to participate or not based on the expected utility of participating, which is:

$$\mathbb{E}_{i \in \tilde{x}^H} \left[u_{ij} \right] = \mathbb{E}_{i \in \tilde{x}^H} \left[\max(\boldsymbol{\alpha}_j \cdot type_i + \boldsymbol{\beta}_j \cdot style_i + \eta_i + \epsilon_{ij}) \right], \tag{3}$$

where $\tilde{x}^H$ is the vector of borrowers that are active on the platform under the policy H, and ϵ_{ij} is a random utility parameter, which is independent across lenders and borrowers, GEV distributed. If $\mathbb{E}_{i \in I_P} > \delta$, where δ is the cost of participating, the lender participates; otherwise, the lender stays out.

Lenders decide to enter or not based on their preference parameters and the expectation of the utility from participating, which depends on the policy $\mathbb{H}$. Thus, the distribution of preference parameters of participating lenders will differ depending on the policy.

Lenders who decide to participate observe the realized choice set of borrowers and choose the option that maximizes their utility: a borrower or the outside option. The utility associated with choosing one of the borrowers is written:

$$u_{ij} = \boldsymbol{\alpha}_j \cdot type_i + \boldsymbol{\beta}_j \cdot style_i + \eta_i + \epsilon_{ij}. \tag{4}$$

[31] $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are vectors with $\boldsymbol{\alpha}_j$ having two elements for *male* and *old* and $\boldsymbol{\beta}_j$ having four elements for *smile, bodyshot, sunglasses*, and *dark hair*.

The utility from choosing the outside option is $u_{oj} = \omega + \epsilon_{oj}$. Lenders choose an option that maximizes their utility.

To summarize, we assume the following timing: (1) The pool of available borrowers is exogenously determined; (2) The platform chooses policy $\mathbb{H}$; (3) Lenders arrive, their preference parameters are realized, and they decide whether or not to participate (they know which policy the platform chose); (4) Lenders who decided to participate, choose between borrowers or an outside option.

6.2 Implementation

Markets. We consider a pool of 22 borrowing campaigns. The distribution of *style* features conditioned on considered *type* features and overall profile attractiveness is based on *Kiva data*.[32] First, campaigns' fixed effects take values of fixed effects estimated in the recruited experiments. We assign these values randomly. Second, we train a GBM model of *cash per day*, and predict the outcome net of the *type* and *style* features of interest. We compute the distribution of *types* in *Kiva data* across deciles of predicted *cash per day*, and match the deciles of the fixed effect to the decile of the predicted *cash per day*:

$$\mathbf{E}_G \left[type | D(\hat{\eta}) \right] = \mathbf{E}_K \left[type | D(\eta_k) \right],$$

where K stands for distribution in *Kiva data*, $D(\cdot)$ is the decile of the fixed effect and η_k is the fixed effect from *Kiva data*. $\mathbf{E}_K \left[type | D(\eta_k) \right]$ is the conditional distribution of *type* profiles in *Kiva data*; thus, for example, the share of *male* borrowers with the fixed effect in the first decile of fixed effects estimated from the recruited experiment equals the share of *male* borrowers in the lowest decile of *Kiva data* fixed effects. This way, we get the distribution of *types* across residual profile attractiveness. Next, we compute the distribution of *style* in *Kiva data* across the *types* and the deciles of predicted *cash per day*, and match on both the deciles and the *types*.

$$\mathbf{E}_G \left[style | type, D(\hat{\eta}) \right] = \mathbf{E}_K \left[style | type, D(\eta_k) \right].$$

Thus, we allow the distribution of the *style* to differ across fixed effects and *type*.

[32] At the time our data was collected, Kiva's policy was based on when the borrower posted the campaign. Thus, assuming that arrival time is independent of characteristics, a lender sees each borrower in the pool with equal probability. In reality, this is an approximation because campaigns that reach their funding outcomes are removed from the platform. Thus, the less attractive campaigns stay longer on the platform, so lenders have a higher chance of observing them. As a consequence, the distribution of *type* and *style* features that we observe in *Kiva data* might differ from the distribution of the pool of borrowers that arrive to Kiva.

The number of borrowers in a market will depend on the policy, but in all cases, it will be a subset of the pool of borrowers.

Lenders' preferences. We assume that lenders' preferences (α_j, β_j) are parameters drawn from distributions estimated using experimental data, such that $\alpha_j \sim N(\alpha, sd_\alpha)$, where α is the estimate of the average treatment effect and sd_α is its standard error, and ϵ_{ij} is a random utility parameter, which is iid across lenders and borrowers, GEV distributed. We set the utility from choosing the outside option to one (the highest FE estimated in the experiment is 0.64).

We assume that the cost of participating δ is fixed and the same for all borrowers. We set the cost at 2.5, resulting in approximately half of the lenders choosing to participate.

Outcome metrics. We propose two metrics of disparities: first, to capture the overall distribution of outcomes, we use the Gini coefficient defined as

$$Gini = \frac{\sum_{j=1}^{n} \sum_{j'=1}^{n} |x_j - x_{j'}|}{2n\bar{x}},$$

where x_j is the outcome of borrower j and $x_{j'}$ of borrower j', n is the number of borrowers and $\bar{x}$ the average outcome. We consider all borrowers in the pool. Second, we consider the gender disparity, defined as the share of lenders that choose a *male* borrower amongst the lenders who decided to participate and did not choose an outside option. We standardize this metric by the share of *male* borrowers in the pool. We measure efficiency as the share of lenders that chose a borrower instead of an outside option.

The type of policy, and thus the expected set of borrowers shown to lenders, will impact the number and the type of active lenders. To capture this, we report the number of active lenders that have high *style* preference parameters. We define high as above the mean of the distribution for each parameter. Thus, a *high smiling* type is the lender who cares more than a typical lender that a borrower smiles in the profile image.

Market outcomes. To determine market outcomes, we simulate markets and choices by lenders. Based on the distribution of outcomes, we compute disparity and efficiency metrics. Each simulation proceeds in three steps: first, we simulate the pool of borrowers. Then, we construct markets from the pool of borrowers. A policy determines $h(\eta_i, typei, style_i)$, the probability that a borrower in a pool appears in the market. A market is constructed per lender. This means that in one simulation there is one pool of borrowers, from which borrowers are sampled for each lender.

Finally, we simulate lenders' preferences, their entry decisions and their choices.. We perform 100 simulations of 2000 lenders' choices for each policy. We use the outcomes to compute our metrics of disparity and efficiency. We consider all borrowers in the pool, irrespective of whether they were shown to lenders or not. Appendix N presents the algorithm that we used.[33]

6.3 Counterfactual policies

Baseline. The baseline policy represents the existing policy on Kiva. In the baseline policy, the platform shows 10 borrowers to each lender, and each borrower in the pool is assigned an equal probability of being included in the market.

Naive Recommendation. In this policy, we show what happens when a platform oversamples profiles with attractive *style*. The platform selects all borrowers with attractive *style* features and randomly selects 10 to include in the market. The selection is done such that the platform starts by selecting profiles that match all 4 style criteria (*smile*, no *sunglasses*, *Dark Hair*, and no *Bodyshots*). If there are fewer than 10 such profiles, the platform relaxes the requirements until there are 10 profiles. If there are more than 10 profiles, the platform selects 10 of them at random.

Style Recommendation. In this policy, the platform recommends that all borrowers follow *style* guidelines. In practice, we assume that previously non-compliant borrowers become compliant with a probability of 75%.[34] After a pool of borrowers is determined, the platform assigns all borrowers an equal probability of being included in the market.[35]

[33] In this analysis, we assume that lenders' preferences are stable across different platform policies. In Appendix K, we exploit a natural experiment in the form of Kiva landing page redesign to provide support of this assumption. The website redesign introduced borrowers categories in place of a list where all borrowers would be displayed together. We find that the impact of *smile* on *cash per day* was similar before and after the website redesign; the difference is statistically insignificant.

[34] Such a profile feature recommendation can be implemented in various ways, for example, through behavioral nudges or a script requiring that several images need to be uploaded from which platform selects the ones to be shown to lenders.

[35] Note that this policy requires that borrowers comply with the policy recommendation. In the analysis, we assume that 25% of the borrowers do not adhere to the recommendation. A particular type of non-compliance in the case of *smile* might be that borrowers attempt to create an image with *smile*, but they do not succeed; for example, the *smile* does not appear genuine. In Appendix O, we develop an additional algorithm that distinguishes between fake and genuine smiles and apply it to the Kiva observational data. We show that only genuine smiles lead to higher outcomes. Consequently, the policy will be less effective if some of the newly added *smile*'s are perceived as non-genuine. This analysis highlights the importance of clear instructions and a well-designed system that supports borrowers when they create profiles.A computer vision algorithm showcased in Appendix O could be a component of such a system, where borrowers could be prompted if their smile is at risk of being perceived as not genuine.

Low-type Support. This policy promotes borrowers predicted to have low funding outcomes based on their *types*, by ensuring they are always included in the market. We focus on *gender* in this application. Practically, the approach is analogous to *Naive*: when the number of *male* campaigns is above ten, the platform samples randomly from them. Otherwise, the platform includes all *male* profiles and fills in other slots by randomly selecting from available profiles. In expectation, there are some *female* profiles included in the market.

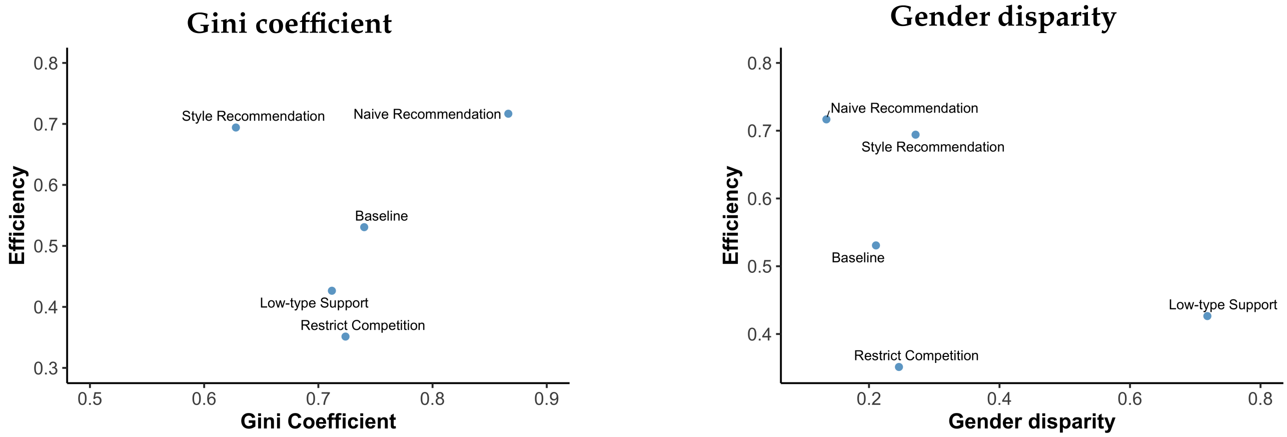
Restrict Competition. In this policy, the platform promotes fairness by reducing the competition between borrowers. To implement this, the platform randomly selects five borrowers from the pool to form the market (instead of ten).

All policies that we propose in expectation give non-zero probabilities of being included in the market to any borrower.

6.4 Results

Figure 11 presents the results from simulations of the proposed policies. In the left panel, on the horizontal axis, we show the mean of Gini coefficients across all simulations of each policy. On the vertical axis, we show the mean of lenders' shares choosing a borrower rather than the outside option.

Figure 11: Disparity-Efficiency Tradeoff



Overall notes: Each point represents the mean of 100 simulations with 2000 lenders each. The vertical axis reports the share of lenders choosing an outside option. Left panel notes: The horizontal axis presents Gini coefficients. Right panel notes: The horizontal axis presents the share of lenders choosing a borrower with a male profile, adjusted for the share of male profiles in the borrower pool.

We find that the proposed policies considerably impact both metrics, disparity, and efficiency. First, in the *Baseline* policy, the Gini coefficient is around 0.74 (the higher, the less equal distribution of funds), and efficiency is 0.53. Both the *Naive Recommendation* and *Style Recommendation* substantially increase the share of lenders choosing one of the borrowers rather than the outside option. The *Naive Recommendation* policy has a slightly higher impact on efficiency. *Low-type Support* and *Restrict Competition* have a strong negative impact on efficiency. In particular, restricting the number of alternatives substantially increases the share of lenders choosing an outside option.

Naive Recommendation is the only policy that increases disparities; Gini coefficient increases to 0.87 under this policy. *Low-type support* and *Restrict Competition* lead to a small reduction in disparities. *Style Recommendation* substantially reduce disparities.

The right panel of Figure 11 shows the impact of our counterfactual policies on the gender gap and efficiency. *Low-type support* stands out, actively including more *male* profiles in a market, has a strong effect on reducing the gender gap. Both *Style Recommendation* and *Restrict Competition* have minor effects on reducing the gender gap. *Naive Recommendation* increases the gender gap.

Naive Recommendation and *Low-Type Support* operate by altering the conditional probability of borrower inclusion in the market. *Naive Recommendation* aims to increase efficiency by prioritizing attractive profiles, while *Low-Type Support* seeks to reduce disparities by increasing impressions for underperforming borrowers. However, both policies have unintended consequences. Favoring attractive profiles exacerbates disparities, as desirable *style* features often coincide with advantageous *type* characteristics, compounding inequalities. Conversely, prioritizing borrowers from an underperforming *type* results in more frequent exposure to profiles with less attractive *style* features.

Style Recommendation, in contrast, adjusts the conditional distribution of *style* across *type*, allowing it to simultaneously enhance efficiency and reduce disparities.

When desirable *style* features are positively correlated with *type* characteristics that improve funding outcomes, platform policies that amplify these features reinforce inequities. In contrast, policies that redistribute attractive *style* features among borrowers with high *type* characteristics promote a more equitable allocation of funding. Moreover, increasing the overall prevalence of desirable *style* features enhances both equity and efficiency.

6.5 Impact of counterfactual policies on type of lenders active in the market

Table 6 presents the impact of different platform policies on the type of active lenders, specifically those with a high preference for certain borrower features. The top section of the table reports the share of active lenders with a high preference for each feature within the population of active lenders.

Table 6: Impact of Platform Policies on the Type of Active Lenders

	Baseline	**Style Rec.**	**Naive Rec.**	**Restrict Comp.**	**Low-Type Supp.**
Share of Active Lenders with High Preference for Feature					
Share Smile	0.36	0.48	0.37	0.96	0.32
Share Bodyshot	0.24	0.39	0.25	0.64	0.20
Share Sunglasses	0.25	0.40	0.25	0.67	0.21
Share Hair	0.28	0.45	0.28	0.77	0.26
The Number of Active Lenders with High Preference for Feature					
Smile	19.00	33.25	26.20	33.65	13.75
Bodyshot	12.55	27.40	17.85	22.60	8.60
Sunglasses	13.10	27.70	18.20	23.55	9.15
Hair	15.05	31.50	20.40	27.20	10.90

Note: The top section shows the share of lenders that have high preference for the specific feature in the population of active lenders. High preference is defined as the preference that is above the mean in the distribution. Active lenders are those that choose one of the borrowers. The bottom section shows the total number of active lenders who have a high preference. These numbers are out of 100 lenders. Thus, 19 in the case of Baseline and Smile preference means that out of 100 lenders who arrived on the platform, 19 decided to participate, selected one of the borrowers, and have a high preference for a smile.

A lender is classified as having a high preference if their preference score is above the mean in the distribution. The bottom section of the table reports the absolute number of active lenders with a high preference for a given feature, out of a total of 100 potential lenders arriving on the platform. For example, in the Baseline condition, 19 out of 100 lenders who arrived on the platform actively participated and exhibited a high preference for smiling borrowers. The table allows us to compare how different platform interventions influence lender engagement and the distribution of preferences among active participants.

The results suggest that platform policies significantly alter the composition of active lenders and their feature preferences. Under the Style Recommendation policy, the share of active lenders with a high preference for any given feature increases across all categories. For example, the share of lenders with a high preference for smiles increases from 0.36 in the baseline to 0.48, while the absolute number of lenders with a high preference for smiles rises from 19 to 33.25 out of 100. This suggests that highlighting certain stylistic features not only influences borrower selection but also affects the composition of lenders who choose to participate, attracting those who place greater emphasis on visual features.

Conversely, the Low-Type Support policy, which increases the visibility of borrowers from underperforming groups, results in a decline in the number of active lenders across all features. For instance, the number of lenders with a high preference for smiles drops to 13.75, compared to 19 in the baseline. This suggests that actively promoting borrowers with less conventionally attractive features may reduce overall engagement from lenders who are more visually selective, potentially influencing platform efficiency. Meanwhile, the Restrict Competition policy, which limits borrower competition, leads to a substantial increase in engagement across all feature preferences, particularly for smiles (33.65 active lenders, up from 19 in the baseline). This indicates that reducing borrower competition can make participation more attractive to lenders, likely by reducing the cognitive load of selection.

Overall, the findings indicate that platform policies shape not just which borrowers receive funding but also which types of lenders engage with the platform. Policies that increase visibility for borrowers with certain stylistic features attract lenders with stronger preferences for those features, while interventions that reduce borrower competition generally increase engagement. However, policies aimed at reducing disparities by promoting borrowers from underrepresented groups appear to disincentivize participation among lenders with strong visual preferences. These insights underscore the trade-offs platforms face when designing interventions that balance improving efficiency and reducing disparities in lender engagement.

7 Conclusion

This paper examines how *type* and *style* features in profile images influence funding outcomes and disparities in an online micro-lending marketplace. Using observational data from Kiva, a large peer-to-peer microfinance platform, we document systematic disparities in funding outcomes across borrower *types* and show that *style* features play a significant role in shaping these disparities.

Certain *style* choices, such as *smiling* and avoiding *body-shots*, are associated with higher funding success, and these choices are correlated with *type*. For example, *men* are less likely to *smile* and more likely to wear *sunglasses* than *women*, while *young* borrowers are more likely to exhibit these traits than *older* borrowers. Decomposing *type*-based disparities, we show that differences in *style* contribute to funding gaps between demographic groups.

Our empirical results based on observational data may suffer from omitted variable bias. To address this, we use the observational results to prioritize features for a sequence of two randomized experiments with recruited subjects. The experiments produce internally valid estimates at the potential cost of external validity (although the funding decisions lead to real-world loans in our second experiment). Subjects make funding decisions based on controlled variations in borrower profile images. The results confirm that *style* features significantly impact selection probabilities: *smiling* increases the likelihood of selection, while *wearing sunglasses* or *glasses* reduces it. The experiment also corroborates *type*-based disparities observed in the Kiva data, with subjects preferring *women* borrowers over *men*, while the difference in selection probabilities between *young* and *old* borrowers is small and statistically insignificant.

Exploring the mechanisms behind these effects, we find that *style* features do not improve predictions of loan repayment probability, suggesting that rational lenders should not use these cues to assess creditworthiness. Instead, they are correlated with perceived attributes such as *trustworthiness* and *dominance*, implying that psychological considerations drive lender decisions more than financial considerations.

Our findings have important implications for platform design. Because *style* features can be modified, they offer a potential lever for reducing disparities while maintaining or improving efficiency. Counterfactual simulations suggest that *style* recommendations—such as encouraging borrowers to *smile* while avoiding *body-shots*—can reduce disparities and increase transaction volume. However, increasing the visibility of borrowers with preferred *style* traits may improve efficiency but exacerbate disparities by disproportionately benefiting borrowers from advantaged *types*. These findings highlight an equilibrium effect, where platform policies influence both borrower funding outcomes and the composition of active lenders.

More broadly, our results underscore that disparities in online marketplaces arise not only from fixed socio-demographic characteristics but also from malleable *style* choices correlated with *type*. Encouraging the adoption of beneficial *style* features among underrepresented groups can mitigate disparities while maintaining efficiency. However, the appropriateness of *style* recommendations must be carefully considered in context.

While our study provides a framework for identifying platform policies that balance fairness and efficiency, the effectiveness of specific interventions depends on lender responses and borrower compliance. Our findings can inform the design of future randomized experiments testing *style*-based interventions in real-world settings. More broadly, they highlight a challenge in algorithmic decision-making: when predictive models incorporate *style* features correlated with *type*, they may reinforce disparities even if the features themselves appear neutral. Future research should examine whether similar mechanisms operate in other online marketplaces, such as hiring platforms and social networks, where *style* choices may systematically influence economic and social outcomes.

References

Abbey, J. D. and Meloy, M. G. (2017). Attention by Design: Using Attention Checks to Detect Inattentive Respondents and Improve Data Quality. *Journal of Operations Management*, 53:63–70.

Abeler, J., Becker, A., and Falk, A. (2014). Representative Evidence on Lying Costs. *Journal of Public Economics*, 113:96–104.

Aggarwal, R., Goodell, J. W., and Selleck, L. J. (2015). Lending to Women in Microfinance: Role of Social Trust. *International Business Review*, 24(1):55–65.

Alesina, A. F., Lotti, F., and Mistrulli, P. E. (2013). Do Women Pay More for Credit? Evidence from Italy. *Journal of the European Economic Association*, 11:45–66.

Andreoni, J. and Petrie, R. (2008). Beauty, Gender and Stereotypes: Evidence from Laboratory Experiments. *Journal of Economic Psychology*, 29(1):73–93.

Asplund, J., Eslami, M., Sundaram, H., Sandvig, C., and Karahalios, K. (2020). Auditing Race and Gender Discrimination in Online Housing Markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 24–35.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2021). Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations. *Journal of Econometrics*.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2024). Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations. *Journal of Econometrics*, 240(2):105076.

Athey, S., Karlan, D., Palikot, E., and Yuan, Y. (2022). Smiles in Profiles: Improving Fairness and Efficiency Using Estimates of User Preferences in Online Marketplaces. Technical report, National Bureau of Economic Research.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics*, 47(2):1148–1178.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A Convex Framework for Fair Regression. *arXiv preprint arXiv:1706.02409*.

Bertrand, M. and Mullainathan, S. (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.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2023). Inaccurate Statistical Discrimination: An Identification Problem. *Review of Economics and Statistics*, pages 1–45.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, 29.

Brock, J. M. and De Haas, R. (2021). Discriminatory Lending: Evidence from Bankers in the Lab. *CentER Discussion Paper*.

Burke, R., Ragothaman, P., Mattei, N., Kimmig, B., Voida, A., Sonboli, N., Kathait, A., and Fabros, M. (2022). A Performance-Preserving Fairness Intervention for Adaptive Microfinance Recommendation. In *KDD Workshop on Online and Adaptive Recommender Systems at the 28th SIGKDD Conference on Knowledge Discovery and Data Mining*.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. (2023). Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Transactions on Information Systems*, 41(3):1–39.

Cook, C., Diamond, R., Hall, J. V., List, J. A., and Oyer, P. (2021). The Gender Earnings Gap in the Gig Economy: Evidence from Over a Million Rideshare Drivers. *The Review of Economic Studies*, 88(5):2210–2238.

Dash, A., Ye, J., and Wang, G. (2023). A Review of Generative Adversarial Networks (GANs) and Its Applications in a Wide Variety of Disciplines: From Medical to Remote Sensing. *IEEE Access*.

Davidson, A. and Gleim, M. R. (2023). The Gender Earnings Gap in Sharing Economy Services: The Role of Price, Number of Stays, and Guests Accommodated on Airbnb. *Journal of Marketing Theory and Practice*, 31(4):490–501.

Dupas, P., Fafchamps, M., and Hernandez-Nunez, L. (2024). Keeping Up Appearances: An Experimental Investigation of Relative Rank Signaling. Technical report, National Bureau of Economic Research.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

D'Espallier, B., Guérin, I., and Mersland, R. (2011). Women and Repayment in Microfinance: A Global Analysis. *World Development*, 39(5):758–772.

Edelman, B., Luca, M., and Svirsky, D. (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*, 9(2):1–22.

Ert, E., Fleischer, A., and Magen, N. (2016). Trust and Reputation in the Sharing Economy: The Role of Personal Photos in Airbnb. *Tourism Management*, 55:62–73.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, pages 1189–1232.

Ge, Y., Knittel, C. R., MacKenzie, D., and Zoepf, S. (2016). Racial and Gender Discrimination in Transportation Network Companies. Technical report, National Bureau of Economic Research.

Gelbach, J. B. (2016). When Do Covariates Matter? And Which Ones, and How Much? *Journal of Labor Economics*, 34(2):509–543.

Glynn, A. N. and Quinn, K. M. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis*, 18(1):36–56.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. *arXiv preprint arXiv:1406.2661*.

Haferkamp, N., Eimler, S. C., Papadakis, A.-M., and Kruck, J. V. (2012). Men Are from Mars, Women Are from Venus? Examining Gender Differences in Self-Presentation on Social Networking Sites. *Cyberpsychology, Behavior, and Social Networking*, 15(2):91–98.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2014). Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, 22(1):1–30.

Han, L., Fang, J., Zheng, Q., George, B. T., Liao, M., and Hossin, M. A. (2024). Unveiling the Effects of Livestream Studio Environment Design on Sales Performance: A Machine Learning Exploration. *Industrial Marketing Management*, 117:161–172.

Hapek, K. (2021). *A Fairness-Based Recommender System for Charitable Lending Platform Kiva Using Classification and ϵ-Greedy Policy*. PhD thesis, Dublin, National College of Ireland.

Herzenstein, M., Sonenshein, S., and Dholakia, U. M. (2011). Tell Me a Good Story and I May Lend You Money: The Role of Narratives in Peer-to-Peer Lending Decisions. *Journal of Marketing Research*, 48(SPL):S138–S149.

Jabbar, A., Li, X., and Omar, B. (2021). A Survey on Generative Adversarial Networks: Variants, Applications, and Training. *ACM Computing Surveys (CSUR)*, 54(8):1–49.

Johannemann, J., Hadad, V., Athey, S., and Wager, S. (2019). Sufficient Representations for Categorical Variables. *arXiv preprint arXiv:1908.09874*.

Kajackaite, A. and Gneezy, U. (2017). Incentives and Cheating. *Games and Economic Behavior*, 102:433–444.

Kamiran, F. and Calders, T. (2012). Data Preprocessing Techniques for Classification Without Discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Karras, T., Laine, S., and Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410.

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

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent Trade-offs in the Fair Determination of Risk Scores. *arXiv preprint arXiv:1609.05807*.

Kristof, N. D. and WuDunn, S. (2010). *Half the Sky: Turning Oppression into Opportunity for Women Worldwide*. Vintage.

Kung, F. Y., Kwok, N., and Brown, D. J. (2018). Are Attention Check Questions a Threat to Scale Validity? *Applied Psychology*, 67(2):264–283.

Lambin, X. and Palikot, E. (2022). The Impact of Online Reputation on Ethnic Discrimination. Technical report, Working Paper.

Lambrecht, A. and Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, 65(7):2966–2981.

Lee, E. L., Lou, J.-K., Chen, W.-M., Chen, Y.-C., Lin, S.-D., Chiang, Y.-S., and Chen, K.-T. (2014). Fairness-Aware Loan Recommendation for Microfinance Services. In *Proceedings of the 2014 International Conference on Social Computing*, pages 1–4.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Li, X., Zhu, C., Guo, H., Yu, Y., Tang, R., et al. (2023). How Can Recommender Systems Benefit from Large Language Models: A Survey (2023). *arXiv preprint arXiv:2306.05817*.

Lin, Y., Yao, D., and Chen, X. (2021). Happiness Begets Money: Emotion and Engagement in Live Streaming. *Journal of Marketing Research*, 58(3):417–438.

Luca, M., Pronkina, E., and Rossi, M. (2024). The Evolution of Discrimination in Online Markets: How the Rise in Anti-Asian Bias Affected Airbnb During the Pandemic. *Marketing Science*.

Ludwig, J. and Mullainathan, S. (2024). Machine Learning as a Tool for Hypothesis Generation. *The Quarterly Journal of Economics*, 139(2):751–827.

Luo, L. E. and Toubia, O. (2024). Using AI for Controllable Stimuli Generation: An Application to Gender Discrimination with Faces. *Available at SSRN 4865798*.

Marchenko, A. (2019). The Impact of Host Race and Gender on Prices on Airbnb. *Journal of Housing Economics*, 46:101635.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Mendes, W. B. and Koslov, K. (2013). Brittle Smiles: Positive Biases Toward Stigmatized and Outgroup Targets. *Journal of Experimental Psychology: General*, 142(3):923.

Naghiaei, M., Rahmani, H. A., and Deldjoo, Y. (2022). Cpfair: Personalized Consumer and Producer Fairness Re-Ranking for Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 770–779.

Netzer, O., Lemaire, A., and Herzenstein, M. (2019). When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. *Journal of Marketing Research*, 56(6):960–980.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., et al. (2007). Pervasiveness and Correlates of Implicit Attitudes and Stereotypes. *European Review of Social Psychology*, 18(1):36–88.

Ozer, G. T., Greenwood, B. N., and Gopal, A. (2023). Digital Multisided Platforms and Women's Health: An Empirical Analysis of Peer-to-Peer Lending and Abortion Rates. *Information Systems Research*, 34(1):223–252.

Park, J., Kim, K., and Hong, Y.-Y. (2019). Beauty, Gender, and Online Charitable Giving. *Available at SSRN 3405823*.

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., and Suchow, J. W. (2022). Deep Models of Superficial Face Judgments. *Proceedings of the National Academy of Sciences*, 119(17):e2115228119.

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

Ravina, E. (2019). Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets. *Available at SSRN 1107307*.

Reuben, E., Sapienza, P., and Zingales, L. (2014). How Stereotypes Impair Women's Careers in Science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866.

Shen, B., RichardWebster, B., O'Toole, A., Bowyer, K., and Scheirer, W. J. (2021). A Study of the Human Perception of Synthetic Faces. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE.

Shishido, J., Narasimhan, J., and Haller, M. (2016). Tell Me Something I Don't Know: Analyzing OkCupid Profiles. In *SciPy*, pages 75–81.

Sisodia, A., Burnap, A., and Kumar, V. (2024). Generative Interpretable Visual Design: Using Disentanglement for Visual Conjoint Analysis. *Journal of Marketing Research*, page 00222437241276736.

Stigler, M. (2018). dec_covar: R Implementation of Gelbach Covariate Decomposition. *https://github.com/MatthieuStigler/Misconometrics/tree/master/Gelbach_decompo*.

Theseira, W. (2009). *Competition to Default: Racial Discrimination in the Market for Online Peer-to-Peer Lending*. PhD thesis, Dissertation, Wharton.

Wang, Y., Tao, L., and Zhang, X. X. (2024). Recommending for a Multi-Sided Marketplace: A Multi-Objective Hierarchical Approach. *Marketing Science*.

Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., and Chi, E. H. (2021). Understanding and Improving Fairness-Accuracy Trade-Offs in Multi-Task Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1748–1757.

Williams, B. A., Brooks, C. F., and Shmargad, Y. (2018). How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy*, 8(1):78–115.

Yang, S., Jiang, L., Liu, Z., and Loy, C. C. (2023). StyleGANEX: StyleGAN-Based Manipulation Beyond Cropped Aligned Faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21000–21010.

Younkin, P. and Kuppuswamy, V. (2018). The Colorblind Crowd? Founder Race and Performance in Crowdfunding. *Management Science*, 64(7):3269–3287.

Yuan, Y., Liu, X., Zhang, S., and Srinivasan, K. (2024). Gender and Racial Price Disparities in the NFT Marketplace. *International Journal of Research in Marketing*.

Zhang, L., Xiong, S., Zhang, L., Bai, L., and Yan, Q. (2022a). Reducing Racial Discrimination in the Sharing Economy: Empirical Results from Airbnb. *International Journal of Hospitality Management*, 102:103151.

Zhang, S., Lee, D., Singh, P. V., and Srinivasan, K. (2022b). What Makes a Good Image? Airbnb Demand Analytics Leveraging Interpretable Image Features. *Management Science*, 68(8):5644–5666.

Zhang, S., Mehta, N., Singh, P. V., and Srinivasan, K. (2021). Can an AI Algorithm Mitigate Racial Economic Inequality? An Analysis in the Context of Airbnb. *Working Paper*.

Appendix

A Feature Detection Algorithms

This Appendix A provides a detailed explanation of how we use computer vision algorithms to extract features from our image data. Specifically, we employ different algorithms, outlined in Appendix A, for object and feature detection using the Mask-RCNN model. The image feature enriched the dataset provided by Kiva to enhance our analysis.

Mask-RCNN. To systematically extract image features, we use Mask R-CNN, an object detection algorithm developed by Facebook. As illustrated in Figure 12, Mask R-CNN processes an input image and produces a “package” for each detected object, which includes the object’s class label, bounding box, and mask. These predictions are jointly optimized through a single loss function.

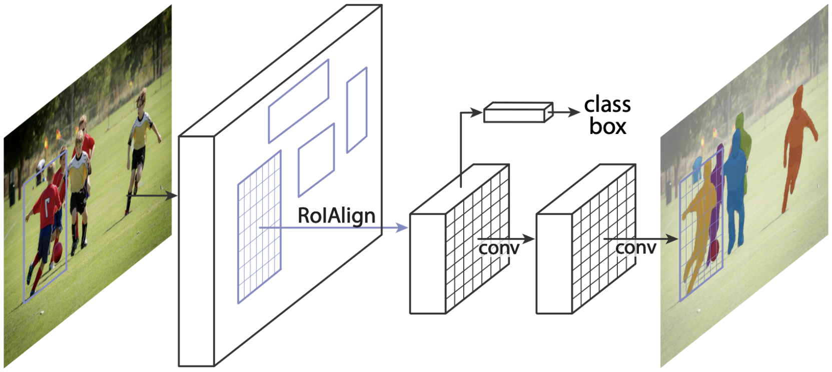


Figure 12: The Mask R-CNN framework: https://arxiv.org/pdf/1703.06870.pdf

Object detection. We apply this pre-trained model to generate a confidence score for each detected object, ranging from 0 to 1. The score represents the algorithm’s confidence in the presence of specific features, such as a tree, person, animal, or digital item. Figure 13 illustrates the resulting output. We also use this algorithm to detect full-body human figures.[36]

Facial feature classification. We detect facial features using the *face-classification* algorithm that takes in one face image and outputs a face embedding vector, evaluated by a pre-trained neural network.[37] Then, the embedding vector, as well as the feature labels, enter another neural network model (Multi-

[36] https://github.com/facebookresearch/detectron2

[37] https://github.com/wondonghyeon/face-classification

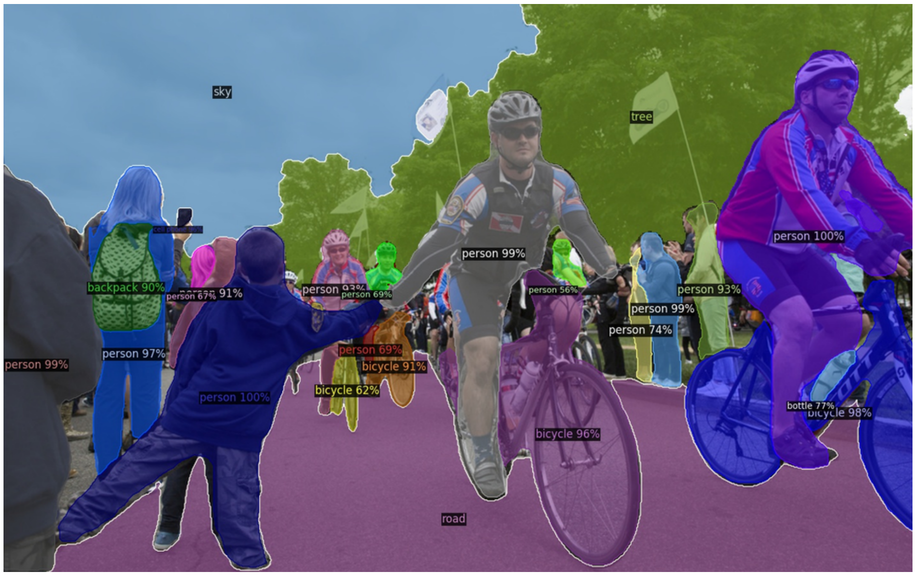


Figure 13: An example outcome of image detection using Mask-RCNN. Each detected object was given a label, put on a mask, and given the corresponding probability score: https://github.com/facebookresearch/detectron2.

layer Perceptron). This model takes in one facial embedding vector and assigns a score for each unique facial feature such as *race*, *gender*, *smile*, etc. It is a supervised learning process, and the training label is pre-annotated.

The features that we obtain from images can be informally classified into three categories: (i) technical aspects of the image (e.g., *blurry*, *flash*, *harsh light*), (ii) personal characteristics (e.g., *straight hair*, *eyes open*, *pale skin*), (iii) objects in the image (e.g., *chair*, *clock*).

Image and personal characteristics (e.g., race, age, hair color, facial shape, eyes/nose characteristics) are detected by FaceNet model which was pre-trained and tested on the large dataset CelebA with over 200,000 facial images. The algorithm detects the person's face and then identifies its features.

B Auditing the Feature Detection Algorithm

The Appendix B covers an audit study aiming to test whether the features detected by our algorithm correspond to the human perception of the image. The audit study proceeded in two steps; first, we recruited human raters and asked them to label a sample of Kiva images. Second, we compared these labels with the prediction of the feature detection algorithm. We focused on two features: smiling and gender.

We carry out three analyses; first, we analyze the overall correlation between how human raters annotate the image and the model's prediction of that annotation. Second, we consider the correlation of the model's prediction of whether the person in the image smiles or not with human labels, separately for images labeled by humans as men and women. Finally, we divide the model's errors into false positives and false negatives. The false positives are instances when human raters indicated that the feature is not present, while the model prediction was that it is; false negatives are when the model predicts that the feature is not present, even though, according to raters, it is. The two types of errors create different types of bias, false positives lead to overestimation of the impact of the feature on outcomes, while false negatives lead to underestimation.

To create human-made labels, we randomly drew 100 images from the Kiva dataset and organized them into ten protocols, so each protocol had ten images. After that, we recruited 30 subjects per protocol on Prolific. Subjects were asked about the gender of the person in the image and whether the person smiles or not. We carried out attention checks and asked about the level of confidence the person had in the response. To compute a label, we average subjects' responses per image.

Table 7 presents the comparison of the mean label per feature in the sample with the mean predicted probability from the feature-detection algorithm.

Table 7: Comparison of mean scores from the audit and CNN output.

Feature	Mean prob. CNN	Mean score audit	Difference	CI low	CI high
Man	0.48	0.56	-0.08	-0.20	0.05
Smiling	0.45	0.45	0.00	-0.09	0.09
Smiling amongst man	0.30	0.31	-0.01	-0.12	0.11
Smiling amongst woman	0.58	0.58	0.00	-0.13	0.13

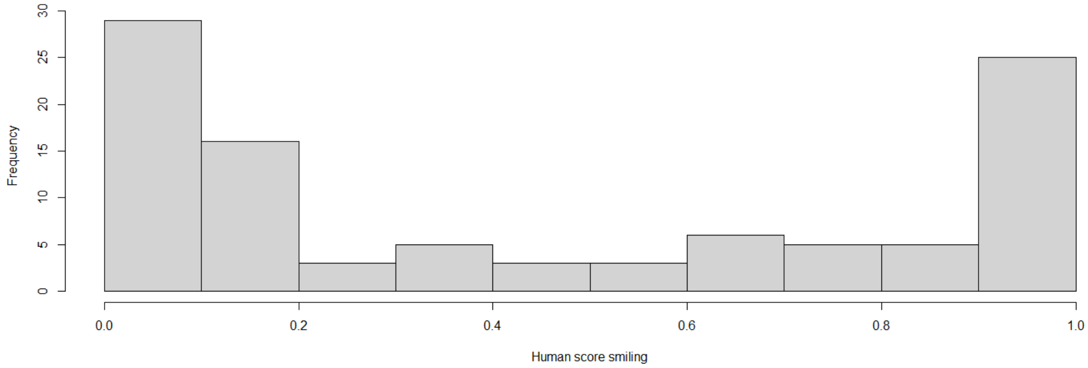
Note: Average CNN is the average of predicted probabilities per image using the feature-detection algorithm, and mean score audit is the average label by human raters. Rows one and two show the values in the entire sample. Rows three and four present the values when considering the subsamples based on gender determined by human subjects. For example, the third row shows the values amongst the images labeled as a man by at least 6 out of 10 human subjects.

Table 7 shows that human raters and the algorithm detect similar frequencies of the selected features. Additionally, as shown in rows three and four, the frequency of smiling across images labeled by humans as man or woman is consistent across the two methods. Thus, the two methods lead to the same conclusion that in the Kiva images, men are less likely to smile than women. [38] Additionally, the Pearson correlation coefficient between the human label and the algorithm's prediction is 0.92 for gender and 0.71 for smiling.

Finally, we transform the labels and the algorithm's predictions into binary indicators, taking the

[38] We reach the same conclusions when considering only subjects that have passed all attention checks and when excluding images that less than 7 out of 10 subjects rated similarly.

Figure 14: Histogram of human scores for smiling



Note: Score is the average per image label by human raters.

value of 1 when the continuous variable is above 0.5 and 0 otherwise. We analyze how often the two indicators coincide. We find that human scores and the algorithm prediction lead to the same gender classification in 93% of cases and for smiling in 81% of images.

In the case of many images, human raters disagree about whether the person in the image smiles or not. Figure 14 documents this. Many of the prediction errors are concentrated in these intermediate cases. When considering the subsample of images that seven or more human raters labeled similarly, the error rate declines to 14%.

We also consider the error rates of smile detection separately across genders as defined by human raters. The error rate amongst images labeled as man is 81.2% and woman 80.8%. We conclude that the algorithm performance is similar across genders. Finally, we group errors into false positives (the algorithm detects a feature that does not exist according to human raters) and false negatives (the algorithm does not detect an existing feature). We find that the rate of false positives is 7% and of false negatives is 34%.

To conclude, we find that the algorithm's predictions highly indicate the feature's presence in an image as perceived by human raters. The main concern is the high rate of false negatives for the smile detection task. We find that 34% of images labeled as smiling by humans are predicted as not smiling by the algorithm. Thus, we might be underestimating the impact of smiling on outcomes in the observational data. To assess the extent of the bias caused by these false negatives, we consider the following example: (i) there is a population in which half of the individuals have images with a *smile* feature, (ii) the outcome is a random variable normally distributed with a mean of 1 and a standard deviation of 1 for individuals that do not have an image with *smile*, and with a mean of 1.3 and a standard deviation of 1, for those that do *smile* in an image, (iii) we assume that the algorithm which detects features has the false negative error rate of 34%, (iv) we estimate a linear probability model using the OLS estimator. We simulate this example 1000 times and estimate the degree of bias caused by the false negatives. We find that the false negative error rate of 34% results in the negative bias of the smiling coefficient of 22% (s.e. 0.16%); the average coefficient, across simulations, is 0.23 instead of 0.3. Consequently, the estimates used in the observational data understate the impact of *smile* on outcomes.

C Generative Adversarial Networks

C.1 Background

Generative artificial intelligence (AI) models learn the patterns and probability distributions of training data, then use that understanding to generate new samples. Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), are a class of deep generative models designed to produce data that resemble real samples — such as realistic images (Ludwig and Mullainathan, 2024) and synthetic datasets (Athey et al., 2024). GANs, although do not directly produce estimates of the density or distribution function at a particular point, can be thought of as implicitly estimating the distribution of latent features, and they can be used to generate or output new examples that plausibly could have been drawn from the original dataset.

GANs are composed of two deep models: a generator G and a discriminator D. As illustrated by Goodfellow et al. (2014), the objective is to learn the training data's distribution p_g over data x. We define a prior $p_z(z)$ on input noise variables, then represent a mapping to data space as $G(z;\theta_g)$. The Generator receive the input vector sampled from $p_z(z)$ and outputs the image. Discriminator $D(x;\theta_d)$ takes in the image and outputs a single scalar, representing the probability that x the image is real (came from the dataset rather than p_g the generation). D and G play the two-player minimax with the objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[1 - logD(G(z))]$$

C.2 Variation of images in latent space

C.2.1 Stimuli generation

GANs are often employed to modify images and create "Deep Fakes" — fabricated images altered along specific dimensions. In our study, we apply StyleGAN (Karras et al., 2019) and StyleGANEX (Yang et al., 2023) to generate images that differ by a particular feature of interest. We use a pre-trained GAN generator G and encoder, which transforms an image x_i into its latent embedding $v_i \in \mathcal{Z}$ (in the latent space $\mathcal{Z}$). From there, we obtain a direction vector Δv_w (e.g., Δv_{smile} for a "smile" feature), many of which are available from online open-source codebase. In cases where a direction vector for a specific feature is unavailable off-the-shelf, we propose the procedure in Appendix C.2.2 to obtain the vector.

After encoding each image x into its latent space, we adjust its embedding along Δv_w and feed the result into the generator to generate it back to the image:

$$\tilde{x}_i = G\big(v_i \pm \alpha \, \Delta v_w\big),$$

where α is a continuous scale parameter. By increasing or decreasing α, we audit the output images $\tilde{x}$ at both extremes and select those reflecting the desired feature alterations.

In Experiments 1 and 2, we apply this technique to manipulate features such as *gender* and *smile*, as well as *age*, *hair color*, and *glasses*. Figures 15 and 16 illustrate examples of these modifications using this approach.

Figure 15: Experiment 1: Example of facial attributes alternations (Gender and Smile) via the corresponding gradient

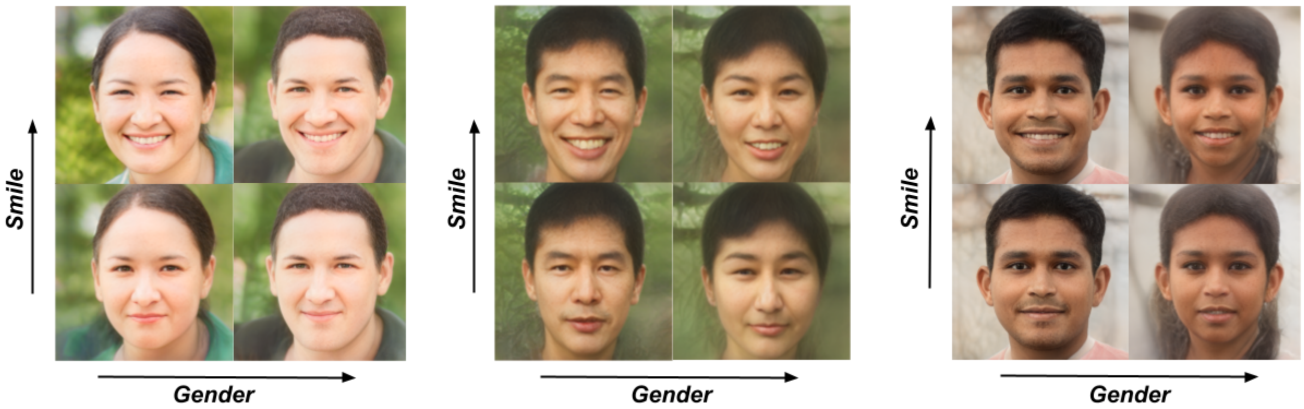


Figure 16: Experiment 2: Example of facial attributes alternations (*Age*, *Hair Color*, and *Glasses*) via the corresponding gradient

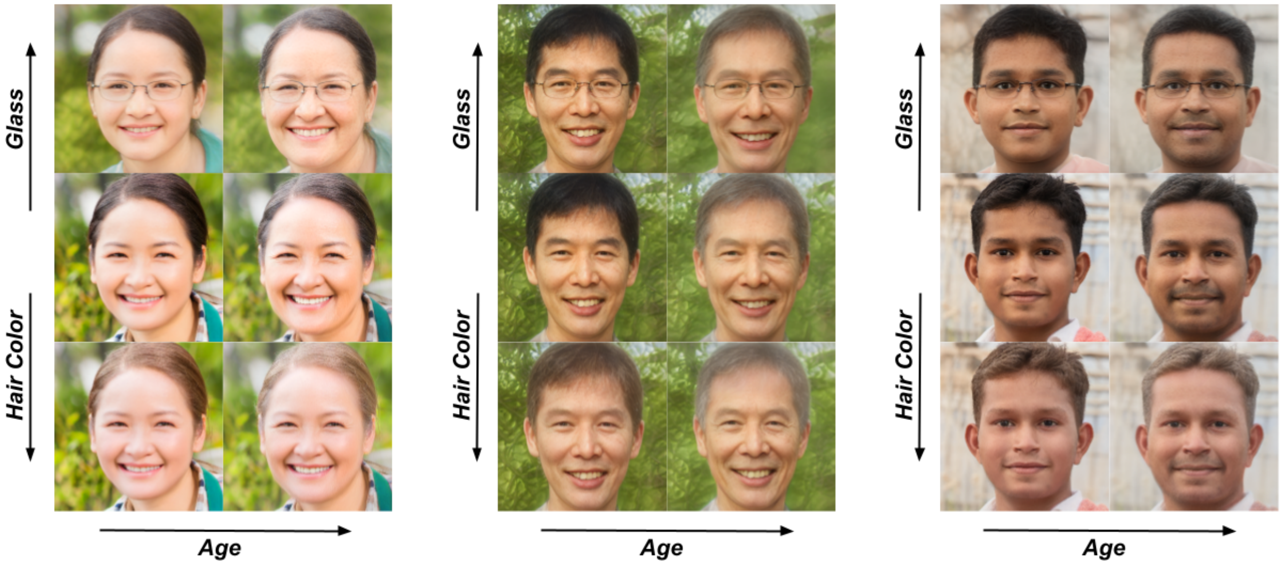


Image features usually have correlations, and this method works even in the presence of correlation between the visual features in the images. We introduce hyper-parameters to control the degree of alteration in the desired direction, and we fine-tune these parameters on a per-image basis. Once the attribute is modified, it is integrated back into the original (unmodified) image. To ensure the result appears realistic and seamless, we apply a series of post-processing steps, including deblurring, inpainting, and auto-blending. After post-processing, the margins around the human headshot are inpainted and blended with the background. Figure 10 in our manuscript shows the full images after post-processing.

C.2.2 Identify direction vector

Since not all direction vectors are readily available off-the-shelf, we adapt recent work and propose a method to derive the direction vector e.g. Δv_{gender} for image manipulation in latent space.

We assume there exists a bijection mapping between the feature of interest in the image and the relevant dimensions of its corresponding latent embedding vector. Consequently, we can use the direction vector to obtain the "potential outcome" of the image with that feature altered, while holding other attributes constant. This parallels the concept of Average Treatment Effects in causal inference, and by using it in the opposite way, we can find the relevant latent dimensions by computing the difference in mean latent embeddings between images that exhibit the feature and those that do not, with other features randomly selected.

Consider the *race* feature as an example (though it is ultimately outside this paper's scope). No published, open-source pre-trained models provide an off-the-shelf direction vector for *race* partially due to its complexity, especially when multiple racial categories (*Asian*, *Black*, *White*) must be toggled. To manipulate between *Asian* and *Black* in the latent space, for instance, we randomly sample about 200 images identified as *Asian* and 200 identified as *Black*, with other characteristics randomly chosen[39]. Let $w_i = 1$ indicate *Black* and $w_i = 0$ otherwise, with s_i indicating other features. Each image x_i yields a latent embedding v_i depending on (w_i, s_i) in the image, and we assume any change in some relevant latent dimension arises solely from the feature of interest, without other unmeasured confounders. We then identify the relevant dimension in the latent embedding that corresponds to the change of the interested feature via the following approach, under our assumption:

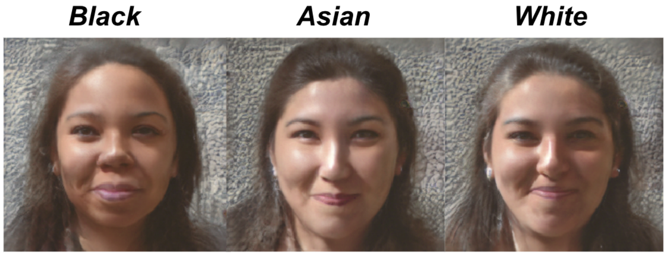
$$\Delta v_w = \mathbb{E}[v_i(1) - v_i(0)] = \mathbb{E}[\mathbb{E}[v_i|s_i, w_i = 1] - \mathbb{E}[v_i|s_i, w_i = 0]] \tag{5}$$

$$\Delta \hat{v}_w = \frac{1}{n_1} \sum_{i \in \{w_i = 1\}} v_i - \frac{1}{n_0} \sum_{i \in \{w_i = 0\}} v_i \tag{6}$$

where $n_1 = |\{w_i = 1\}|$ and $n_0 = |\{w_i = 0\}|$.

In this case, the approximated $\Delta \hat{v}_w$ serves as a direction vector that is, ideally, orthogonal to other image attributes, allowing for targeted manipulation of the specified feature without affecting unrelated characteristics. We showcase below how our proposed method alters *race* in a randomly selected image. Although *race* is beyond the scope of our study, it serves as an example of a feature for which the direction vector is particularly challenging to find.

Figure 17: Showcase of facial attribute alterations (*Race*) using the corresponding direction vector created by the above method.



[39] We used images labeled by CNN and verified by human audit

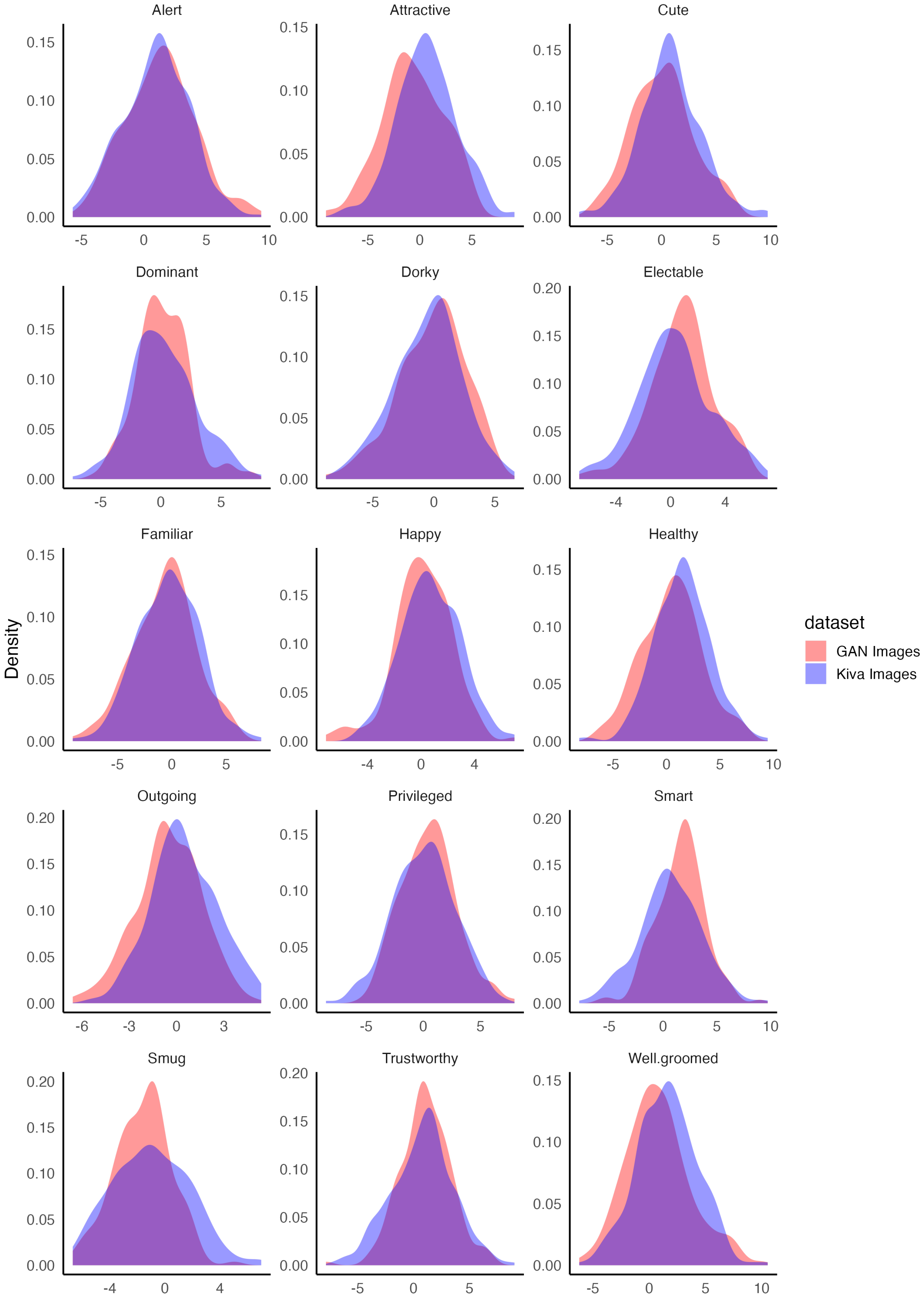
D Comparison statistics of *GANs* generated images and *Kiva* images

Recognizing the importance of ensuring our GAN-generated images reflect real-world variation in borrowers' profiles, the Appendix D documents the comparison statistics for both *GAN*-generated and *Kiva* images. To evaluate this alignment, again we use an external API provided by Peterson et al. (2022) to estimate psychological traits in each set of images.

We estimated psychological traits in both GAN-generated images and sub-sampled Kiva images. We include nearly 200 images from each set, obtain the scores of the psychological traits, and then plot out the score distribution. When comparing the trait distributions, we find that they substantially overlap, with similar means and standard deviations for the majority of traits.

While these checks do not provide definitive proof, they increase our confidence that our GAN-generated images effectively capture key perceived psychological attributes in a way that aligns with real borrower profiles from Kiva. It also demonstrates that any modifications in the generated images correspond to similar effects in psychological traits as perceived by lenders, aligning with the natural variations observed in the Kiva images.

Figure 18: Traits score density comparison between sub-sampled *GAN* (red) generated images and *Kiva* images.



E Summary statistics of *Kiva data*

In Appendix E, we present summary statistics for the complete set of Kiva variables.

Table 8: Summary statistics of *Kiva data*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Loan amount	420,765	800.107	993.370	25	275	950	50,000
Cash per day	420,765	123.522	270.186	1	25	116.7	8,750
Days to raise	420,765	13.427	11.667	1	5	20	83
Total number of lenders	420,765	0.012	0.015	0.001	0.005	0.015	0.967
default	420,765	0.050	0.218	0	0	0	1
Male	420,765	0.198	0.398	0	0	0	1
Number of borrowers	420,765	1.958	3.171	1	1	1	50
No. competitors	420,765	0.091	0.173	0.003	0.006	0.075	1.000
Same race gender share	420,765	0.665	0.294	0	0.4	1	1
Asian	420,765	0.191	0.261	0.0001	0.016	0.266	0.995
White	420,765	0.218	0.265	0.001	0.031	0.323	0.999
Black	420,765	0.167	0.281	0.0001	0.006	0.148	0.990
Baby	420,765	0.004	0.003	0.0001	0.002	0.006	0.067
Child	420,765	0.073	0.056	0.001	0.034	0.095	0.609
Youth	420,765	0.264	0.211	0.0002	0.092	0.391	0.982
Middle.Aged	420,765	0.084	0.093	0.0004	0.026	0.104	0.898
Senior	420,765	0.041	0.079	0.0001	0.004	0.039	0.950
Black.Hair	420,765	0.388	0.242	0.0005	0.171	0.589	0.970
Blond.Hair	420,765	0.007	0.029	0.00000	0.001	0.004	0.943
Brown.Hair	420,765	0.405	0.156	0.012	0.288	0.517	0.919
Bald	420,765	0.037	0.073	0.0001	0.004	0.030	0.835
No.Eyewear	420,765	0.865	0.148	0.007	0.830	0.959	1.000
Sunglasses	420,765	0.017	0.014	0.001	0.008	0.020	0.327
Mustache	420,765	0.072	0.160	0.00003	0.004	0.046	0.998
Smiling	420,765	0.549	0.177	0.013	0.424	0.685	0.966
Chubby	420,765	0.339	0.190	0.012	0.185	0.466	0.972
Blurry	420,765	0.162	0.095	0.006	0.090	0.214	0.758
Harsh.Lighting	420,765	0.339	0.165	0.031	0.217	0.430	0.930
Flash	420,765	0.245	0.126	0.010	0.148	0.322	0.855
Soft.Lighting	420,765	0.677	0.090	0.222	0.623	0.742	0.943
Outdoor	420,765	0.447	0.140	0.045	0.343	0.545	0.914
Curly.Hair	420,765	0.394	0.155	0.031	0.275	0.499	0.932
Wavy.Hair	420,765	0.226	0.170	0.004	0.095	0.312	0.991
Straight.Hair	420,765	0.606	0.178	0.034	0.489	0.741	0.982
Receding.Hairline	420,765	0.205	0.235	0.0004	0.039	0.282	0.995
Bangs	420,765	0.171	0.171	0.001	0.052	0.229	0.993
Sideburns	420,765	0.145	0.195	0.001	0.025	0.168	0.977
Partially.Visible.Forehead	420,765	0.094	0.090	0.001	0.032	0.125	0.834
Arched.Eyebrows	420,765	0.451	0.213	0.004	0.282	0.618	0.978
Narrow.Eyes	420,765	0.588	0.204	0.031	0.432	0.755	0.992
Eyes.Open	420,765	0.871	0.073	0.338	0.834	0.925	0.991
Big.Nose	420,765	0.730	0.190	0.042	0.606	0.886	0.998
Big.Lips	420,765	0.586	0.215	0.014	0.425	0.766	0.986
Mouth.Closed	420,765	0.303	0.146	0.018	0.193	0.390	0.944
Mouth.Wide.Open	420,765	0.057	0.040	0.002	0.030	0.072	0.516
Square.Face	420,765	0.019	0.041	0.00005	0.002	0.015	0.759
Round.Face	420,765	0.201	0.155	0.002	0.078	0.287	0.908
Color.Photo	420,765	0.948	0.026	0.632	0.935	0.966	0.997
Posed.Photo	420,765	0.486	0.132	0.069	0.391	0.581	0.925
Attractive.Woman	420,765	0.125	0.151	0.001	0.028	0.158	0.989
Indian	420,765	0.061	0.098	0.00002	0.009	0.066	0.962
Bags.Under.Eyes	420,765	0.586	0.170	0.016	0.468	0.717	0.967
Rosy.Cheeks	420,765	0.122	0.069	0.011	0.072	0.155	0.729
Shiny.Skin	420,765	0.215	0.121	0.004	0.121	0.288	0.808
Pale.Skin	420,765	0.334	0.171	0.014	0.192	0.460	0.908
Strong.Nose.Mouth.Lines	420,765	0.611	0.172	0.026	0.496	0.746	0.966
Flushed.Face	420,765	0.102	0.050	0.009	0.067	0.126	0.573
Top	420,765	157.544	106.715	0	80	204	1,598
Right	420,765	410.062	174.165	29	271	534	960
Bottle	420,765	0.503	2.259	0	0	0	99
Chair	420,765	0.125	0.498	0	0	0	24
Person	420,765	2.119	3.002	1	1	2	39
Bodyshot	420,765	0.406	0.491	0	0	1	1

F Choice of the predictive model

In this section, we present a comparison of the accuracy of the GBM to other predictive models. We consider several predictive models over three specifications and determine the model to be used in the baseline analysis.

We analyze the performance of models predicting *cash per day*. We consider the following models: Linear Regression, LASSO, Random Forrest (grf), and Boosted Random Forrest (grf and gbm). All models (except for LM) are tuned for the task at hand, we report the performance of the selected best (lowest MSE) model. All models are trained using a 70% sample of *Kiva data* and tested on the 30%.

We consider three specifications differing by the number of covariates: (A) covariates include: details of the loan including amount, repayment scheme, *sector*, *country*, etc. and weekly dummies, (B) details of the photo including both *type* and *style* characteristics, (C) total number of active lenders in this *week*sector*, total number of competitors in this *week*sector*, number of competitors of the same *race* and *gender*, and interaction of *week* and *sector*, and interaction of *week* and *country*. For boosted Forrest we also add a 4th specification where we have a sufficient representation of *week* sector* (D) (Johannemann et al., 2019). Table 9 presents results.

Table 9: Comparison of the test-set predictive performance of selected models

Model	Specification	MSE	SE
Linear regression	A	13840	159
Linear regression	B	13466	155
Linear regression	C	13565	166
LASSO	A	13797	161
LASSO	B	13379	157
LASSO	C	13183	156
Random forest	A	13930	163
Random forest	B	13530	145
Random forest	C	13099	157
Boosted forest (gbm)	A	12235	156
Boosted forest (gbm)	B	11477	141
Boosted forest (gbm)	C	10929	157
Boosted forest (gbm)	D	11406	173
Boosted forest (grf)	A	12665	147
Boosted forest (grf)	B	12003	149
Boosted forest (grf)	C	11777	139
Boosted forest (grf)	D	11962	177

Note: Test set performance of selected predictive models with different sets of covariates.

We conclude that Boosted Forrest has the best test-set predictive performance across all specifications and we decide to use it as a baseline model for the predictive tasks throughout the paper. *GBM* implementation of the Boosted Forrest has better performance than *GRF*, the difference is moderately small. Sufficient representation does not improve models' performance and will not be used in the predictive tasks.

G Model Diagnostics

In Appendix G, we expand the set of outcome variables and consider a constructed variable, which adjusts for systematic differences in cash per day and total loan amounts requested across business categories. It also provides diagnostics for the GBM models.

G.1 Alternative outcome variable

To account for heterogeneity in the supply-side motivations of lenders on Kiva, we introduce a new outcome variable that captures the cash collected by a borrower relative to the total funds requested, adjusted for differences within and across categories. This adjustment acknowledges that lenders' preferences and funding behaviors vary not only by borrower characteristics but also by the categories Kiva uses to segment loans. Specifically, we compute the outcome variable as:

$$\text{Adjusted Outcome} = \left(\frac{\text{Cash per Day}}{\text{Category Average Cash per Day}}\right) \cdot \left(\frac{\text{Category Average Requested}}{\text{Platform Average Requested}}\right).$$

This formula incorporates two key components. First, the ratio of the cash collected per day to the category average cash collected per day accounts for differences within categories, capturing how well a specific loan is performing relative to others in the same category. Second, the ratio of the category average requested to the overall platform average requested adjusts for differences across categories, recognizing that funding needs and typical loan amounts can vary significantly across different types of projects (e.g., drilling a town's well versus supporting a seed entrepreneur). Together, this outcome variable provides a more nuanced measure of lender behavior, addressing both intra-category and inter-category differences and accounting for potential supply-side heterogeneity.

G.2 Diagnostics for the GBM models

Table 10 shows test-set MSE for the three GBM models: a constant model, a model with *style* features, and *full model*. Note that differences in outcomes across categories are already partly accounted for in

Table 10: Image features as predictors of *adjusted outcome*.

Specification	MSE	SE
Mean model	1.435	0.018
Style features	1.244	0.016
Full model	1.232	0.016

Note: Test set performance of a gradient boosted machine (GBM) trained using all available covariates (full model) and simplified model using image style features (Style features) and a model with only an intercept (Mean model). Models trained on 70% of data and tested on 30%. Mean squared errors are in the second column. Standard errors of MSE are in the third column.

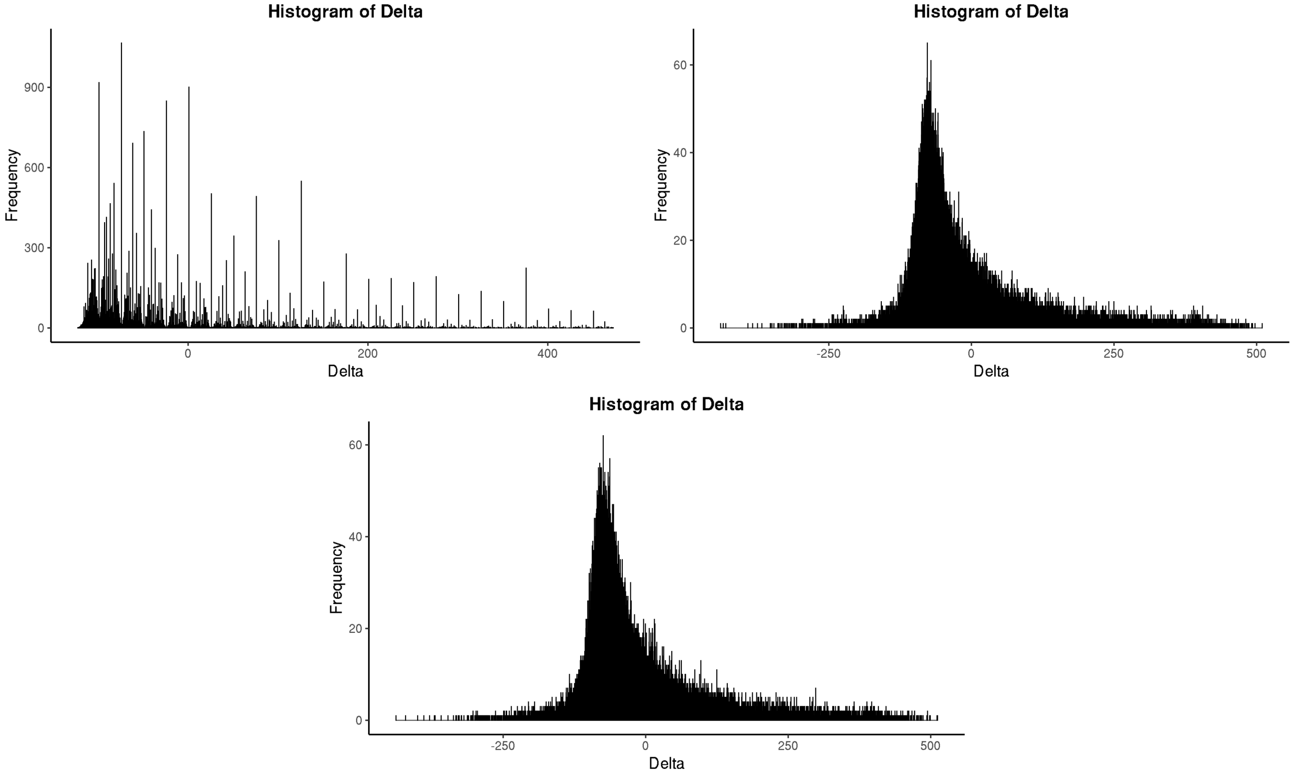
the outcome variable.

We find that including *style* features in the predictive model of the adjusted outcome improves the predictive accuracy as measured by the MSE. The difference between the MSE of the mean model and the model with *style* features is statistically significant. The improvement due to inclusion of all other covariates in the *full model* from the *style* model is statistically insignificant.

Now we show diagnostic plots from the three models of Table 10. Figure 19 shows histograms of error terms from the three models of cash per day evaluated in Table 2. We can notice that error terms from models adjusting for *style* features and then for all other features are similar.

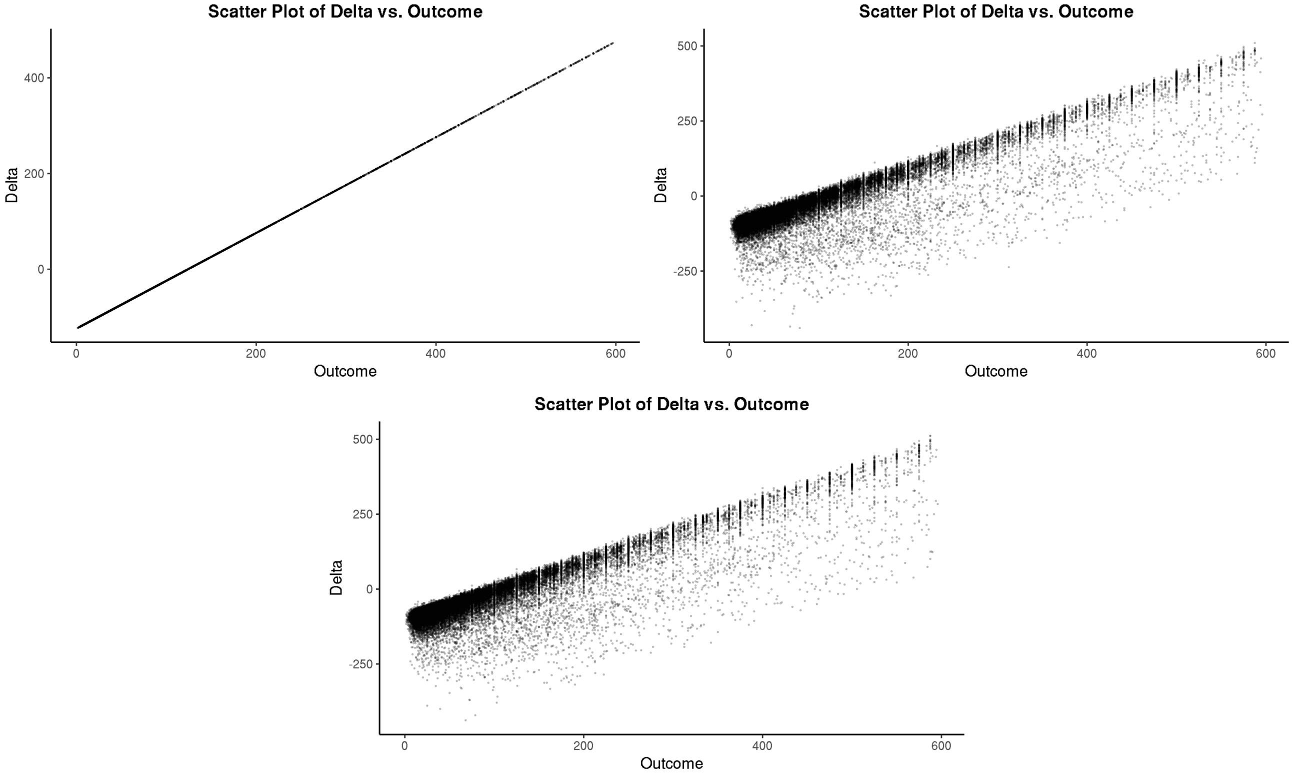
We start with the histograms of the standard errors in Figure 21, and Figure 20 plots fitted values and observed values for cash per day. Additionally, we also report scatter plots of predicted and observed outcomes in Figure 22.

Figure 19: Histograms of error terms for predictive models of cash per day



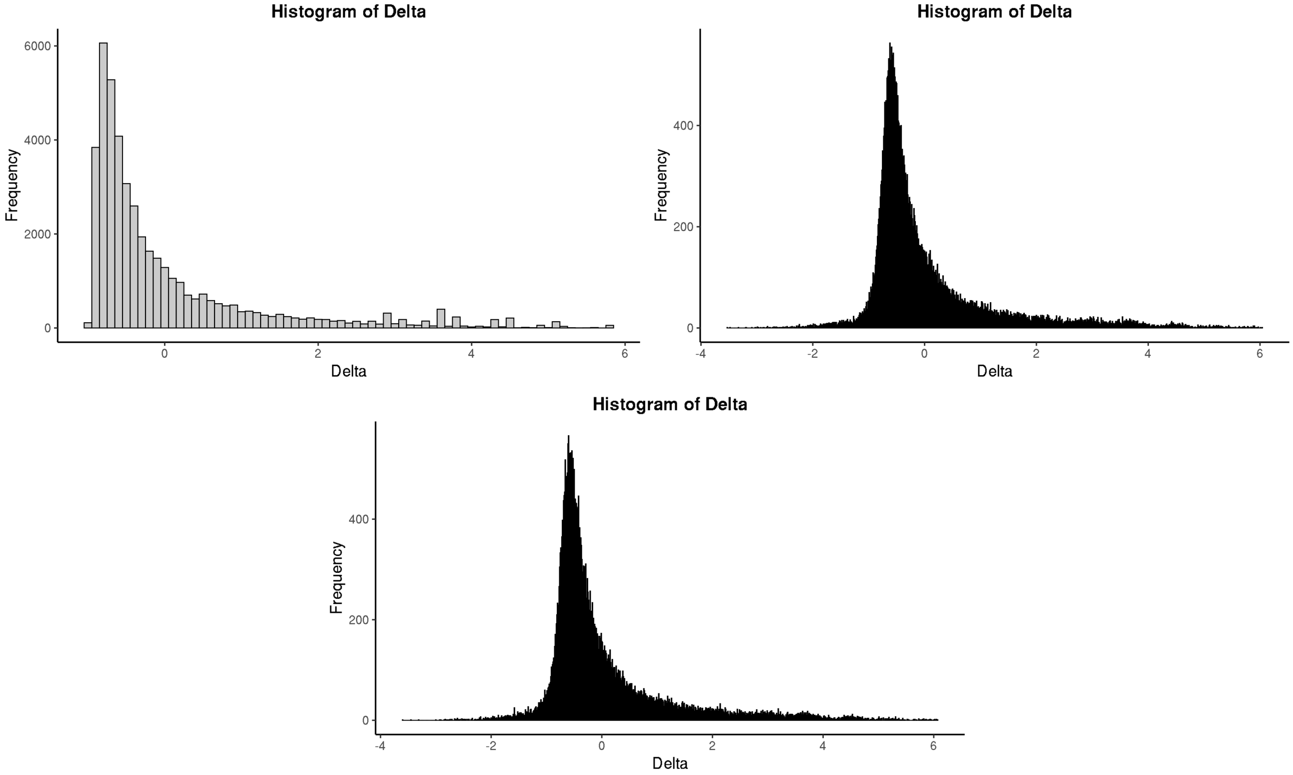
Note: The first figure shows the histogram of errors from the model with the constant term; the second figure shows the histogram of the errors from model with style features only, and the last figure shows the histogram of the errors from the full model.

Figure 20: Scatter plots of observed and predicted values



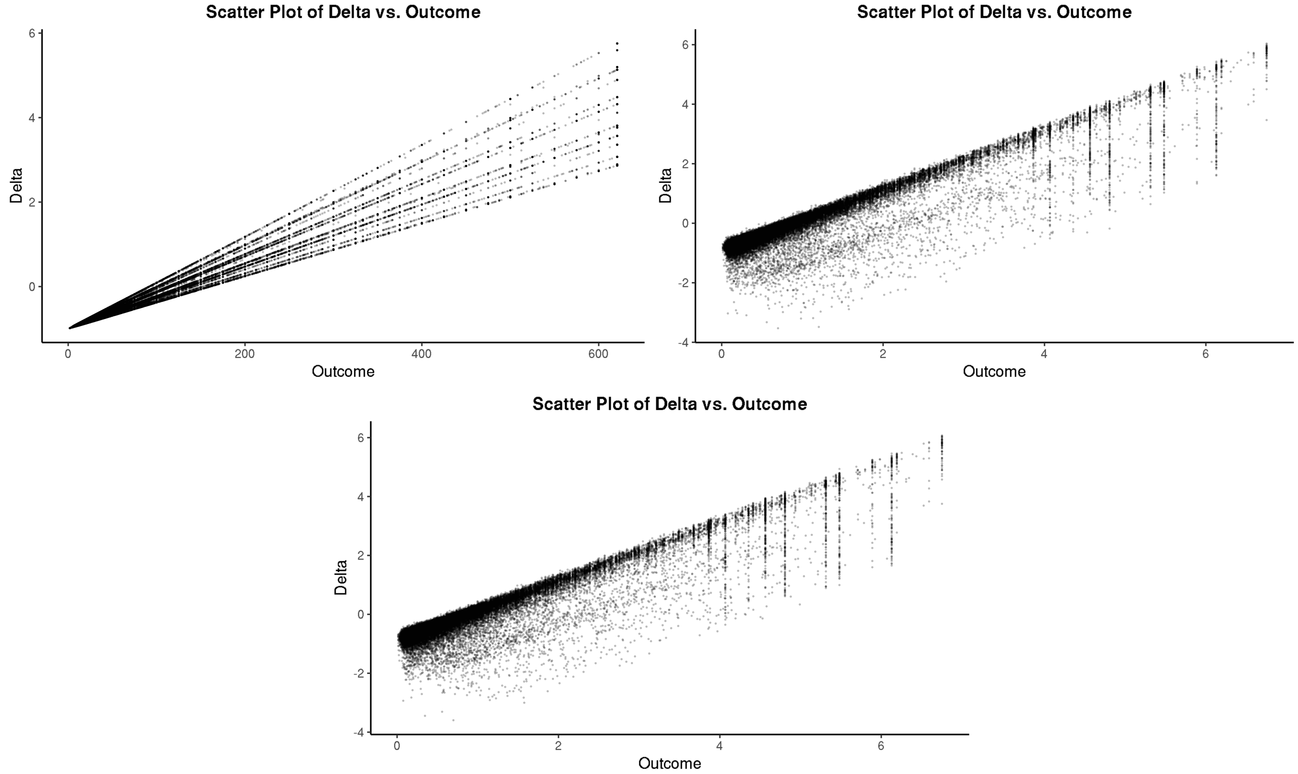
Note: Y-axis in all figures shows the predicted value from the model and the x-axis the the observed value. The first figure shows values form the model with the constant term; the second figure from model with style features only, and the last figure from the full model.

Figure 21: Histograms of error terms for predictive models of cash per day



Note: The first figure shows the histogram of errors from the model with the constant term; the second figure shows the histogram of the errors from model with style features only, and the last figure shows the histogram of the errors from the full model.

Figure 22: Scatter plots of observed and predicted values of the *Adjusted outcome*



Note: Y-axis in all figures shows the predicted value from the model and the x-axis the the observed value of the adjusted outcome. The first figure shows values form the model with the constant term; the second figure from model with style features only, and the last figure from the full model.

H Analysis of defaults across default types

In this section, we analyze alternative outcomes related to loan repayment. The non-repayment of a Kiva loan is typically attributable to a borrower's default. However, in certain instances, a microfinance organization's default might precipitate the non-repayment. It is plausible that the borrower's image features might be indicative of the borrower's propensity for repayment but not necessarily predictive of a default by the microfinance organization. To investigate this, we define a new outcome variable, *default by borrower*. This variable takes the value of 0 in scenarios where the loan has been repaid or the microfinance organization has defaulted; otherwise, it takes the value of 1.

In some cases, a borrower defaults not on the entirety of the loan but only on a part of it. To capture this we introduce an outcome *share not repaid* that takes values between 0 and 1 and represents the proportion of funds left unpaid by the borrower.

First, we analyze whether *style* features are predictive of these new outcomes. We train a Boosted Forrest (GBM) on 70% of data and report the predictive performance on the 30% test set. We consider three model specifications: a constant model, a model incorporating all *style* features, and a full model which includes all covariates. The results are presented in Table 11. We find that *style* features do not improve the predictive performance of models of either of the outcomes.

Table 11: Comparison of the test-set predictive performance of default models with and without image features.

Outcome variable	Model specification	MSE	Standard error
Default borrower	Constant	0.0058	0.0004
Default borrower	Style	0.0057	0.0003
Default borrower	Full	0.0039	0.0002
Share not repaid	Constant	0.0030	0.0002
Share not repaid	Style	0.0033	0.0002
Share not repaid	Full	0.0031	0.0002

Note: Test set performance of selected predictive models with different sets of covariates.

I Supplementary analysis for AIPW estimates of style features

This Appendix I demonstrates that propensity scores estimated via our machine learning method yield better covariate balance and produce a more closely matched treatment and control group. In Appendix I.1, we detail the balance assessment after weighting the treatment and control observations using these propensity scores. Appendices I.2 and I.3 illustrate the density plots for the propensity score distributions across different features, provide examples of features removed through this procedure, and show those retained for further analysis.

I.1 Diagnostics for selected *style* features

In observational settings, the covariate distributions can differ substantially between treated and untreated individuals, potentially biasing estimates of the average treatment effect (ATE). By adjusting each observation's weight based on its inverse propensity score, we aim to make these distributions more comparable. This section presents the balance check results using the absolute standardized mean difference (ASMD) of covariates in the treatment group (e.g. with *Bodyshot*) versus the control group, before and after propensity-score weighting. We find that applying propensity-score weighting improves the ASMD values, indicating better balance between the treated and untreated groups.

Diagnostics *bodyshot*. Figure 23 shows the ASMD of covariates across the treatment group (with *Bodyshot*) and control. We see that the adjusted values (yellow) are well-balanced. Specifically, the left-hand side of Figure lists the covariates, revealing that the variables we introduced — along with correlated variables — were far from balanced before weighting (blue dots). After applying propensity-score weighting, however, the absolute standardized mean differences (ASMDs) move closer to zero, indicating improved balance.

Diagnostics *smile*. Figure 24 shows standardized absolute mean differences of covariates across the treatment group (with *smile*) and control. We see that the adjusted values (yellow) are well-balanced.

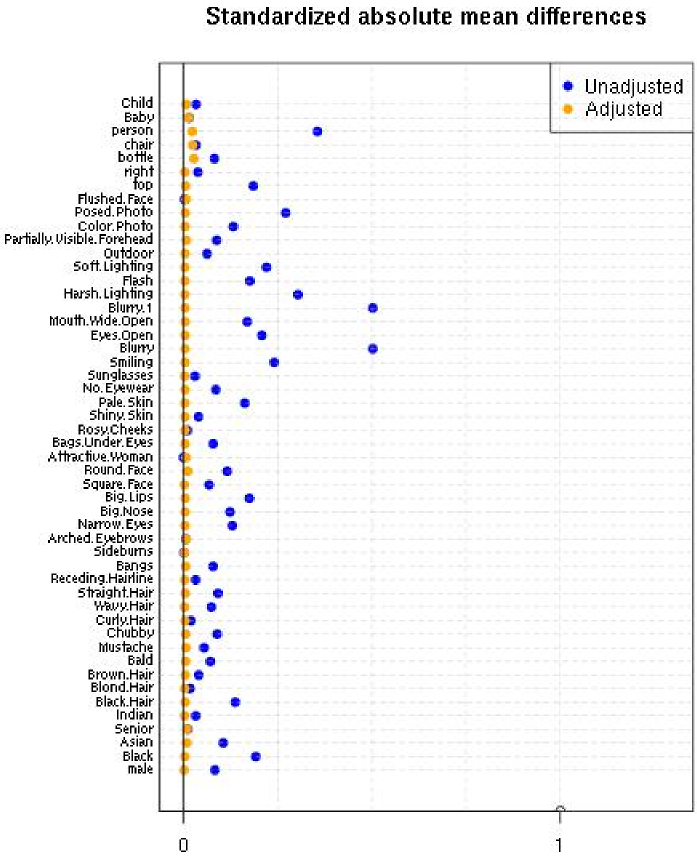
I.2 Density Plots of Dropped Features

This section provides density plots illustrating the propensity score distributions for features that were excluded from the analysis due to insufficient overlap between the treatment and control groups. As discussed in the main text, features were dropped if either the treatment or control group had propensity score mass below 0.1 or above 0.9, indicating limited comparability between the groups. Figure 25 highlights the lack of overlap in the propensity score distributions for specific *style* features.

I.3 Density Plots of Considered Features

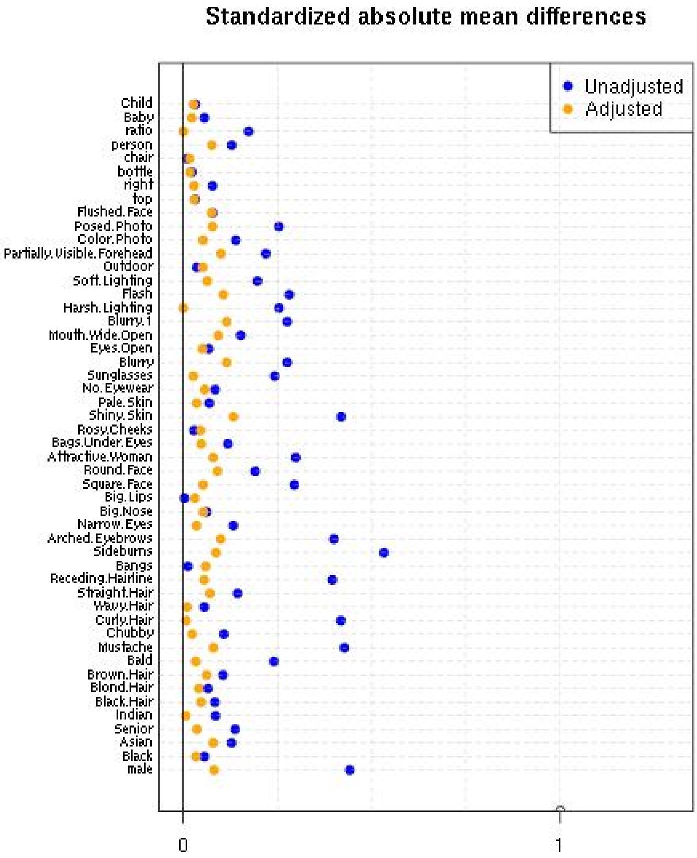
This section shows the propensity density plots for selected *style* features included in the analysis. Figure 26 presents the density plots of propensity scores for the non-dropped *style* features among

Figure 23: Diagnostics for *Bodyshot*



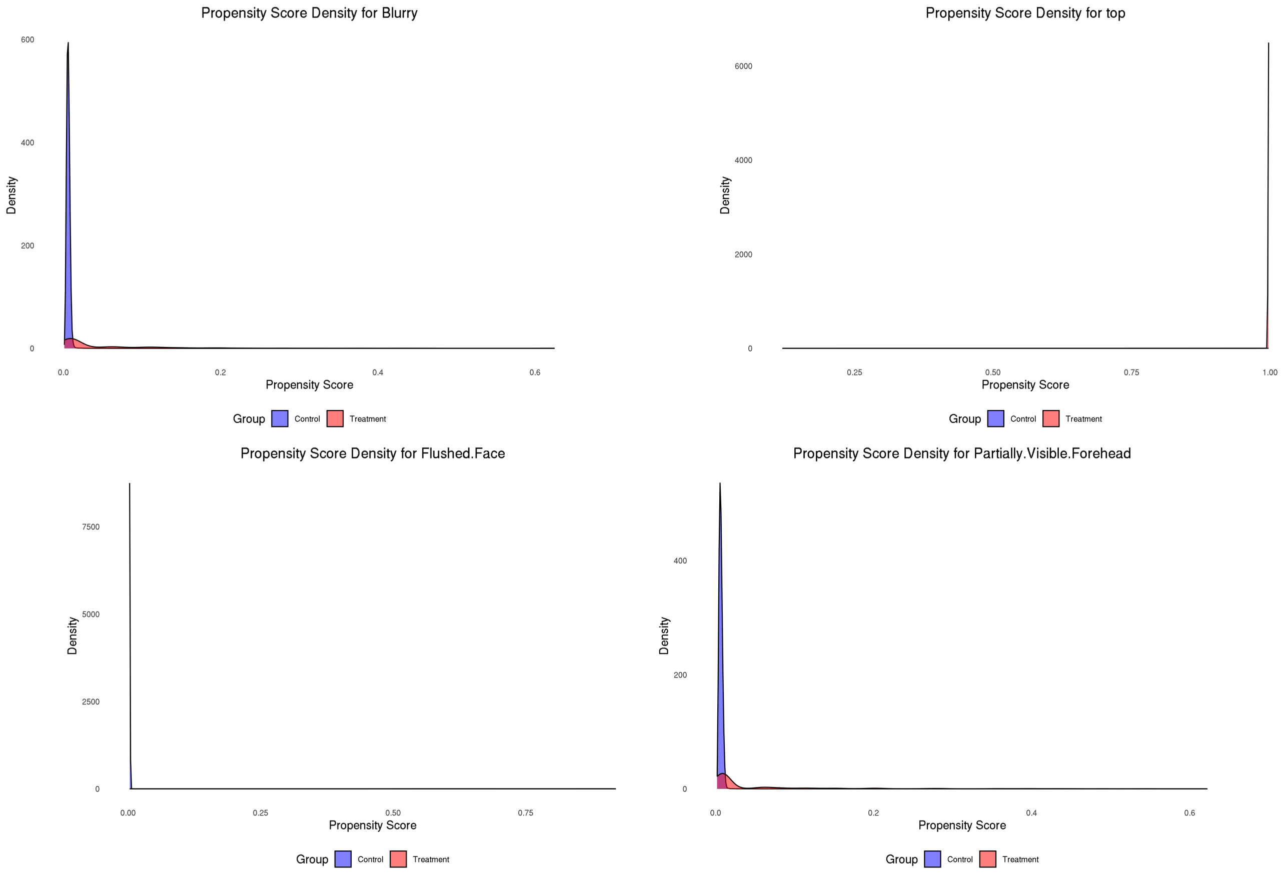
Note: Standardized absolute mean differences of a selected subset of other covariates across profiles with and without bodyshot. *Propensity score used for reweighing obtained using GBM model trained on all covariates in* Kiva data.

Figure 24: Diagnostics for *Smile*



Note: Standardized absolute mean differences of a selected subset of other covariates across profiles with and without smile*. Propensity score used for reweighing obtained using GBM model trained on all covariates in* Kiva data*.*

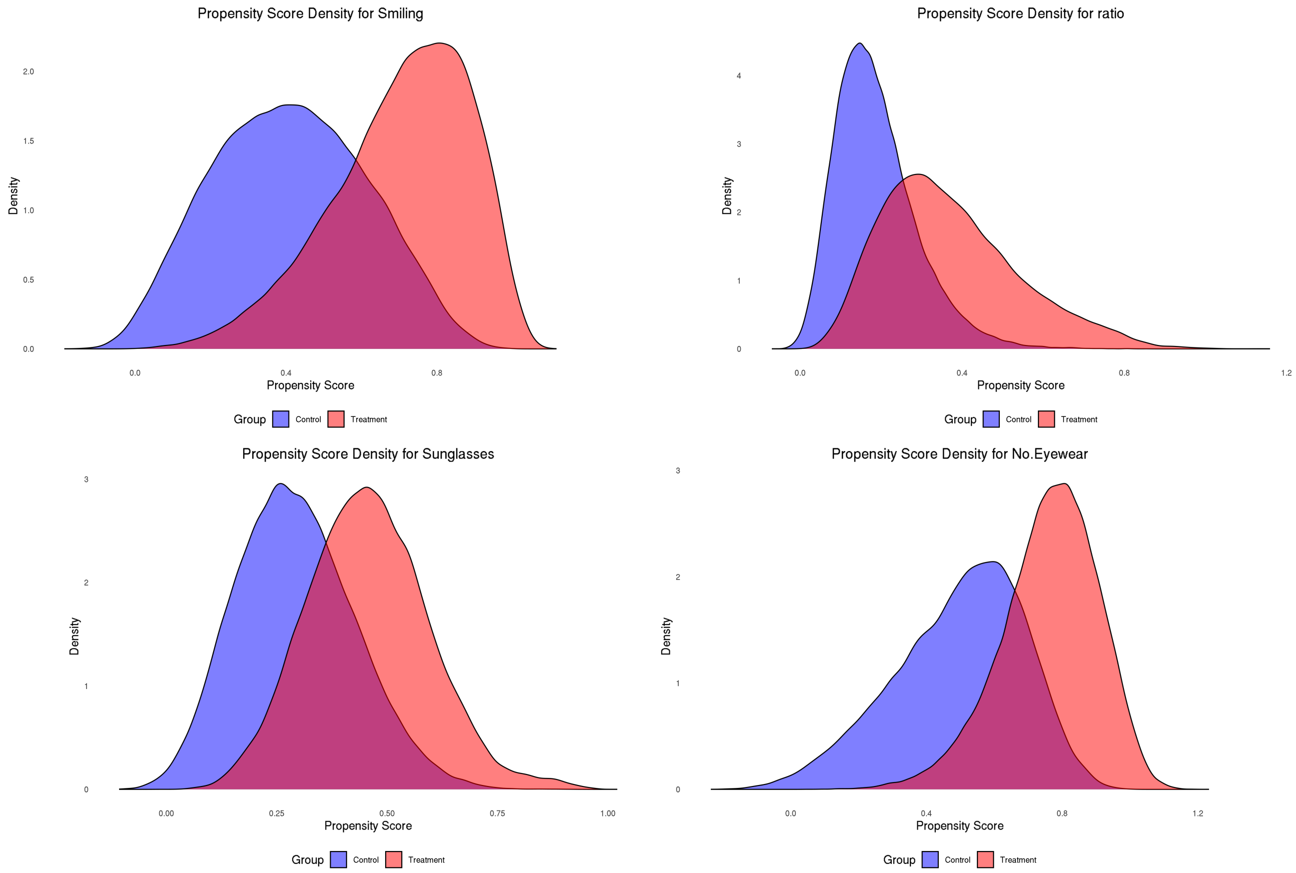
Figure 25: Propensity Estimates of Dropped Style Features



Note: These density plots show the gradient boosted machine (GBM) model estimates of the propensity scores for the features Blurry, Top, Flushed Face, *and* Partially Visible Forehead. *The lack of overlap between treatment and control groups is evident, justifying their exclusion from the analysis.*

profiles with and without such features. As the figure illustrates, there is substantial overlap between the two groups, indicating common support in the distribution of these features.

Figure 26: Propensity Estimates of Selected Style Features



Note: These density plots show the gradient boosted machine (GBM) model estimates of the propensity scores for the features Smiling, Bodyshot, Sunglasses, *and* No Glasses. *The lack of overlap between treatment and control groups is evident, justifying their exclusion from the analysis.*

J Excluding potential mediators in ATE estimation

In this section , we examine whether certain image-extracted features could, in principle, be "caused" by a given *treatment feature*, making them potential moderators rather than appropriate controls in an AIPW estimator. For example, if the treatment feature is *Smiling*, then a feature like *Crooked Teeth* would only be observed when *Smiling* = 1, implying that it should not be included as a control.

We focus on three treatment features: *Sunglasses*, *Bodyshot*, and *Smiling*. To assess the potential impact of problematic controls, we estimate two AIPW models: one with the full set of covariates and one in which we exclude potentially endogenous features. Specifically, we remove *Glasses*, *Narrow Eyes*, *Eyes Open*, and *Bags Under Eyes* for *Sunglasses*; *Outdoor*, *Bottle*, *Chair*, *Harsh Lighting*, *Flash*, and

Soft Lighting for *Bodyshot*; and *Mustache*, *Mouth Closed*, *Mouth Wide Open*, and *Strong Nose-Mouth Line* for *Smiling*. Table 12 presents the results. We find that excluding these features does not substantially alter the estimates.

Table 12: Estimated Average Treatment Effects (ATE)

	ATE Estimates (Model 1)		**ATE Estimates (Model 2)**	
	Estimate	**Std. Err.**	**Estimate**	**Std. Err.**
Sunglasses	-7.42	0.64	-7.54	0.68
Bodyshot	-21.27	0.72	-21.43	0.64
Smiling	15.54	0.74	14.16	0.67

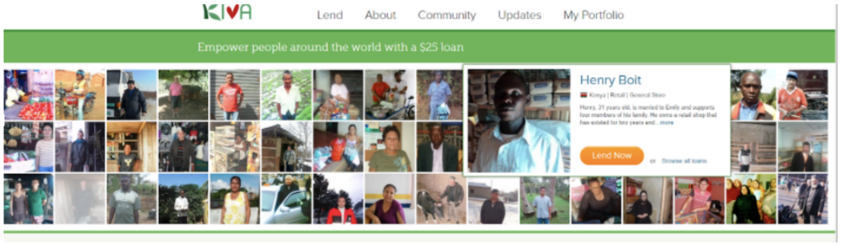
Note: This table presents the Average Treatment Effects (ATE) estimates for two models. Model 1 has the full set of controls; in Model 2 we are excluding potential mediators from the list of covariates. Model 1 estimates are reported in the second and third columns, while Model 2 estimates appear in the fourth and fifth columns. Standard errors are provided in parentheses.

K Stability of coefficient estimates

In the analysis of counterfactual platform policies, we assume that the lenders' preferences for image features are stable across different market structures. However, the impact of specific image features might vary across different ways of organizing the marketplace, particularly when they're used as quality signals. The extent to which image features affect beliefs about quality depends on lenders' beliefs about how common it is for high-quality borrowers to have images with those features, and how frequent these features are in general. If a change in the market design alters the set of borrowers lenders consistently see, it's plausible that it will consequently shift both their prior beliefs about the borrowers' quality and their perception of how image features impact their posterior beliefs about quality. In the extreme case, when all borrowers a lender sees on the platform have a certain image feature, that image feature will not affect the beliefs about quality. In contrast, when image features impact lenders' utility from selecting a borrower, they will impact the outcomes irrespective of lenders' beliefs about how informative image features are of the borrowers' quality.

To test this, we exploit a natural experiment in the form of the Kiva landing page redesign. On the 28th of May 2016, Kiva carried out a major website change; before that, all borrowers were displayed on the same page (see Figure 27). In the updated design, borrowers were sorted into categories. Figure 28 shows the available categories as displayed on Kiva's new landing page. After the change, lenders can quickly select categories, and in doing so, they'll see a different pool of borrowers than before the website update. If *style* features are used to compare the available borrowers and mostly act as signals

Figure 27: Kiva website prior to 28th of May 2016

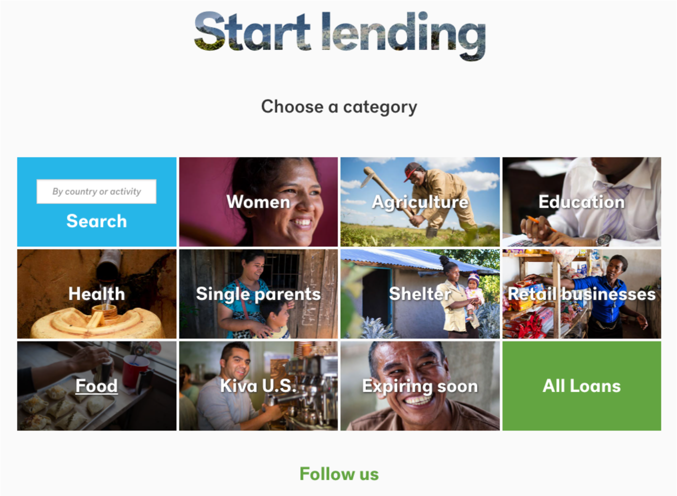


Henry, 31 years old, is married to Emily and supports four members of his family. He owns a retail shop that has existed for two years and... more

Note: Screenshot from https://kiva.org landing page. Source - https://archive.org/

of underlying quality, then it's plausible that the change in the way borrowers are displayed should change the impact of *style* on lenders' choices.

Figure 28: Borrowers' categories introduced after 28th of May 2016



Note: Screenshot from https://kiva.org landing page. Source - https://archive.org/

To evaluate the stability of *style* coefficients, we consider two periods: period *before*, which starts on 5/28/2015 and ends on 4/28/2016, and the period *after*, which starts on 5/28/2016 and ends on 8/28/2017. Within these periods, the website was organized following the same logic, but across the periods, the display of borrowers differed. We end the *before* period one month before the change so that most of the borrowing campaigns posted in the *before* period would have ended before the introduction of the new system.

We estimate the average treatment effect of *smile* on *cash per day*. We use the AIPW estimator, where the propensity and outcome models are estimated using Gradient Boosted Trees. Table 13 presents the estimated average treatment effects and their difference. We find that the impact of *smile* on *cash per day* was statistically significant and positive in each of these two periods. Their difference is statistically insignificant. However, we also note that either of these estimates is lower than the estimate from Table 5. This can, for example, reflect the change in the user base of Kiva over time. In the main analysis, we restricted attention to borrowing campaigns posted between 2006 and 2016, as we only had information on loan repayment for this period.

Table 13: Comparison of the impact of *Smile* before and after website redesign.

Period	Estimate	Std.err
Before period	2.842	0.339
After period	2.781	0.506
Difference	0.061	0.609

Note: Estimates of the impact of smile on cash-per-day before and after the website redesign. Estimates from the AIPW estimator. Last row estimate before less estimate after.

L Attention checks in the experiment

The Appendix L shows the details on the design and outcomes of these checks. To check the quality of experimental data, we included attention checks in the survey. Attention checks are questions designed explicitly to detect inattentive subjects through additional questions (Abbey and Meloy (2017)). There are three purposes of the attention checks in our experimental setting: first, attention checks provide a signal of whether a recruited subject is paying attention to the information on the screen. Second, attention checks encourage the subjects to make thoughtful decisions. In addition, attention checks also give us the flexibility to filter the data in order to have high-quality ones, depending on whether we would like to tighten or loosen our criteria.

In order to avoid the attention checks themselves inducing a deliberative mindset and becoming a threat to the validity, we try to ask the subjects to recall detail in a previous image after they make the choice and the correct answer to that gives us the reason to believe that people have been paying rational attention to their choices.[40]

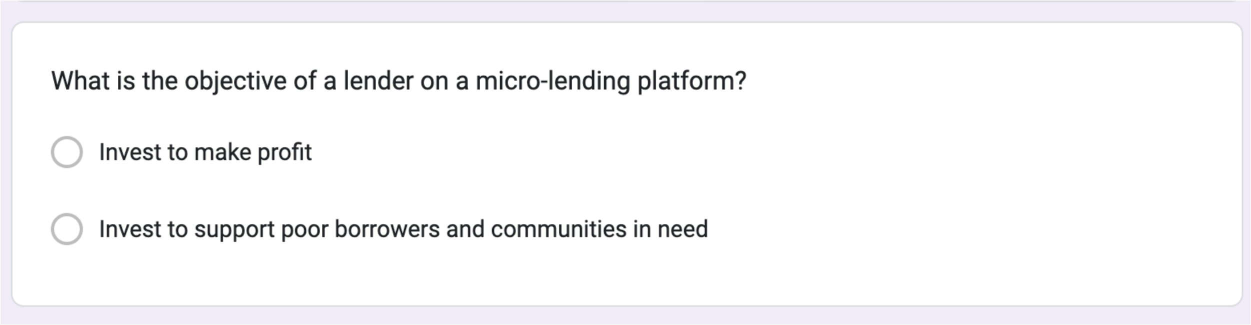
The attention check in Figure 29 asks *What is the objective of a lender on a micro-lending platform?* This question asks about information provided on the first slide of each protocol.

Attention checks in Figures 30 and 31 are conducted in the format of a quiz. Attention check 2 is an open-ended query asking the subject for the reason for their decisions.[41] The last check is a multiple choice query asking about the occupation of the borrower on the previous slide.

[40] Kung et al. (2018) encourage researchers to justify the use of attention checks without compromising scale validity

[41] Abbey and Meloy (2017) uses this type of attention checks and manipulation validations to detect inattentive respondents in primary empirical data collection

Figure 29: Attention check 1

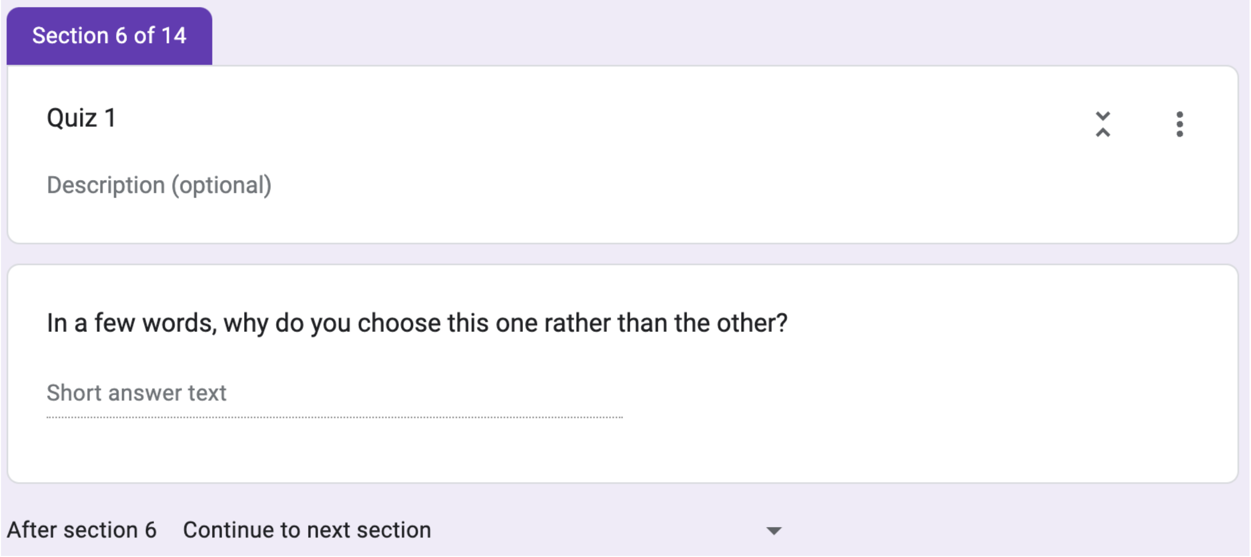


What is the objective of a lender on a micro-lending platform?

○ Invest to make profit

○ Invest to support poor borrowers and communities in need

Figure 30: Attention check 2

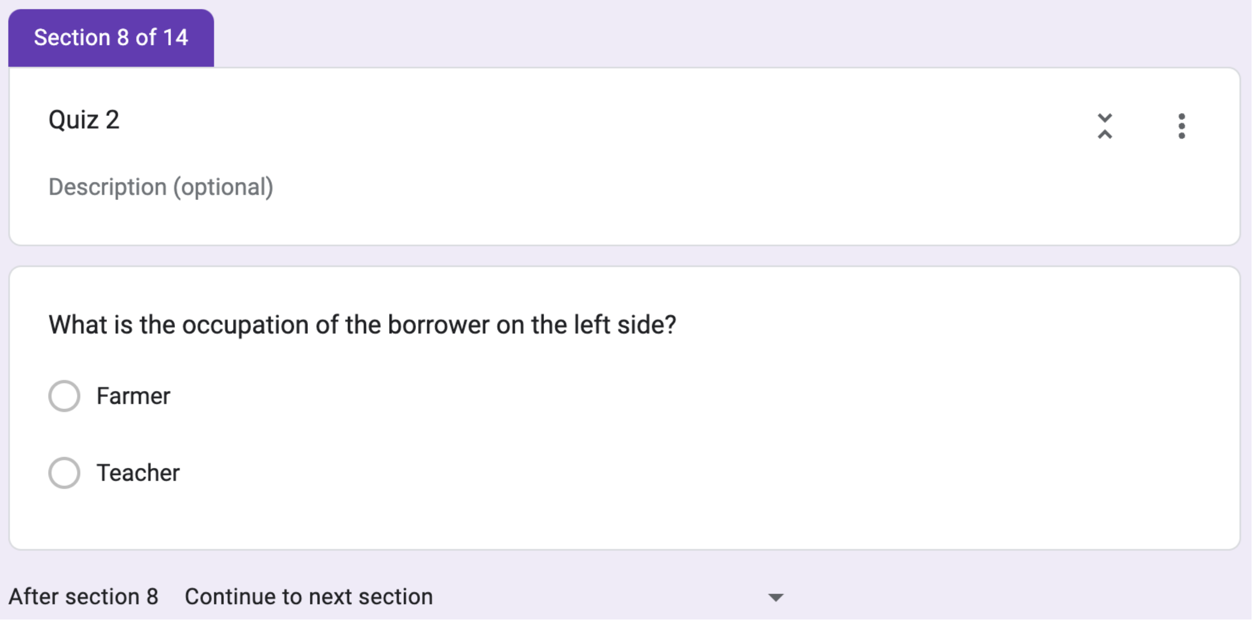


Quiz 1

In a few words, why do you choose this one rather than the other?

After section 6 Continue to next section

Figure 31: Attention check 3



Quiz 2

What is the occupation of the borrower on the left side?

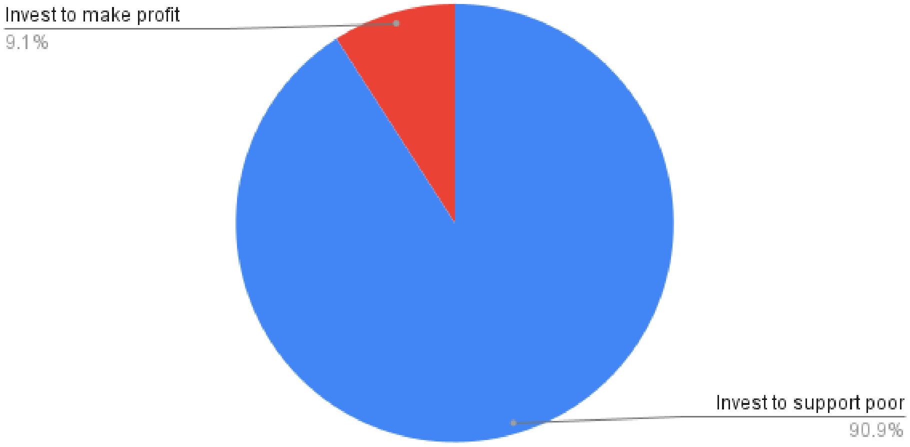
○ Farmer

○ Teacher

After section 8 Continue to next section

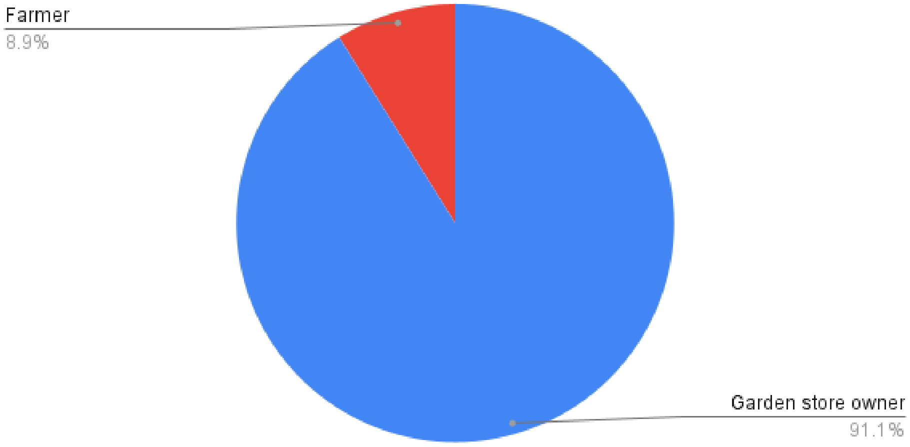
Figures 32 and 33 show shares of subjects that responded correctly to Attention check 1 and 3. In both cases, correct response rates are above 90%. We take this as an indication that subjects were generally paying attention to their choices.

Figure 32: The proportion of correct answers (blue) to the object of a lender



Note: Count of responses to the question: "What is the objective of a lender on a micro-lending platform?"

Figure 33: The proportion of correct answers (blue) to the borrower's occupation is shown on the previous page



Note: Count of borrower's occupation shown in the previous page

In experiment 2, to introduce financial incentives and improve realism and ensure that subjects had a tangible stake in their decisions, for each participant, we allocated $10 in individual-specific, dedicated real funds to one of the borrowers selected by the participant. Additionally, we include a comprehension check related to the purpose of micro-lending to confirm that respondents understood the context of the platform and the role of lenders. Almost all participants answered the question correctly.

M Summary statistics of the recruited subjects

In Appendix M, we present summary plots of Country of Residence, Employment Status, Sex, and Socioeconomic Status. We recruited 400 subjects who had donated at least USD 1 to charity in the previous year; 60% contributed less than USD 75. Figure 35 shows the summary histogram.

We considered subjects from developed countries with high socioeconomic status (self-reported); we required subjects to have at least a score of 3 in their self-assessed socio-economic status. The United Kingdom is our subjects' most common current country of residence, with 40% of subjects, followed by Spain with 30%, and France with 9%. Majority of the subjects hold a full-time job or are onboarding a new position. 51% of the subjects are women and the mean age of subjects in the experiment is 33 years, and with the most subjects at the student age. This aligns with what is reported online ragarding Kiva lenders that the most frequent occupation of kiva lenders is Student.

In Experiment 2, we collected fewer demographic details because Prolific.co changes its format and drops certain items. However, the demographics of Experiment 2 are very similar to those in Experiment 1.

Figure 34: Experiment Demographics (1)

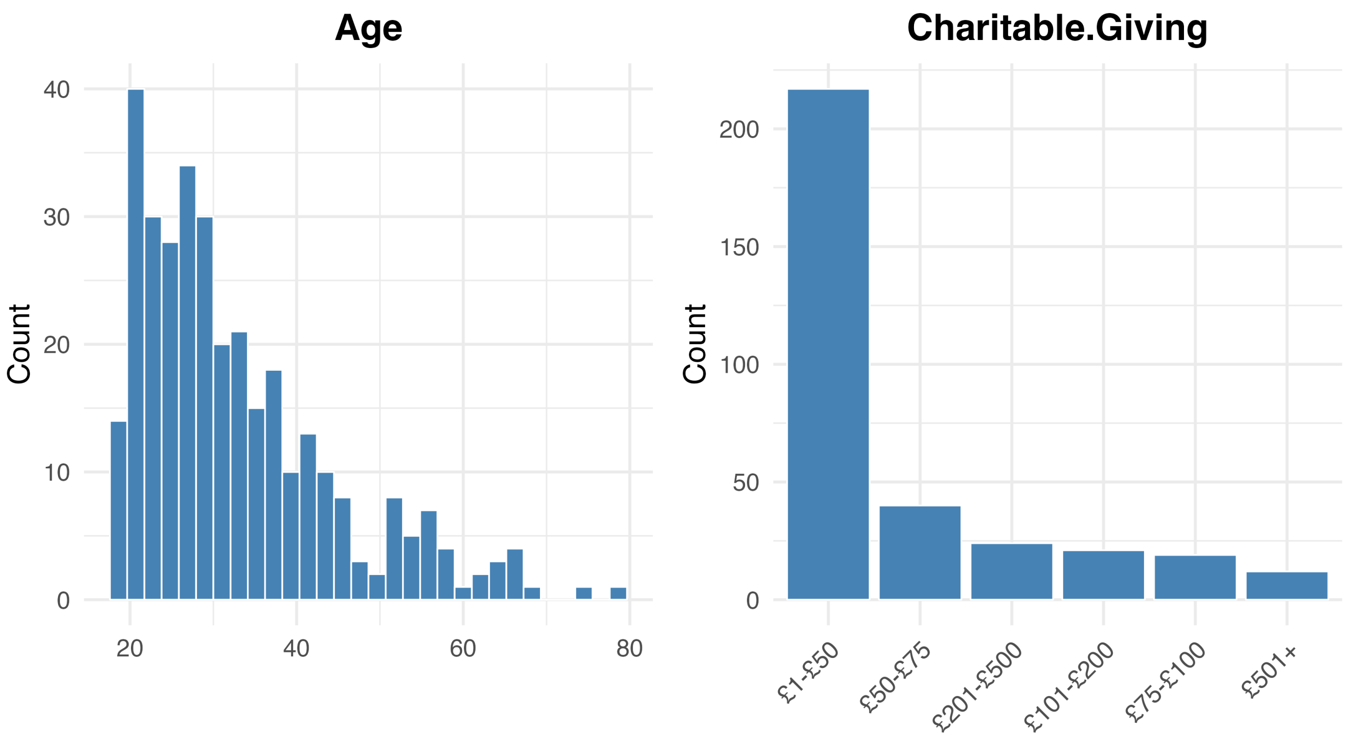
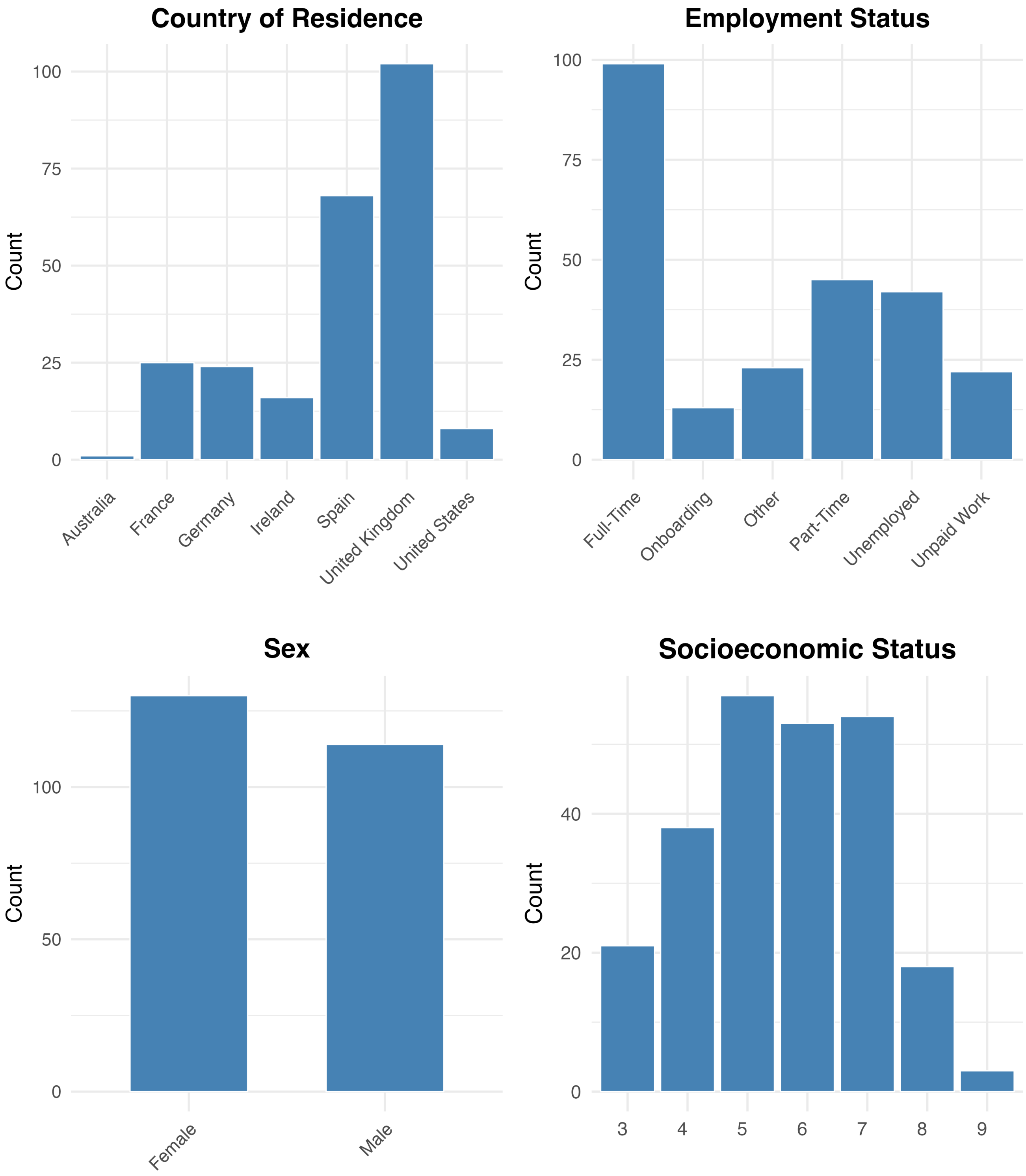


Figure 35: Experiment Demographics (2)



Note: Both socioeconomic status and employment status are self-reported. We drop observations where the data is unavailable (29%). A lot of missing data, to a large extent, is due to the employment information being expired. We required subjects to have at least a score of 3 in their self-assessed socio-economic status.

N Algorithm for counterfactual simulations

In this section N, we describe the algorithm for generating outcomes under counterfactual policies in more detail. We divide the algorithm into two parts: (i) simulation of a market, and (ii) simulation of lenders' choices. We focus on a simplified case in which we consider only *male*, *bodyshot*, and *smile*.

Algorithm 1 Simulation of a market

```
$\tilde{\eta} \leftarrow U(\mathcal{N};22)$    ▷ Draw 22 fixed effects uniformly from the set of estimated fixed effects
$\widetilde{male} \leftarrow \mathbf{E}_G[male|D(\tilde{\eta});22]$    ▷ Draw 22 gender realizations
$\widetilde{bodyshot} \leftarrow \mathbf{E}_G[bodyshot|D(\tilde{\eta}),\widetilde{male};22]$
$\widetilde{smile} \leftarrow \mathbf{E}_G[smile|D(\tilde{\eta}),\widetilde{male};22]$
if $H \in \{Partialcompliance\}$ then
    if $\widetilde{bodyshot} == 1$ then
        $\widetilde{bodyshot} = B_{0.25}$    ▷ Bernoulli trial with $p = 0.25$
    end if
    if $\widetilde{smile} == 0$ then
        $\widetilde{smile} = B_{0.75}$
    end if
end if
$x \leftarrow (\tilde{\eta},\widetilde{male},\widetilde{bodyshot},\widetilde{smile})$

if $H \in \{RestrictCompetition\}$ then
    $\mathcal{M} \leftarrow h(x;5)$
else
    $\mathcal{M} \leftarrow h(x;11)$    ▷ Draw borrowers from the pool following the probability function $h$
end if
$\mathcal{M} \leftarrow (\mathcal{M},\omega)$    ▷ add outside option
return $\mathcal{M}$
```

Algorithm 1 proceeds in two steps, first, simulates the pool of borrowers and, second, samples from the pool to construct the market. Policies impact the distribution of the features in the pool (*partial compliance*), the size of the market (*Restrict competition*), and the probability of being sampled into the market (through the function h).

Once a market is simulated we determined lenders' choices with Algorithm 2. We first simulate the preferences of a lender, then compute the utility associates from different borrowers, and, finally, determined which borrower is selected.

Algorithm 2 Simulation of a lender choice

```
$(\tilde{\alpha},\tilde{\beta},\tilde{\gamma}) \leftarrow (N(\alpha,sd_\alpha),N(\beta,sd_\beta),N(\gamma,sd_\gamma))$    ▷ draw preference parameters
$\tilde{\epsilon} \leftarrow GEV$    ▷ draw random utility parameters for each borrowing campaign
$u \leftarrow U(\mathcal{M};\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\epsilon})$    ▷ compute utilities from choosing any of the borrowers
$choice \leftarrow max(u)$
return $choice$
```

O Fake and genuine smiles distinction

In Appendix O, we develop an algorithm to distinguish fake from genuine smiles and apply it to the Kiva observational data. We show that only genuine smiles increase funding outcomes, implying that a policy promoting smiles may be less effective if new smiles are perceived as non-genuine, underscoring the importance of providing clear instructions and designing systems that help borrowers create effective profiles.

The effectiveness of policies that encourage borrowers to change facial expressions, specifically to smile, relies on the premise that the previously non-compliant borrowers can create images with desired features and that these newly added features impact lenders' choices; for example, the platform policy might be ineffective if lenders perceive the facial expressions in new images as not genuine. This section argues that this concern is legitimate by showing that non-genuine smiles do not increase funding rates.

To introduce a distinction between genuine and fake (forced) smiles, we train an algorithm that classifies the type of smile. We develop this algorithm using a dataset of 6442 images classified by human annotators as fake or genuine smiles. [42]

We use the algorithm in a random sample of 45 thousand profiles from the Kiva observational dataset. First, we predict whether the person in the image smiles and whether the smile is genuine or fake. Next, we group the borrowers by the predicted type of smile and compute the average cash collected per day. Finally, we estimate the average impact of each type of smile on cash collected per day; to do that we use the AIPW estimator (we follow the same methodology as in Section 4). Table 14 shows the results.

Results presented in Table 14 indicate that only smiles that our algorithm predicted to be genuine lead to higher outcomes. Specifically, we estimate that a genuine smile increases the cash collected per day by $15, while a fake smile has no statistically significant impact.

This analysis showcases an important limitation of the policy based on facial expressions. If lenders perceive some of the smiles created in response to the new policy as not genuine, they might not increase funding rates. In the simulation exercise, we assumed that 75% of the previously non-

[42] The dataset and the original model structure are referred here: https://github.com/vviveks/FakeSmileDetection; we modified the original algorithm to the task of binary prediction - genuine or fake. Our algorithm predicts the fakeness of smiles using three different detected components of each face: whole face, eyes, and mouth. We train three deep neural networks (ResNet, DenseNet, and AlexNet) jointly and concatenate the learned latent vectors to make a joint prediction of whether the smiling is fake in the last layer. The cross-entropy loss of the prediction in the test set of 0.67 (0.66 in the train set), and the precision (f1-score) of 0.70 in the test set (also 0.71 in the train set).

Table 14: The impact of different types of smile on cash per day.

Estimand	Not smiling	Any smile	Genuine smile	Fake smile
Mean outcome in group	131.8 (0.7)	115.0 (0.7)	136.5 (0.8)	116.4 (0.7)
Average treatment effect	-	7.3 (1.0)	15.3 (1.2)	0.8 (1.2)

Note: The first row shows the mean cash per day across four groups of borrowers: not smiling, having any type of smile, having a genuine smile, and a fake smile. The second row shows the average effect of having a smile on cash collected per day. We estimate the effect using the AIPW estimator, which adjusts for all other observable characteristics. The comparison group includes borrowers that do not have images with a smile. Standard errors are in parentheses.

compliant borrowers become compliant under the new policy. The policy becomes less effective when the share of borrowers that create images with genuine smiles decreases.

To mitigate the risk that a new policy is not effective, a platform might design a system that gives borrowers instant feedback on their images, helping them create profile photos that are impactful. An algorithm similar to the one developed in this section can be a part of such a policy.

P Diagnostics of the Recruited Experiments

Table 15 and Table 16 show tests for covariate differences in the two experiments. We compare characteristics of subjects across all the treatments. We report standard errors not-adjusted for multiple hypotheses testing. We find that subjects' characteristics are balanced across treatments in the first experiment. In the second experiment, there is one statistically significant difference – there is a difference in the probability of being full time employed between subjects who saw borrowing campaigns with glasses and without glasses. Note, that each subject saw all potential combinations of features, however, some subjects saw specific features multiple times.

Table 15: Covariate Balance Across Treatments

Covariate	**Smile (1 vs 0)**	**Bodyshot (1 vs 0)**	**Male (1 vs 0)**
Full-Time Employed	-0.001 (0.014)	-0.022 (0.014)	0.002 (0.013)
High Charity	0.002 (0.007)	-0.004 (0.007)	-0.003 (0.006)
High Status	0.010 (0.014)	0.020 (0.014)	-0.001 (0.013)
Male Subject	0.007 (0.015)	0.005 (0.015)	-0.002 (0.014)
Student	-0.005 (0.013)	0.012 (0.013)	0.002 (0.012)

Notes: Each cell reports the difference in means of the covariates for the specified treatment comparison. Standard errors are shown in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

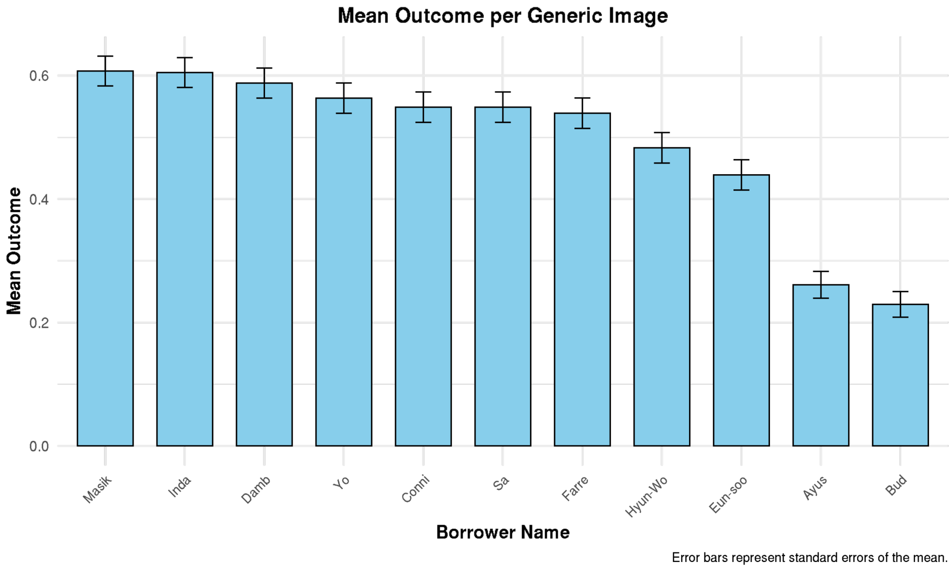
Figures 36 and 37 show average outcomes per profile. We find that there is substantial variation in the overall attractiveness of borrowing campaigns; however, all profiles were selected by some

Table 16: Covariate Balance Across Groups

Covariate	**Old (1 vs 0)**	**SE (Old)**	**p-value (Old)**	**Dark Hair (1 vs 0)**	**SE (Dark Hair)**	**p-value (Dark Hair)**
Full-Time Employed	0.005	0.012	0.678	-0.002	0.013	0.868
High Charity	0.000	0.006	0.957	-0.005	0.006	0.378
Male Subject	0.012	0.013	0.365	-0.023	0.013	0.064
Student	-0.001	0.009	0.928	-0.006	0.008	0.502
Covariate	**Glasses (1 vs 0)**	**SE (Glasses)**	**p-value (Glasses)**	**Sunglasses (1 vs 0)**	**SE (Sunglasses)**	**p-value (Sunglasses)**
Full-Time Employed	-0.031	0.013	0.020	0.015	0.013	0.261
High Charity	-0.005	0.006	0.453	0.003	0.006	0.623
Male Subject	-0.010	0.013	0.446	-0.000	0.013	0.976
Student	-0.001	0.009	0.868	0.002	0.009	0.819

Notes: The table reports the differences in covariates across groups. Columns show the difference in means, standard errors (SE), and p-values for each covariate and treatment comparison. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

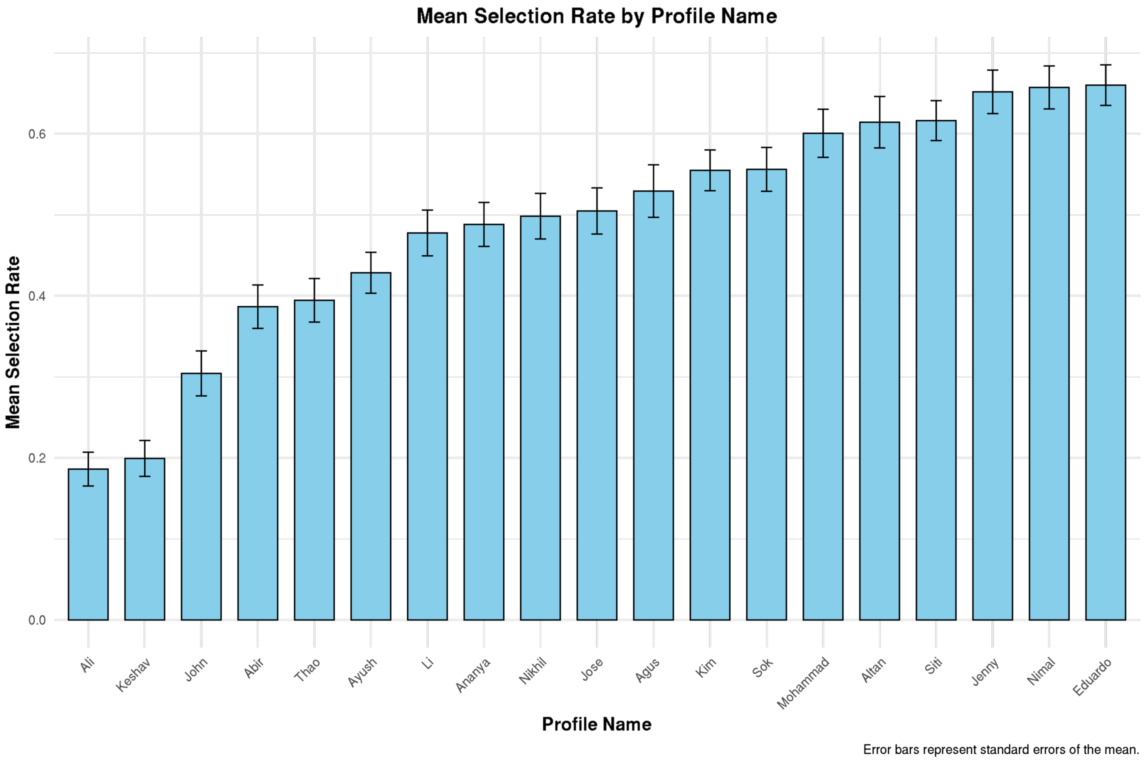
Figure 36: Mean Outcomes Experiment 1.



Note: Average outcome per profile.

subjects.

Figure 37: Mean Outcomes Experiment 2.



Note: Average outcome per profile.