

Can customer ratings be discrimination amplifiers? Evidence from a gig economy platform

Fei Teng

National University of Singapore, fei.teng@nus.edu.sg

Tristan L. Botelho

Yale School of Management, tristan.botelho@yale.edu

K. Sudhir

Yale School of Management, k.sudhir@yale.edu*

This paper investigates whether rating systems generate discriminatory spillovers and act as “discrimination amplifiers.” When platforms display aggregated customer ratings, these “quality metrics” also serve as anchors for future evaluations. Because they embed ratings from biased customers, displayed averages memorialize past discrimination, transmitting it to otherwise unbiased customers and amplifying bias among discriminatory ones. We formalize this mechanism using a stylized analytical model and test it with data from an online labor platform. Allowing for unobserved heterogeneity, we identify three customer segments: a neutral segment, and two biased segments, one that cancels more minority-accepted jobs and another that both cancels more and rates minorities lower. All segments are impacted strongly by displayed ratings. Customer discrimination generates a rating gap of 33.9% and an earnings gap of 6.5% between minority and majority workers. Notably, the unbiased segment produces discriminatory outcomes through spillovers. Adjusting displayed ratings reduces the gaps by mitigating the spillovers to the unbiased segment and amplification of the biased segment, but aggregate rating and earnings disparities persist because biased segments constitute a large share of the market.

Key words: statistical discrimination, taste-based discrimination, prejudice, minorities, earnings gap, customer rating systems, online platforms, gig economy

* The authors are grateful for insightful discussions and comments from Khai Chiong, Soheil Ghili, Vineet Kumar, Jiwoong Shin, Kosuke Uetake, Dennis Zhang, Zhenling Jiang and participants in the Yale Quantitative Marketing Brown Bag Seminar, the 2023 AI/ML Conference at Temple, and the 2024 JHU Social Impact Conference.

1. Introduction

Digital commerce platforms (e.g., Amazon, eBay) and mediated marketplaces across industries, including travel (e.g., Tripadvisor), dining (e.g., Yelp), healthcare (e.g., Healthgrades), education (e.g., Coursera), and professional services (e.g., Avvo), routinely aggregate customer evaluations into visible “quality” metrics. Whether expressed as 1–5 star ratings or thumbs-up/down votes, these summaries shape consumer trust and reduce information frictions, helping platforms facilitate transactions and discipline supplier quality (Chevalier and Mayzlin 2006, Ananthakrishnan et al. 2023). Customer ratings also play an increasingly central role in evaluating and partially automating the management of disaggregated workforces on gig platforms, such as TaskRabbit, Uber, and Upwork (Rosenblat et al. 2016). Viewed this way, rating systems appear to offer a “win–win”: greater customer confidence, higher transaction volume, and improved earnings for workers and platforms.

This optimistic view implicitly relies on a statistical discrimination framework, in which customers substitute group-level beliefs for missing individual-level information (Arrow 1973, Phelps 1972). When post-transaction ratings are informative and publicly observable, they can shift decisions away from group priors toward realized performance, thereby reducing reliance on stereotypes. Consistent with this logic, prior work shows that experience-based ratings mitigate discrimination by Airbnb hosts against minority customers by disciplining inaccurate beliefs and focusing them towards individual ratings (Cui et al. 2020). A key aspect of this mechanism (and the Airbnb setting) is that statistical discrimination is costly to the decision maker, as rejecting minority guests entails foregone revenue. Hence, informative ratings make biased priors expensive to sustain.

But rating systems may function very differently when discrimination operates through the evaluation process itself. A large literature documents systematic racial and gender disparities in customer-generated ratings (e.g., Botelho and Gertsberg 2022, Hannák et al. 2017), indicating that ratings may reflect biased perceptions or preferences in addition to performance. Further, in settings where discrimination is low-cost, customers may face little immediate economic or convenience

penalty when assigning lower ratings or rejecting a minority worker on a gig economy platform. In such settings, ratings cease to be exogenous measures of quality and instead become an endogenous channel through which bias is encoded and propagated.¹

In this paper, we propose that rating systems can do more than merely transmit discrimination; they can act as discrimination amplifiers. When platforms publicly display aggregated ratings, those metrics do not function only as summaries of past experiences; they also shape future decisions. Building on the fundamental anchoring effect in psychology ([Tversky and Kahneman 1974](#)), a large body of work in marketing, organizational behavior, and information systems shows that prior ratings or evaluations not only serve as quality metrics but also act as anchors for subsequent evaluations (e.g., [Bohnet et al. 2025](#), [Northcraft and Neale 1987](#), [Wang et al. 2022](#)). When early ratings capture preferences shaped by discrimination, even from only a subset of customers, publicly displayed aggregates can amplify discrimination over time through two distinct channels. First, a selection (opportunity) channel: Customers rely on displayed ratings when deciding whether to proceed with a transaction (e.g., accept or cancel a job), magnifying initial differences in perceived quality into persistent differences in realized opportunities and earnings. Second, an evaluation (anchoring) channel: Displayed averages anchor subsequent evaluations, causing even otherwise neutral (i.e., non-discriminatory) customers to adjust insufficiently away from biased historical ratings. Together, these mechanisms both intensify discriminatory outcomes among biased customers and generate spillovers among otherwise unbiased customers. As a result, platforms' reliance on publicly displayed ratings can perpetuate and amplify disparities in both ratings and earnings among workers/suppliers, even when they do not differ in true underlying quality.

To clarify and build intuition for our hypothesis, we develop a simple, stylized analytical model that formalizes how customer rating systems can act as discrimination amplifiers when past ratings

¹Beyond the focus of this paper on discrimination in reviews, prior work has studied other issues in review based rating scores, including review manipulation due to conflicts of interest (e.g., [Mayzlin et al. 2014](#), [Ham et al. 2021](#)), social influence and relational effects independent of actual experiences (e.g., [Aral 2013](#), [Kim et al. 2019](#)), and selection in review writing (e.g., [Chakraborty et al. 2022](#)).

serve both as “quality metrics” and as anchors for future evaluations. The model demonstrates that displaying aggregate customer ratings generates two reinforcing outcomes: (i) it allows the discriminatory preferences of one customer segment to spill over to another segment that would not otherwise discriminate, and (ii) it reinforces and amplifies the discriminatory behavior of customers who already hold biased preferences. We also examine how displaying the number of past jobs affects the disadvantaged group. When job volume is interpreted as an additional quality indicator that influences customer acceptance decisions, showing this metric, along with the past ratings, amplifies the earnings penalty faced by the disadvantaged group by reducing the flow of jobs they receive. Importantly, these inequities do not stem from true performance differences, which are assumed to be identical across groups in the model. Instead, they arise from how displayed ratings and volume metrics shape customers’ job acceptance decisions and their subsequent ratings. The model thus illustrates how systemic disadvantages can emerge endogenously across groups due to biases among a subset of customers, even in the absence of any discriminatory intent from the platform.

Next, we examine the empirical relevance of our framework using data from an online labor market platform that connects customers with home service workers. As discussed, unlike prior research that primarily documents the presence of discriminatory ratings, our objective is to study how the display of ratings can generate spillovers and amplify discriminatory behavior across customer segments and at multiple stages of the customer journey, ultimately leading to persistent inequities in workers’ earnings.

Given our objective, we develop a flexible model that allows for customer discrimination at multiple stages of the consumer journey, specifically cancellation and rating decisions, while accounting for unobserved heterogeneity in customer behavior. We exploit the two-sided panel structure of the data, which features repeated interactions for both customers and workers, to identify this heterogeneity. In our setting, customers’ jobs are accepted by both minority and majority workers, and many customers use the platform repeatedly. At the same time, workers complete a large number

of jobs over time, serve diverse customer segments, and accumulate observable metrics such as ratings and total jobs completed. Finally, to isolate the impact of displayed ratings from underlying worker quality, we leverage an institutional feature of the platform: ratings are not displayed before a worker has received five ratings. This feature enables us to separately identify the effect of rating displays from worker quality.

We use the estimates of the model and counterfactual analyses to address three key research questions. First, is there evidence of customers discriminating against minority workers, and if so, how does this discrimination occur across the different segments (e.g., higher cancellations or lower ratings)? Second, to what extent does customer discrimination affect the ratings and earnings of minority workers compared to their majority counterparts? Third, does the display of customer ratings serve as “discrimination amplifiers,” spilling over to those who do not discriminate? If yes, how can we mitigate the effects of discrimination spillovers and amplification?

Our findings yield several key insights. Our model estimates reveal that there are three distinct customer segments: one neutral segment that does not discriminate, and two segments that do so in different ways. One discriminatory segment primarily cancels jobs accepted by minority workers, while the other segment discriminates at both the cancellation and rating stages. Counterfactual analysis shows that, on a rating scale of 1 to 5 stars, compared to the majority workers, the share of sub-5-star ratings for minority workers increases by 33.9% due to discrimination, and this reduces earnings by 6.5%. Moreover, consistent with our conjecture, the public display of customer ratings induces discrimination spillovers among the neutral customer segment and amplifies discrimination among the two discriminatory segments. We then propose a mitigation strategy, where we adjust the displayed ratings to correct for the impact of customer bias. While this cannot fully mitigate the discriminatory impact of bias, it can shrink the spillover and amplification effects of displaying ratings, and the rating gap decreases by 12.2%, and the earnings gap decreases by 9.1%. Notably, the unbiased segment produces the discriminatory outcomes through spillovers. Adjusting displayed ratings mostly eliminates these gaps for the unbiased segment, but aggregate rating and earnings disparities persist because biased segments are a large share of the market.

Though our paper focuses on the empirical setting of online ratings, the fundamental principle we expose, that the memorialization of discrimination-tainted metrics as markers of “merit” or “quality” can spill over to amplify discrimination, has implications far beyond evaluation processes on digital platforms. Our logic is relevant in other societal and economic areas, including organizational hiring and promotions, education, and criminal justice. For example, [Owens and McLanahan \(2020\)](#) show that teacher biases can lead to greater minority student suspensions and expulsions. When these punitive actions become part of a student’s academic record or the implicit impression about a student, they can have long-term negative impacts, both on the student’s subsequent disciplinary actions and academic performance. Similarly, initial disparities in law enforcement regarding drug use between minorities and majorities can create biased criminal records, influencing future legal outcomes for the same illegal behaviors. In the context of the Indian caste system and college admission, [Subramanian \(2019\)](#) argues that relabeling structural disadvantages as differences in observable “merit” not only legitimizes existing inequalities but also portrays efforts to redress these structural inequities as antithetical to a merit-based society. Our analytical approach offers a framework to formally explore and substantiate such phenomena, especially in settings where the evolution of “merit” and outcomes can be fully observed.

From a modeling perspective, our paper contributes to the burgeoning discourse on fairness and algorithmic discrimination within machine learning. A unique contribution is our advancement of the important role of unobserved heterogeneity among discriminatory agents. Although extant research often focuses on “average bias” or bias conditional on observables, our research highlights how unobserved heterogeneity in bias dynamically influences the “observable metrics” that are subsequently presented to customers. This detail is crucial for machine learning applications and for the study of the conditions under which bias occurs in evaluations more broadly, as it underscores the need to account for such heterogeneity in efforts to mitigate bias. To the extent that “observable metrics” used by a system carry the imprint of past discrimination, they will not only embed but also amplify existing structural inequities and disparities over time.

From a social standpoint, our focus on unobserved heterogeneity challenges the notion of “average discrimination.” Framing discrimination as an average phenomenon often leads to public resistance to research findings. It is common for individuals to consider themselves merit-oriented and non-discriminatory, which is likely accurate. By allowing for unobserved heterogeneity in discrimination, our model presents a more accurate representation of market dynamics that allows for interaction between discrimination and “merit.” Furthermore, our modeling approach offers a valuable lens when customer choices may be influenced not only by discriminatory attitudes but also by evolving control variables, including the lingering effects of lagged discrimination.

The remainder of this paper is structured as follows: Section 2 situates our work within the broader literature on discrimination and online rating systems. Section 3 presents a stylized analytical model that formalizes our core hypothesis, elaborating how rating systems can adversely affect minorities through bias spillovers and amplification. Section 4 describes the empirical setting and the data. Section 5 outlines the empirical model we use to analyze customer cancellations and ratings. Section 6 presents our empirical results. Section 7 delves into our counterfactual results. Finally, Section 8 offers concluding remarks.

2. Related Literature

We situate our paper within the broad literature on discrimination in marketplaces and online rating (reputation) systems. Both areas are expansive; therefore, the references we provide are intended to be representative rather than exhaustive.

Discrimination in Marketplaces. We organize the literature on marketplace discrimination into three streams (see Table 1): (i) discrimination against consumers by firms or workers; (ii) discrimination against suppliers by consumers; and (iii) discrimination against workers by consumers or firms. Work on the first stream documents discriminatory treatment across a wide range of markets, including housing (e.g., [Yinger 1995](#)), automobile (e.g., [Ayres and Siegelman 1995](#)), credit markets (e.g., [Blanchflower et al. 2003](#)), and lodging platforms (e.g., [Edelman et al. 2017](#)). A related stream covers discrimination against suppliers in online product and sharing-economy markets. On

eBay, products sold by Black (Ayres et al. 2015) and female (Kricheli-Katz and Regev 2016) sellers receive lower prices than similar products sold by White and male sellers. Similarly, Black and female hosts on Airbnb earn less than White and male hosts for comparable properties (Marchenko 2019).

Table 1 Illustrative Sample of Literature on Discrimination in Marketplaces

	Example Papers	Measured Outcome of Bias
<i>Discrimination against consumers by market type</i>		
Housing	Yinger (1995)	Homes shown for buying and rental
Consumer Products	Ayres and Siegelman (1995)	Bargained car prices
Credit	Blanchflower et al. (2003)	Credit approval for small business
Lodging	Edelman et al. (2017)	Guest acceptance on Airbnb
<i>Discrimination against suppliers by setting</i>		
eBay	Ayres et al. (2015)	Prices for products by race
	Kricheli-Katz and Regev (2016)	Prices for products by gender
Airbnb	Marchenko (2019)	Prices for hosts by gender and race
<i>Discrimination against workers by source of bias</i>		
Employer	Bertrand and Mullainathan (2004)	Hiring of workers
Co-worker	Bodvarsson and Partridge (2001)	NBA team composition
Consumer (Our focus)	Lynn and Sturman (2011)	Ratings for servers
	Brewster and Lynn (2014)	Tips for servers
	Bar and Zussman (2017)	Hiring in labor-intensive services
	Botelho and DeCelles (2025)	Job cancellations on a labor platform

Our primary focus is the third stream: discrimination against workers (service providers) by consumers. Foundationally, Becker (1957) distinguishes discrimination rooted in employer, co-worker, and customer preferences. He notes that while competitive pressures may reduce discrimination driven by the first two sources, customer-driven discrimination may be especially persistent because firms can be rewarded for catering to biased demand. Discrimination against workers or service providers by customers is an issue of particular importance for marketers across a wide range of service industries. Although the literature in this area is relatively limited, existing studies provide evidence that customer preferences can contribute to unequal treatment of workers, consistent with Becker (1957) framework. Existing empirical evidence documents unequal treatment of service

workers through customer evaluations and purchasing decisions, including ratings of restaurant servers (Lynn and Sturman 2011), tipping behavior (Brewster and Lynn 2014), and hiring or contracting outcomes in service markets (Bar and Zussman 2017, Botelho and DeCelles 2025). In platform-mediated service settings, this form of discrimination is particularly consequential because consumers often influence both (i) whether an exchange occurs and (ii) how the provider is subsequently evaluated.

Discrimination against workers has far-reaching economic implications, leading to significant lifetime earnings gaps and stunted wealth accumulation for disfavored groups. Although differences in social and economic backgrounds contribute to wealth disparities (Altonji and Doraszelski 2005), they cannot fully explain the observed gaps. Even after controlling for a wide range of demographic, marital, and socioeconomic factors, more than 70% of the racial gap in wealth remains unexplained (Oliver and Shapiro 2013, Shapiro et al. 2004). Moreover, even after controlling for relevant human capital characteristics, minority workers are less likely to be hired, experience longer job searches, accumulate less work experience and tenure, and earn lower wages compared to majority workers (Pager and Shepherd 2008, Tomaskovic-Devey et al. 2005).

Methodologically, most evidence on labor-market discrimination relies on reduced-form comparisons that quantify gaps in hiring, wages, or job-search outcomes (e.g., Pager and Shepherd 2008, Tomaskovic-Devey et al. 2005). Our approach is closer to the case-study tradition in industrial organization: we develop and estimate a process model of customer decisions within a specific marketplace and use counterfactuals to quantify how platform policies shape disparities. This approach is especially relevant for firms and platforms because it directly maps design choices, such as whether and how to display ratings, into predicted effects on opportunities and earnings within their operating environment.

Online Rating Systems. Online rating systems play a foundational role in modern digital marketplaces. For consumers, ratings reduce information frictions and enhance consumer welfare (Chevalier and Mayzlin 2006, Fang 2022, Luca 2016, Reimers and Waldfogel 2021). For suppliers, ratings

shape demand, search visibility, and platform-imposed rewards or penalties, thereby creating strong incentives to invest in service quality (Ananthakrishnan et al. 2023, Donati 2025, Tadelis 2016). Furthermore, Cui et al. (2020) shows that online rating systems can help mitigate discrimination against marginalized groups. The argument is that by providing reviews as observable quality metrics, rating systems can reduce statistical discrimination.

A parallel stream of research has raised concerns about the reliability and fairness of consumer-generated ratings. Although platforms often treat ratings as objective evaluations of underlying performance, empirical evidence suggests that such evaluations frequently incorporate biases unrelated to actual quality. Prior work documents discrimination against minority and female workers on gig-economy platforms (Botelho et al. 2025, Greenwood et al. 2022, Hannák et al. 2017), as well as systematic biases in ratings against female-led movies (Aguiar 2024). Together, these findings underscore the limitations of interpreting customer ratings as unbiased measures of quality. We contribute to this literature by showing that when ratings embed such biases, rating systems can do more than merely serve as “vehicles for discrimination.” By aggregating and displaying biased evaluations, platforms can amplify existing disparities and generate discrimination spillovers that influence even neutral customers who would otherwise not discriminate.

The mechanism we identify builds on the anchoring effect, a cognitive heuristic first documented by Tversky and Kahneman (1974) and extensively studied and validated in psychology and behavioral economics (Furnham and Boo 2011). Anchoring shapes individuals’ perceptions and influences how they make evaluation decisions, which generate dependence on prior evaluations across different contexts. Northcraft and Neale (1987) show that anchor values affect property evaluations. Price anchors influence perceived value and willingness to pay (Adaval and Wyer Jr 2011, Whitley et al. 2025), and anchoring affects performance evaluations as well (Thorsteinson et al. 2008). In the context of online marketplaces, prior ratings can serve as anchors that affect subsequent evaluations. Empirical studies find that online reviews exhibit dependence on earlier reviews (Ma et al. 2013, Park et al. 2021, Schlosser 2005, Sunder et al. 2019). Wang et al. (2022) finds that prior average ratings positively affect subsequent ratings and attributes this dependence to anchoring.

Our paper bridges these two streams of research: bias in consumer-generated ratings and anchoring in sequential evaluations. We demonstrate that anchoring amplifies the downstream impact of discriminatory ratings, creating spillovers that shape the behavior of otherwise neutral customers. Using an empirical model, we estimate the extent to which past ratings influence subsequent evaluations and quantify the degree to which rating systems can magnify discrimination and contribute to earnings disparities. In doing so, we provide a behavioral foundation and a novel framework for understanding how rating systems can unintentionally perpetuate inequality even when the platform design itself is race-neutral.

3. Stylized Analytical Model

We formalize the intuition of how displaying worker metrics can endogenously amplify discrimination and lead to discrimination spillovers even among unbiased, neutral customers using a stylized analytical model.

3.1. Environment

Workers. Consider a labor market platform with two workers who differ only in their observable labels: advantaged (A) and disadvantaged (D). The D worker is disadvantaged in the sense that a subset of customers gives lower ratings to the D worker for the same underlying quality of work relative to the A worker. To rule out quality differences from driving differences in outcomes, we assume both workers provide service with the same quality level of 1.

Customers. Customers are of two types $c \in \{N, P\}$. Given equal service quality, neutral customers (N) rate both A and D workers identically, while partial customers (P) down-weight D 's rating by a certain factor, as we elaborate in the rating formulation process below. The share of partial customers is α , where $0 < \alpha < 1$.

Timing and Jobs. Time is discrete and indexed by $t = 1, 2, \dots$. In each period, each worker is matched with a unit mass of potential jobs, with customer types drawn in proportion to their population shares. Before each job begins, customers may accept the matched worker or cancel the job. After each job is completed, we assume that the customer always submits a rating for the

worker. Accordingly, for each worker, total rating volume equals total job volume. As we show later, rating volume and job volume are highly correlated in our empirical setting, and the likelihood of leaving a rating does not depend on workers' race. Therefore, in this stylized model, we abstract from rating submission and focus on rating outcomes.

Worker Metrics. The platform may choose to display worker metrics to customers, including cumulative job volume and cumulative rating. These metrics summarize a worker's past performance and influence customers' job acceptance and rating decisions.

In each period t , let n_t^{wc} denote the volume of jobs completed by worker w for customers of type c . The total job volume for worker w in period t is $n_t^w = \sum_{c \in \{N, P\}} n_t^{wc}$, and the cumulative job volume at the end of period t is $N_t^w = \sum_{t'=1}^t n_{t'}^w$.

Let r_t^{wc} denote the rating given by a type- c customer to worker w in period t . As we describe in the rating formation process below, r_t^{wc} always lies in $(0, 1]$. The period- t aggregate rating for worker w is defined as the geometric mean² of the ratings across all jobs and customer types:

$$r_t^w = \left(r_t^{wN}\right)^{\frac{n_t^{wN}}{n_t^w}} \times \left(r_t^{wP}\right)^{\frac{n_t^{wP}}{n_t^w}}. \quad (1)$$

The cumulative displayed rating for worker w at the end of period t is:

$$R_t^w = \left(\prod_{t'=1}^t (r_{t'}^w)^{n_{t'}^w}\right)^{\frac{1}{\sum_{t'=1}^t n_{t'}^w}}. \quad (2)$$

3.2. Job Acceptance and Earnings

In each period t , after being matched to a worker, if the platform displays worker metrics, customers observe rating valence (R_{t-1}^w) and rating volume (N_{t-1}^w) and use these metrics as signals of worker quality to decide whether to accept the worker or cancel the job. A higher rating and a larger number of completed jobs indicate that the worker has consistently demonstrated high performance and reliability, which increases the likelihood of acceptance. We model this acceptance probability as:

$$P(\text{accept job at } t) = (R_{t-1}^w)^{1/N_{t-1}^w}. \quad (3)$$

² Results hold numerically under arithmetic aggregation; the geometric form aids tractability.

This specification ensures that the acceptance probability increases in both rating valence and rating volume. The volume of completed jobs in period t (i.e., earnings) is therefore $n_t^w = (R_{t-1}^w)^{1/N_{t-1}^w}$. When worker metrics are hidden, or in the initial period, we normalize $R_0^w = 1$ and $N_0^w = 1$ so that the acceptance probability equals 1.

3.3. Anchoring and Rating Formation

We now describe how ratings are generated conditional on job completion. Past research has shown that early reviews can influence subsequent reviews by customers (Muchnik et al. 2013, Park et al. 2021, Wang et al. 2022, Botelho 2024). One theoretical justification for this is based on the classic “anchoring and adjustment” theory of Tversky and Kahneman (1974). In digital platform contexts, the publicly displayed average rating serves as the anchor because it is typically the first and most salient number a user encounters, thereby framing the evaluation task.

We incorporate this idea in the model by allowing the ratings a worker receives from a new job to be correlated with their displayed rating. We provide model-free evidence later in our empirical analysis to support this key assumption. Specifically, when past ratings are displayed, we assume that the rating r_t^{wc} is the product of the worker’s true quality and the displayed rating. In addition, partial customers discount the rating of the disadvantaged worker by a factor of $1 - \delta$, where $\delta \in (0, 1)$ and a larger δ reflects greater partiality. These assumptions are summarized in Table 2.³ When ratings are *not* displayed, we normalize $R_0^w = 1$, which implies that $\forall t$, $r_t^{AN} = r_t^{AP} = 1$, $r_t^{DN} = 1$, and $r_t^{DP} = 1 - \delta$.

Table 2 r_t^{wc} when past rating R_{t-1}^w is displayed

	worker $w = A$	worker $w = D$
$c = N$	R_{t-1}^A	R_{t-1}^D
$c = P$	R_{t-1}^A	$(1 - \delta)R_{t-1}^D$

³ We assume that customers are not aware of the rating generating function. In our empirical setting, each customer offers only a small number of jobs on the platform and does not observe the ratings of other workers except for those with whom they interact. Knowledge of the rating generating function would matter if customers were repeatedly exposed to a mix of advantaged and disadvantaged workers and could detect mismatches between their own experiences and the displayed ratings across workers (e.g. Bohren et al. 2019).

3.4. Discrimination Spillover and Amplification

Since the displayed rating serves as an anchor and affects the current rating, it generates a discrimination spillover onto neutral customers and induces partial customers to behave even more discriminatorily. Over time, the displayed rating amplifies the disadvantages faced by the D worker, leading to larger rating differences between the two workers. Early differences in displayed ratings also reduce the job acceptance rate for the D worker; with fewer accepted jobs, the D worker is less likely to accumulate future acceptances. Consequently, displaying worker metrics widen the earnings gap, as displaying either rating or job volume amplifies the consequences of biased evaluations. This analysis leads to the following proposition.

Proposition 1. *Let superscripts o and h denote outcomes when ratings are observable (displayed) and hidden (not displayed). Then:*

- (i) **Spillover.** *When ratings are displayed, neutral customers rate disadvantaged workers lower than when ratings are hidden:*

$$\forall t \geq 2, \quad r_t^{DN,o} < r_t^{DN,h}.$$

- (ii) **Amplification.** *Displaying ratings causes disadvantaged workers' ratings to decline over time for both customer types, widening the rating gap:*

$$\forall t \geq 2, \quad r_t^{DN,o} < r_{t-1}^{DN,o}, \quad r_t^{DP,o} < r_{t-1}^{DP,o}, \quad R_t^{A,o} - R_t^{D,o} > R_{t-1}^{A,o} - R_{t-1}^{D,o}.$$

- (iii) **Earnings Gap (Valence–Volume Interaction).** *Because job acceptance depends on both rating valence and accumulated volume, the disadvantaged worker faces a dual penalty: lower ratings reduce future opportunities, and fewer jobs further depress job flow. Consequently, earnings diverge, and the gap expands over time:*

$$\forall t \geq 3, \quad N_t^{A,o} - N_t^{D,o} > N_{t-1}^{A,o} - N_{t-1}^{D,o} > 0.$$

This stylized model formalizes the intuition developed in the introduction within a minimal analytical structure.⁴ Even when only a fraction of customers are partial, anchoring through displayed

⁴The model abstracts from workers' supply-side responses, as our objective of the model is to isolate the rating-based behavioral mechanism rather than model full equilibrium dynamics.

ratings propagates bias to neutral customers and, through job acceptance behavior, compounds it into structural earnings inequity. Volume further reinforces this dynamic: reduced demand for the disadvantaged worker due to lower ratings depresses their acceptance rates across both the partial and neutral segments and diminishes their future job flow. Having shown analytically that discrimination spillovers and amplification arise endogenously from the feedback between ratings and job volume, we next quantify the magnitude of these effects using data from a real-world online labor platform.

4. Empirical Setting and Descriptive Analysis

We first describe the empirical setting of the labor market platform. We then provide model-free evidence in support of the analytical model and the subsequent empirical models.

4.1. The Online Labor Market Platform

Platform Overview. Our empirical setting is a North American online labor market platform. We use a pseudonym, ServicesConnect (“SC” for short), to refer to the platform as the platform wishes to remain anonymous. The platform connects customers and workers for a range of home-service jobs and manages the entire transaction process.

Workers on SC are mostly small-business owners who specialize in certain types of home-service jobs. Before taking any jobs on SC, workers must go through SC’s screening process, which includes skill verification, criminal background check, and interview. Customers are reminded of this screening in various ways. Therefore, the screening process helps alleviate concerns about significant discrepancies in worker quality, for both customers and researchers. Jobs on SC typically require only a few hours. Examples of services include appliance repair, electrical work, and plumbing.

Job Journey on SC. A job starts when a customer submits a service request by selecting from a predetermined list of service categories. A minimum cost and an hourly rate are then generated.⁵

⁵The platform charges a commission as a share of the transaction price rather than a separate usage fee. Although some transactions may occur off-platform to avoid this commission, this concern is limited compared to settings with frequent repeat interactions, such as Uber. On our platform, services are infrequent and diverse, making unreported transactions uncommon.

Labor costs are uniform within each service category and are non-negotiable. Material costs (e.g., a new circuit breaker) are charged separately and monitored by the platform. The customer then selects from available time slots for the job to be completed.

After a job is submitted, SC provides workers with only basic information for them to consider whether to accept the job, including a brief task description, an approximate location, and the customer's preferred time. For example, for an appliance repair job, workers know only the type of repair (e.g., some issue with a refrigerator), the scheduled time, and a certain metropolitan area. No exact address, customer name, or demographic information about the customer is disclosed. SC also does not collect demographic information on workers or customers.

SC offers jobs to workers through a two-stage process. In the first stage, SC employs an algorithm that selects a small set of workers using two worker metrics: (1) the number of jobs the worker has completed and (2) the worker's average rating to date. The algorithm does not use any worker demographic information or customer characteristics as inputs. The selected workers can then accept the job during an exclusive 15-minute window on a first-come, first-served basis. If no worker accepts the job within the window, in the second stage, the job becomes available to all eligible workers (based on job expertise and location), again on a first-come, first-served basis.⁶

The design of SC's job acceptance process ensures that the matching between workers and customers is effectively random with respect to worker race. Customers have no control over which workers accept their jobs, and workers do not observe any customer information. Moreover, jobs are accepted on a first-come, first-served basis, leaving workers only a brief window to decide whether to accept a request. Together, limited information and limited decision time make strategic behavior to counteract potential discrimination unlikely and also prevent workers from discriminating against

⁶ The rationale for employing this algorithm, rather than allowing customers to directly select workers from a list, is related to SC's concern that granting customers full discretion could result in highly rated, experienced workers consistently receiving jobs, making it difficult for newer workers to obtain opportunities. Such an imbalance could discourage new workers from joining the platform and ultimately hinder the platform's growth.

customers.⁷ Consistent with this, SC reports that workers base their acceptance decisions primarily on availability at the requested time. Taken together, these features support our assumption that worker–customer matching is effectively random with respect to worker race.

Once a worker accepts a job and agrees to the customer’s requested time, SC sends the customer a notification that displays the worker’s name, photo, average rating to date, and total number of completed jobs. The average rating is masked, however, if the worker has received fewer than 5 ratings. It becomes visible once the worker has received 5 ratings. After receiving the notification, customers may cancel the job at any time before the worker indicates that they are en route. After the job is completed, SC sends another message prompting customers to rate their experience with the worker on a scale from 1 to 5 stars. Fig. A.1 in Appendix A provides examples of the two messages sent to customers.

Data Overview. We obtain data on all jobs completed on the platform in one metropolitan area in North America over a four-year period, from its launch in 2016 through 2019. We exclude the 11 female workers in the dataset because the platform workers are primarily male. Additionally, since the platform does not collect demographic information on workers, we infer customers’ perceived worker race from worker photos. Therefore, we exclude jobs in the very rare case that a worker photo is unavailable. The final dataset consists of 86157 jobs, involving 34110 customers and 633 workers.

SC classifies jobs into several service categories, such as maintenance, plumbing, appliance, and electrical. Most categories can be distinguished by whether the jobs require worker credentials. For example, plumbing and HVAC jobs require credentials, whereas snow removal does not. Appliance and electrical work are exceptions, as some jobs within these categories require credentials while others do not. As such, we split jobs of these two categories further into credential-required and non-credential-required categories. Maintenance is the most frequently requested service category,

⁷ In SC’s surveys of its workers, concerns regarding discrimination by customers have not been an issue. Moreover, workers do not observe ratings received by other workers, making it difficult for them to detect any potential discriminatory behavior. We therefore believe that workers are unaware of discrimination and cannot counteract it.

followed by plumbing and appliance services. Cancellation rates and average ratings differ across categories. More urgent service types, such as locksmith services and appliance repair, exhibit the lowest cancellation rates, whereas upholstery services show the highest. Average ratings also vary meaningfully across categories, ranging from 4.57 to 4.87. These differences motivate the inclusion of service category fixed effects in our empirical model.

We also observe changes in job volume, cancellations, and ratings over time and across locations. The number of jobs on the platform increased steadily from 2016 to 2019. As the platform expanded, cancellation rates also rose, perhaps because early adopters tended to be more risk-tolerant. Meanwhile, average ratings increased over the years, possibly because more capable workers remained on the platform. In addition, customers across locations exhibit variation in both cancellation rates and average ratings. These patterns motivate the inclusion of year and customer location fixed effects in our model.

4.2. Job Outcome and Worker Race

Because SC does not collect demographic information on workers, we infer perceived worker race based on customers' likely perceived race/ethnicity of the worker. Specifically, two independent coders reviewed workers' profile photographs and classified the perceived race of each worker using a predefined set of categories. Details of the coding protocol and validation procedures are provided in [Botelho and DeCelles \(2025\)](#). Given that the vast majority of workers are classified as White, and that our theoretical framework focuses on whether workers are perceived as belonging to a marginalized group rather than on differences across specific minority categories, we define a worker as a minority if the coders did not perceive the worker as White, and as a majority otherwise. This classification reflects perceived race, the relevant construct for customer decision-making in our setting, rather than workers' self-identified demographic characteristics.

Table 3 reports average job outcomes by worker race. The summary statistics suggest that minority workers are disadvantaged in both cancellations and ratings. Overall, minority workers experience higher cancellation rates. Once a job is completed, there is no significant difference in

Table 3 Job Outcome and Perceived Worker Race

	All Jobs	Majority	Minority	Minority Gap	t-test
% Workers		66.35%	33.65%	32.70%	
Cancellation Rate	18.93%	17.60%	21.70%	-4.10%	-14.01
Rating Submission Rate	69.06%	69.11%	68.97%	0.14%	0.36
Ratings:					
Avg. Rating	4.76	4.78	4.71	0.07	9.10
Perc. Rating=5	85.20%	86.26%	82.88%	3.38%	9.38
% for Ratings below 5:					
Rating=4	9.58%	9.09%	10.65%	-1.56%	-5.26
Rating below 4	5.22%	4.65%	6.46%	-1.82%	-7.86

Note: The last column of the table presents the t-statistics from the two-sample t-test of the corresponding outcome variables between the minority and majority workers.

the likelihood of submitting a rating between minority and majority workers. However, conditional on rating submission, minority workers receive lower ratings on average.

We also observe in Table 3 that the distribution of the ratings is highly skewed: Most customers give 5-star ratings, while ratings below 4-star are rare. Therefore, in the analyses below, we focus on a binary rating outcome that indicates whether a job receives a 5-star rating, rather than modeling ratings on the full five-point scale. This approach aligns with the customer satisfaction literature, where “top-box” metrics are widely used and have strong predictive power for outcomes such as customer retention and sales growth (Baehre et al. 2022, De Haan et al. 2015, Botelho et al. 2025). In our setting, 5-star ratings account for 85% of all ratings and 4-star ratings account for another 10%, implying that nearly 95% of ratings fall within the top two categories.

4.3. Job Outcomes and the Displayed Rating

Table 3 seems to be suggestive of discrimination, where minority workers are faced with higher cancellation rates and receive lower ratings. Because the platform displays workers’ average ratings to customers, we aim to move beyond this estimate to explore whether the displayed ratings reduce customer uncertainty and narrow the gap between minority and majority workers. For each job, we compute the real-time average rating to be displayed to the customer, provided that the worker has received at least 5 ratings. Since ratings in our context are highly skewed, and consistent with prior work (e.g., Abdulsalam et al. 2025, Lee et al. 2018), we winsorize the displayed average ratings

and rescale them to a 0–1 scale to reduce the influence of outliers and enhance interpretability. Specifically, let r_o denote the original displayed ratings on a 1-5 scale, and let r_w denote the normalization level, i.e., the winsorization cutoff. The normalized rating, r_n , is given by

$$r_n = \frac{(\mathbf{1}_{\{r_o \geq r_w\}} \times r_o + \mathbf{1}_{\{r_o < r_w\}} \times r_w) - r_w}{5 - r_w} \quad (4)$$

In our empirical analysis, we set $r_w = 4.4$, since 95% of the average ratings are above 4.4.⁸

Table 4 reports how job outcomes correlate with the normalized displayed ratings. First, job cancellation rates are negatively associated with displayed ratings, suggesting that higher displayed ratings may signal better worker quality and, in turn, reduce cancellations. However, within any given range of displayed ratings, minority workers still experience significantly higher cancellation rates than majority workers. This indicates that although higher ratings may reduce uncertainty, customers do not attribute the same credibility to the ratings of minority and majority workers.

Second, job ratings are positively correlated with displayed ratings. This could potentially be driven by two mechanisms. First, higher displayed ratings capture higher underlying worker quality, which leads to better job performance and therefore higher ratings. Alternatively, customers' ratings may be influenced by the ratings previously assigned by others, serving as an anchor that shifts subsequent ratings upward when the displayed average is higher. Regardless of the underlying mechanism, the evidence suggests that minority workers receive lower job ratings on average, even conditional on their displayed ratings.

Table 4 Job Outcomes by Normalized Displayed Ratings

Rating	Cancellation Rate				Perc. 5-Star Rating			
	All	Majority	Minority	t-stat	All	Majority	Minority	t-stat
= 0	20.7%	18.9%	23.7%	-3.16	72.87%	74.69%	69.72%	2.15
(0, 0.5]	20.47%	18.89%	23.02%	-7.89	79.0%	79.56%	78.05%	2.10
[0.5, 1]	17.75%	16.77%	20.16%	-8.92	88.62%	89.43%	86.52%	6.80

Note: The t-stats are for two-sample t-tests of the corresponding outcome variables between the majority and minority workers within the range of the normalized displayed ratings.

⁸ Our model estimates remain robust when using different normalization levels (e.g., the 10th percentile of the ratings at 4.5 and the 1st percentile at 4.2) and are also robust to alternative normalization methods (e.g., mean-centered normalization). The estimation results for these alternative normalizations are reported in Online Appendix ??.

4.4. Expansion in Minority Rating Gap

Section 4.2 documents the overall minority gap in ratings. Here, we explore whether this gap expands over time and discuss potential explanations behind the expansion. To gain insight into how displaying ratings may influence the evolution of the rating gap, we exploit the institutional feature that ratings are not displayed until the worker has received five ratings. We focus on the percentage of sub-5-star ratings received by workers, as most ratings are 5-star, so receiving even one additional sub-5-star rating can substantially affect a worker’s overall rating. Fig. 1a plots the share of sub-5-star ratings for minority and majority workers at different points in a worker’s tenure, specifically after receiving 5, 10, 15, and up to 40 ratings. The figure reveals a notable increase in the rating gap once ratings begin to be displayed. Ratings for majority workers remain relatively stable over time, whereas the percentage of sub-5-star ratings for minority workers increases substantially after their ratings are displayed. Before ratings are shown, the ratings received by minority and majority workers are not significantly different. This pattern suggests that, in addition to the effect of worker quality on current ratings, the display of ratings not only influences customers’ current rating decisions but also amplifies any potential differential impact between minority and majority workers over time.

Fig. 1b presents the pattern for a subset of workers who receive all 5-star ratings in their first five ratings. Although these workers are likely to be of higher quality than those who do not begin with an initial perfect average, a minority–majority rating gap still emerges and widens over time, as minority workers receive lower ratings in later periods. This pattern suggests that even when workers start with identical displayed ratings, subsequent bias can affect their ratings, and this impact becomes amplified over time.

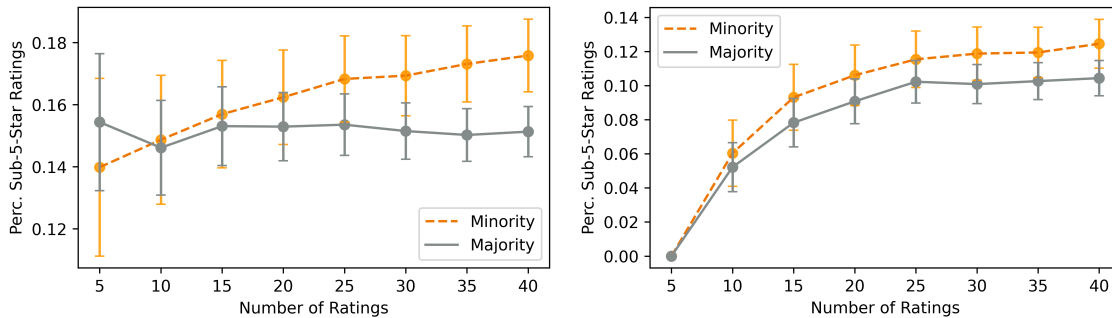
5. Empirical Model and Estimation

In this section, we bring the intuition from the analytical model and the model-free evidence and assess its implications by calibrating an empirical model of customer behavior. Our focus is on customer behaviors after the workers’ job acceptance. Specifically, we model three customer

Figure 1 Rating Gap between Minority and Majority Workers

(a) All Workers

(b) Workers with All 5-star before Display



Note: The left panel shows the rating trend for all workers, and the right panel shows the trend for workers who received all 5-star ratings in their first five ratings. We restrict the sample to workers with at least 40 total ratings; the pattern remains robust for workers with at least 20 ratings.

decisions: whether to cancel the job, whether to submit a rating, and what rating to submit. After describing the empirical model, we outline our estimation approach. Finally, we clarify the data features that identify the key parameters of the model.

5.1. The Empirical Model

The model consists of three parts, corresponding to the three customer decisions we model. The first part models the binary decision of whether to cancel the job after the job is accepted by a worker. The second part captures the binary decision of whether to submit a rating to the platform once the job has been completed. The third part models whether to give a 5-star rating to the job, conditional on the customer deciding to submit a rating. The model allows for unobserved heterogeneity as latent segments across customers. We summarize the notation used in the empirical model in Appendix C.

Let i denote the customer. Each customer belongs to a segment $g \in \{1, \dots, G\}$. We denote a job by j and a worker by k . Suppose customer i requests job j , and the job is accepted by worker k . We first model the decision of job cancellation as a binary logit. The variable C_{ijk} takes the value of 1 if customer i cancels job j conditional on worker k accepting the job. Let X_{ijk}^C denote the observables that affect the cancellation decision. For each customer segment g , we specify the cancellation probability as a binary logit:

$$P_{ijk}^{Cg} = P^g(C_{ijk} = 1) = \frac{\exp(\kappa^g X_{ijk}^C)}{1 + \exp(\kappa^g X_{ijk}^C)},$$

where κ^g denotes the parameters for segment g .

For job j that is not canceled and thus completed, let $W_{ijk} \in \{0, 1\}$ denote whether the customer submits a rating to the platform. We model W_{ijk} using a binary logit specification. For a customer who belongs to segment g , the probability of submitting a rating, conditional on the job not being canceled, is given by

$$P_{ijk}^{Wg} = P^g(W_{ijk} = 1 | C_{ijk} = 0) = \frac{\exp(\lambda^g X_{ijk}^W)}{1 + \exp(\lambda^g X_{ijk}^W)},$$

where X_{ijk}^W denotes the relevant observables and λ^g denotes the parameters for segment g .

Finally, conditional on a rating being submitted, we model the rating choice of the customer. As shown in Table 3, the very small number of one- to three-star ratings makes it infeasible to model the rating decision using an ordinal specification and to estimate the corresponding parameters reliably across the latent customer segments. Moreover, focusing on whether a rating is 5-star is consistent with the customer satisfaction literature, as discussed in Section 4.2. As such, we model the rating choice as a binary decision. Specifically, the variable R_{ijk} takes the value of 1 if the customer gives a 5-star rating and 0 if the customer gives a rating below 5 stars. For a customer who belongs to segment g , the probability of giving a 5-star rating, conditional on submitting a rating, is

$$P_{ijk}^{Rg} = P^g(R_{ijk} = 1 | W_{ijk} = 1) = \frac{\exp(\rho^g X_{ijk}^R)}{1 + \exp(\rho^g X_{ijk}^R)},$$

where X_{ijk}^R denotes the observables and ρ^g denotes the parameters for segment g .

Let q^g denote the share of customers who belong to segment g , where $\sum_g q^g = 1$. The parameters to be estimated are $\Theta = \{\Theta_1, \dots, \Theta_G\}$, with $\Theta_g = \{\kappa^g, \lambda^g, \rho^g, q^g\}$. Let $J(i)$ denote the set of jobs requested by customer i , and let $S_i = \{(C_{ijk}, W_{ijk}, R_{ijk}) : j \in J(i)\}$ denote the set of observed choices made by customer i . The likelihood function of an individual customer i conditional on segment g is given by

$$\begin{aligned} L_i^g(S_i; \Theta_g) = & \underbrace{\left[\prod_{j \in J(i): C_{ijk}=1} (P_{ijk}^{Cg}) \right]}_{\text{for jobs canceled}} \times \underbrace{\left[\prod_{j \in J(i): C_{ijk}=0, W_{ijk}=0} (1 - P_{ijk}^{Cg}) \times (1 - P_{ijk}^{Wg}) \right]}_{\text{for jobs not canceled but not rated}} \\ & \times \underbrace{\left[\prod_{j \in J(i): C_{ijk}=0, W_{ijk}=1} (1 - P_{ijk}^{Cg}) \times P_{ijk}^{Wg} \times (P_{ijk}^{Rg})^{R_{ijk}} \times (1 - P_{ijk}^{Rg})^{1-R_{ijk}} \right]}_{\text{for jobs not canceled and rated}}, \end{aligned}$$

where the first component accounts for jobs that are accepted by a worker and subsequently canceled by customer i , the second component accounts for jobs not canceled and not rated, and the third component accounts for jobs not canceled and rated. By summing over all unobserved segments $g \in \{1, \dots, G\}$, we obtain the overall likelihood for customer i :

$$L_i(S_i; \Theta) = \sum_g q^g L_i^g(S_i; \Theta_g)$$

The log-likelihood over all customers is then given by

$$\sum_i \log(L_i(S_i; \Theta)) \quad (5)$$

5.2. Estimation

We estimate the model using the EM algorithm, which iteratively maximizes the expected log-likelihood in the equation below:

$$\sum_i \sum_g q_i^g \log(L_i^g(S_i; \Theta_g)),$$

where q_i^g denotes the probability that customer i belongs to segment g given parameters values Θ , conditional on all observed jobs of customer i :

$$q_i^g = Pr(g|S_i; \Theta) = \frac{q^g L_i^g(S_i; \Theta_g)}{L_i(S_i; \Theta)} \quad (6)$$

The EM algorithm is implemented as follows:

1. Initialize Θ .⁹
2. For each customer and each segment, calculate the probability of being in the segment conditional on the customer's cancellation and rating decisions given Θ , following Eq. (6).
3. Update the segment probabilities: $q^g = \frac{\sum_i q_i^g}{\sum_i \sum_{g'} q_i^{g'}}$.

⁹ The initial values of the segment probabilities q^g are set equally across segments. For κ^g , λ^g , and ρ^g , we estimate the model without unobserved heterogeneity and then perturb the resulting parameter estimates by one-tenth of their standard errors to obtain the initial values.

4. Using the updated q_i^g from step 2, update the segment-specific parameters $\{\kappa^g, \lambda^g, \rho^g\}$ for each segment g by maximizing the weighted log-likelihood:

$$\sum_i q_i^g \log(L_i^g(S_i; \Theta_g))$$

5. Repeat steps 2-4 until convergence.

5.3. Identification

Our empirical strategy for identifying discrimination across customer segments exploits how customers' cancellation and rating behaviors vary with a worker's membership in a disadvantaged group. Identification is enabled by the panel structure of the data, which contains multiple job requests from the same customers and performance histories for individual workers. In addition, because workers accept customer jobs independently of customer characteristics and perceived worker race, we treat the worker-customer matching as exogenous.

At the cancellation stage, after a worker accepts a job, the customer observes worker attributes, including the number of past jobs, ratings, and perceived race, and then decides whether to cancel the job. Observing multiple job decisions per customer allows us to identify how cancellation behavior responds to worker and job characteristics, and in particular to whether the worker is perceived as a minority. Evidence of discrimination arises if minority status has a negative main effect on cancellation or if minority status interacts negatively with displayed ratings, consistent with either taste-based or statistical discrimination.

At the rating stage, customers observe the worker's realized performance, which is unobserved by the researcher. To proxy for worker quality, we use the average of the worker's first five ratings as a control. This choice exploits a feature of the platform: ratings are not displayed until a worker has received five ratings. As a result, these initial evaluations are not subject to anchoring from previously displayed ratings and therefore provide a cleaner measure of underlying performance. That said, these early ratings may still reflect discriminatory behavior if minority workers receive systematically lower evaluations. To account for this possibility, we include a minority indicator in

the model. Because quality for minority workers may be under-estimated with the first five ratings, this positive quality effect is absorbed by the minority coefficient. Hence, the estimated negative main effect of minority status is biased toward zero and should be interpreted as a conservative estimate of discrimination.

Conditional on this quality control, anchoring effects are identified through the relationship between a customer's rating for a given job and the worker's cumulative average past rating, once ratings become observable. If the sensitivity of current ratings to past ratings differs by minority status, specifically, if it is weaker for minority workers, we interpret this pattern as additional evidence of discriminatory behavior towards minorities.

Given our interest in spillovers across customer segments, we allow for unobserved heterogeneity in customers' cancellation and rating behavior. Identification of this heterogeneity exploits the two-sided panel structure of the data, which contains multiple observations on both customers and workers. Workers perform many jobs over time and are matched with different customers, while customers interact with multiple workers whose ratings and job histories evolve. Because worker acceptance are quasi-random with respect to customer characteristics and perceived worker race, each customer is exposed to both minority and majority workers and to variation in worker attributes. This allows us to distinguish customer segments by how perceived worker race moderates the influence of past ratings on cancellation decisions and current ratings, including segments that exhibit no discrimination. At the same time, all segments may respond to overall past ratings (anchoring effect), independent of race.

6. Empirical Results

We begin with a brief discussion of the sample adjustments made to allow for service category fixed effects and customer unobserved heterogeneity. We then report the model estimates, showing evidence of anchoring effects and discrimination in some latent segments but not in others. Finally, we describe the characteristics of the latent segments to facilitate interpretation.

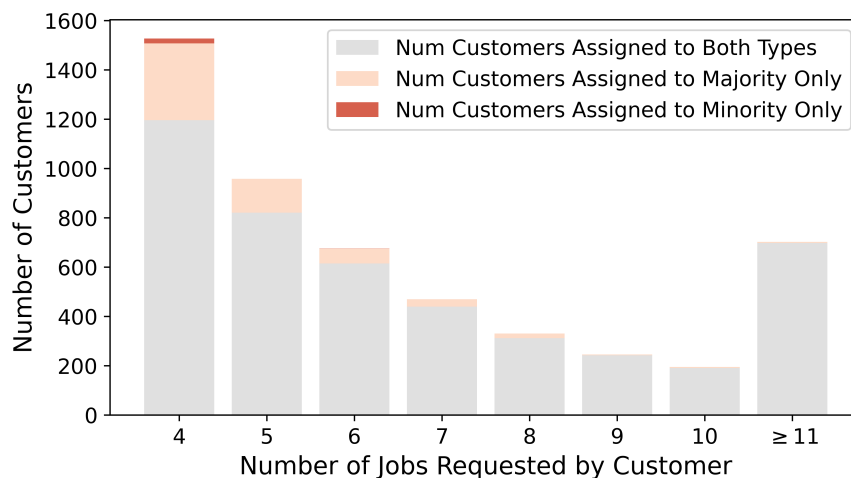
6.1. Sample Adjustments for Estimation and Empirical Model Specification

First, because average cancellation rates and ratings differ across service categories, we control for service category fixed effects in the model. To ensure reliable estimation of these fixed effects across all equations, we include only service categories that account for at least 1.5% of the jobs.

Second, to identify customers' unobserved heterogeneity in cancellation and rating behaviors, we require multiple jobs per customer. We therefore restrict the sample to customers who have requested at least four jobs in our analysis. After imposing these restrictions, the final estimation sample includes 36161 jobs requested by 5105 customers and accepted by 555 workers.

To identify heterogeneity in biases across minority and majority workers, it is also important that customers have been exposed to both types of workers. Under our sample restriction of customers with at least 4 job requests, most customers are indeed exposed to both groups. Fig. 2 reports the number of customers exposed to both minority and majority workers, only majority workers, and only minority workers. Among customers with 4 job requests, 78.3% have interacted with both groups. This share increases to 85.7% for customers with five requests and continues to increase with the number of job requests. For estimation, we use all customers with at least four job requests.

Figure 2 Customer Distribution by Number of Jobs Requested and Worker Composition



We briefly describe the explanatory variables included in the model. We include variables that capture the worker's past history, including the average rating and the logarithm of cumulative

job count up to each focal job. We use job count rather than rating volume because job count is the information displayed to customers on the platform, as shown in Fig. A.1 (Appendix A). Moreover, job count and rating volume are highly correlated in our data, with a correlation of approximately 0.99. Given our focus on how consumers make differential choices for minorities, we include a minority indicator and its interactions with both the normalized rating measure (defined in Section 4.3), and job count. As discussed, we control for the average of the first five ratings as a proxy for worker quality. We also include a dummy variable, *no rating*, indicating whether the worker has received fewer than 5 ratings and therefore has no displayed rating.¹⁰ Finally, we include service category, year, and customer location fixed effects.

6.2. Model Estimates

We estimate the model with two, three, and four customer segments, respectively. Based on the highest log-likelihood and lowest AIC, we choose the three-segment model as the best-fitting model.¹¹ We adopt the three-segment model for our main empirical analysis. Table 5 reports the estimates for the consumer model of cancellation decisions, rating submission, and rating choice.

For all three segments, the displayed rating enters the cancellation equation with a negative coefficient, indicating that higher ratings reduce cancellations and increase job earnings. Further, we find evidence of anchoring effects across all segments, as the coefficient on displayed past ratings in the rating equation is positive and statistically significant for each segment. These results support the two key assumptions of our theoretical model regarding the effects of past ratings on cancellations and current ratings. Notably, these anchoring effects are identified after controlling

¹⁰ We also considered including an interaction between *no rating* and the minority indicator, but the coefficient is not statistically significant and the results are robust to its inclusion. We therefore exclude it from the main specification. Estimates with this interaction are reported in ?? in Online Appendix ??.

¹¹ The log-likelihood for two-segment (see model estimates in Online Appendix ??) and three-segment models are -39629 and -39053 with corresponding AIC values of 79615 and 78643. The four-segment model does not converge due to over-parameterization from the large number of segment-level fixed effects.

Table 5 Consumer Model Parameter Estimates

	Segment 1: Unbiased Customers	Segment 2: Minority Avoiders	Segment 3: Minority Avoiders & Under-Raters
Segment Probability	0.20 (0.02)	0.24 (0.01)	0.56 (0.02)
Cancellation			
Rating	-0.66 (0.22)**	-0.63 (0.18)**	-0.65 (0.12)**
logJobCount	-0.07 (0.04)	0.04 (0.04)	-0.03 (0.02)
Minority	0.41 (0.32)	0.61 (0.24)*	0.37 (0.17)*
Minority × Rating	0.06 (0.31)	0.65 (0.26)*	-0.11 (0.17)
Minority × logJobCount	-0.04 (0.06)	-0.14 (0.05)**	-0.05 (0.03)
NoRating	-0.54 (0.25)*	-0.13 (0.20)	-0.58 (0.14)**
Constant	-0.66 (0.39)	-2.31 (0.51)**	-1.08 (0.20)**
Service FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Customer Location FE	Yes	Yes	Yes
Rating Submission			
Rating	0.16 (0.33)	0.66 (0.18)**	0.58 (0.21)**
logJobCount	-0.11 (0.05)*	-0.07 (0.03)*	-0.02 (0.03)
Minority	0.17 (0.48)	0.37 (0.22)	0.38 (0.30)
Minority × Rating	-0.74 (0.44)	-0.25 (0.24)	-0.07 (0.29)
Minority × logJobCount	0.02 (0.08)	-0.03 (0.04)	-0.06 (0.05)
WorkerQuality	0.09 (0.16)	-0.16 (0.10)	0.01 (0.10)
NoRating	-0.61 (0.38)	0.40 (0.18)*	0.75 (0.27)**
Constant	4.77 (0.84)**	-0.64 (0.31)*	4.19 (0.41)**
Service FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Customer Location FE	Yes	Yes	Yes
5-Star Rating			
Rating	1.18 (0.23)**	1.10 (0.48)*	1.12 (0.19)**
logJobCount	0.08 (0.04)	-0.04 (0.09)	0.01 (0.04)
Minority	-0.19 (0.32)	0.22 (0.59)	-0.01 (0.23)
Minority × Rating	0.15 (0.32)	0.08 (0.64)	-0.88 (0.25)**
Minority × logJobCount	0.00 (0.06)	-0.10 (0.13)	0.05 (0.05)
WorkerQuality	0.04 (0.13)	0.12 (0.26)	0.38 (0.11)**
NoRating	1.43 (0.27)**	0.61 (0.50)	0.19 (0.20)
Constant	-1.66 (0.44)**	1.83 (1.03)	0.99 (0.29)**
Service FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Customer Location FE	Yes	Yes	Yes
N: 36161			
log-likelihood: -39053			
AIC: 78643			

Note: *p<0.05; **p<0.01. Coefficients are reported with standard errors in parentheses. Estimates for the service category, year, and customer location fixed effects are reported in Online Appendix ??.

for worker quality using the first five ratings, which are not displayed to customers and therefore do not generate spillover effects.¹²¹³

¹² As we noted in the identification section, the magnitude of the coefficient on worker quality will be underestimated (and therefore lead to conservative estimates of bias) due to minority bias embedded in this initial quality estimate.

Next, we turn to the discrimination effects across the three segments. Overall, only a subset of the minority related coefficients is statistically significant. This pattern is consistent with our theory, which emphasizes that discrimination arises only in some segments, while the presence of rating systems may generate spillovers and amplify discrimination across segments. Specifically, for Segment 1 (20% of customers), none of the minority coefficients or minority interactions are significantly different from zero. Any differences in their behaviors between minorities and majorities arise only through observable variables. We therefore concluded that this segment does not engage in either taste or belief-based (i.e., statistical) discrimination between minority and majority workers. We call this the “unbiased” (or no-discrimination) segment.

For Segment 2 (24% of customers), both the minority coefficient and its interaction term with the displayed rating are statistically significant in the cancellation model. That is, customers in this segment are more likely to cancel jobs accepted by minority workers, and although higher ratings reduce cancellations for minority workers, customers in this segment do not give as much credit to minority workers based on the displayed rating. Although minorities may partially offset discrimination by accumulating more jobs, the median value of *logJobCount* is 4.82, making it difficult to fully mitigate the effect. On the other hand, neither the minority main effect nor its interaction terms are statistically significant in the rating equation for this segment. We call this the “minority avoider” segment.

Finally, Segment 3 (56% of the customers) discriminates in both the cancellation and rating stages. In the cancellation stage, customers in this segment are more likely to cancel jobs accepted by minorities regardless of the displayed rating. In the rating stage, compared to customers in segments 1 and 2, they systematically assign lower ratings to minority workers than to majority workers with the same level of past ratings. This reflects δ in the analytical model. We refer to this group as the “minority avoider and under-rater” segment.

¹³ While we cannot test whether performance varies over time given the available data, a demotivation mechanism where minority workers who receive lower ratings may reduce effort over time would predict anchoring effects primarily for minority workers. Instead, we find anchoring effects for both groups, suggesting this mechanism is unlikely.

6.3. Latent Segment Characteristics

Table 6 presents descriptive statistics on cancellations and ratings by customer segment, based on both model predictions and the observed data. The share of jobs requiring credentials is similar across all three segments, suggesting no significant differences in job types. In the cancellation stage, the minority gap is largest for Segment 2, the minority avoider segment, and smaller for Segment 3, the minority avoider and under rater segment. Consistent with both our model estimates and the data descriptives, rating submission rates do not differ significantly between minority and majority workers in any segment. Finally, all segments exhibit rating differences between minority and majority workers, while Segment 1 are much harsher customers who give significantly fewer 5-star ratings than the other two segments.

Table 6 Customer Segments Description: Predicted vs. Actual

	Segment 1 Neutral Customers		Segment 2 Minority Avoiders		Segment 3 Minority Avoiders & Under-Raters	
	Pred.	Actual	Pred.	Actual	Pred.	Actual
total job count	6025		8277		21859	
% jobs need credential	41.00%		41.28%		42.15%	
<i>Cancellation Rate</i>						
majority	19.12%	19.12%	14.55%	14.55%	15.28%	15.28%
minority	23.03%	23.03%	20.31%	20.31%	18.44%	18.44%
<i>Rating Submission Rate</i>						
majority	82.68%	82.62%	29.70%	29.69%	90.92%	90.94%
minority	80.01%	80.11%	31.04%	31.09%	89.90%	89.87%
<i>Perc. Rating=5</i>						
majority	59.26%	59.24%	89.00%	89.00%	91.36%	91.35%
minority	52.55%	52.48%	85.09%	84.89%	88.99%	88.99%

Note: For each metric and each segment, given customers' posterior segment probabilities, we first compute the average choice probabilities from the model, and then compute the corresponding average from the observed data. The close correspondence between model predictions and the data indicates good model fit.

What is particularly noteworthy is that the rating gap between minority and majority workers arises in Segment 1, the unbiased segment. As shown in Table 5, while customers in Segment 1 do not appear to be intrinsically discriminatory, their ratings nonetheless exhibit “discriminatory” patterns due to anchoring on previously displayed ratings. This substantial gap stems also from

the lower baseline of ratings in Segment 1, as shown by the estimated constant in the rating model in Table 5. In other words, although these customers do not discriminate themselves, they are influenced by the lower ratings that minority workers have received from the discriminatory segment, leading them to give lower ratings due to anchoring. As a result, even a proportion of discriminatory customers can generate spillover effects that affect how neutral customers rate minority workers. Without a model that explicitly controls for displayed ratings and accounts for latent customer heterogeneity, such neutral customers could be misclassified as “discriminatory,” leading to misleading interpretations and counterfactual conclusions.

7. Counterfactuals

Using the model estimates, we conduct counterfactual analysis to answer our second and third research questions: (i) what is the magnitude of the minority rating and earnings gap due to customer discrimination; and (ii) how much of the rating and earnings gaps can be mitigated by adjusting the displayed ratings to reduce discrimination spillover and amplification.

7.1. Quantify the impact of customer discrimination on the minority gap

To examine the effect of customer discrimination on minority workers’ rating and earnings gaps, we compare ratings and earnings between minority and majority workers within each service category. More specifically, for each service category, we randomly sample 1000 majority and 1000 minority workers by drawing the quality metric, defined as the average of the first five ratings, from its empirical distribution. This ensures that workers differ only in race and quality, while the quality distribution remains identical across the two groups. For each matched pair of minority and majority workers with the same quality, we simulate 300 jobs from 75 customers, each offering four jobs. Each customer is randomly drawn from one of the three estimated customer segments, with segment shares matching the estimates in Table 5. Customer location is then drawn from the empirical distribution conditional on the service category.

To align the simulations with data, we apply the corresponding service category and customer location fixed effects from the model estimates so that the baseline cancellation, rating submission,

and rating outcomes match the empirical patterns.¹⁴ Overall, these simulation choices imply that the jobs are all homogeneous within a service category, except for customer segment and location.

Given each service category, we simulate the cancellations, ratings submission, and rating outcomes for the 300 jobs for 1000 majority workers and 1000 minority workers.¹⁵ We then compute average ratings and job counts across all simulated workers over the full job sequence, separately for majority and minority workers. The resulting ratings and job counts by service category are reported in Table 7.

To measure the rating gap, we report the percentage of sub-5-star ratings for the majority and minority workers separately in Columns 2 and 3, as well as the percentage gap between them in Column 4. Columns 5 and 6 show the average job count for the two groups, and Column 7 reports the earnings gap, measured as the percentage difference in total job count of majority workers relative to minority workers. Across all service categories, majority workers receive both higher ratings and more jobs, though the magnitude of these differences varies by category. The overall rating gap, which is the average of service-specific rating gap, weighted by the empirical share of jobs, is 33.9%. The overall minority earnings gap is 6.5%. These results indicate that even though minority and majority workers start with identical quality distributions, bias among customers and

¹⁴ We use the 2019 year fixed effect, so the simulation results should be interpreted as reflecting conditions in 2019, the most recent year in our sample. In addition, the coefficient on *logJobCount* variable for Segment 2 in the rating equation is small and insignificant, but has a negative sign (opposite of what is theoretically expected). To avoid introducing noise into the simulation results, we set this coefficient to zero. Similarly, we set all insignificant coefficients on minority-related variables to zero.

¹⁵ In this analysis, given the institutional features of the setting, we assume that a canceled job is a lost job. This focuses the analysis on demand-side responses and yields a back-of-the-envelope estimate of discrimination's economic impact. If workers could easily substitute a canceled job with a new one, as on platforms such as Uber, the earnings gap could be smaller. However, in our setting, cancellations typically occur well after job acceptance and closer to the scheduled time. Specifically, canceled jobs are typically scheduled about two days in advance, and cancellations most often occur more than one day after scheduling. As a result, it is difficult for workers to replace a canceled job, and our simulated earnings reflect this institutional feature.

Table 7 Minority Rating and Earnings Gap by Service Category

Service Category	% Sub-5 Star Rating			Job Count (Earnings)		
	Majority	Minority	%Gap	Majority	Minority	%Gap
Maintenance	8.4	11.5	36	245	228	7.5
Plumbing	10.1	13.4	33	259	245	5.6
Appliance (Non Cred.)	10.5	13.6	29	262	249	5.0
Landscaping	14.1	18.8	33	255	240	6.3
Electrical	9.4	12.5	33	256	243	5.6
HVAC	9.6	12.8	34	241	223	8.2
Gutters	11.9	16.0	35	262	248	5.4
Snow	13.5	17.7	32	256	241	6.2
Moving	7.8	11.0	41	250	233	7.1
Upholstery	15.2	19.7	30	218	195	11.7
Appliance (Cred.)	6.1	9.5	56	247	230	7.1
Locksmith	7.0	9.9	42	266	255	4.5
Misc.	11.8	15.9	35	243	225	7.9

Note: The service categories are listed in descending order by the empirical share in the data.

the spillover and amplification effects of displaying ratings can produce gaps in both ratings and earnings.

7.2. Adjust ratings to mitigate the impact of customer discrimination

The counterfactual simulations in Section 7.1 quantify how much customer discrimination and display of ratings lead to ratings and earnings gaps. In this section, we propose a rating adjustment approach based on empirical inference of worker quality and customer segment. This approach reduces the effects of discrimination spillovers and amplification while preserving the informational value of customer ratings and thus remaining realistic from a platform design perspective.

Specifically, the platform takes as priors the empirical distribution of worker quality (measured by the average of the first ratings observed in the data) and the customer segment distribution reported in Table 5. After each job is completed, the platform updates its beliefs about both the worker’s quality and the customer’s segment based on the observable job outcomes (cancellation, rating submission, and rating outcome). Using the updated distributions, for each job completed before, the platform then computes the probability that the customer would have given a 5-star rating had the worker been from the majority group, holding all other observables (e.g., number of jobs completed) constant. A new rating is subsequently drawn according to this probability. The details on the implementation of this counterfactual are provided in Appendix D.

Table 8 reports the gaps before and after adjusting the rating display by service category. Columns 2 and 5 show the ratings and earnings gaps when ratings are unadjusted, as reported in Table 7. Columns 3 and 6 show the ratings and earnings gaps when ratings are adjusted, and Columns 4 and 7 show the percentage reduction in the gaps after adjustment. Overall, when weighted by the proportion of jobs within each service category, with adjusted ratings, the rating gap and earnings gap are 29.7% and 5.9% respectively. That is, adjusting ratings reduces the rating gap by 12.2% and the earnings gap by 9.1%. Interestingly, for the neutral segment, the adjusted rating virtually wipes out the minority rating gap from 4.23% to 0.11%, a decline of 97%. The minority earnings gap also falls from 0.73% to 0.06%, corresponding to a reduction of 91%.¹⁶

Table 8 Minority Rating and Earnings Gap with/without Ratings Adjusted by Service Category

Service Category	% Rating Gap			% Earnings Gap		
	No Adjust	Adjust	% ↓ in Gap	No Adjust	Adjust	% ↓ in Gap
Maintenance	36	32	12.2	7.5	6.9	8.6
Plumbing	33	29	13.6	5.6	5.1	10.4
Appliance (Non Cred.)	29	26	12.1	5.0	4.6	8.8
Landscaping	33	28	15.9	6.3	5.6	13.2
Electrical	33	29	12.3	5.6	5.1	10.3
HVAC	34	30	14.0	8.2	7.4	9.6
Gutters	35	30	17.3	5.4	4.9	11.7
Snow	32	27	18.7	6.2	5.5	12.2
Moving	41	37	12.5	7.1	6.6	8.0
Upholstery	30	26	15.5	11.7	10.7	9.5
Appliance (Cred.)	56	52	7.5	7.1	6.5	10.1
Locksmith	42	38	10.2	4.5	4.2	8.9
Misc.	35	30	15.2	7.9	7.1	11.4

Note: The service categories are listed in descending order by the empirical share in the data.

Our simulations condition on both the minority and majority workers having accepted the same number of jobs. Because the priority algorithm for job acceptance itself incorporates ratings and job

¹⁶ Another potential and “naive” approach to mitigating the effects of rating display on discrimination spillovers and amplification would be to remove ratings altogether. However, because rating systems provide valuable information and facilitate transactions, the adjustment counterfactual is more realistic from the platform’s perspective. Nevertheless, we conduct an additional set of simulations in which the platform does not display ratings. Without modeling the impact of displaying ratings and focusing only on discrimination effects, we find that the rating and earnings gaps are broadly similar to those reported in Table 8. We provide the details in Appendix E.

counts as inputs, being agnostic about the job acceptance procedure can lead to an underestimation of the total effect of adjusting rating displays on worker earnings by omitting its impact on job acceptance. Moreover, in our empirical setting, worker performance plays a relatively limited role in shaping customer demand, as workers may accept jobs on a first-come, first-served basis, and customers can subsequently cancel. On platforms where customers actively select workers based on performance metrics, such as when they can view a list of workers and choose whom to hire, the impact of rating adjustments on earnings gaps could therefore be larger. Given these caveats about not accounting for the supply side, and the fact that customers do not choose workers based on ratings, it is important to note that the above estimates of the ratings and earnings gaps are likely to be a lower bound, and in other settings, the gaps may be larger.

8. Conclusion

Customer rating systems have become a defining feature of online commerce and service platforms. Consumers use ratings to select among options and manage perceived risk, while platforms rely on the same metrics to govern suppliers and service providers. Nevertheless, a common concern is that rating systems may embed biased evaluations. We propose a novel idea: when ratings embed biased evaluations from even a subset of customers, publicly displaying aggregated ratings can *amplify* discrimination over time and generate spillovers to customers who would not otherwise discriminate. The core mechanism is that displayed ratings memorialize discrimination-tainted ratings as differences in “quality,” and these displayed metrics then affect both subsequent opportunity (via job acceptance/cancellation) and subsequent evaluation (via anchoring).

We formalized the idea in a stylized analytical model and then empirically investigated it on an online labor market platform. Allowing for unobserved heterogeneity in customer behavior, we find three segments of customers: an unbiased segment (that does not discriminate), a minority avoider segment (that discriminates on cancellations), and a minority avoider and under-rater segment (that discriminates on both cancellations and ratings). Overall, we find that customer discrimination leads to a 33.9% rating gap and a 6.5% earnings gap between minorities and majorities.

Further, under the proposed approach for adjusting the ratings displayed, the rating gap decreases by 12.2%, and the earnings gap decreases by 9.1%. We note that though the earnings gap of 6.5% may seem small, to the extent that online labor market workers depend on this income, the cumulative impact of the 6.5% earnings gap in terms of savings and wealth gaps will be much larger.

We conclude with a discussion of limitations and suggestions for future work. First, we abstract from workers' dynamic responses to cancellation rates and review outcomes. Future work should investigate the supply side (worker) responses to discrimination in terms of effort or their decisions to participate on the platform.

Second, because our analysis focuses on a single online labor market platform, an important direction for future research is to assess the generalizability of our findings across platforms and to examine how platform design features shape discrimination and earnings gaps. For example, the anonymity of customer actions such as cancellations and ratings may facilitate discriminatory behavior. Moreover, in settings where discrimination is costly, the information content of reviews may mitigate discrimination more effectively, as in [Cui et al. \(2020\)](#), who show that even mildly negative reviews substantially reduce discrimination against minority guests on Airbnb. One interpretation is that when discrimination entails forgone revenue, even limited information can be sufficient to overturn negative priors. In contrast, in settings such as ours, customers may face little economic cost or inconvenience from canceling a minority worker or assigning a lower rating, even when reviews are informative. Future work could formally investigate how the costs of discrimination and decision-makers' incentives interact with review information to determine when reviews attenuate, rather than amplify, discriminatory behavior.

Third, the small number of jobs per customer limits our ability to identify temporal changes in discriminatory behavior within consumers based on their exposure to minorities. Although priors about minorities likely arise from a lifetime of experience and may not change significantly based on limited experience within a platform, when discriminatory behavior can be mitigated through greater exposure to minority workers remains an open question.

In general, the discrimination literature has focused on average effects and typically has not considered unobserved heterogeneity. As we show, accounting for unobserved heterogeneity provides a richer description of discrimination. This is relevant not just in marketing settings, but also in settings such as education and criminal justice, where there is much work on bias and discrimination, and there is likely heterogeneity among teachers and police officers in whether they discriminate and the magnitude of discrimination. Not only does accounting for unobserved heterogeneity provide an accurate description, but it can also lead to greater acceptance by society of the research findings, as it fits the lay notion that not all people discriminate, and equally. In particular, we hope that our work inspires the literature on fairness and biases in machine learning to account for unobserved heterogeneity, and the impacts of cross-segment spillovers and amplification.

Our modeling approach is also a useful lens to study questions of structural inequity in settings such as education and criminal justice. For example, it can be used to quantify how biases in early disciplinary actions in school or interactions with the criminal justice system can lead to a “record” that later justifies harsher actions against minorities, leading to worse life outcomes. More broadly, we hope our case-based modeling approach allows managers and scholars not only to measure the presence of discrimination, but also to quantify how early discriminatory outcomes, when translated into merit, lead to structural inequity over time and to assess potential mitigation strategies.

Funding and Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

References

- Abdulsalam K, Christensen DM, Graffin SD, Li J (2025) Do boards reward and punish ceos based on employee satisfaction ratings? *Organization Science* 36(2):881–902.
- Adaval R, Wyer Jr RS (2011) Conscious and nonconscious comparisons with price anchors: Effects on willingness to pay for related and unrelated products. *Journal of Marketing Research* 48(2):355–365.

- Aguiar L (2024) Bad apples on rotten tomatoes: critics, crowds, and gender bias in product ratings. *Marketing Science* .
- Altonji JG, Doraszelski U (2005) The role of permanent income and demographics in black/white differences in wealth. *Journal of Human Resources* 40(1):1–30.
- Ananthakrishnan U, Proserpio D, Sharma S (2023) I hear you: does quality improve with customer voice? *Marketing Science* 42(6):1143–1161.
- Aral S (2013) The problem with online ratings. *MIT Sloan Management Review* .
- Arrow KJ (1973) The theory of discrimination. *Discrimination in Labor Markets*, 1–33 (Princeton NJ: Princeton University Press).
- Ayres I, Banaji M, Jolls C (2015) Race effects on ebay. *The RAND Journal of Economics* 46(4):891–917.
- Ayres I, Siegelman P (1995) Race and gender discrimination in bargaining for a new car. *The American Economic Review* 304–321.
- Baehre S, O'Dwyer M, O'Malley L, Story VM (2022) Customer mindset metrics: A systematic evaluation of the net promoter score (nps) vs. alternative calculation methods. *Journal of Business Research* 149:353–362.
- Bar R, Zussman A (2017) Customer discrimination: evidence from israel. *Journal of Labor Economics* 35(4):1031–1059.
- Becker G (1957) *The economics of discrimination* 1st ed (university of chicago press, chicago).
- Bertrand M, 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.
- Blanchflower DG, Levine PB, Zimmerman DJ (2003) Discrimination in the small-business credit market. *Review of Economics and Statistics* 85(4):930–943.
- Bodvarsson ÖB, Partridge MD (2001) A supply and demand model of co-worker, employer and customer discrimination. *Labour Economics* 8(3):389–416.
- Bohnet I, Hauser OP, Kristal AS (2025) Can gender and race dynamics in performance appraisals be disrupted? the case of social influence. *Journal of Economic Behavior & Organization* 235:107032.

- Bohren JA, Imas A, Rosenberg M (2019) The dynamics of discrimination: Theory and evidence. *American economic review* 109(10):3395–3436.
- Botelho TL (2024) From audience to evaluator: When visibility into prior evaluations leads to convergence or divergence in subsequent evaluations among professionals 35(5):1682–1703, ISSN 1047-7039.
- Botelho TL, DeCelles KA (2025) The customer cancellation gap: The drivers of racial/ethnic disparities in on-demand work. *Working Paper* .
- Botelho TL, Gertsberg M (2022) The disciplining effect of status: Evaluator status awards and observed gender bias in evaluations. *Management Science* 68(7):5311–5329.
- Botelho TL, Jun S, Humes D, DeCelles KA (2025) Scale dichotomization reduces customer racial discrimination and income inequality. *Nature* 1–9.
- Brewster ZW, Lynn M (2014) Black–white earnings gap among restaurant servers: A replication, extension, and exploration of consumer racial discrimination in tipping. *Sociological Inquiry* 84(4):545–569.
- Chakraborty I, Kim M, Sudhir K (2022) Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes. *Journal of Marketing Research* 59(3):600–622.
- Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43(3):345–354.
- Cui R, Li J, Zhang DJ (2020) Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on airbnb. *Management Science* 66(3):1071–1094.
- De Haan E, Verhoef PC, Wiesel T (2015) The predictive ability of different customer feedback metrics for retention. *International Journal of Research in Marketing* 32(2):195–206.
- Donati D (2025) The end of tourist traps: The impact of review platforms on quality upgrading. *Marketing Science* .
- Edelman B, Luca M, Svirsky D (2017) Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics* 9(2):1–22.
- Fang L (2022) The effects of online review platforms on restaurant revenue, consumer learning, and welfare. *Management Science* 68(11):8116–8143.

- Furnham A, Boo HC (2011) A literature review of the anchoring effect. *The journal of socio-economics* 40(1):35–42.
- Greenwood B, Adjerid I, Angst CM, Meikle NL (2022) How unbecoming of you: Online experiments uncovering gender biases in perceptions of ridesharing performance. *Journal of Business Ethics* 175(3):499–518.
- Ham SH, Koch I, Lim N, Wu J (2021) Conflict of interest in third-party reviews: an experimental study. *Management Science* 67(12):7535–7559.
- Hannák A, Wagner C, Garcia D, Mislove A, Strohmaier M, Wilson C (2017) Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 1914–1933.
- Kim K, Chung K, Lim N (2019) Third-party reviews and quality provision. *Management Science* 65(6):2695–2716.
- Kricheli-Katz T, Regev T (2016) How many cents on the dollar? women and men in product markets. *Science advances* 2(2):e1500599.
- Lee J, Lim Y, Oh HI (2018) Does customer satisfaction matter to managers' earnings forecasts and stock returns? *European Journal of Marketing* 52(9/10):2026–2051.
- Luca M (2016) Reviews, reputation, and revenue: The case of yelp. com. *Com (March 15, 2016)*. *Harvard Business School NOM Unit Working Paper* (12-016).
- Lynn M, Sturman M (2011) Is the customer always right? the potential for racial bias in customer evaluations of employee performance. *Journal of Applied Social Psychology* 41(9):2312–2324.
- Ma X, Khansa L, Deng Y, Kim SS (2013) Impact of prior reviews on the subsequent review process in reputation systems. *Journal of Management Information Systems* 30(3):279–310.
- Marchenko A (2019) The impact of host race and gender on prices on airbnb. *Journal of Housing Economics* 46:101635.
- Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* 104(8):2421–2455.
- Muchnik L, Aral S, Taylor SJ (2013) Social influence bias: A randomized experiment. *Science* 341(6146):647–651.

- Northcraft GB, Neale MA (1987) Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational behavior and human decision processes* 39(1):84–97.
- Oliver M, Shapiro T (2013) *Black wealth/white wealth: A new perspective on racial inequality* (Routledge).
- Owens J, McLanahan SS (2020) Unpacking the drivers of racial disparities in school suspension and expulsion. *Social Forces* 98(4):1548–1577.
- Pager D, Shepherd H (2008) The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annu. Rev. Sociol* 34:181–209.
- Park S, Shin W, Xie J (2021) The fateful first consumer review. *Marketing Science* 40(3):481–507.
- Phelps ES (1972) The statistical theory of racism and sexism. *The american economic review* 62(4):659–661.
- Reimers I, Waldfogel J (2021) Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings. *American Economic Review* 111(6):1944–1971.
- Rosenblat A, Levy K, Barocas S, Hwang T (2016) Discriminating tastes: Customer ratings as vehicles for bias. *Data & Society* 1–21.
- Schlosser AE (2005) Posting versus lurking: Communicating in a multiple audience context. *Journal of Consumer Research* 32(2):260–265.
- Shapiro TM, et al. (2004) *The hidden cost of being African American: How wealth perpetuates inequality* (Oxford University Press, USA).
- Subramanian A (2019) *The caste of merit: Engineering education in India* (Harvard University Press).
- Sunder S, Kim KH, Yorkston EA (2019) What drives herding behavior in online ratings? the role of rater experience, product portfolio, and diverging opinions. *Journal of Marketing* 83(6):93–112.
- Tadelis S (2016) Reputation and feedback systems in online platform markets. *Annual review of economics* 8(1):321–340.
- Thorsteinson TJ, Breier J, Atwell A, Hamilton C, Privette M (2008) Anchoring effects on performance judgments. *Organizational Behavior and Human Decision Processes* 107(1):29–40.
- Tomaskovic-Devey D, Thomas M, Johnson K (2005) Race and the accumulation of human capital across the career: A theoretical model and fixed-effects application. *American Journal of Sociology* 111(1):58–89.

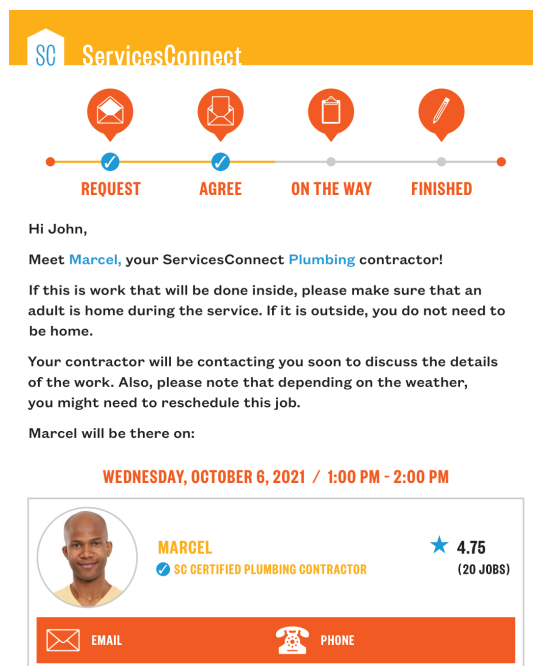
- Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185(4157):1124–1131.
- Wang Q, Chau M, Peng CH, Ngai EW (2022) Using the anchoring effect and the cultural dimensions theory to study customers' online rating behaviors. *Information Systems Frontiers* 24(5):1451–1463.
- Whitley SC, Sevilla J, Isaac MS (2025) Units or pounds? how anchoring on salient price information influences perceptions of product value. *Journal of Marketing Research* 62(5):876–894.
- Yinger J (1995) *Closed doors, opportunities lost: The continuing costs of housing discrimination* (Russell Sage Foundation).

Appendix

A. Communications with Customers

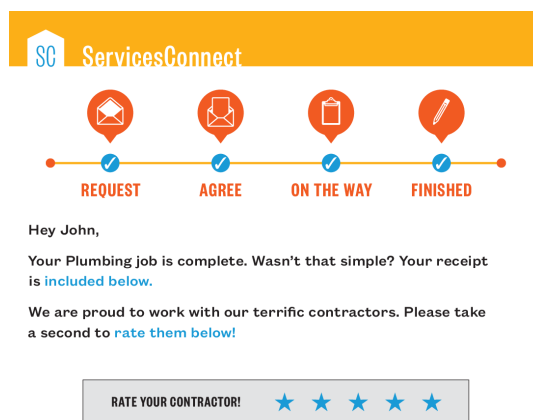
Figure A.1 Screenshots of Emails

(a) Confirmation Email after a Worker Accepts Job



If you need to adjust the timing or have questions, please contact Marcel or use the [links](#) provided, or use the [help desk](#).

(b) Rating Reminder Email after a Job is Completed



If you have questions or complaints, please visit our [help desk](#) or [reply to this email](#).

B. Proof for the Stylized Analytical Model

We prove Proposition 1 in the following steps.

Step 1: We first derive the ratings for both types of workers when ratings are hidden. For brevity, we omit the superscript h here.

When ratings are hidden, given the normalization of $R_0^w = 1$ and Table 2, for the advantaged worker, we have $\forall t \geq 1$, $r_t^{AN} = 1$ and $r_t^{AP} = 1$; for the disadvantaged worker, we have $\forall t \geq 1$, $r_t^{DN} = 1$ and $r_t^{DP} = 1 - \delta$. For both workers, given the normalization, $\forall t$, $n_t^w = 1$ and thus $\forall t$, $n_t^{wN} = 1 - \alpha$, $n_t^{wP} = \alpha$. Therefore, from Eq. (1) and Eq. (2), we have $\forall t \geq 1$, $r_t^A = 1$, $R_t^A = 1$, $r_t^D = (1 - \delta)^\alpha$, $R_t^D = (1 - \delta)^\alpha$.

Step 2: Next, we consider the case where ratings are displayed. For brevity, we omit the superscript o here.

For the disadvantaged worker, given R_{t-1}^D and N_{t-1}^D , in period t , we have

$$\begin{aligned} n_t^{DN} &= (1 - \alpha) \times (R_{t-1}^D)^{1/N_{t-1}^D}, \quad n_t^{DP} = \alpha \times (R_{t-1}^D)^{1/N_{t-1}^D} \\ r_t^D &= (r_t^{DN})^{1-\alpha} \times (r_t^{DP})^\alpha = (1 - \delta)^\alpha \times R_{t-1}^D \end{aligned}$$

The last equation follows from Eq. (1) and Table 2.

The cumulative rating at the end of period t is given by

$$\begin{aligned} R_t^D &= \left(\prod_{t'=1}^t (r_{t'}^D)^{n_{t'}^D} \right)^{\frac{1}{\sum_{t'=1}^t n_{t'}^D}} = \left(\left(\prod_{t'=1}^{t-1} (r_{t'}^D)^{n_{t'}^D} \right) \times (r_t^D)^{n_t^D} \right)^{\frac{1}{N_t^D}} \\ &= \left(\prod_{t'=1}^{t-1} (r_{t'}^D)^{n_{t'}^D} \right)^{\frac{1}{N_{t-1}^D} \times \frac{N_{t-1}^D}{N_t^D}} \times \left((r_t^D)^{n_t^D} \right)^{\frac{1}{N_t^D}} \\ &= (R_{t-1}^D)^{\frac{N_{t-1}^D}{N_t^D}} \times (r_t^D)^{\frac{n_t^D}{N_t^D}} \\ &= (R_{t-1}^D)^{\frac{N_{t-1}^D}{N_t^D}} \times ((1 - \delta)^\alpha \times R_{t-1}^D)^{\frac{n_t^D}{N_t^D}} \\ &= (1 - \delta)^{\alpha \times \frac{n_t^D}{N_t^D}} \times R_{t-1}^D \end{aligned} \tag{7}$$

Given $R_0^D = 1$, following the recursion above, we have $R_1^D = (1 - \delta)^\alpha$. Since, $0 < 1 - \delta < 1$ and $0 < \alpha \times \frac{n_t^D}{N_t^D} < 1$, we have $R_t^D < R_{t-1}^D \forall t \geq 1$. That is, the disadvantage worker's rating declines over time. Therefore, $\forall t \geq 2$, $R_t^D < (1 - \delta)^\alpha$ and $\forall t \geq 3$, $r_t^{DN} = R_{t-1}^D < (1 - \delta)^\alpha$.

For the advantaged worker, the recursion in Eq. (7) also applies if we plug in $\delta = 0$ and $\alpha = 1$. Therefore, $R_t^A = R_{t-1}^A$. Hence, $\forall t$, $R_t^A = 1$.

Step 3: Given Steps 1 and 2, we prove the three results respectively.

The spillover result: From Step 1, we know that $\forall t$, $r_t^{DN,h} = 1$. From Step 2, we have $r_1^{DN,o} = 1$, $r_2^{DN,o} = R_1^{D,o} = (1-\delta)^\alpha$ and $\forall t \geq 2$, $r_t^{DN,o} = R_{t-1}^{D,o} \leq R_1^{D,o} = (1-\delta)^\alpha$. Therefore, $\forall t \geq 2$, $r_t^{DN,o} \leq (1-\delta)^\alpha < 1 = r_t^{DN,h}$.

The amplification result: When ratings are displayed and observed by customers, from Step 2, we know that $\forall t \geq 1$, $R_t^{D,o} < R_{t-1}^{D,o}$. Thus, $\forall t \geq 2$, $r_t^{DN,o} = R_{t-1}^{D,o} < R_{t-2}^{D,o} = r_{t-1}^{DN,o}$ and $r_t^{DP,o} = (1-\delta) \times R_{t-1}^{D,o} < (1-\delta) \times R_{t-2}^{D,o} = r_{t-1}^{DP,o}$. Also, $R_t^{A,o} - R_t^{D,o} = 1 - R_t^{D,o} > 1 - R_{t-1}^{D,o} = R_{t-1}^{A,o} - R_{t-1}^{D,o}$.

The earnings gap: When ratings are displayed, from $R_t^{A,o} = 1$, we have $n_t^{A,o} = 1$. Then the earnings gap is given by

$$N_t^{A,o} - N_t^{D,o} = \sum_{t'=1}^t (n_{t'}^{A,o} - n_{t'}^{D,o}) = \sum_{t'=1}^t (1 - n_{t'}^{D,o})$$

For $t' \geq 2$, $n_{t'}^{D,o} = (R_{t'-1}^{D,o})^{\frac{1}{N_{t'-1}^{D,o}}}$. From Step 2, we have $\forall t' \geq 1$, $R_{t'}^{D,o} \leq (1-\delta)^\alpha < 1$. Hence, $\forall t' \geq 2$, $n_{t'}^{D,o} < 1$.

Thus, $\forall t \geq 2$, $N_{t+1}^{A,o} - N_{t+1}^{D,o} > N_t^{A,o} - N_t^{D,o} > 0$

C. Table of Notations

Table A.1 summarizes the notations we use in the empirical model.

Table A.1 Table of Notations

Symbol	Category	Meaning
i	Index	Customer index
j	Index	Job index
k	Index	Worker index
g	Index	Index of the latent segment
C_{ijk}	Data	Customer i 's binary decision of whether to cancel job j
W_{ijk}	Data	Customer i 's binary decision of whether to submit a rating of job j
R_{ijk}	Data	Customer i 's binary decision of whether to submit a 5-star rating for job j
X_{ijk}^C	Data	Observables affecting the cancellation decision
X_{ijk}^W	Data	Observables affecting the rating submission decision
X_{ijk}^R	Data	Observables affecting the decision of whether to submit a 5-star rating
q^g	Parameter	Share of customers who belong to segment g
κ^g	Parameter	Parameters of segment g 's cancellation decision
λ^g	Parameter	Parameters of segment g 's rating submission decision
ρ^g	Parameter	Parameters of segment g 's rating decision

D. Implementation Details for Rating Adjustment in Section 7.2

Given a specific service category, the worker's true quality, and a sequence of customer draws, each simulated job trajectory with J jobs is generated as follows:

1. Initialize the platform's prior beliefs on worker quality and customer segments:
 - (i) Let π_K denote the platform's belief about the worker's quality. Initialize $\pi_K = \pi_K^0$.
 - (ii) For each job $j = 1, \dots, J$, let π_{G_j} denote the platform's belief about the segment of the customer requesting job j . For all j , initialize the beliefs $\pi_{G_j} = \pi_G^0$.
2. For each job $j = 1, 2, \dots$,
 - (i) Conditional on the worker's true quality and the segment of the customer requesting job j , simulate the observable outcomes for job j , including cancellation, rating submission and the realized rating.
 - (ii) Let S_j denote all observable decisions made by the customer requesting job j up to and including job j . Given S_j , update the joint posterior probability that the worker has quality k and that the customer requesting job j belongs to segment g using the Bayes' rule:

$$Pr(K = k, G_j = g | S_j) = \frac{Pr(S_j | K = k, G_j = g) Pr(K = k, G_j = g)}{\sum_{k', g'} Pr(S_j | K = k', G = g') Pr(K = k', G_j = g')},$$

where $Pr(K = k, G_j = g) = Pr(K = k) \times Pr(G_j = g)$, with $Pr(K = k) = \pi_K$ and $Pr(G_j = g) = \pi_{G_j}^0$ representing the platform's current priors.

- (iii) Update the platform's beliefs, π_K and π_{G_j} , according to the resulting posterior probabilities above.
- (iv) For each previously completed job $j' = 1, 2, \dots, j$ with a rating, compute the adjusted probability of receiving a 5-star rating under the counterfactual where the worker is majority, integrating over the posterior beliefs about the worker quality and customer segment. Specifically, for each job, we compute the probability of 5-star rating as follows:

$$p_{j'} = \sum_{k, g} Pr(K = k, G_{j'} = g) \times p_{j'}^{kg}$$

where $p_{j'}^{kg}$ is the probability that a customer of segment g would assign a five star rating to a *majority* worker of quality k , holding all other observables fixed. A new rating is then drawn from Bernoulli($p_{j'}$) so that the adjusted rating for job j' is 5-star if the draw is 1 and 4-star otherwise.

- (v) Using the adjusted ratings, the platform updates the displayed ratings for the next customer requesting job $j + 1$.

We set π_K^0 to be the empirical distribution of the average of the first five ratings across all workers in the data, and π_G^0 to be the estimated segment probabilities reported in Table 5.

E. Minority Gaps with Ratings Hidden

Here we compare the ratings and earning differences between minority and majority workers for a sequence of jobs when (i) ratings are displayed versus (ii) ratings are not displayed. For the first case, the simulation is identical to Section 7.1. For the second case, the simulation steps are largely similar but with one difference: when the rating is not displayed, starting from the 6th rating, we need an assumption about what the consumer will impute. To make the imputations comparable between the two cases, we assume that customers impute the average rating across all jobs in the corresponding service category in the display case.

Table A.2 reports the gaps with and without displayed ratings by service category. The second and the 5th columns show ratings and earnings gaps when ratings are displayed, where the numbers are identical to those in Table 7. The 3rd and 6th columns show the gaps when ratings are not displayed, and the 4th and 7th columns show the percentage gap closed when ratings are not displayed compared to the case when ratings are displayed. Overall, when weighted by the proportion of jobs within each service category, when ratings are hidden, the rating gap and earnings gap are 27.6% and 5.9% respectively. That is, removing rating display reduces the rating gap by 18.4% and the earnings gap by 8.6%.

Table A.2 Minority Rating and Earnings Gap with/without Displayed Ratings by Service Category

Service Category	% Rating Gap			% Earnings Gap		
	Display	No Display	% ↓ in Gap	Display	No Display	% ↓ in Gap
Maintenance	36	31	18.8	7.5	6.9	8.2
Plumbing	33	28	20.6	5.6	5.1	9.5
Appliance (Non Cred.)	29	24	23.9	5.0	4.6	8.1
Landscaping	33	26	27.6	6.3	5.6	12.0
Electrical	33	27	20.2	5.6	5.1	9.7
HVAC	34	28	22.4	8.2	7.5	9.1
Gutters	35	27	27.6	5.4	4.9	10.5
Snow	32	25	27.5	6.2	5.6	10.7
Moving	41	35	16.4	7.1	6.6	7.8
Upholstery	30	23	28.4	11.7	10.5	11.5
Appliance (Cred.)	56	50	11.2	7.1	6.5	9.8
Locksmith	42	36	14.6	4.5	4.2	8.2
Misc.	35	27	27.8	7.9	7.1	11.2

Note: The service categories are listed in descending order by the empirical share in the data.