Gender Disparity in Online Reputation: Evidence from an Online Freelance Platform*

Mimansa Bairathi, Anja Lambrecht, Xu Zhang

September 2024

Abstract

Online ratings aim to reduce information asymmetry between buyers and sellers, assisting buyers in making more informed purchase decisions. However, a problem arises if ratings are biased and do not reflect true buyer satisfaction. We analyze online ratings for services by male and female freelancers using data from an online freelance platform that elicits both public and private ratings from buyers post job completion. Public ratings are displayed on the website, while private ratings remain undisclosed. Using private ratings as a proxy for buyer satisfaction, we find that, conditional on the same private rating, public ratings are higher for male freelancers. We propose this is a result of buyers' gender stereotypes about confrontational behavior of freelancers following a critical review. Buyers are more concerned about a threat of confrontation with male than female freelancers and, thus, give male freelancers a higher public rating. We present two supporting pieces of evidence. This gender disparity in public ratings is most pronounced for buyers from countries with greater gender inequality and for buyers from US states with a lower percentage of women in the labor force. These results matter for freelancers and platforms because systematic differences in evaluation by gender can lead to unfair outcomes for platform participants.

Keywords: online ratings, gender disparity, digital platforms

Total Word Count: 6,282

*Mimansa Bairathi is Assistant Professor of Marketing at University College London (email: *m.bairathi@ucl.ac.uk*). Anja Lambrecht is Professor of Marketing at London Business School (email: *alambrecht@london.edu*). Xu Zhang is Assistant Professor of Marketing at London Business School (email: *xzhang@london.edu*). This paper is based on one of the essays that comprise the first author's dissertation. The authors thank Stephan Seiler and Naufel Vilcassim for their feedback, and Cheng Lu for their excellent research assistance. They also thank seminar participants at Marketing Science Doctoral Consortium 2023, Choice Symposium 2023, Keystone Strategy, LIEPP Workshop 2023, BASS Forms Conference 2024, Paris Conference on Digital Economics, UCLA, EMAC Conference 2024, and Customer Journeys in Digital World Conference 2024.

1 Introduction

Digital platforms are typically characterized by a large number of sellers offering a wide variety of items and users' uncertainty about item quality. In order to reduce information asymmetry, platforms have implemented online reputation systems such as displaying customers' reviews and ratings that inform buyers about product or service quality (Tadelis, 2016). Such reputation systems are particularly important when quality cannot easily be assessed ex-ante and products cannot be returned, such as on online freelance platforms.

Reputation systems have been shown to build trust between buyers and sellers (Abrahao et al., 2017) and reduce discrimination (Cui et al., 2020) because they allow sellers to signal their quality (Kanat et al., 2018). A concern arises, however, if, rather than presenting true service quality, ratings are themselves biased. This is a particular worry when the identity of the seller is visible to the buyer and the assessment of seller quality might be subject to buyers' biases. Initial evidence suggests that in online freelance markets where seller identity is known to buyers, individuals from underrepresented or protected groups (e.g., racial minorities, women) receive lower ratings than others (Hannák et al., 2017; Botelho and Gertsberg, 2022). However, lacking data on true buyer satisfaction, it is challenging to determine whether differences in ratings across groups are a result of racial or gender biases or instead reflect different levels of buyer satisfaction.

We use data from one of the world's largest online freelance platforms where freelancers offer services (e.g., app development, translation) for buyers to purchase and ask whether buyers evaluate male and female freelancers differently. Importantly, we use a buyer's private rating as a measure of their satisfaction to determine the degree to which differences in ratings are due to buyers' biases. Such private ratings are only shared with the platform, not the seller or the wider public, and are seen as a more reliable indicator of true buyer satisfaction (Filippas et al., 2022). This is because a buyer's fear of confrontation with the seller in response to a less favorable evaluation may lead them to inflate the rating that is publicly displayed. Because private ratings are not visible to the freelancer or other buyers, there is

little incentive to inflate them and so they typically reflect buyer satisfaction more closely.

We examine differences in public ratings for male and female freelancers, conditional on private ratings. Our results demonstrate consistently that male freelancers receive higher public ratings than female freelancers, after controlling for private ratings. We propose that this gender disparity is a result of gender stereotypes. In the context of digital platforms, previous research has demonstrated that aggressive behavior can alter ratings (Proserpio et al., 2021). Further, there is evidence that men are stereotypically considered to be more aggressive than women (Broverman et al., 1972; Ruble, 1983). Since people rely on stereotypes when forming expectations about others' behavior (Ridgeway, 2001), such stereotypes can shape users' ratings. If a buyer is more concerned about a confrontational response when submitting a public rating for a male – but not for a female – freelancer, they would inflate the public rating of a male freelancer to a greater degree than that of a female freelancer. We conduct two tests to examine whether gender stereotypes cause greater rating inflation for male freelancers.

First, we rely on the fact that gender stereotypes are typically more pronounced in geographic regions with lower gender equality (Eagly and Mitchell, 2004). We use measures of gender equality from the United Nations Development Report and expect that gender differences in public ratings, after controlling for private ratings, should be more pronounced for buyers from countries with lower gender equality. Our data confirm that this is the case.

Second, we propose that within a country, the extent of gender stereotypes varies with how active women are in the labor force. We use US data on female labor force participation to proxy the extent of gender stereotypes across states (Akhtar et al., 2023). We expect gender disparity in ratings to be less pronounced for buyers from states with higher rates of female labor force participation. Indeed, conditional on private ratings, gender differences in public ratings decrease as female labor force participation increases.

Our results demonstrate that while reputation systems are designed to reduce information asymmetry to help buyers make more informed purchase decisions, ratings that enter these

systems may reflect stereotypes by those submitting them. As a result, at least in online labor markets where the buyer is typically aware of the freelancer's gender, ratings may be less effective than expected in leveling the playing field between male and female freelancers and in helping buyers make informed purchase decisions.

Our findings have implications for freelancers and platforms. Ratings that disadvantage women relative to men may affect future purchases in at least three ways. First, buyers who rely on ratings to make purchase decisions may be more likely to choose a service from a male than a female freelancer if men are rated more highly than women. Second, since online ratings are often an input into ranking algorithms and an item's rank in search results directly impacts purchase probabilities, platforms relying on public ratings may inadvertently disadvantage female freelancers. Our back of the envelope calculation shows that gender disparity in ratings against female freelancers would lower their ranking in search results and reduce their discoverability by 12.36%, and consequently impact their sales. Third, gender disparities in online ratings may impact the long-term retention of female freelancers and the diversity of a platform's offering.

Fortunately, platforms can counter the negative effects of gender disparity in ratings in at least two ways. Platforms can solicit private ratings and ensure that ranking algorithms take private ratings into consideration, instead of public ratings. Platforms can reduce buyers' concerns about confrontation with freelancers by anonymizing buyers' identity in public ratings (Chen et al., 2023). For policy makers, our results emphasize that differences in ratings between social groups may not be a result of platform policy, but a reflection of stereotypical beliefs prevalent in society that cannot be addressed through platform regulation. On a broader level, our findings contribute to the evaluation of the trade-offs around data collection by firms and privacy policies. The platform in our context was able to identify the differential evaluation for male and female freelancers as they collected data on gender. Restrictions on data collection may hamper a platform's ability to uncover uneven outcomes among platform participants and develop measures to mitigate these.

2 Related Literature

Our paper relates to three streams of literature. First, our paper extends the literature on the gender gap across various empirical contexts such as academia (Card et al., 2020), entrepreneurship (Brooks et al., 2014) and offline labor markets (Correll et al., 2007) in that we demonstrate that after controlling for a proxy of buyer satisfaction, publicly available evaluations in online labor markets favor men. In contrast to Chan and Wang (2018) who in the context of an online freelance platform focus on buyers' selection of freelancers at the purchase stage, our analysis focuses on post-purchase evaluations, a dimension that has received little attention.

Second, our paper extends the literature on inequities in online reviews. Prior research has studied buyers' reluctance to leave negative reviews (Dellarocas and Wood, 2008), concerns about retaliation from sellers (Bolton et al., 2013), and discrimination against minority groups (Teng et al., 2023). Our study provides evidence for a gender gap in ratings of freelancers and suggests gender stereotypes as the underlying reason for this gender disparity.

Last, our paper contributes to research that identifies gender disparity in online markets. Lambrecht and Tucker (2019) reveal that women may be exposed less to STEM ads than men. Botelho and Gertsberg (2022) find that users tend to be more critical when reviewing restaurants on Yelp after being served by female servers. Chan and Wang (2018) identify reverse discrimination in an online labor marketplace such that women are more likely to be hired relative to men. Bohren et al. (2019) and Proserpio et al. (2021) document that users are more hostile towards women in online settings. We not only show that buyers inflate ratings of male freelancers significantly more than of female freelancers, but also provide evidence that this disparity is a result of buyers' gender stereotypes.

3 Empirical Setting and Data

3.1 Empirical Setting

Our data come from a major freelance platform where freelancers across the world offer services that are completed remotely. Services offered span multiple categories including logo design, language translation, and programming. When buyers search on the platform, they are presented with a list of search results. These search results show an icon representing the service, the freelancer's username, as well as the average rating by past buyers of the service. By clicking on an individual service listing, buyers can see details regarding the freelancer and the service. The description of the freelancer includes the name of the freelancer, country, date of registration, and a profile picture that also informs buyers of the freelancer's gender.[1] Service details include a description of the service, delivery time, and public ratings and reviews submitted. Users can place an order on the service page.

After an order is delivered, buyers are prompted via email to provide a public rating and written feedback that is then displayed on the service page along with the buyer's username. Freelancers can respond to buyers' feedback. Figure 1 shows examples of buyers' reviews and freelancers' responses illustrating confrontational responses by freelancers.[2]

After receiving a public rating, the platform prompts buyers to leave private feedback that is shared neither with the freelancer nor the public. The platform believes buyers to be more frank and truthful in private as buyers do not need to worry about potential confrontation with the freelancer or hurting the future prospects of the freelancer (Filippas et al., 2022).

[1] We conducted a survey on Prolific to assess whether users can accurately infer the gender of freelancers based on a service's page. Participants correctly inferred the gender in 94.4% of cases. See Online Appendix A for details regarding the survey.

[2] Since we cannot reveal the identify of the online freelance platform, the examples are mock-ups of reviews similar to those on the platform.

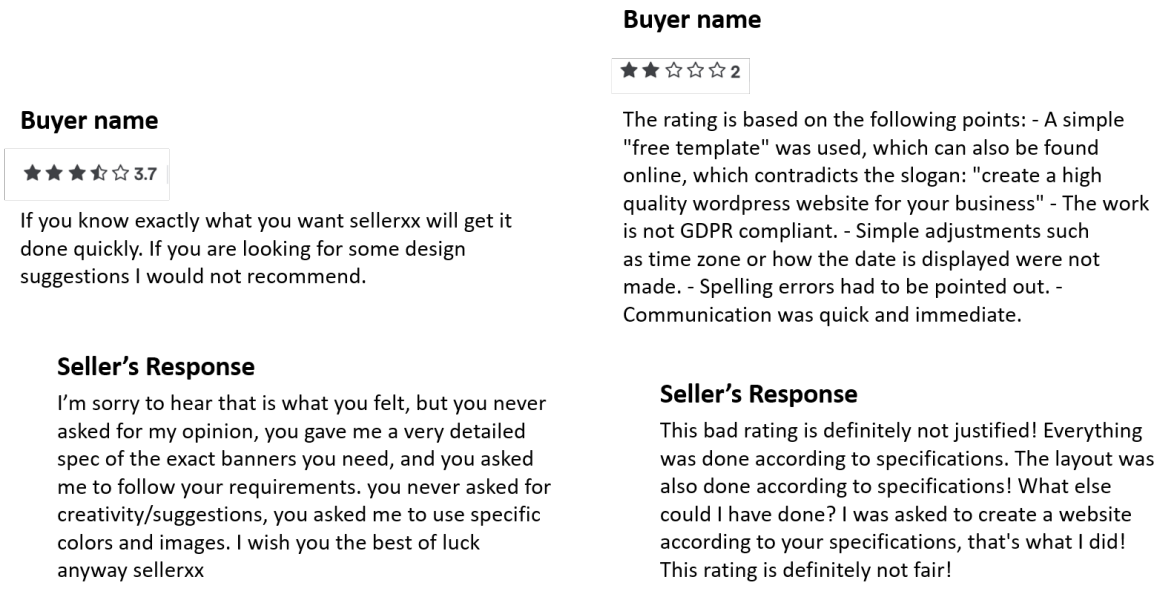


Figure 1: Mock-ups of freelancers' confrontational response to public feedback

3.2 Data

We obtain three data sets. The first data set records all purchases made and completed between January 2018 and October 2019. For each order, we observe the freelancer, the buyer, the date the order was placed, the service category, order price, delivery time, and the buyer's public and private ratings for this service. The second data set captures the country of the buyer, the state of the buyer if the buyer is from the United States, and when the buyer registered on the platform. The final data set captures a freelancer's country, registration date on the platform, and gender if the freelancer disclosed this information. The data also records for each day a freelancer's 'status' level on the platform, an aggregate measure of experience and quality of a freelancer based on a freelancer's tenure on the platform, total orders, earnings, and average rating. Out of 1,047,126 orders where we observe both freelancer gender and public rating, buyers gave a private rating for 319,857 orders which is the focus of our analysis.

To protect the confidentiality of the platform, we scale public and private ratings to be in the range of 1 to 5. On average, relative to female freelancers, male freelancers receive

a significantly higher public rating (4.92 vs. 4.91, see Table 1), marginally higher private rating (4.85 vs. 4.846), have lower order price ($31.85 vs. $32.77), have quicker delivery time (3.28 vs. 3.36 days), are more experienced on the platform (693.70 vs. 598.47 days), and attract more experienced buyers (649.86 vs. 630.90 days). In addition, the distribution of orders across freelancer status levels, where 1 represents the lowest and 4 the highest status level, differs between male and female freelancers. Table C1 in the Online Appendix C shows descriptive statistics across the full data set, rather than split by freelancer's gender.

Table 1: Summary statistics based on gender of freelancers

	Male		Female		
	Mean	SD	Mean	SD	p-value
Public rating	4.915	0.390	4.909	0.396	0.000
Private rating	4.850	0.547	4.846	0.552	0.068
Order price (USD)	31.850	82.009	32.766	62.539	0.000
Delivery time (days)	3.276	4.787	3.356	4.414	0.000
Freelancer tenure (days)	693.698	599.571	598.474	604.562	0.000
Buyer tenure (days)	649.855	704.985	630.900	701.452	0.000
Seller Status					
If(Status level 1)	0.373	0.484	0.400	0.490	0.000
If(Status level 2)	0.208	0.406	0.207	0.405	0.383
If(Status level 3)	0.364	0.481	0.328	0.469	0.000
If(Status level 4)	0.054	0.225	0.065	0.247	0.000

Notes: Number of orders for male freelancers is 238,363 and for female freelancers is 81,494.

4 Effect of Freelancer's Gender on Ratings

4.1 Initial Evidence

Figure 2 demonstrates that the average public rating for female freelancers is significantly lower than that for male freelancers (4.909 vs 4.915, $p < 0.01$). This pattern could be a result of differences in buyer satisfaction with the service provided by male and female freelancers. To identify the extent to which differences in ratings are due to freelancer gender rather than underlying buyer satisfaction, we need to account for buyer satisfaction.

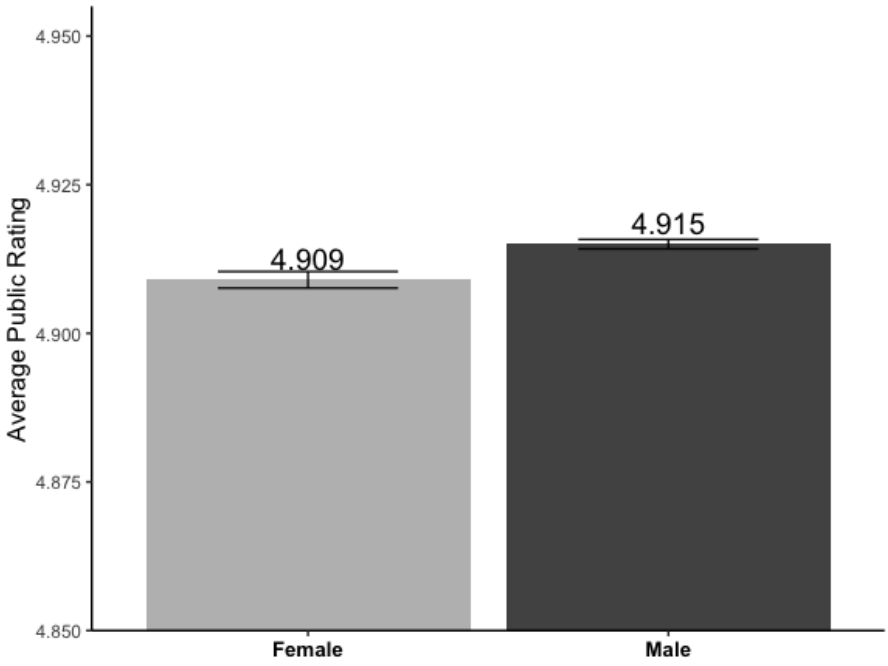


Figure 2: Public ratings for female vs male freelancers

On online freelance platforms buyers are hesitant to give low public ratings because of a fear of confrontation with an angry freelancer who may respond in an aggressive manner (Filippas et al., 2022). Buyers may further worry about hurting a freelancer's future prospects. However, since private ratings are neither shared with the freelancer nor available to other buyers, they are unlikely to be influenced by these issues and a more accurate measure of buyer satisfaction (Filippas et al., 2022).

We find that buyers indeed inflate their public ratings relative to private ratings (4.849 vs 4.913, $p < 0.1$, Figure 3). While the average private rating for female freelancers is marginally lower than that of male freelancers (4.846 vs 4.850, $p < 0.1$, Figure 4), this difference is less pronounced relative to the difference in their public ratings (Figure 2). Since the difference between public and private ratings between male and female freelancers could be driven by factors correlated with a freelancer's gender, such as the price of an order, category of the service, or freelancer's broader experience as measured by their status, we estimate a specification that controls for such factors.

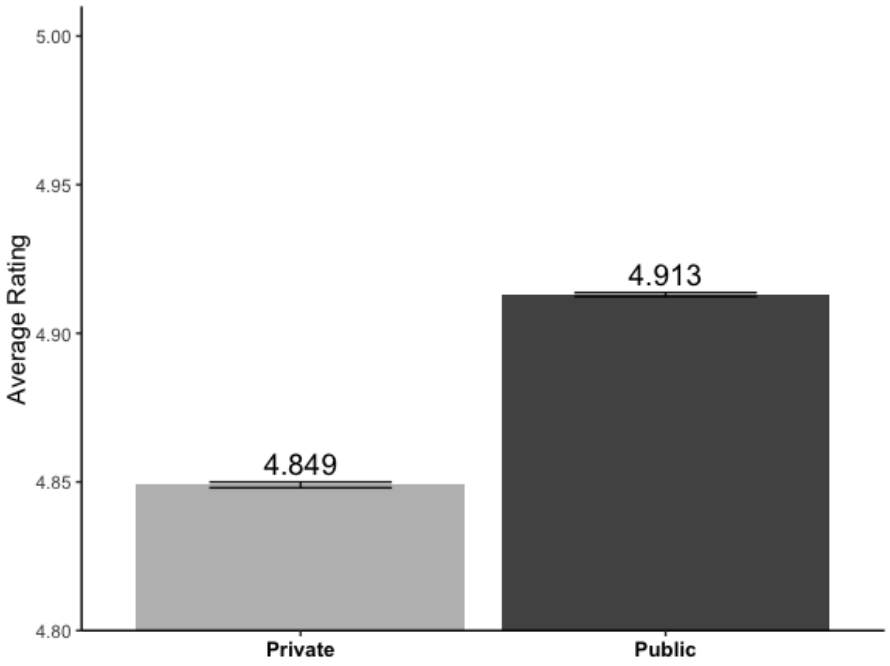


Figure 3: Private ratings vs public ratings

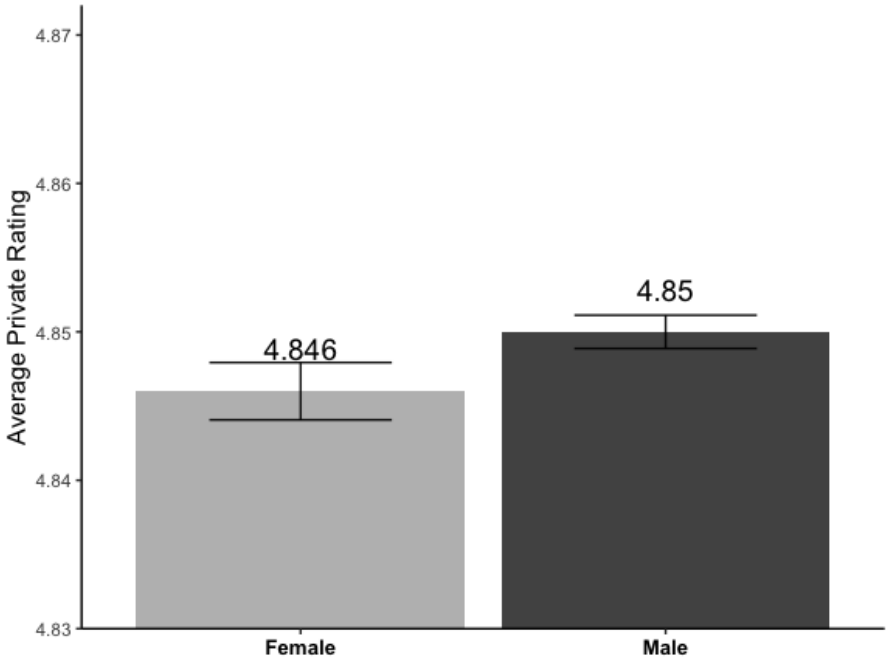


Figure 4: Private ratings for female vs male freelancers

4.2 Empirical Specification and Results

To examine gender disparity in public ratings, conditional on buyer satisfaction, we estimate the following specification:

$$\text{Public Rating}_{ijst} = \beta_0 + \beta_1 \text{Male}_s + \beta_2 \text{Private Rating}_{ijst} + \alpha O_i + \delta S_{st} + \gamma B_{jt} + \tau_t + \lambda_t + \epsilon_{ijst} \quad (1)$$

where the dependent variable Public Rating$_{ijst}$ captures the public rating of order i placed by buyer j from freelancer s at time t. Male$_s$ indicates whether freelancer s is male. Private Rating$_{ijst}$ controls for buyer j's satisfaction with order i from freelancer s at time t. O_i controls for order characteristics such as the price, category, and delivery time. Since previous research has shown that high status and more experienced candidates might be evaluated more favorably relative to low status and less experienced candidates (Bowles and Gelfand, 2010), S_{st} controls for a freelancer's status and a freelancer's tenure since registration on the platform. In addition, B_{jt} controls for buyer j's tenure since registration on the platform. τ_t are month-year and λ_t are day-of-the-week fixed effects.

Conditional on a freelancer's private rating, an OLS estimation confirms that male freelancers receive higher public ratings than female freelancers (Table 2, Column (1)). Our results hold if, instead of controlling for the private rating, we use the difference between the public and private rating (Column (2)) or the ratio between the public and private rating as the dependent variable (Column (3)).

4.3 Implications for Freelancers

The estimated disparity in Table 2, Column (1) in ratings translates into a 0.07% (0.0035 stars) higher rating for male relative to female freelancers. Though this difference might appear small at first sight, there are at least three reasons why such small differences in public ratings can significantly impact freelancers.

First, public ratings are crucial inputs into platforms' ranking algorithms, which deter-

Table 2: Gender disparity in public ratings

	(1) Public rating	(2) (Public rating - Private rating)	(3) Public rating / Private rating
Male	0.0035***	0.0038**	0.0019**
	(0.0012)	(0.0016)	(0.0009)
Private rating	0.4990***		
	(0.0039)		
Order controls	Yes	Yes	Yes
Freelancer controls	Yes	Yes	Yes
Buyer tenure	Yes	Yes	Yes
Month-year F.E.	Yes	Yes	Yes
Day of the week F.E.	Yes	Yes	Yes
Observations	319,857	319,857	319,857
R sq.	0.4931	0.0028	0.0019

Notes: Order controls include price, category, and log(delivery time). Freelancer controls include freelancer status and tenure. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

mine the order in which services are displayed following a search. Given the large choice of items available on platforms, potential buyers typically examine only a fraction of the available items, a lower public rating can lead to an item being ranked less prominently, reducing the likelihood that buyers will discover a service.

We examine the impact of a drop in average rating by 0.0035 on service discoverability. We rely on a separate dataset from the same platform that records results for over 800,000 randomly drawn searches. For each item listed in a set of search results, the data record the average star rating and the row in which it was displayed in the search results. We compute the average star rating per row and the difference between the average star ratings across adjacent rows. We also compute the likelihood of a row of search results being displayed to a user, and so an item receiving an impression. Table 3 demonstrates that on average across all search results, the difference between the average star rating in rows three and four was 0.0027 and that an item being displayed in row four instead of row three receives 12.36% fewer impressions. A drop in average star rating of 0.0035 thus implies that a service otherwise displayed in row three is now displayed in row four and, as a result will receive 12.36% fewer impressions. Assuming a constant conversion probability this translates into a

12.36% decrease in a freelancer's revenues.

Table 3: Effect on discoverability

	Δ average rating	p-value	Δ (-) impressions	Obs.
Row 1-2	0.0074	0.00	15.95%	5,117,530
Row 2-3	0.0098	0.00	17.34%	4,150,995
Row 3-4	0.0027	0.00	12.36%	3,257,943
Row 4-5	0.0012	0.00	8.73%	2,679,631
Row 5-6	0.0108	0.00	6.89%	2,145,086

Second, on many online platforms, ratings are strongly skewed towards 5-stars and any rating below 5 stars is often perceived as negative (Athey et al., 2019). Consequently, even small deviations from the highest possible rating can considerably affect future purchases. This likewise holds for the focal freelance platform, where the average public rating is 4.91 and the median is 5 (Table C1). In fact, 65.6% of freelancers have an average rating of 5-stars. Table 4 explores the relationship between a freelancer's gender and the probability of receiving a 5-star public rating, conditional on their private rating. Column (1) shows for a logit specification that after controlling for private ratings, male freelancers are significantly more likely to receive a 5-star public rating than their female counterparts. Column (2) confirms the robustness of this result using a probit model. Given the critical role of 5-star ratings, slight deviations such as the drop by 0.0035 stars we observe for female relative to male freelancers may notably impact a freelancer's revenue.

Finally, the effect of a single less favorable rating goes beyond its impact on the average rating displayed on the platform. Vana and Lambrecht (2021) demonstrate that the most prominently visible review can directly and strongly influence future purchases, above and beyond its effect on the average rating. Furthermore, Cabral and Hortacsu (2010) show that an individual less favorable rating can accelerate the receipt of subsequent negative ratings for a seller, potentially leading to earlier market exit. Additionally, Teng et al. (2023) reveal that a biased rating could have spillover effects for users who do not discriminate and exacerbate discrimination among those who do, thereby increasing inequity. Consequently, a single less favorable public rating towards female freelancers, as documented in our study,

Table 4: Gender disparity in getting 5-star public rating

	(1) Logit	(2) Probit
Male	0.0790***	0.0379***
	(0.0184)	(0.0089)
Private rating	2.1368***	1.0685***
	(0.0171)	(0.0084)
Order controls	Yes	Yes
Freelancer controls	Yes	Yes
Buyer tenure	Yes	Yes
Month-year F.E.	Yes	Yes
Day of the week F.E.	Yes	Yes
Observations	319,848	319,848
Log likelihood	-60411.81	-60615.09

Notes: Dependent variable is an indicator for whether the public rating is 5-star (or not). Order controls include price, category, and log(delivery time). Freelancer controls include freelancer status and tenure. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

may have a prolonged negative impact on the freelancers' revenue.

4.4 Potential Selection Concerns

Our analysis focuses on orders where buyers gave both public and private ratings and so conditions on buyers submitting a public rating.[3] Given that we examine the relationship of freelancers' gender to the rating they receive, this focus on orders for which buyers submit a public rating is appropriate.

Our analysis further conditions on buyers leaving a private in addition to their public rating. A potential concern could be that buyers self-select into giving a private rating. For instance, it is possible that buyers are more likely to leave a private rating after inflating their public rating and use private ratings to correct for this. If this were the case, since our analysis focuses on orders where buyers give both a public and a private rating, the gender disparity we detect in public ratings might only apply to instances when a private rating is

[3]As for any study of reviews, our study is also conditional on buyers purchasing the service from the specific freelancer. Unlike Chan and Wang (2018), we do not focus on the choice of the freelancer but the evaluation of this freelancer after the task completion.

submitted rather than the full population of public ratings.

To check the robustness of our results, we focus on all orders where buyers gave public ratings, regardless of whether they gave a private rating or not, and the gender of the freelancer is available. This means the data include 319,857 orders with private ratings and 727,269 orders without private ratings. We consider an extreme hypothetical scenario where for all buyers who did not leave a private rating the value of their public rating mirrors their level of satisfaction and replace all missing private ratings with their corresponding public ratings. We continue to find a significant gender disparity in public ratings under this scenario, though as one would expect the effect size is lower (Table 5).

Table 5: Robustness: Gender disparity in public ratings

	(1)	(2)	(3)
	Public rating	(Public rating - Private rating)	Public rating / Private rating
Male	0.0010**	0.0013**	0.0006**
	(0.0005)	(0.0005)	(0.0003)
Private rating	0.8319***		
	(0.0017)		
Order controls	Yes	Yes	Yes
Freelancer controls	Yes	Yes	Yes
Buyer tenure	Yes	Yes	Yes
Month-year F.E.	Yes	Yes	Yes
Day of the week F.E.	Yes	Yes	Yes
Observations	1,047,126	1,047,126	1,047,126
R sq.	0.8289	0.0064	0.0049

Notes: Order controls include price, category and log(delivery time). Freelancer controls include freelancer status and tenure. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5 Why Do Female Freelancers Receive Lower Public Ratings?

We propose that gender disparity in public ratings are a result of gender stereotypes. On digital platforms, Proserpio et al. (2021) have demonstrated that aggressive behavior can alter users' rating behavior. Men are considered to be more aggressive than women (Brover-

man et al., 1972; Ruble, 1983) and such stereotypes shape people's choices as people rely on stereotypes to form expectations about others' behavior (Ridgeway, 2001). Since on the focal freelance platform, public ratings are visible to freelancers and freelancers can respond to public ratings given by buyers, gender stereotypes may lead buyers to be more concerned about potential confrontation with male than with female freelancers in response to their public ratings. As a result, they may evaluate male freelancers relatively more positively than female freelancers in public ratings. We conduct two sets of analyses to explore this possibility.

5.1 Analysis Across Countries

We rely on findings that gender stereotypes are stronger in societies exhibiting lower gender equality (Eagly and Mitchell, 2004). We measure gender equality using the Human Development Index (HDI) from the United Nations Human Development Report (UNDP, 2019), a good measure for the level of gender equality in a country (see Online Appendix D for details and robustness).

The UN classifies countries with HDI below 0.55 to be low in human development (UNDP, 2019) so we use this cut-off for grouping countries into those with higher or lower human development. Countries with low levels of gender equality are concentrated in central Africa and south Asia (Figure 5). We classify 27 countries as having a low and 152 countries as having a high level of gender equality. If indeed gender differences in ratings stem from stereotypes, the pattern we document should be more pronounced for buyers from regions with lower gender equality.

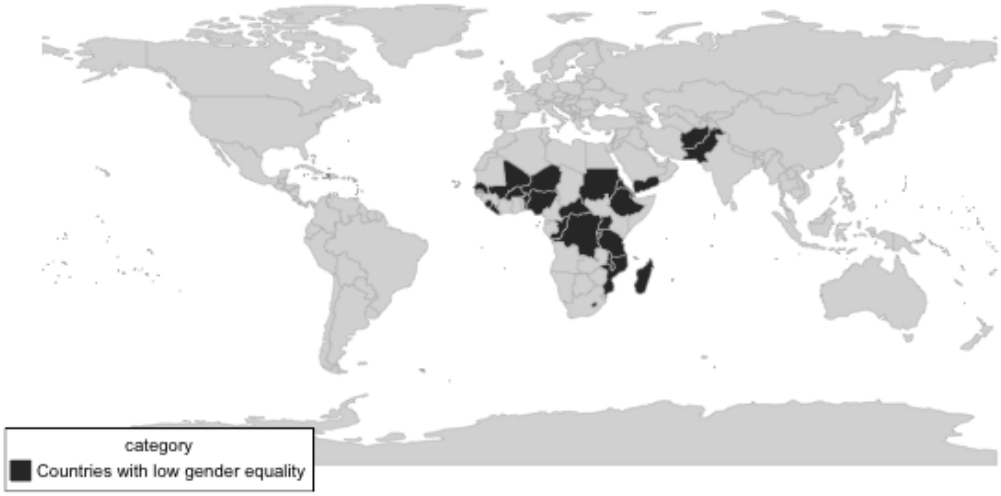


Figure 5: Countries with low gender equality (based on HDI)

To provide initial descriptive evidence, we compute for each order the difference between public and private ratings for male and female freelancers. We find a significantly higher reputation inflation in public ratings for male compared to female freelancers for buyers from countries with lower gender equality (0.112 vs 0.072, $p < 0.01$, Figure 6). We find no significant difference in reputation inflation between male and female freelancers when buyers are from countries with high gender equality. To provide further model-free evidence on gender disparity across buyers' countries, we examine the variation within freelancers. Specifically, Figure E1 in Online Appendix focuses on the 4,848 freelancers who serviced orders from both low and high equality countries. This figure reports the difference in public and private ratings for each order within the same freelancer, demonstrating consistent results.

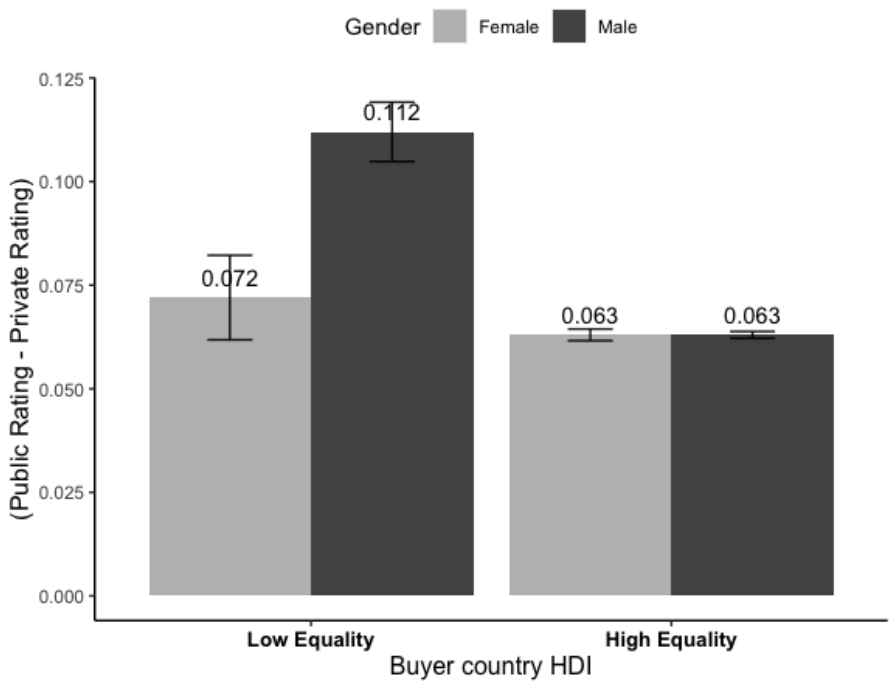


Figure 6: Difference in public ratings and private ratings (based on HDI)

Column (1) in Table 6 shows that, after controlling for private ratings, reputation inflation in public ratings in favor of male freelancers is significantly more pronounced for buyers from countries with lower gender equality. Columns (2) and (3) show that our results are consistent if we use as dependent variable the difference between the public and the private rating or, alternatively, their ratio.

5.2 Analysis Across US States

We next focus our analysis on the 52% of buyers in our data located in the United States and leverage variation in the female labor force participation rate across buyers' states. The decision of women to participate in the labor force is driven by multiple factors, including socio-economic constraints like pressure to conform to traditional gender roles of being a homemaker and caregiver. Regions that exhibit higher female participation in the labor force are considered to have a higher level of gender equality relative to regions with lower female participation in the labor force. Female participation in the labor force has been used

Table 6: Gender disparity in ratings based on gender equality in buyer country (HDI)

	(1)	(2)	(3)
	Public rating	Public rating- Private rating	Public rating/ Private rating
Male	0.0034***	0.0005	-0.0001
	(0.0012)	(0.0017)	(0.0009)
Buyer country: low equality (HDI)	0.0241***	-0.0042	-0.0066
	(0.0072)	(0.0103)	(0.0055)
Male × Buyer country: low equality (HDI)	0.0204**	0.0422***	0.0267***
	(0.0084)	(0.0126)	(0.0075)
Private rating	0.5001***		
	(0.0039)		
Order controls	Yes	Yes	Yes
Freelancer controls	Yes	Yes	Yes
Buyer tenure	Yes	Yes	Yes
Month-year F.E.	Yes	Yes	Yes
Day of the week F.E.	Yes	Yes	Yes
Seller country HDI	Yes	Yes	Yes
Observations	309,838	309,838	309,838
R sq.	0.4950	0.0038	0.0025

Notes: Order controls include price, cateogry, and log(delivery time). Freelancer controls include freelancer status and tenure. The number of observations are lower compared to Table 2 because data on HDI is only available for a subset of countries. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

as a measure of gender equality (Akhtar et al., 2023).

We use state-level data released by the Bureau of Labor Statistics in the US on labor force participation in 2022.[4] Figure 7 shows how female participation in the labor force varied across states. We expect gender disparity in ratings to be more pronounced for buyers from states with lower female labor force participation compared to buyers from states with higher female labor force participation, after controlling for male labor force participation. We find that gender disparity in public ratings reduces as female labor force participation increases (Column (1), Table 7). The results hold if we instead use the difference in public and private ratings (Column (2)) or the ratio between the two (Column (3)) as the dependent variable.

[4]Source: Bureau of Labor Statistics, U.S. Department of Labor, The Economics Daily, Labor force participation rate for women highest in the District of Columbia in 2022 at `https://www.bls.gov/opub/ted/2023/labor-force-participation-rate-for-women-highest-in-the-district-of-columbia-in-2022.htm` (visited September 15, 2023)

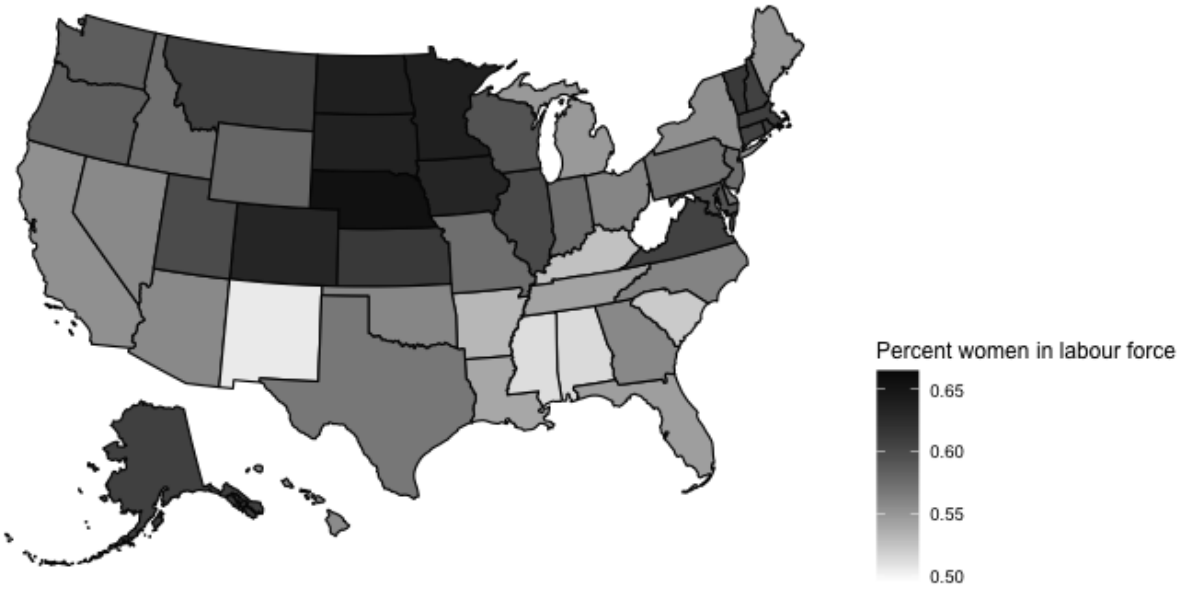


Figure 7: Female labor participation rate for states

Together, our results support that the difference in public ratings between male and female freelancers after controlling for their private ratings is a result of gender stereotyping. This insight is important since it demonstrates that gender disparity is driven by societal factors and not by platform policy. It further suggests that gender disparity in ratings is related to fear of confrontation. A platform could implement measures that reduce this fear of confrontation, for example by hiding a buyer's identity or by listing reviews in an order other than recency, which would make the identification of which buyer had submitted a rating more difficult.

Table 7: Gender disparity in ratings for buyers from the US

	(1)	(2)	(3)
	Public rating	Public rating - Private rating	Public rating/ Private rating
Male	0.0809**	0.1607***	0.0825***
	(0.0359)	(0.0497)	(0.0276)
%Women in labor force	0.1818***	0.3959***	0.2053***
	(0.0621)	(0.0881)	(0.0510)
Male × %Women in labor force	-0.1385**	-0.2782***	-0.1435***
	(0.0633)	(0.0879)	(0.0488)
Private rating	0.5175***		
	(0.0055)		
%Males in labor force	-0.0501	-0.1430***	-0.0889***
	(0.0346)	(0.0479)	(0.0266)
Order controls	Yes	Yes	Yes
Freelancer controls	Yes	Yes	Yes
Buyer tenure	Yes	Yes	Yes
Month-year F.E.	Yes	Yes	Yes
Day of the week F.E.	Yes	Yes	Yes
Observations	157,573	157,573	157,573
R sq.	0.5120	0.0025	0.0020

Notes: Order controls include price, category, and log(delivery time). Freelancer controls include freelancer status and tenure. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

6 Conclusion

Our research demonstrates that public ratings for male freelancers on online freelance platforms are inflated to a greater degree than those of female freelancers, disadvantaging female freelancers. Relying on variation across countries and US states, we demonstrate that male freelancers' ratings are more positive due to gender stereotypes: men are considered to be more confrontational and so more likely to respond aggressively to negative public ratings. To avoid such confrontation, buyers are more likely to inflate public ratings for male freelancers.

Our findings are important for freelancers and platforms. For freelancers, divergence in ratings can impact their earnings because lower ratings are visible to future buyers and lower ratings can impact the algorithmically-determined rank in search results. Such disparities may impact the long-term retention of female freelancers and the diversity on online freelance

platforms. To mitigate this disparity, platforms can solicit private ratings and ensure that ranking algorithms take these private ratings into consideration. Similarly, they can reduce disparity generated from stereotypical beliefs in ratings by anonymizing buyer identity. For policy makers, our results imply that differences in ratings between social groups reflect stereotypical beliefs in society rather than being a result of platform policy. More broadly, our findings contribute to a recent discussion on the trade-offs around privacy policies. If a platform is discouraged from collecting information such as gender, this may hamper its ability to uncover uneven outcomes among platform participants and develop measures to mitigate these.

There are limitations to our findings. First, we are not aware of the gender of the buyer and so cannot evaluate whether rating inflation varies by whether ratings have been submitted by men or women. Second, while our results are consistent with buyers avoiding confrontational behavior by male freelancers, we are unable to pin down in more detail the specific gender stereotypes driving our results.

Declarations

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

Abrahao, B., Parigi, P., Gupta, A., and Cook, K. S. (2017). Reputation offsets trust judgments based on social biases among airbnb users. *Proceedings of the National Academy of Sciences*, 114(37):9848–9853.

Akhtar, R., Masud, M. M., Jafrin, N., and Shahabudin, S. M. (2023). Economic growth, gender inequality, openness of trade, and female labour force participation: a nonlinear ardl approach. *Economic Change and Restructuring*, 56(3):1725–1752.

Athey, S., Castillo, J. C., and Chandar, B. (2019). Service quality in the gig economy: Empirical evidence about driving quality at uber. *Available at SSRN*.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10):3395–3436.

Bolton, G., Greiner, B., and Ockenfels, A. (2013). Engineering trust: Reciprocity in the production of reputation information. *Management science*, 59(2):265–285.

Botelho, T. L. and Gertsberg, M. (2022). The disciplining effect of status: Evaluator status awards and observed gender bias in evaluations. *Management Science*, 68(7):5311–5329.

Bowles, H. R. and Gelfand, M. (2010). Status and the evaluation of workplace deviance. *Psychological Science*, 21(1):49–54.

Brooks, A. W., Huang, L., Kearney, S. W., and Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences*, 111(12):4427–4431.

Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., and Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current appraisal 1. *Journal of Social issues*, 28(2):59–78.

Cabral, L. and Hortacsu, A. (2010). The dynamics of seller reputation: Evidence from ebay. *The Journal of Industrial Economics*, 58(1):54–78.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2020). Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics*, 135(1):269–327.

Chan, J. and Wang, J. (2018). Hiring preferences in online labor markets: Evidence of a

female hiring bias. *Management Science*, 64(7):2973–2994.

Chen, X., Zeng, X., and Zhang, C. (2023). Does concealing gender identity help women win the competition? An empirical investigation into online video games. *Marketing Science*, 42(3):551–568.

Correll, S. J., Benard, S., and Paik, I. (2007). Getting a job: Is there a motherhood penalty? *American journal of sociology*, 112(5):1297–1338.

Cui, R., Li, J., and Zhang, D. J. (2020). Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on airbnb. *Management Science*, 66(3):1071–1094.

Dellarocas, C. and Wood, C. A. (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management science*, 54(3):460–476.

Eagly, A. H. and Mitchell, A. A. (2004). Social role theory of sex differences and similarities: Implications for the sociopolitical attitudes of women and men.

Filippas, A., Horton, J. J., and Golden, J. M. (2022). Reputation inflation. *Marketing Science*, 41(4):733–745.

Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. (2017). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1914–1933.

Kanat, I., Hong, Y., and Raghu, T. (2018). Surviving in global online labor markets for it services: A geo-economic analysis. *Information systems research*, 29(4):893–909.

Lambrecht, A. and Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65(7):2966–2981.

Proserpio, D., Troncoso, I., and Valsesia, F. (2021). Does gender matter? The effect of management responses on reviewing behavior. *Marketing Science*, 40(6):1199–1213.

Ridgeway, C. (2001). Small-group interaction and gender. In Smelser, N. J. and Baltes, P. B.,

editors, *International Encyclopedia of the Social Behavioral Sciences*, pages 14185–14189. Pergamon, Oxford.

Ruble, T. L. (1983). Sex stereotypes: Issues of change in the 1970s. *Sex roles*, 9:397–402.

Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8:321–340.

Teng, F., Botelho, T., and Sudhir, K. (2023). Can customer ratings be discrimination amplifiers? Evidence from a gig-economy platform. *working paper*.

UNDP (2019). Technical notes: Calculating the human development indices—graphical presentation.

Vana, P. and Lambrecht, A. (2021). The effect of individual online reviews on purchase likelihood. *Marketing Science*, 40(4):708–730.

Online Appendix

A Identifying the Gender of Freelancers

We conducted a survey on the Prolific platform to determine whether users can identify the gender of freelancers based on the information available on a service's page on the focal platform.

We selected a total of 160 unique services across seven service categories offered on the focal platform, each provided by a different freelancer. In each service category, we selected the top eight rows from the search results. Services marked as promoted were excluded, as the platform did not promote services during our data collection period. We then recruited a sample of 48 participants located in the United States via Prolific who accessed the platform through a desktop. We provided each participant with links to ten freelancers' service pages.

For each freelancer, we asked participants to determine based on the information available on the service page (that includes the freelancer's username and profile picture) whether this freelancer is a man, a woman, or whether they were unable to determine whether the freelancer is a man or a woman. Each freelancer was classified by at least three survey respondents. Further, one of the authors classified each of the 160 freelancers based on the information available on the service page.

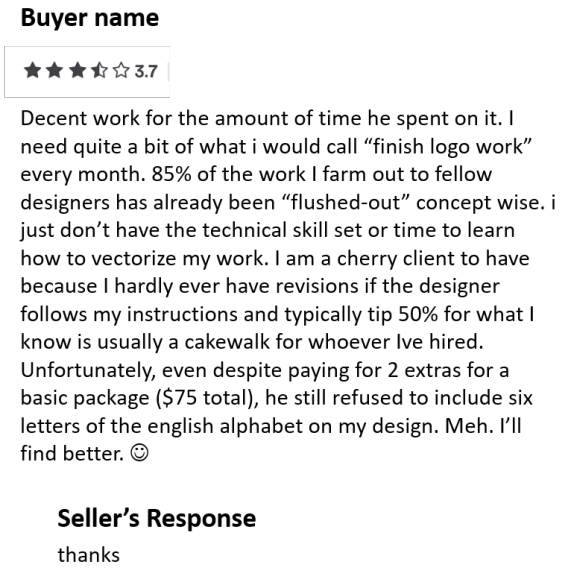
Out of a total of 480 classifications, 453 agree with our assessment i.e. 94.4% agreement. These patterns confirm that users of the platform can reliably infer the gender of freelancers based on the information provided on services' pages.

Further details on the survey are available from the authors upon request.

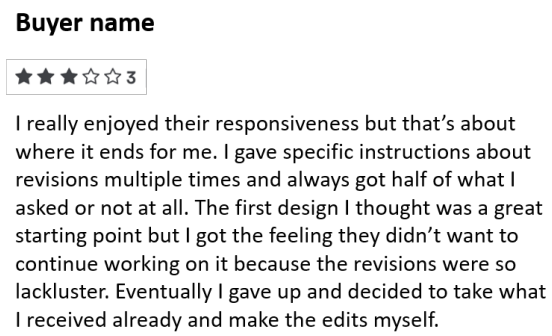
B Freelancer Response to Buyer Ratings

Figure B1 shows examples of freelancer responses to public ratings by buyers on the focal freelance platform. Figures B1 (c) and (d) show examples where the freelancers chose not

to confront the buyer for a low rating.



(a) Non-confrontational response



(b) No response

Figure B1: Mock-ups of freelancers’ response to public feedback

C Summary Statistics

Table C1: Summary statistics

	N	Mean	SD	Median	Min	Max
Public rating	319,857	4.91	0.39	5	1	5
Private rating	319,857	4.85	0.55	5	1	5
Order price (USD)	319,857	32.08	77.52	15	5.00	22,118
Delivery time (days)	319,857	3.30	4.70	2	1.00	337
Freelancer tenure (days)	319,857	669.44	602.28	467	0.00	3,323
Buyer tenure (days)	319,857	645.02	704.13	389	0.00	3,339
If freelancer male	319,857	0.74	0.44	1	0.00	1
Seller Status						
If(Status level 1)	319,857	0.38	0.48	0	0.00	1
If(Status level 2)	319,857	0.21	0.40	0	0.00	1
If(Status level 3)	319,857	0.36	0.48	0	0.00	1
If(Status level 4)	319,857	0.06	0.23	0	0.00	1

D Analysis Across Countries Using GII

HDI is a composite indicator of education, health and income and regarded as a good measure for the level of gender equality in a country. We focus on HDI instead of on an alternative measure, the Gender Inequality Index (GII) as HDI is strongly correlated with GII (-0.92) but the latter is available for fewer countries. The HDI lies on a scale of 0 to 1 where a higher score indicates higher human development and greater gender equality.

As an alternate measure, we further use the Gender Inequality Index (GII) from the United Nations Human Development Report (UNDP, 2019). The GII is a composite indicator of inequality in countries along dimensions such as education, health, and income. GII lies in the range of 0-1 such that a higher GII score implies a higher level of gender inequality within a country. We classify countries with a GII above 0.5 as countries with a low level of

gender equality. In our data, we classify 36 countries as having a low and 121 countries as having a high level of gender equality. While the GII measures more directly the construct of interest, it is available for fewer countries.

The results from OLS in Table D2 demonstrate throughout that the results we obtained when measuring gender equality using the HDI are robust to instead using the GII which measures gender equality more directly.

Table D2: Gender disparity in ratings based on gender equality in buyer country

	(1)	(2)	(3)
	Public rating	Public rating- Private rating	Public rating/ Private rating
Male	0.0031***	0.0004	-0.0003
	(0.0012)	(0.0017)	(0.0009)
Buyer country: low equality (GII)	0.0363***	0.0218**	0.0105
	(0.0065)	(0.0098)	(0.0065)
Male × Buyer country: low equality (GII)	0.0160**	0.0302***	0.0197**
	(0.0075)	(0.0116)	(0.0079)
Private rating	0.5006***		
	(0.0039)		
Order controls	Yes	Yes	Yes
Freelancer controls	Yes	Yes	Yes
Buyer tenure	Yes	Yes	Yes
Month-year F.E.	Yes	Yes	Yes
Day of the week F.E.	Yes	Yes	Yes
Seller country GII	Yes	Yes	Yes
Observations	309,085	309,085	309,085
R sq.	0.4957	0.0040	0.0028

Notes: Order controls include price and log(delivery time). Freelancer controls include freelancer status and tenure. The number of observations are lower compared to Table 2 because data on GII is available for a subset of countries. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

E Gender Disparity in Public Ratings Across Countries Within Freelancer

We focus on the 4,848 freelancers in our data who serviced orders from buyers from both countries with low gender equality and from countries with high gender equality. We then compute the difference between public and private ratings for each order within a freelancer. Figure E1 plots this difference separately for whether the rating came from a buyer in a country with low or with high gender equality, based on the HDI. It demonstrates that inflation in public relative to private ratings is higher for the male than the female freelancer if the buyer comes from a country with low gender equality (0.024 vs -0.002, $p < 0.05$).

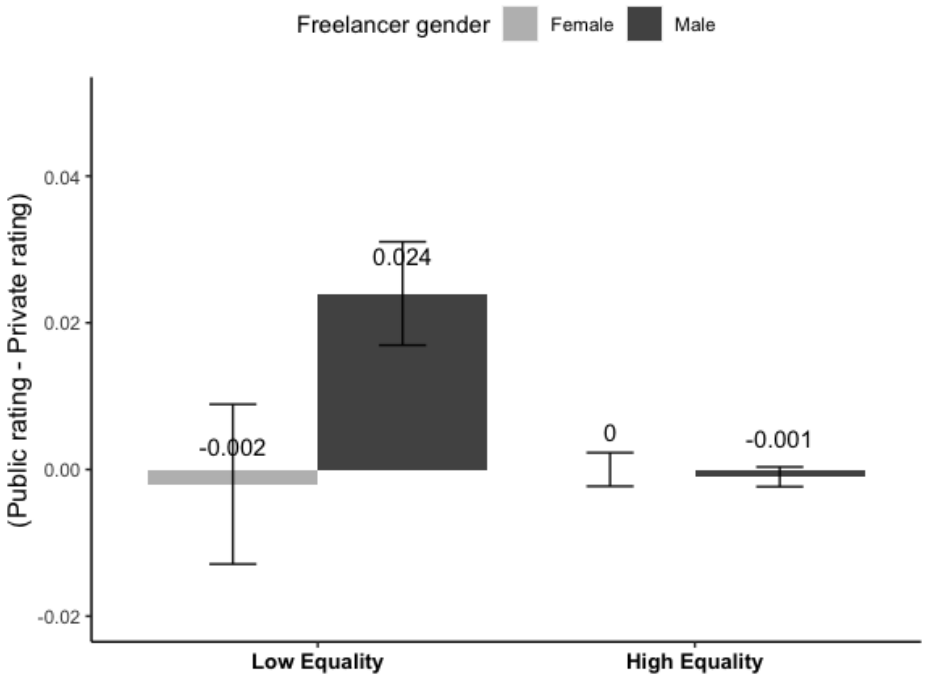


Figure E1: Difference in public ratings and private ratings within freelancer (based on HDI)