ONLINE APPENDIX: NOT FOR PUBLICATION

Appendices

A Additional Tables and Figures

Table A1: Composition across Gender and Sectors, Excluding Family-owned Firms

	Log(L)		Frac. female emp.	
	1998	2005	1998	2005
	(1)	(2)	(3)	(4)
Panel A: Without industry fixed effects				
Female	0.0484	-0.0473***	0.326***	0.331***
	(0.0449)	(0.00773)	(0.0225)	(0.0115)
Formal	2.200***	2.475***	0.120***	0.125***
	(0.0348)	(0.0334)	(0.00915)	(0.00988)
Female × Formal	0.0149	0.229***	-0.184***	-0.151***
	(0.0853)	(0.0444)	(0.0289)	(0.0162)
R^2	0.226	0.305	0.231	0.210
Panel B: With industry fixed effects				
Female	-0.00646	-0.0770***	0.264***	0.266***
	(0.0279)	(0.00811)	(0.0169)	(0.00811)
Formal	1.889***	2.306***	0.0763***	0.0941***
	(0.0303)	(0.0365)	(0.00757)	(0.00855)
Female × Formal	0.0815	0.250***	-0.145***	-0.116***
	(0.0632)	(0.0480)	(0.0231)	(0.0139)
R^2	0.378	0.378	0.368	0.294
N	5.23m	9.88m	5.23m	9.88m
Male, Informal	1.192	1.059	0.0855	0.126
Firm controls	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes

Notes: The sample is restricted to firms that are not "family-owned". Family-owned firms are defined as those firms where more than half the employees are not hired on wage contracts. Female and Formal are dummy variables that take the value 1 if the firm is female-owned or if it is in the formal sector and 0 otherwise. All regressions control for district fixed effects, along with whether the firm has access to power, dummy variables for different forms of financial access, and whether the firm is in the rural or urban area. Industry fixed effects are at the four-digit level using the NIC98 for 1998 and NIC04 for 2005. Standard errors are clustered at the district level.

Table A2: Derivatives of Moments to Parameters

Moment	A_I	A_F	τ_I^f	τ_F^f	τ_I	τ_F	λ
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Sample from the 1998 Round of the Economic Census							
$R_{mI,j}/R_{mI,Serv.}$	**0.67**	0.00	0.00	0.00	0.00	0.00	0.00
$R_{mF,j}/R_{mF,Serv.}$	0.00	**0.66**	0.00	0.00	0.00	0.00	0.00
$R_{fI,j}/R_{mI,j}$	-0.00	0.00	**-2.18**	0.00	0.00	0.00	0.00
$R_{fF,j}/R_{mF,j}$	0.00	0.00	0.00	**-2.25**	0.00	0.00	0.00
$\bar{l}_{fI,j}/\bar{l}_{mI,j}$	0.03	0.04	-0.83	0.09	**-1.34**	0.31	-0.04
$\bar{l}_{fF,j}/\bar{l}_{mF,j}$	-0.13	0.15	-0.23	-0.32	-0.38	**-1.27**	-2.46
$\bar{l}_{mI,j}/\bar{l}_{mF,j}$	0.25	-0.11	-0.00	-0.01	-0.03	-0.09	**3.42**
Panel B: Sample from the 2005 Round of the Economic Census							
$R_{mI,j}/R_{mI,Serv.}$	**0.67**	0.00	0.00	0.00	0.00	0.00	0.00
$R_{mF,j}/R_{mF,Serv.}$	0.00	**0.65**	0.00	0.00	0.00	0.00	0.00
$R_{fI,j}/R_{mI,j}$	0.00	0.00	**-2.19**	0.00	0.00	0.00	0.00
$R_{fF,j}/R_{mF,j}$	0.00	0.00	0.00	**-2.26**	0.00	0.00	0.00
$\bar{l}_{fI,j}/\bar{l}_{mI,j}$	0.01	0.03	-0.86	0.09	**-1.39**	**0.19**	-1.05
$\bar{l}_{fF,j}/\bar{l}_{mF,j}$	-0.02	0.01	-0.58	0.01	-0.99	-0.01	-0.30
$\bar{l}_{mI,j}/\bar{l}_{mF,j}$	0.20	-0.09	0.00	-0.00	-0.03	-0.01	**1.75**

Notes: This table reports the derivatives of each moment with respect to each parameter. Each row is a moment calculated from the model simulation. Each number in the table indexed by row R and column C, is the percent change in the moment in row R, when a parameter in column C is increased by 1 p.p. The largest value in each column is bold faced. Panel A (B) reports the results from the 1998 (2005) Round of the Economic Census. R_{gsj} and $\bar{l}_{gsj}$ are the ratio of female-male workers and the average size of a firm owned by an entrepreneur of gender g in sector s and industry j.

Table A3: Model Fit I

	Male		Female	
	Data	Model	Data	Model
	(1)	(2)	(3)	(4)
Panel A: Occupational choice of individuals				
1-LFP	0.43	0.43	0.70	0.69
	(0.04)	(0.04)	(0.08)	(0.08)
Frac. Wage Emp.	0.31	0.31	0.25	0.25
	(0.04)	(0.04)	(0.07)	(0.07)
Frac. Self Emp.	0.15	0.14	0.03	0.03
	(0.02)	(0.02)	(0.03)	(0.03)
Frac. Inf. Entrp.	0.11	0.11	0.02	0.02
	(0.01)	(0.01)	(0.01)	(0.01)
Frac. Formal Entrp.	0.001	0.001	0.000	0.000
	(0.0005)	(0.0005)	(0.0001)	(0.0001)
Panel B: Ratio of female-male workers in a firm				
Informal	0.95	0.95	1.07	1.06
	(0.06)	(0.06)	(0.08)	(0.08)
Formal	0.77	0.77	0.87	0.87
	(0.15)	(0.15)	(0.36)	(0.36)

Notes: Each row reports the average value across regions with the standard deviation in parentheses. Columns (1)-(2) report the moments for men, while (3)-(4) report those for women. Columns (1) and (3) report the moments in the Data, while (2) and (4) report their simulated counterparts from the Model. Panel A reports the allocation of men/women in the economy with the fraction of individuals who are (a) not in the labor force; (ii) in wage employment; (iii) informal entrepreneurship and (iv) formal entrepreneurship. Panel B reports the ratio of female to male workers in an informal and formal male-owned (Coumns 1-2) and female-owned firm (Columns 3-4).

Table A4: Model Fit II

	Male		Female	
	Data	Model	Data	Model
	(1)	(2)	(3)	(4)
Panel A: Ratio of average firm size				
$\bar{l}_{gI}/\bar{l}_{mI}$	1.00	1.00	1.06	1.04
	(0)	(0)	(0.18)	(0.17)
$\bar{l}_{gF}/\bar{l}_{mF}$	1.00	1.00	1.18	1.05
	(0)	(0)	(0.62)	(0.29)
$\bar{l}_{gF}/\bar{l}_{gI}$	22.69	28.70	26.15	28.66
	(9.39)	(7.55)	(20.64)	(8.99)
Panel B: Average firm size				
Informal	4.21	6.83	4.37	7.11
	(0.70)	(0.88)	(0.40)	(1.39)
Formal	95.05	193.54	113.38	199.02
	(41.85)	(45.90)	(87.40)	(59.45)
Panel C: Std. Deviation of firm size				
Informal	3.60	3.63	3.58	3.35
	(1.35)	(1.23)	(1.16)	(1.55)
Formal	184.70	191.89	160.68	200.95
	(108.70)	(92.96)	(172.76)	(102.24)

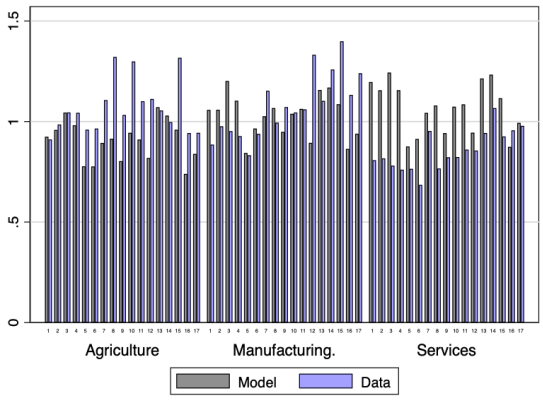
Notes: Each row reports the average value across regions with the standard deviation in parentheses. Columns (1)-(2) report the moments for men, while (3)-(4) report those for women. Columns (1) and (3) report the moments in the Data, while (2) and (4) report their simulated counterparts from the Model. Panel A reports the ratio of the average firm size for: (i) firms of gender g relative to male-owned firms in the informal sector; (ii) firms of gender g relative to male-owned firms in the formal sector and (iii) firms of gender g in the formal relative to the informal sector. Panel B reports the average firm-size in the informal and formal sector and Panel C reports the standard deviation for those firms.

Table A5: Model Fit III

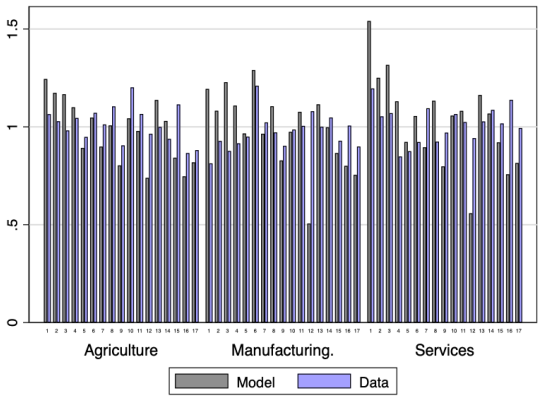
	Male		Female	
	Data	Model	Data	Model
	(1)	(2)	(3)	(4)
Panel A: Share of Firms in the Informal Sector				
Agriculture	0.15	0.16	0.25	0.15
	(0.08)	(0.09)	(0.22)	(0.17)
Manf.	0.26	0.27	0.33	0.22
	(0.06)	(0.07)	(0.16)	(0.11)
Services	0.59	0.56	0.43	0.62
	(0.07)	(0.08)	(0.17)	(0.14)
Panel B: Share of Firms in the Formal Sector				
Agriculture	0.06	0.06	0.15	0.24
	(0.05)	(0.05)	(0.14)	(0.2)
Manf.	0.58	0.59	0.35	0.37
	(0.12)	(0.11)	(0.12)	(0.17)
Services	0.37	0.35	0.50	0.38
	(0.11)	(0.1)	(0.12)	(0.16)

Notes: Each row reports the average value across regions with the standard deviation in parentheses. Columns (1)-(2) report the moments for men, while (3)-(4) report those for women. Columns (1) and (3) report the moments in the Data, while (2) and (4) report their simulated counterparts from the Model. Panel A reports the allocation of men/women in the economy with the fraction of individuals who are (a) not in the labor force; (ii) in wage employment; (iii) informal entrepreneurship and (iv) formal entrepreneurship. Panel B reports the ratio of female to male workers in an informal and formal male-owned (Coumns 1-2) and female-owned firm (Columns 3-4).

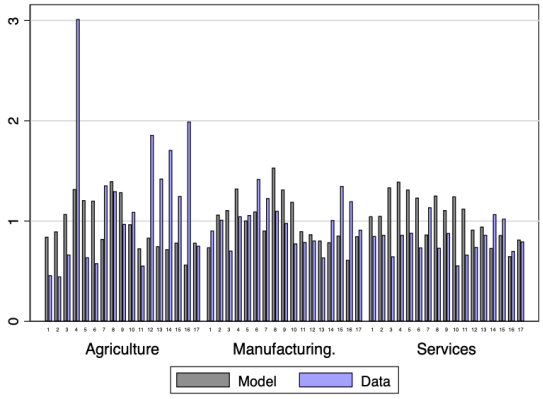
Figure A1: Model Fit: Average Firm Size



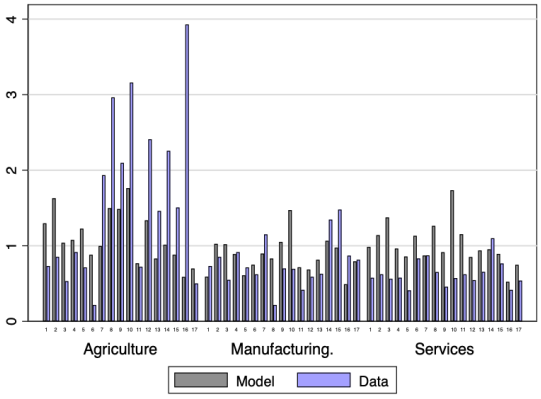
(a) Avg. Firm Size: Male, Informal



(b) Avg. Firm Size: Female, Informal



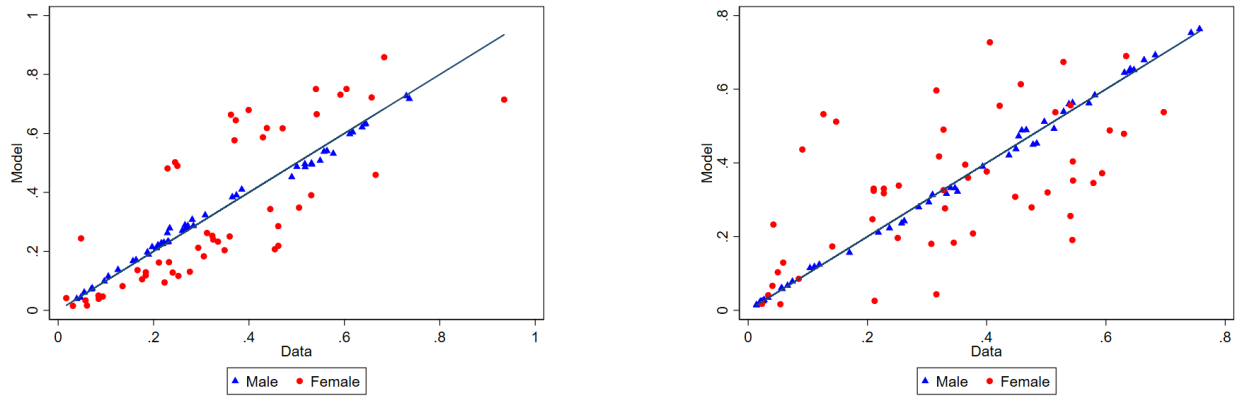
(c) Avg. Firm Size: Male, Formal



(d) Avg. Firm Size: Female, Formal

Notes: Figures (a)-(b) plot the distribution of hiring barriers faced by women entrepreneurs (relative to men) across regions and industries in the informal and formal sectors in 2005 i.e., $1 - \tau_{fs}$. Figures (c)-(d) plot the distribution of barriers faced by women entrepreneurs (relative to male entrepreneurs) in hiring female workers (relative to male workers) i.e., i.e., $1 - \tau_{fs}^{f}$.

Figure A2: Model Fit: Share of Firms in Each Industry



(a) Informal Sector

(b) Formal Sector

Notes: Figures (a)-(b) plot the distribution of hiring barriers faced by women entrepreneurs (relative to men) across regions and industries in the informal and formal sectors in 2005 i.e., $1 - \tau_{fs}$. Figures (c)-(d) plot the distribution of barriers faced by women entrepreneurs (relative to male entrepreneurs) in hirng female workers (relative to male workers) i.e., i.e., $1 - \tau_{fs}^{f}$.

B Mathematical Proofs

B.1 Incumbent Firm Decisions

The problem of a firm with productivity z in a sector s (dropping gender and industry for notational ease) is given by:

$$\max \; p_s z l^{\rho_s} - \left[w^m l^m + w^f l^f \right]$$

where $\{\rho_I, \rho_F\} = \{\lambda\rho, \rho\}$ and $\{p_I, p_F\} = \{p, (1-t)p\}$. Define:

$$w = \left[\sum_g A^g (w^g)^{(1-\gamma)} \right]^{\frac{1}{1-\gamma}}$$

We can rewrite the maximization problem as a two-step problem where in the first step, the firm chooses labor l to maximize profits: $\max p_s z l^{\rho_s} - wl/T$ and then minimizes expenditure on male and female workers, given this choice of l. Taking the FOC and solving we get:

$$l_I^*(z) = \left[\rho_s \times \frac{z}{w/p_s} \right]^{\frac{1}{1-\rho_s}}$$
$$\pi_I^*(z) = \frac{1-\rho_s}{\rho_s} \times w l_I^*(z)$$

Cost minimization in the second stage implies:

$$\min \; w^m l^m + w^f l^f$$
$$\text{s.t.} \left[\sum_g A^g (l^g)^{\frac{\gamma-1}{\gamma}} \right]^{\frac{\gamma}{\gamma-1}} = l_I^*$$

Taking the first order conditions and rearranging, we get:

$$w^g l^g(z) = A^g \left(\frac{w^g}{w} \right)^{1-\gamma} \times w l^*(z)$$

B.2 Penalty of Operating in the Informal Sector

An alternative way to present the model is to allow for a size-dependent penalty of operating in the informal sector. Let $\tau(l)$ be the penalty function such that $\tau(l) > 0$, $\tau'(l) < 0$ and $\tau(\infty) \to 0$. One can think of $t_I(l)$ as a per-unit size-dependent tax of operating in the informal sector, such that $\tau(l) = 1 - t_I(l)$. Accordingly, the maximization problem of the firm can be written as:

$$\max_l \tau(l) pzl^\rho - wl$$

Taking the first order condition and rearranging:

$$\left[\rho\tau(l) + l\tau'(l)\right] pzl^\rho = wl \tag{10}$$

Compared to the baseline model, we have:

$$\left[\widetilde{\rho} \times l^{\tilde{\rho}-\rho}\right] pzl^\rho = wl \tag{11}$$

Equations (10) and (11) are therefore connected through the $\tau(l)$ function, so that:

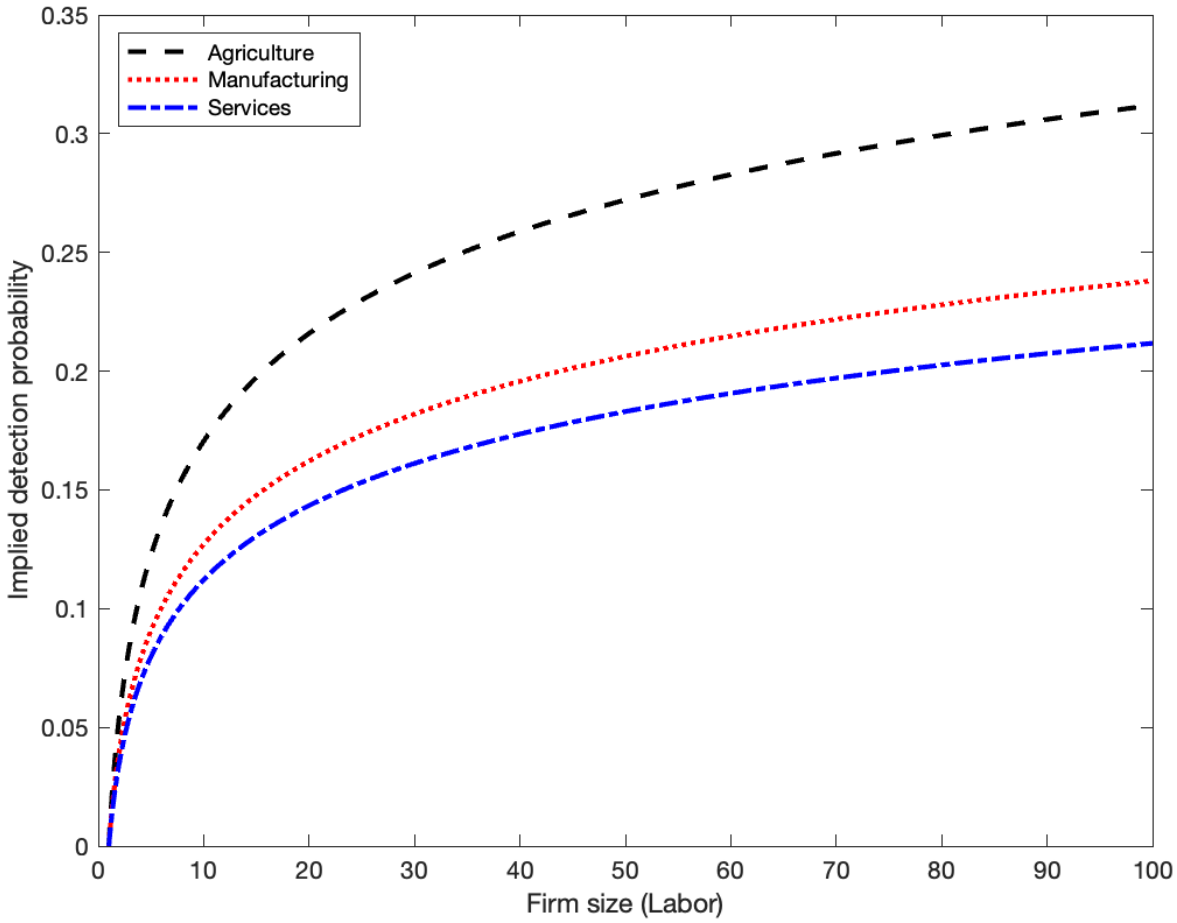
$$\rho\tau(l) + l\tau'(l) = \widetilde{\rho} \times l^{\tilde{\rho}-\rho} \tag{12}$$

This is a differential equation of the form $ay + xdy/dx = bx^c$, where $y = f(x)$. This has a general solution of the form $y = \frac{bx^c}{a+c} + \frac{k}{x^a}$ where k is an integration constant. Therefore the general solution to $\tau(l)$ is given by:

$$\tau(l) = \left[l^{\tilde{\rho}} + k\right] l^{-\rho} \tag{13}$$

To restrict $0 < \tau(l) < 1$, we assume $k = 0$ and plot $t_I(l) = 1 - \tau(l)$ in Figure B1.

Figure B1: Size-based Penalty Function



Notes: The above graph plots the size-based penalty function of operating in the informal sector as a function of firm size.

B.3 Allocation of entrepreneurs across industries

From Equations (2), (3), (5) and (6), the general form of the profit function and wage bill for a firm in sector s (dropping g for notational convenience) is given by:

$$b_{sj} \equiv \frac{w_{sj} l_{sj}}{T_{sj}} = \eta_{L,sj} \times \varepsilon^{\frac{1}{1-\rho_s}}$$

$$\pi_{sj} = \eta_{\pi,sj} \times \varepsilon^{\frac{1}{1-\rho_s}}$$

where:

$$\eta_{L,sj} = \frac{w_{sj}}{T_j}\left[\rho_s \frac{T_j}{w_{sj}/p_{sj}} \times x\right]^{\frac{1}{1-\rho_s}}$$

$$\eta_{\pi,sj} = \frac{1-\rho_s}{\rho_s} \times \eta_{L,sj}$$

Let $\widetilde{\theta}_s = \theta(1-\rho_s)$. Dropping s for notational ease, the distribution of π_j within a sector s will follow a Frechet distribution given by $\pi_j \sim Frechet(\widetilde{\theta}, \eta_{\pi,j})$ with a CDF given by:

$$F(\pi) = exp\left[-\left(\frac{\pi}{\eta_\pi}\right)^{-\widetilde{\theta}}\right]$$

Note that the share of firms in an industry k will be the probability that the profits in industry k are higher than in all other industries. This implies that:

$$\begin{aligned}
\varphi_k &= Pr(\pi_k = \max\{\pi_j\}_{\forall j}) \\
&= \int \prod_{j \neq k} F(\pi_k) \times dF(\pi_k) d\pi_k \\
&= \int \prod_{j \neq k} e^{-(\pi_k/\eta_{\pi,j})^{-\widetilde{\theta}}} \times e^{-(\pi_k/\eta_{\pi,k})^{-\widetilde{\theta}}} \times \widetilde{\theta}\eta_{\pi,k}^{\widetilde{\theta}} \times \pi_k^{-\widetilde{\theta}-1} d\pi_k \\
&= \int e^{-(\sum_j \eta_{\pi,j}^{\widetilde{\theta}})\pi_k^{-\widetilde{\theta}}} \times \widetilde{\theta}\eta_{\pi,k}^{\widetilde{\theta}} \times \pi_k^{-\widetilde{\theta}-1} dx \\
&= \frac{\eta_{\pi,k}^{\widetilde{\theta}}}{\sum \eta_{\pi,j}^{\widetilde{\theta}}} \times \int \underbrace{e^{-\sum \eta_{\pi,j}^{\widetilde{\theta}} \pi_k^{-\widetilde{\theta}}} \times \widetilde{\theta}(\sum \eta_{\pi,k}^{\widetilde{\theta}}) \pi_k^{-\widetilde{\theta}-1}}_{\text{Frechet distribution}} dx \\
&= \frac{\eta_{\pi,k}^{\widetilde{\theta}}}{\sum_j \eta_{\pi,j}^{\widetilde{\theta}}}
\end{aligned}$$

Substituting the values in the expression above, we have:

$$\begin{aligned}\eta_{\pi,j} &= \frac{1-\rho_s}{\rho_s} \times \frac{w_{sj}}{T_j}\left[\rho_s \frac{p_{sj}}{w_{sj}/T_j} \times x\right]^{\frac{1}{1-\rho_s}} \\ &= \left\{\frac{1-\rho_s}{\rho_s} \times (\rho_s x)^{\frac{1}{1-\rho_s}}\right\} \times \left[\frac{p_{sj}}{(w_{sj}/T_j)^{\rho_s}}\right]^{\frac{1}{1-\rho_s}} \\ \Rightarrow \frac{\eta_{\pi,j}^{\widetilde{\theta}}}{\sum_j \eta_{\pi,k}^{\widetilde{\theta}}} &= \frac{\left[\frac{p_{sj}}{(w_{sj}/T_j)^{\rho_s}}\right]^{\theta}}{\sum_k \left[\frac{p_{sk}}{(w_{sk}/T_k)^{\rho_s}}\right]^{\theta}}\end{aligned}$$

Note that since $\pi_k \sim Frechet(\widetilde{\theta}, \eta_{\pi,k})$, the distribution of maximum profits $\pi_j = \max\{\pi_k\}_j$ will also follow a Frechet distribution where $\pi_j \sim Frechet(\widetilde{\theta}, (\sum \eta_{\pi,k}^{\widetilde{\theta}})^{1/\widetilde{\theta}})$, so that:

$$\begin{aligned}E(\pi_j | \pi_j = \max\{\pi_k\}_{\forall k}) &= (\sum \eta_{\pi,k}^{\widetilde{\theta}})^{1/\widetilde{\theta}} \Gamma_{\widetilde{\theta}} \\ &= \Gamma_{\widetilde{\theta}} \times \varphi_j^{-1/\widetilde{\theta}} \eta_{\pi,j}\end{aligned}$$

where $\Gamma_{\widetilde{\theta}} = \Gamma(1 - 1/\widetilde{\theta})$. Lastly, turning to the wage bill (b_j), note that similar to profits, $b_k \sim Frechet(\widetilde{\theta}, \eta_{L,k})$. Note that since $\pi_k = (\frac{1-\rho}{\rho})b_k$, $\pi_j = \max\{\pi_k\}_{\forall k}$ implies that $b_j = \max\{b_k\}_{\forall k}$. This implies that similar to the profits above,

$$E(b_j | \pi_j = \max\{\pi_k\}_{\forall k}) = \Gamma_{\widetilde{\theta}} \times \varphi_j^{-1/\widetilde{\theta}} \eta_{L,j}$$

Substituting in the values for η_π and η_L, we get:

$$(a)\varphi_{sj} = \frac{\left[\frac{p_{sj}}{(w_{sj}/T_j)^{\rho_s}}\right]^{\theta}}{\sum_k \left[\frac{p_{sk}}{(w_{sk}/T_k)^{\rho_s}}\right]^{\theta}}$$

$$(b)E[l_{sj}(x)] = \varphi_{sj}^{-1/\widetilde{\theta}_s} \Gamma_{\widetilde{\theta}_s} \left[\rho_s \frac{T_j p_j}{w_{sj}}\right]^{\frac{1}{1-\rho_s}} \times x^{\frac{1}{1-\rho_s}}$$

$$(c)E[\pi_{sj}(x)] = \frac{1-\rho_s}{\rho_s} \frac{w_{gsj}}{T_j} \times \left\{ \underbrace{\varphi_{gsj}^{-1/\widetilde{\theta}_s} \Gamma_{\widetilde{\theta}_s} \left[\rho_s \frac{T_j p_{sj}}{w_{gsj}}\right]^{\frac{1}{1-\rho_s}} \times x^{\frac{1}{1-\rho_s}}}_{=E[l_{sj}(x)]} \right\}$$

C Robustness of Model Estimation and Results

C.1 Model Estimation at a More Disaggregate Industry Level

In the baseline empirical exercise (Section 4), we aggregate industries to agriculture, manufacturing and services. As we discuss in Section 3 (and Table 2), it is unlikely that sorting into more disaggregate industries drives our results given that the main patterns of the data are also present at the NIC-4 classification level, including the fact that women hire more women. Nevertheless, to examine the robustness of our conclusions, we also conduct the analysis at a more disaggregate level – to the extent permitted by data constraints.

Specifically, we re-estimate our model using data at the NIC 1-digit level instead of the three aggregate industries. To facilitate comparison of the new estimates with the baseline case, we aggregate each industry-region specific estimate across regions and NIC 1-digit industries (weighted by the total individuals in that industry-region) to the three industries we consider in our baseline analysis (agriculture, manufacturing and services), and report the results in Table C1.

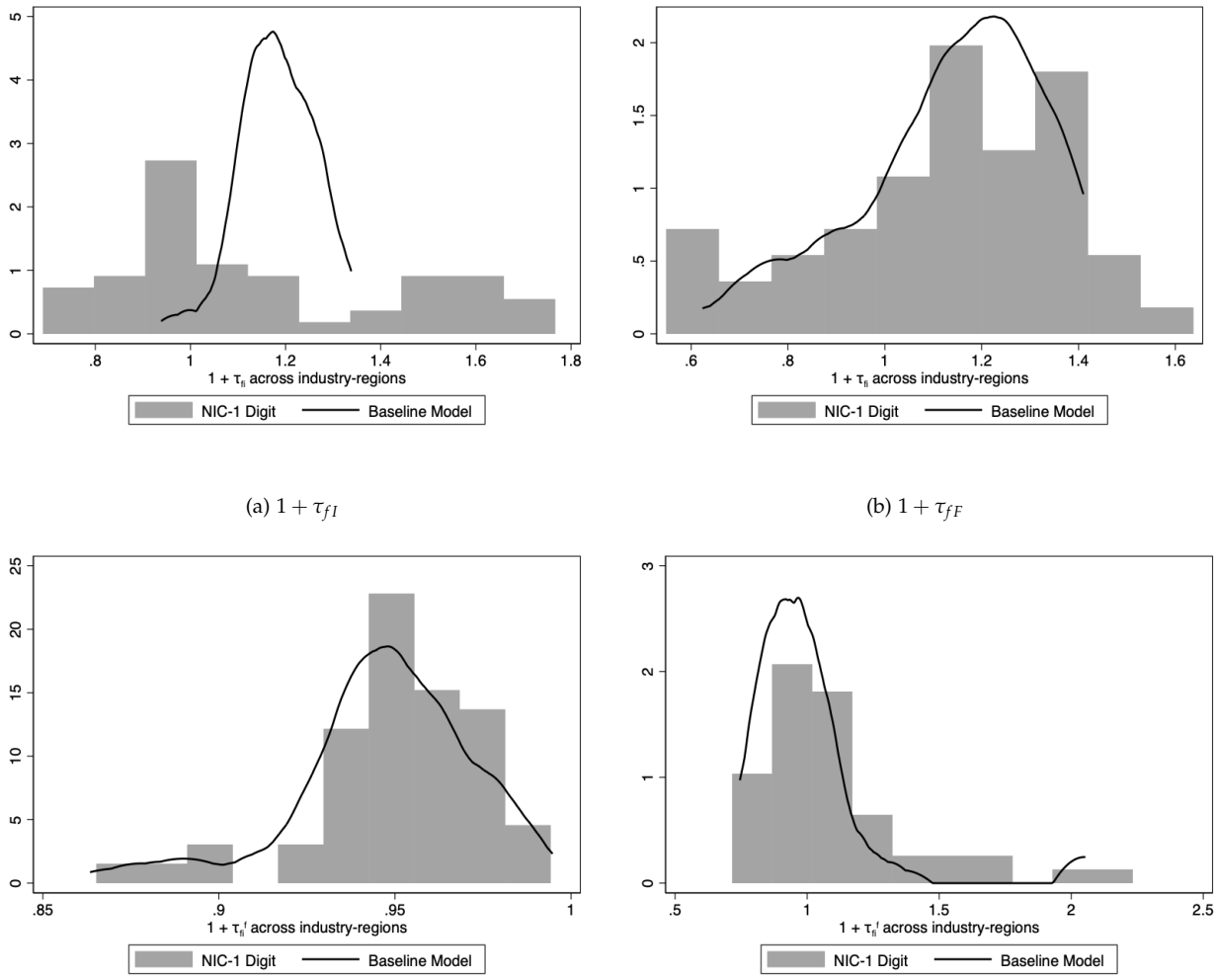
Table C1: Parameter Estimates based on NIC-1 Digit Classification of Industries

	Baseline Model	NIC 1-Digit	(1) - (2)
	(1)	(2)	(3)
τ_{fI}	1.18	1.15	0.04
	[0.08]	[0.31]	
τ_{fF}	1.14	1.12	0.01
	[0.19]	[0.25]	
τ^{f}_{fI}	0.95	0.95	-0.00
	[0.03]	[0.02]	
τ^{f}_{fF}	1.00	1.11	-0.11
	[0.25]	[0.30]	

Column (1) reports the values from the baseline model (Table 5), while Column (2) reports the values obtained by aggregating the NIC 1-digit estimates. Column (3) reports the difference between the two columns. The numbers in Column (2) are very similar to those in Column (1). Figure C1 compares the entire distribution of hiring barriers estimated based on NIC 1-digit level data (gray bars) to the distribution from the baseline

model (from Figure 3). The distributions overlap greatly, except for τ_{fI}. Importantly, the distributions of the τ^f_{fs}'s, which reflect the comparative advantage of females in hiring females, and which play an important role in our counterfactual exercises, are very similar in the two cases.

Figure C1: Comparing Parameter Estimates for Baseline Model and NIC 1-digit Level



(a) $1 + \tau_{fI}$

(b) $1 + \tau_{fF}$

(c) $1 + \tau^f_{fI}$

(d) $1 + \tau^f_{fF}$

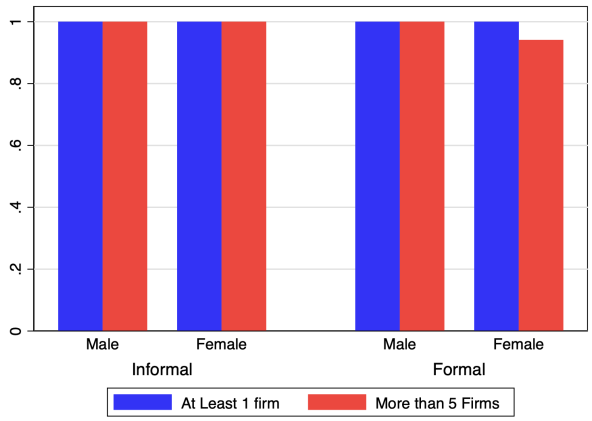
Notes: The above figures report the parameter estimates for the hiring barriers faced by women entrepreneurs. The histogram shows the estimates at the NIC 1-digit, while the solid line shows the density for the aggregate industries as reported in the paper (Figure 3).

In theory, one could re-estimate the model at even more disaggregate levels (NIC 2- or 3-digit levels). However, the poor representation of female-owned firms in several industries limits this exercise. We illustrate the problem in Figure C2 below.

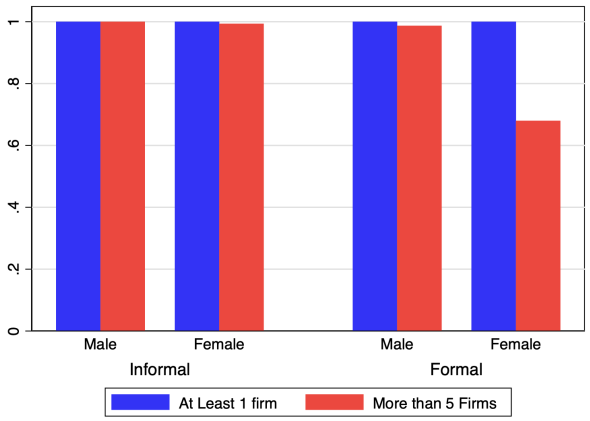
Specifically, we conduct the following exercise: consider a particular level of industry classification (NIC 1-digit, 2-digit, etc.). We calculate the fraction of industry-region pairs in 2005 that have - at that level of industry classification - at least one (five) firms of gender g in a sector s (for example, male-owned firms in the informal sector). In Figure C2(a), in which we define an industry at the aggregate level (agriculture, manufacturing and services), all industry-region pairs have at least 5 male-owned firms in both the informal and formal sectors, and 100% (95%) of industry-regions have at least 5 female-owned firms in the informal (formal) sector. At the NIC 1-digit level (Figure C2(b)), only two-thirds of industry-regions have at least 5 female-owned firms in the formal sector. At the NIC 2-digit level (Figure C2(c)), the coverage of firms drops even more. Only 85% of industry-regions have at least 5 female-owned firms in the informal sector. In the formal sector, only 79% and 31% of industry-regions have at least five male-owned and female-owned firms respectively. Finally, at the NIC 3-digit level (Figure C2(d)), only 87% (63%) of industry-regions have at least 1 (5) female-owned firm in the informal sector. In the formal sector, only 80.5% (54.5%) of industry-regions have at least 1 (5) male-owned firms and 34% (10%) of industry-regions have at least 1 (5) female-owned firms.

Having no or very few firms - especially owned by women - in several industry-region pairs does not allow us to estimate the fixed costs of entry into these industry-region pairs, prohibiting us for conducting the analysis at a more disaggregate level. However, the fact that the estimated barriers are virtually unchanged when we estimate the model at the 1-digit level (instead of the more aggregate level in the baseline case), is reassuring.

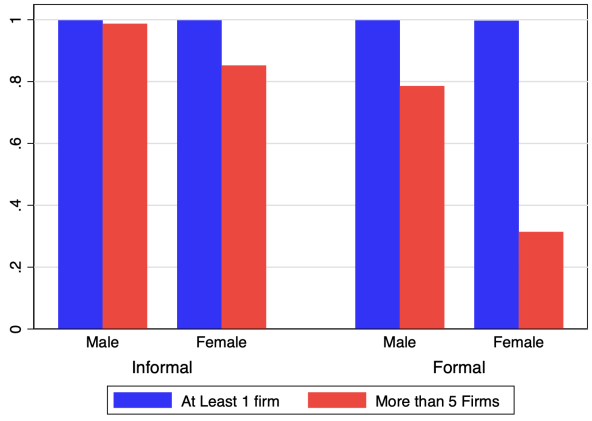
Figure C2: Fraction of Male-Owned and Female-Owned Firms at NIC 1, 2 and 3-digit Industries



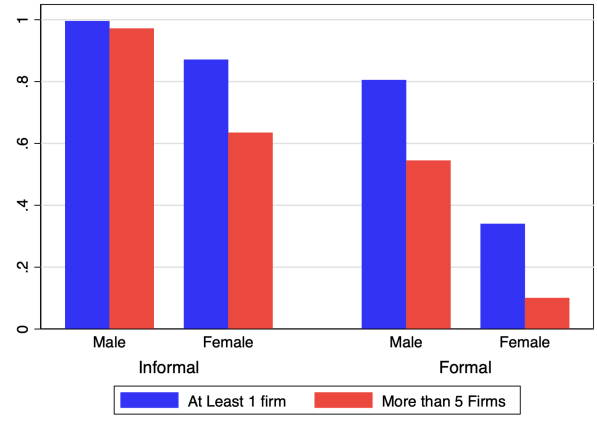
(a) Aggregate Industries



(b) NIC 1-digit Industries



(c) NIC 2-digit Industries



(d) NIC 3-digit Industries

Notes: The above figures report the fraction of industry-region pairs that have at least one firm (green bars) or five firms (orange bars) of gender g (Male, Female) and sector s (Informal, Formal). Figure (a) defines an industry at the aggregate level (agriculture, manufacturing and services). Figures (b)-(d) define an industry at the NIC 1-digit, 2-digit, 3-digit respectively.

C.2 The Role of Non-Hired Individuals

Figure 2 shows that the fixed costs of entering wage work or starting informal entrepreneurship are very low (relative to self-employment), for both men and women. This may seem surprising at first, given that wage work is considered highly desirable in many low-income countries, and women have been shown to be reluctant entrepreneurs (Jensen, 2022; Schoar, 2010).

As noted earlier, these estimates may reflect heterogeneity in wage employment and informal entrepreneurship. Many wage jobs are low-paying and provide no benefits. Similarly, some informal enterprises barely differ from self-employment (in the sense that they may employ two, instead of just one, people, but are otherwise similar in size and productivity to the self-employed). Such options may not seem particularly desirable relative to self-employment. Hence, they may not entail the high fixed costs of entry one typically associates with "good" wage jobs or successful enterprises.

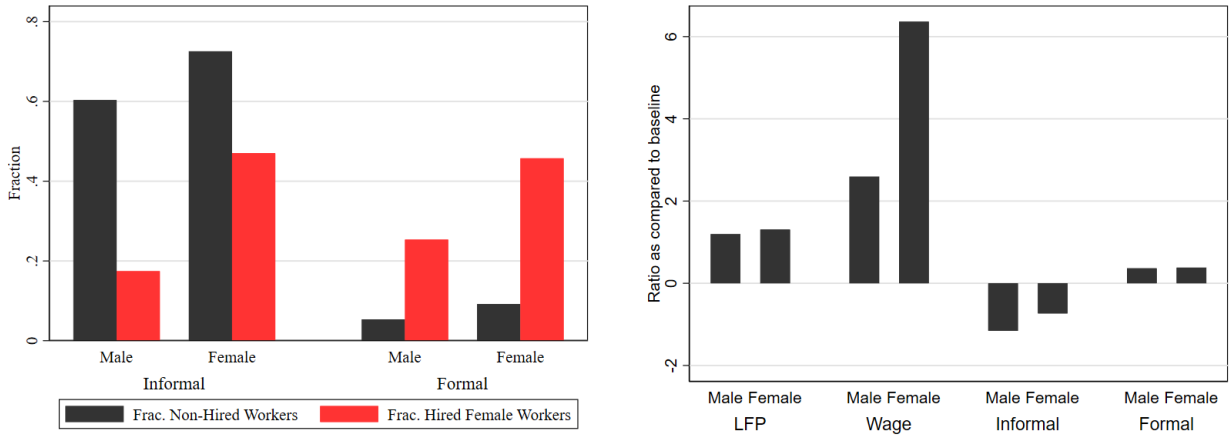
In this section, we explore one particular source of heterogeneity: the employment of "non-hired" workers. The Economic Census separately reports the number of "hired" and "non-hired" workers (by gender) within a firm. Non-hired workers are typically household members working in smaller firms and/or apprentices. Such workers are classified as "wage workers" in our baseline framework (since we do not distinguish between hired and non-hired workers). Given that they do not go through a formal hiring process, they presumably face lower fixed costs of entering wage employment.

Figure C3(a) reports the non-hired workers and the hired female workers as fractions of total workers across firms of gender g in sector s. Two observations stand out.

First, non-hired labor is pervasive in the informal sector for both male- and female-owned firms (60-70% on average), but less so in the formal sector (around 5%). The high incidence of non-hired labor could rationalize the low fixed costs of wage employment we estimate, shown in Figure 2. Second, the fraction of hired female workers is higher (around 40%) in female-owned firms than in male-owned firms (around 20%), indicating that the comparative advantage of female entrepreneurs in employing females is not driven by the use of "non-hired" labor, but is present among hired workers as well.

To understand the role of non-hired labor in the fixed cost estimation, we classify non-hired workers as self-employed, and then re-estimate the model to obtain new fixed cost estimates. This scenario, though extreme, is useful as a benchmark because classifying

Figure C3: Fixed Cost Estimates after Reclassifying Non-Hired Workers



(a) Frac. of Hired Female Workers

(b) Ratio of Fixed Costs

Notes: Both figures use data from the 2005 Economic Census. Figure (a) reports the non-hired workers and hired female workers as fractions of total workers in firms owned by gender g in sector s. Figure (b) reports the average ratio of the fixed costs for LFP, entry into wage work, entry into the informal, and entry into the formal sector for male-owned and female-owned firms, when non-hired workers are classified as self-employed, to the fixed costs as estimated in our baseline model.

non-hired workers as self-employed implies that they earn an income λw^g, which is lower than the market wage w^g in expectation. Figure C3(b) reports the ratio of the new gender-specific fixed costs in LFP, wage work, informal and formal entrepreneurship to the gender-specific fixed costs in our baseline framework.

The results are intuitive and confirm the hypothesis that the low estimates of the fixed costs of wage employment are driven by non-hired labor. When non-hired labor is treated as being self-employed, the big change is in the fixed costs of wage work which increase substantially for both men (2.6x) and women (6.3x) relative to the baseline. Correspondingly (and perhaps unsurprisingly), the fixed cost of informal entrepreneurship decreases slightly for both men and women (by around 1x), indicating the emergence of "reluctant" informal entrepreneurs now that the fixed costs of wage employment are higher.

Given that the focus of the paper is on entrepreneurship, and not on wage work, our baseline specification, in which all workers (hired and non-hired) are considered firm employees, remains our preferred specification. In future work, it would be interesting to explore the heterogeneity in wage employment more fully, but this is outside the scope of the present paper.

D Correlation of Parameter Estimates with Measures of Gender Norms

Figure 2 and Table 5 indicate that women face higher costs of participating in the labor force (LFP costs), formalizing their business, and hiring workers. On the other hand, they face an advantage in hiring female workers (in both the formal and informal sectors). This section explores the plausibility of the estimates. Specifically, we use region-specific measures of women empowerment from various sources in the literature to examine whether our implied measures of gender-related barriers correlate with the documented level of women empowerment in these regions.

D.1 Measuring Gender Empowerment

We use three widely used measures of gender inequality and empowerment in India: (a) Women Empowerment Index (Bansal, 2017); (b) Gender Vulnerability Index (Plan International, 2017); and (c) Patriarchy Index (Singh et al., 2021).

The Women Empowerment Index (WEI), proposed by Bansal (2017) at the Hindustan Times (a widely circulated national daily) uses data from the National Family Health Survey (NFHS), a large, nationally representative survey conducted by the Health and Family Welfare Ministry. In particular, it is based on data for eight indicators, such as the participation of women in household decisions, ownership of land, cell phones and bank account, instances of spousal violence, etc., to construct a state-specific Women Empowerment Index.

The Gender Vulnerability Index (GVI), proposed by Plan International (2017), expands the scope of the WEI by using a set of 170 indicators constructed from large nationally representative data like the Population Census of India, National Family Health Survey (NFHS), Health Management Information System, District Information for School Education (DISE), Rapid Survey on Children, Annual Economic Survey, Annual Survey on Education Report and National Achievement Survey to construct a state-specific, comprehensive measure of gender parity along various dimensions, such as Social Protection (26 indicators), Education (68 indicators), Health (57 indicators), Poverty (19 indicators). These are then aggregated to construct a state-level index of Gender Vulnerability.

Lastly, the Patriarchy Index (PI), proposed by Singh et al. (2021), adapts the Patriarchy Index developed by Gruber and Szołtysek (2016) for Europe, to the Indian context. Using the NFHS data as well, the PI uses measures that span five domains: (1) domination of men over women; (2) domination of the older generation over the younger generation; (3) patrilocality; (4) son preference; and (5) socio-economic domination that recognizes the social and economic imbalances between men and women in households in terms of both earning and control over money and education.

D.2 Gender Norms, Fixed Costs and Hiring Barriers

We begin by examining the association between LFP costs and measures of gender norms by estimating the following regression:

$$Y_{st} = \alpha_t + \beta I_s + \gamma X_{st} + \varepsilon_{st} \tag{14}$$

where Y_{st} is the ratio of female to male LFP costs. We pool the 1998 and 2005 estimates, and examine their correlation with state-specific measures of women empowerment $I_s = \{GVI, WEI, PI\}$. All indices are normalized to have mean 0 and standard deviation 1. We control for state-year-specific observables such as GDP and the fraction of SC/ST population (backward castes), as well as year fixed effects that capture all observable and unobservable trends in India over this time period. Given the small sample size, we bootstrap our standard errors.

Our coefficient of interest is β. As reported in Columns (1)-(3) of Panel A in Table D1, a one standard deviation increase (decrease) in WEI/GVI (PI) is correlated with approximately a 0.4-0.5 p.p. or 30-35% (0.24 p.p. or 18%) decrease in the ratio of female to male LFP costs. There is no statistical association between gender empowerment and formalization costs (Panel B), though the coefficients in Panel B have the expected signs.

Next, we examine how hiring distortions (τ_{fs} and τ^f_{fs}) relate to measures of women empowerment. We re-estimate Equation (14), where Y_{jst} is now the hiring distortion in industry j, state s and year t. In addition to the variables described previously, we include industry fixed effects, α_j, to control for time-invariant differences across industries and control for the female labor force participation rate in order to net out the costs to LFP participation that were the focus of Table D1. As reported in Panel A of Table D2, we find a negative association between empowerment indices and hiring distortions in

the informal and formal sectors, indicating that - conditional on entry - the barriers to business expansion for women entrepreneurs are higher in the more gender-conservative areas. Regarding the comparative advantage of female entrepreneurs in the hiring of female workers (Panel B), we find no statistically significant association. A possible interpretation is that - as noted earlier - this comparative "advantage" could itself be the result of gender-related distortions; if women are discouraged from finding work outside the home due to conservative norms, it is possible that they will only take jobs in female-owned firms, giving rise to the documented pattern in the data.

The associations documented above suggest that while the model treats barriers to entry and operation facing women as a black box, our estimates of such barriers do correlate with measures of women empowerment across Indian states.

Table D1: Correlations of Cost Estimates and Measures of Women Empowerment

	WEI	GVI	PI
	(1)	(2)	(3)
Panel A: Relative LFP Costs			
Index	-0.487***	-0.429***	0.242*
	(0.002)	(0.003)	(0.069)
R^2	0.350	0.307	0.229
Panel B: Relative Formal Sector Entry Costs			
Index	-0.185	0.00324	0.0125
	(0.497)	(0.988)	(0.936)
R^2	0.0995	0.0897	0.0897
N	34	34	34

Notes: The dependent variable in Panel A (B) are the relative LFP (Formal Sector Entry) costs, which is the percentage difference between female and male costs. WEI = Women Empowerment Index; GVI = Gender Vulnerability Index; PI = Patriarchal Index. All indices have been normalized to have mean 0 and standard deviation 1. All regressions control for the GDP of the state, fraction of population comprising of SC/ST castes, and year fixed effects. p-values from bootstrapped standard errors are reported in parentheses.

Table D2: Correlations of Hiring Barriers and Measures of Women Empowerment

	Informal			Formal		
	WEI	GVI	PI	WEI	GVI	PI
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Hiring barriers $(1+\tau_{fsj})$						
Index	-0.0258**	-0.0353***	0.00616	-0.0344	-0.0862**	-0.0137
	(0.028)	(0.004)	(0.552)	(0.254)	(0.012)	(0.568)
R^2	0.181	0.203	0.152	0.488	0.521	0.482
Panel B: Hiring barriers for female relative to male workers $(1+\tau^{f}_{fsj})$						
Index	0.0000599	-0.00375	-0.000280	0.0367	0.0124	0.00880
	(0.988)	(0.301)	(0.908)	(0.227)	(0.734)	(0.539)
R^2	0.246	0.252	0.246	0.156	0.143	0.143
N	102	102	102	102	102	102

Notes: The dependent variable in Panel A is $1+\tau_{fsj}$ and $1+\tau^{f}_{fsj}$ in Panel B. Columns (1)-(3) refer to the informal sector, while Columns (4)-(6) refer to the formal sector. WEI = Women Empowerment Index; GVI = Gender Vulnerability Index; PI = Patriarchal Index. All indices have been normalized to have mean 0 and standard deviation 1. All regressions control for the GDP of the state, and fraction of population comprising of SC/ST castes, female labor force participation rates, and fixed effects for year and industry. p-values from bootstrapped standard errors are reported in parentheses.

E A General Model of Production with Many Inputs

Our model in the paper considers labor as the only input in production. This modeling is driven by data constraints as we do not observe any other inputs in the Economic Census that we use for estimation. In this section, we extend our baseline model to allow for multiple inputs and examine its implications for our estimates. We use the extended model for two purposes. First, we derive expressions of the gender distortions in the extended model and compare them to those obtained in the single-input model. Second, we use the NSS data that provides information on multiple inputs for a subset of firms to estimate the distortions in informal manufacturing based on the extended model, and compare the estimates to those obtained using our baseline model.

Consider the following production function:

$$Y = z\left(L^{\alpha_L}\prod_{i\neq L}K_i^{\alpha_i}\right)^{\rho} \tag{15}$$
$$\text{where: } \alpha_L + \sum_{i\neq L}\alpha_i = 1$$

where K_i are a set of i inputs in production with an expenditure share α_i. For now, we abstract from the distinction between the formal and informal sector. Let w_{ig} be the price for input i paid by an entrepreneur of gender g such that $w_{im} = w_i$ and $w_{if} = (1+\tau_i)w_i$, i.e., women entrepreneurs face a potential distortion τ_i on input i. The profit maximization problem of an entrepreneur becomes (we drop the gender script for ease of notation):

$$\pi = \max_{\{L,K_i\}} pz\left(L^{\alpha_L}\prod_{i\neq L}K_i^{\alpha_i}\right)^{\rho} - w_L L - \sum_i w_i K_i$$

E.1 Identification of Gender Barriers and Comparison with the Single-Input Model

To solve the profit-maximization problem of the entrepreneur, we can break the optimization problem into two steps. In the first step, the profit maximization problem can

be written as:

$$\pi = \max_{\{M\}} pzM^{\rho} - w_M M$$
$$\text{where: } M = L^{\alpha_L} \prod_{i \neq L} K_i^{\alpha_i} \tag{16}$$
$$\text{and: } w_M = \left(\frac{w_L}{\alpha_L}\right)^{\alpha_L} \times \prod_i \left(\frac{w_i}{\alpha_i}\right)^{\alpha_i}$$

The first-order condition implies:

$$M^*(z) = \left[\rho \frac{z}{w_M/p}\right]^{\frac{1}{1-\rho}} \tag{17}$$

In the second step, we solve the cost-minimization problem conditional on the choice of $M^*(z)$, which implies:

$$L^*(z) = \frac{\alpha_L}{w_L} \times w_M \left[\rho \frac{z}{w_M/p}\right]^{\frac{1}{1-\rho}} \tag{18}$$

$$K_i^*(z) = \frac{\alpha_i}{w_i} \times w_M \left[\rho \frac{z}{w_M/p}\right]^{\frac{1}{1-\rho}} \tag{19}$$

Equations (17)-(19) provide important insights as to how this extension relates to our baseline model. From Equation (17), note that since $w_{if} = (1+\tau_i)w_i$, for an entrepreneur with ability z,

$$M_f(z) = \left[\underbrace{(1+\tau_L)^{\alpha_L} \times \prod_i (1+\tau_i)^{\alpha_i}}_{=1+\tau_M}\right]^{\frac{-1}{1-\rho}} \times M_m(z) \tag{20}$$
$$\Rightarrow \frac{M_f(z)}{M_m(z)} = (1+\tau_M)^{\frac{-1}{1-\rho}}$$

i.e., if one had information on the other inputs K_i, as we do with labor, then one could identify a composite index of distortions faced by women entrepreneurs as compared to men.

Moreover, from Equations (18) and (19), note that:

$$\frac{L_f(z)}{L_m(z)} = \frac{1+\tau_M}{1+\tau_L} \times (1+\tau_M)^{\frac{-1}{1-\rho}}$$

If we had information on the other inputs, so that we could identify τ_M, then we could separately identify the true distortion in labor hiring $1+\tau_L$, from the distortions affecting other inputs $1+\tau_M$. Instead, what we identify based on the current approach that considers labor as the only input is $(1+\widetilde{\tau_L})$, where:

$$\begin{aligned}
(1+\widetilde{\tau_L})^{\frac{-1}{1-\rho}} &= \frac{1+\tau_M}{1+\tau_L} \times (1+\tau_M)^{\frac{-1}{1-\rho}} \\
1+\widetilde{\tau_L} &= \left[\frac{1+\tau_L}{1+\tau_M}\right]^{1-\rho} \times (1+\tau_M) \\
&= \left[1+\tau_L\right]^{1-\rho} \left[1+\tau_M\right]^{\rho} \qquad (21)
\end{aligned}$$

i.e., we identify a weighted average of the true τ_L and barriers to all inputs (τ_M). This is why we interpret the distortions in hiring as distortions in expanding the business. Note however that this modeling does not affect the finding that female entrepreneurs have a comparative advantage in the hiring of female workers, since this comparative advantage is identified from the ratio of female to male workers in each firm, conditional on firm size.

E.2 Estimating A Model with Multiple Inputs Using the NSS Establishment Surveys

As noted earlier, the Economic Census provides information only on one input, labor. We use the Economic Census because it is the only data set that covers the entire firm distribution. However, if we confine the analysis to a subset of firms, then we can draw on other data sets that contain information on additional inputs. Such a data set is the Survey of Unorganized Manufacturing Firms from the National Sample Survey (Round 62) in 2005. Like the Economic Census, the NSS asks firms to report the gender of the owner as well as the number of employees and their gender. In addition, it asks firms detailed questions on their sales, wage bill, expenditure on raw materials, capital, and loans. We use the NSS to estimate a model with multiple inputs and compare it to our baseline model. However, the NSS surveys only small, informal firms, and only in the manufacturing sector. This implies that we cannot use it to estimate the barriers faced by women in agriculture or services or the formal manufacturing sector. Therefore, we use the NSS only to examine the robustness of our findings.

Gender Differences in Production Technology

A potential concern in our analysis is that distortions in input markets may affect the production technology women use relative to men. As a result, the "barriers" we estimate could reflect underlying differences in the production functions of male-owned versus female-owned firms. For instance, if female entrepreneurs do not have access to capital, they may choose to operate more labor-intensive technologies.

The NSS data allows us to examine this hypothesis. Note that according to the model presented above, the share of expenditure on an input i is equal to $\rho\alpha_i$. This share incorporates the relevant parameters of the production technology. Based on the NSS, we can calculate the expenditure shares for the three key inputs (labor, capital, materials) as follows. We define the firm expenditure on capital to be the total value of assets that are owned or hired by the firm. These include plant and machinery, transport, and expenditure on software and hardware. For expenditure on labor and materials, we use the total wage bill and the expenditure on raw materials respectively. We then calculate the expenditure share of each input (labor, capital and materials) in total sales.

As reported in Table E1, the three expenditure shares are similar across male-owned and female-owned firms. Not only is the raw difference (Column 3) negligible in magnitude,

Table E1: Share of Inputs in Total Sales

	Male	Female	Difference: (2) - (1)	
			Raw	F.E.
	(1)	(2)	(3)	(4)
Labor	0.12	0.13	0.01	-0.011
			[0.66]	[0.36]
Capital	0.15	0.12	-0.025	-0.018
			[0.33]	[0.38]
Raw Material	0.52	0.49	-0.029	0.0091
			[0.38]	[0.75]

Notes: The table shows the share of labor, capital and raw materials in sales averaged across male-owned and female-owned firms in Columns (1) and (2) respectively. Column (3) reports the raw difference between the means in the previous two columns. The discrepancies are due to rounding errors. Column (4) reports the difference based on regressions that control for an entrepreneur's education level, whether the owner works full-time in the firm, whether the firm is registered with any authority, and district and NIC 5-digit industry fixed effects. p-values calculated from robust standard errors are reported in parentheses below.

this difference is similar even after including district and NIC 4-digit industry fixed effects (Column 4), indicating that they are not driven by sorting across space or industries either. We conclude that at least in the NSS data, there is no evidence of men and women using different production technologies.

Estimating Barriers Using Measures of MRPL, MRPK, MRPR

One of the limitations of the Economic Census is that it does not report the expenditure on any input (including labor). However, given that we observe the input expenditures in the NSS, we can follow the methodology of Hsieh and Klenow (2009) to calculate measures of marginal product revenues of labor (MRPL), capital (MRPK) and raw materials (MRPR) and examine their magnitudes across male-owned and female-owned firms. The

model presented above implies that:

$$
\begin{aligned}
MRPL_g \equiv \frac{\rho \alpha_L p Y_g}{L_g} &= (1+\tau_L^g) w_L && \text{(Labor)} \\
MRPK_g \equiv \frac{\rho \alpha_K p Y_g}{K_g} &= (1+\tau_K^g) w_K && \text{(Capital)} \\
MRPR_g \equiv \frac{\rho \alpha_R p Y_g}{R_g} &= (1+\tau_R^g) w_R && \text{(Raw Materials)}
\end{aligned} \tag{22}
$$

Given that there is no evidence (at least in the NSS data) of any differences in production technology between male- and female-owned firms, any deviations of the MRPs of female-owned firms from those of male-owned firms must reflect distortions (Hsieh and Klenow, 2009). We calculate the MRP of each of the three inputs in our data as follows.

In contrast to labor, the NSS does not provide information on the "quantity" of capital or materials. We follow an approach similar to Hsieh and Klenow (2009) to assign "prices" to capital and materials. The NSS asks firms about their total outstanding loans, along with the interest payable on these loans during the reference period. We calculate the interest rate as the ratio of these two values[36] and use it to deflate the total capital expenditure to calculate K. For raw materials, each firm reports the value and quantity for up to 5 specific products used as raw materials. We use this information to calculate the price for each product, and weight it by its share in total expenditure on raw materials to calculate an (expenditure-weighted) price of raw materials for each firm. We then deflate the expenditure on raw materials by this price index to calculate M. Given these measures, we then compute measures of MRPL, MRPK and MRPR for each firm (Equation 22) and estimate the following regression:

$$
\ln MRPx_i = \alpha_x + \beta_{x,s} FemaleOwner_i + \varepsilon_i \tag{23}
$$

where $x = \{K, L, R\}$ and from Equations (22) and (23), τ_x will be equal to $e^{\widehat{\beta_x}} - 1$. This is reported in Table E2. Columns (1)-(3) report the value for τ_x. Columns (4) and (5) use Equations (20) and (21) to calculate τ_M and $\widetilde{\tau_L}$ respectively. Note that this estimate in Column (5) is close to the value that we estimate in Section 5.3 of the paper, which is 0.21 (mean) and 0.23 (median) for informal manufacturing in 2005.

There are two main takeaways from these results. First, the distortion estimates we obtain from the NSS data when we make the assumption of a single-input ($\widetilde{\tau_L}$) are very

[36] We replace missing values with the gender-, registration-status-, and state-specific average.

Table E2: MRPL, MRPK and MRPR

	τ_L	τ_K	τ_R	τ_M	$\widetilde{\tau}_L$
	(1)	(2)	(3)	(4)	(5)
NSS	0.36	1.10	0.20	0.29	**0.30**

Notes: For an input x, $\tau_x = e^{\widehat{\beta_x}} - 1$ using estimates from Equation (23). Columns 1-3 report the estimates for τ_L, τ_K and τ_R respectively. Column 4 uses Equation (20) to calculate τ_M. Column 5 uses Equation (21) and reports the "implied" $\widetilde{\tau}_L$.

similar to those obtained from the Census data for the corresponding sector (informal manufacturing). More importantly, the estimate of the composite gender distortions in the multiple-input model, τ_M, is similar to the one obtained using the approach we described in the baseline model, $\widetilde{\tau}_L$. This gives us confidence that our estimates of "hiring" barriers reflect the combined distortions women face in *expanding* their business.

F Gender Differences in Entrepreneurial Ability

Our baseline model assumes that the entrepreneurial ability distribution is the same for men and women. This section examines the validity of this assumption and its implications for our main conclusions.

Even if men and women have the same innate ability, it is possible that gender-based discrimination leads to differences in other characteristics, most importantly education, which could make women less suitable to entrepreneurship than men. Therefore, in the next subsection, we examine gender differences in educational outcomes in India during our sample period. Education is only one among several characteristics that could affect entrepreneurial performance. Therefore, we next investigate whether surveys of the population and experts show women to have traits that are considered undesirable for entrepreneurship (of course, the survey responses could themselves reflect gender-bias, but this makes responses that do not suggest any innate differences in entrepreneurial suitability even more credible). Finally, we estimate a version of the model in which we allow the variances of the ability distributions of men and women to differ, and show that the results are virtually unchanged.

F.1 Measuring Ability based on Micro Data (IHDS)

We use data from the 2005 round of the India Human Development Survey (Desai et al., 2005) to compare the educational attainment of men and women. The IHDS is a nationally representative, multi-topic survey of 41,554 households in 1,504 villages and 970 urban neighborhoods across India.

The IHDS collects data on the educational attainment of all household members. A key advantage of this data set is that children aged 8-11 had to also complete short reading and arithmetic tests, which were implemented in a way similar to the ASER modules. For example, the reading test (implemented in the local language) had four levels corresponding to being able to recognize letters, words, paragraphs, and read stories respectively. The arithmetic test tested whether a child could recognize numbers, perform elementary operations like addition and subtraction, and more complex ones like multiplication and division.

We use this data to estimate the following regression for an individual i between the ages 18-65, living in a household h of village v:

$$Y_{i(hv)} = \alpha + \beta Female_i + \gamma X_i + \varepsilon_i \tag{24}$$

where Y_i are two outcome variables: (i) a dummy variable that takes the value 1 if the individual is literate and 0 otherwise; (ii) years of education. Female$_i$ takes the value 1 if the respondent is a female and 0 otherwise. We also control a quadratic polynomial for age, and add either village or household fixed effects to take into account unobservable differences across villages or households that could impact the educational attainment of individuals. We cluster standard errors at the village-level.

Table F1: Education Levels

	Literate		Ed. Years		LAYS #1		LAYS #2	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.263***	-0.267***	-2.686***	-2.764***	0.355***	0.355***	0.347***	0.346***
	(0.005)	(0.005)	(0.044)	(0.042)	(0.006)	(0.006)	(0.006)	(0.006)
Male Mean	0.78	0.78	7.05	7.05	-0.00	-0.00	-0.00	-0.00
R2	0.34	0.66	0.43	0.76	0.39	0.63	0.40	0.63
N	113627	112798	113627	112798	113627	112798	113627	112798
Village FE	Yes	No	Yes	No	Yes	No	Yes	No
HH FE	No	Yes	No	Yes	No	Yes	No	Yes

The results are reported in Columns (1)-(4) of Table F1. Approximately 78% of men and 51% of women between the ages of 18-65 are literate. While men have around 7 years of education, women only have around 4.35 years of education. These results suggest that women lag behind men in terms of schooling.

However, as documented by Angrist, Djankov, Goldberg, and Patrinos (2021), enrollment and learning are different measures of educational attainment, and they do not always go hand in had. As discussed earlier, a key advantage of the IHDS is that it measures *learning*, and not just schooling, for children between 8-11 years of age. We use this information to create a measure of "learning-adjusted" years of schooling (LAYS) in the following four steps:

1. For each child c in a household h and village v, we calculate her/his total "learning" score as the sum of the (standardized) reading and math scores.

2. Using the sample of children for whom we observe the learning scores, we estimate Equation (24), where Y_c now indicates the learning score of the child. We add village fixed effects, control for a quadratic polynomial of age, type of school (public, private, convent, madrassa, etc.), and a set of household characteristics such as size, asset index, highest educational level of parents, whether at least one person works in the household or not, (log) household income and poverty status.

3. Based on the estimated coefficients, we then *predict* the learning levels for the sample of adults (between the ages 18-64) and calculate LAYS for an individual i as the product of the years and his/her (predicted) learning level. For ease of interpretation, we standardize this measure to have mean 0 and standard deviation of 1 for men and define it as LAYS #1.

4. We repeat steps 2 and 3, but now add household fixed effects to the regression specification in Step 2 (instead of household characteristics), and calculate a second measure of LAYS #2.

We then estimate Equation (24) with the LAYS measures as our dependent variables and report the results in Columns (5)-(8) of Table F1. As is clear from the table, even though women have lower levels of literacy and schooling years, they have 0.3 standard deviations higher learning. These results are consistent with the cross-country patterns documented by Angrist, Djankov, Goldberg, and Patrinos (2021).

To summarize the above discussion, the analysis in this section indicates that data on education do not provide support for the premise that women may be less suited to entrepreneurship due to lack of education. Women may have fewer years of schooling, but they exhibit higher learning. This pattern may also justify an assumption that we explore later in this section, namely that the variance of the ability distribution is higher for women than for men. Some women have very few years of schooling or are illiterate, and they may make poor entrepreneurs. But there are also other highly competent women, who have made the most of their schooling.

F.2 Entrepreneurial Ability from GEM Surveys

This subsection takes another approach for assessing entrepreneurial ability based on data from the Adult Population Surveys (APS) implemented by the Global Entrepreneurship Monitor GEM (Reynolds et al., 1999). The APS is particularly valuable since it explores the role of the individual in the entrepreneurial process. The questions focus not

only on business characteristics, but also on people's motivation for starting a business, the actions taken to start and run a business, as well as entrepreneurship-related personality traits. The APS is administered to a minimum of 2000 adults in each economy, ensuring that it is nationally representative. We use all rounds of the APS in India between 2001-2007 and restrict the sample to adults between the age 18-65. We estimate the following regression specification, where i denotes a respondent:

$$Y_i = \alpha_t + \beta Female_i + \gamma X_i + \varepsilon_i \tag{25}$$

Y_i are a set of individual beliefs/opinions/outcomes that we will discuss below. Female$_i$ is a binary variable that takes the value 1 if the respondent is a female and 0 otherwise. X_i are individual controls such as age, income category and educational level. We add year fixed effects in all specifications.

Barriers to Entrepreneurship and Differences in Attitudes/Traits

We first explore gender differences in the ownership and firm size. The results are reported in Table F2. In Column (1), the outcome variable is a binary variable that take the value 1 if an individual reports owning a firm. Women (as compared to men) are 12.9 p.p. (44.4%) less likely to own a firm. Columns (2) and (3) report gender differences in the current and expected (in five years) firm size. Female-owned firms hire 1.4 fewer workers (56.4%) on average, and even expect to hire 1.8 fewer workers (27.5%) in the future as well. These patterns confirm those we documented earlier using the Census data.

Next, we examine gender differences in other variables capturing risk appetite, expectations, and other attitudes as measured in the APS. For each outcome variable, Table F5 provides the detailed questions that were asked.

Table F3 examines gender differences in attitudes towards risks associated with entrepreneurship. We do not find any gender differences with respect to: (i) fear of failure that would prevent women from starting a business (Column 1); (ii) competition faced by other businesses who offer similar products and services (Column 2); (ii) optimistic or pessimistic assessment of the novelty of the product/service provided (Column 3) or the novelty of technology (Column 4); (iii) their perception of whether starting new businesses is considered a desirable career choice (Column 5), is respected (Column 6) or reported positively in the news media (Columns 7).

Table F2: Firm Characteristics

	(1) Own	(2) Current L	(3) Expected L
Female	-0.129***	-1.370**	-1.773**
	(0.013)	(0.601)	(0.776)
Male Mean	0.29	4.66	6.45
R2	0.08	0.05	0.06
N	8306	793	793

Notes: See Table F5 for a definition of all the outcome variables. Female takes the value 1 if the respondent is a female and 0 otherwise. Male mean is the average value of the outcome variable for male respondents. All regressions control for respondents' age, education, and income category along with year fixed effects. Robust standard errors are reported in parentheses. * is $p < 0.1$, ** is $p < 0.05$, and *** is $p < 0.001$.

Lastly, Table F4 examines gender differences in reasons individuals give for starting a business. Columns (1) to (4) show no differences between men and women.

To summarize the APS analysis, there is no evidence of innate gender differences in risk appetite or entrepreneurship-related attitudes that would explain the low share of female entrepreneurs and the small size of their businesses.

F.3 Re-Estimating the Model with Gender-Specific Ability Distributions

In a final exercise, we re-estimate the model to allow for a gender-specific ability distribution i.e., $x \sim \log N(0, \sigma_x^g)$. The differences in educational attainment between men and women documented earlier suggest a larger variance for the ability distribution of women (given that some women are illiterate or have very few years of schooling, while at the other end, some women exhibit higher learning than men conditional on the same years of schooling). We assume that the means of the two distributions are the same as we cannot identify differences in means. But we remind the reader that the evidence we have presented so far does not provide any support for the hypothesis that on average, women differ from men in ways that affect their suitability for entrepreneurship and their performance.

We estimate σ_x^f to be 0.37, which is greater than 0.31 – the estimate in our baseline scenario (Table 4). Figure F1 shows however that relaxing this assumption does not

Table F3: Attitudes and Risk

	Risks		Innovation		Perception		
	(1) Failure	(2) Competition	(3) New Prod.	(4) New Tech.	(5) Desirable	(6) Prestige	(7) Media
Female	0.010	-0.007	0.041	0.026	-0.036	-0.041	0.012
	(0.016)	(0.012)	(0.046)	(0.041)	(0.030)	(0.025)	(0.029)
Male Mean	0.31	0.96	0.36	0.65	0.71	0.83	0.73
R2	0.03	0.02	0.03	0.06	0.03	0.02	0.03
N	6819	2045	718	718	1382	1382	1382

Notes: See Table F5 for a definition of all the outcome variables. Female takes the value 1 if the respondent is a female and 0 otherwise. Male mean is the average value of the outcome variable for male respondents. All regressions control for respondents' age, education, and income category along with year fixed effects. Robust standard errors are reported in parentheses. * is $p < 0.1$, ** is $p < 0.05$, and *** is $p < 0.001$.

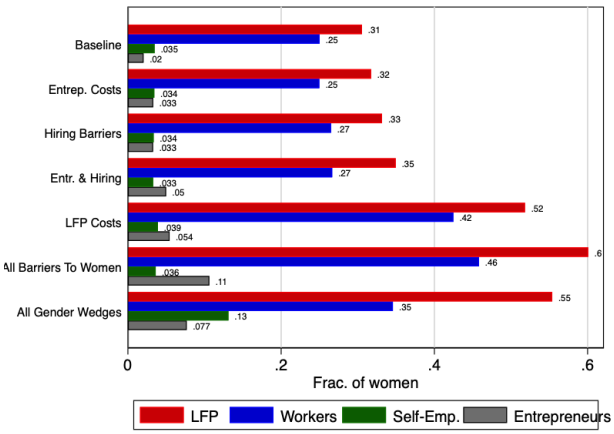
Table F4: Reason for starting business

	(1) Business Opp.	(2) Independence	(3) Higher Income	(4) Maintain Income
Female	0.015	0.050	-0.039	-0.007
	(0.031)	(0.050)	(0.053)	(0.032)
Male Mean	0.50	0.36	0.52	0.12
R2	0.04	0.09	0.08	0.02
N	2397	514	514	514

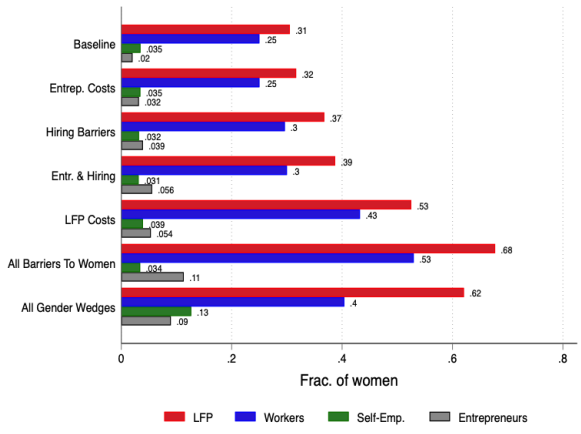
Notes: See Table F5 for a definition of all the outcome variables. Female takes the value 1 if the respondent is a female and 0 otherwise. Male mean is the average value of the outcome variable for male respondents. All regressions control for respondents' age, education, and income category along with year fixed effects. Robust standard errors are reported in parentheses. * is $p < 0.1$, ** is $p < 0.05$, and *** is $p < 0.001$.

impact our results in any meaningful way. In particular, the impact of removing gender barriers has a very similar impact on the allocation of women in the economy (Figures F1(a) and F1(b)), as well as on changes in real income (Figures F1(c) and F1(d)). If anything, the results are quantitatively larger in this case. This is because $\sigma_x^f > \sigma_{x,base}^f$. This in turn implies that when gender barriers are now removed, even more productive women become entrepreneurs, who hire other women, which increases FLFP and real income more than in the baseline.

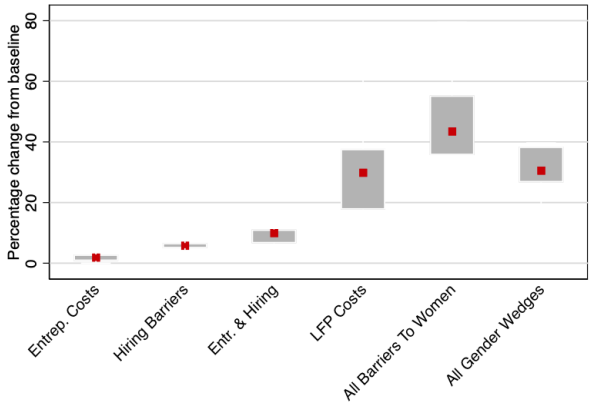
Figure F1: Gender-Specific Ability Distribution and Aggregate Impact



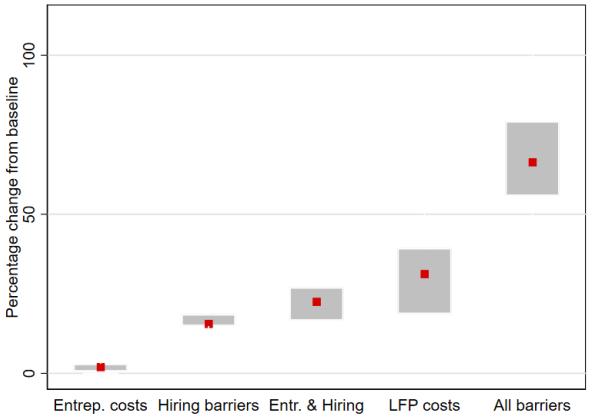
(a) Distribution of Women: Baseline Scenario



(b) Distribution of Women: Gender-Specific σ_x



(c) △ Real Income: Baseline Scenario



(d) △ Real Income: Gender-Specific σ_x

Notes: Figures (a)-(b) compare the distribution of women in the economy when the same σ_x is imposed for men and women (a), and when σ_x is allowed to vary by gender (b). Figures (c)-(d) report the corresponding changes in real income.

Table F5: Questions and Variables

Variable	Definition
Own	You are, alone or with others, currently the owner of a company you help manage, self-employed, or selling any goods or services to others.
Current L	Current firm size
Expected L	Expected firm size in the next 5 years
Risk	Fear of failure would prevent you from starting a business.
Competition	Right now, are there many, some, or no other businesses offering the same products or services to your potential customers? The variable takes the value 1 if there are some/many competitors.
New Product	Will all, some, or none of your potential customers consider this product or service new and unfamiliar? New Product takes the value 1 if "all" or "some" customers consider this product/service new.
New Technology	Have the technologies or procedures required for this product or service been available? The variable takes the value 1 if the technology has been around for less than 5 years.
Desirable	In your country, most people consider starting a new business a desirable career choice.
Prestige	In your country, those successful at starting a new business have a high level of status and respect.
Media	In your country, you will often see stories in the public media about successful new businesses.
Business Opp.	Are you involved in this start-up to take advantage of a business opportunity or because you have no better choices for work?
Independence	Which one of the following, is the most important motive for pursuing this opportunity: to have greater independence
Higher Income	Which one of the following, is the most important motive for pursuing this opportunity: higher income
Maintain Income	Which one of the following, is the most important motive for pursuing this opportunity: maintain income