Firm Size and Internet Adoption in the Indian Informal Sector

Jayati Chatterjee[1] and Subash S[2]

Abstract

India's informal sector is dominated by subsistence-level own-account enterprises that rarely scale up despite growing digital penetration. This paper examines whether internet adoption smooths enterprise's vertical mobility among unincorporated non-agricultural enterprises in India and whether these gains differ by gender. Using nationally representative survey data spanning 2010-11 to 2023-24, we observe transitions across Own Account, Non-Directory, and Directory Enterprises using an ordered probit framework. To address endogeneity in internet adoption, we instrument internet use with distance from the state headquarters. We observe that internet adoption significantly increases the probability of operating above the subsistence threshold. However, decomposing digital access into market, financial, and communication dimensions reveals substantial heterogeneity in returns. Female-owned enterprises benefit significantly less from digital financial access than male-owned firms. In contrast, digital communication tools generate larger gains for female enterprises, consistent with online platforms partially offsetting mobility and network constraints. The pseudo-level results show the differential impact of the BharatNet programme across gender, implying the need for women-specific policies to utilise the available technology. These findings suggest that digitalisation alone cannot ensure inclusive enterprise growth without targeted interventions addressing gendered financial exclusion and sector-specific constraints.

[1] Corresponding author. PhD Research scholar, Department of Humanities and Social Sciences, IIT Madras, India. Jchatterjee95@gmail.com

[2] Professor, Department of Humanities and Social Sciences, IIT Madras, India. subash@iitm.ac.in

21.Introduction

Informal firms contribute a substantial share to the economic activity and employment in developing countries. In India, the informal (or unincorporated) sector accounts for more than 70 per cent of total employment and a significant share of value added (International Labour Organisation, 2018; ASUSE 2021, 2022, 2023, 2024). But the informal sector faces obstacles that often prevent it from growing into larger firms. Within the informal sector, there is a wide disparity in the productivity of male- and female-owned firms, and understanding the mechanisms that facilitate bridging this gap is therefore critical for development policy. One of the main reasons for the productivity disparity is the concentration of female firms in the lowest size category as own-account enterprises. Even across gender, firms in the informal sector are generally less productive and have a higher concentration of own-account enterprises (Meherotra & Giri, 2023). Hsieh and Olken (2014) describe a 'missing middle' in the aggregate size distribution of firms, referring to the very few mid-sized formal enterprises; a similar, and arguably more persistent, stagnation exists within the informal sector itself. Own-Account Enterprises (OAEs) are over-concentrated, and these are subsistence-level units that operate without hired labour. The low transition rate in India's informal sector has been concerning (Rabbani & Raj, 2024; Rabbani & Raj, 2023; Ghani et al., 2013). Although the informal sector is important in the country, these enterprises remain severely constrained by limited access to formal credit, low levels of technology adoption, weak infrastructure, and gaps in managerial capabilities (Banerjee & Duflo, 2014; Raj & Sen, 2016), which prevent them from vertical mobility to reach the formal ‘middle’. Despite their economic importance, transitions in size among these firms are rare. These micro-units rarely cross the 'hiring threshold' to transition into Non-Directory (NDE, 1-5 workers) or Directory Enterprises (DE, 6-20 workers). Thus, the central concern is not bimodality in the aggregate firm-size distribution, but the limited upward mobility among micro-enterprises themselves. Rabbani and Raj (2024) explain the firm transition using financial constraints, utilising data from 2010-2016. Since 2016, India has undergone a rapid expansion of digital infrastructure. This increase occurred with the rollout of affordable mobile internet, Jio's entry in 2016, the BharatNet program connecting rural gram panchayats to optical fibre, and the rapid adoption of digital payment platforms in the post-demonetization and post-pandemic periods. These structural changes raise the question of whether adopting digital technology provides an additional pathway for informal microenterprises to cross the hiring threshold. The post-pandemic period has further accelerated this digital transition, although aggregate firm-size distributions have shifted

toward greater concentration at the subsistence scale, with the number of own-account enterprises increased between 2015-16 and 2022-23. Against this backdrop, this paper enquires whether internet adoption has enabled informal firms to overcome the hiring threshold, and whether this effect differs systematically between male and female-owned enterprises. By focusing on the 0-20 worker, this paper investigates the catalysts that enable micro-enterprises to overcome this 'missing middle' and achieve structural growth. During the time period that the study has been set, there has been a rapid increase in the digitalization in the country. This paper examines whether internet adoption enables informal micro-enterprises to cross the hiring threshold, the critical transition from subsistence own-account status to small employer status that defines upward mobility within India's informal sector. Using five rounds of the NSS Unincorporated Enterprise Survey spanning 2010–11 to 2022–23, covering approximately 1.4 million firm observations across 451 districts in 17 states, we estimate a Two-Stage Residual Inclusion (2SRI) ordered probit model that instruments for internet adoption using the distance of each district from its state headquarters, a measure of administrative remoteness that predicts telecommunications infrastructure availability through the hierarchical rollout logic of India's State Wide Area Network (SWAN) architecture.

We establish four findings. First, internet adoption significantly increased the probability of association among informal micro-enterprises that crossed the hiring threshold. The effect was substantially larger than the OLS estimates, consistent with negative selection among adopters, which caused OLS to understate the true effect. Second, the returns to internet adoption are substantially larger for female-owned firms than for male-owned firms. Internet adoption reduced the probability of female own-account status by 21 percentage points, three times the male effect of 7 percentage points, thus reversing the baseline gender size disadvantage among adopting firms. Third, we observe that OLS overstated internet returns for men and understated them for women. This underestimated result indicates that the true gender gap in vertical mobility is larger. Fourth, using a cohort pseudo-panel with BharatNet gram panchayat coverage as an instrument, we find that supply-side connectivity expansion significantly predicts male, but not female, internet adoption. This reveals that female entrepreneurs face demand-side barriers to digital adoption that physical infrastructure alone probably cannot overcome. Thus, closing the gender gap in digital returns requires demand-side interventions targeting literacy, device access, and social norms, alongside connectivity expansion, especially in the context of a developing country like India.

In developing countries, economies of scale and productivity gains are important to structural transformation. But the inability of firms to grow beyond subsistence scale might have aggregate consequences. With limited vertical mobility, female entrepreneurs face an additional constraint: a gendered "glass ceiling" that restricts their advancement up the size hierarchy. While all micro-enterprises face structural barriers to growth, women-owned firms are predominantly own-account enterprises that operate within households, exhibiting significantly lower probabilities of employing hired labour or transitioning into employer firms (Bruhn, 2009; Campos & Gassier, 2017; World Bank, 2019). In India, female-owned enterprises are more likely to operate from within the household. They depend on family labour and remain concentrated in low-capital, low-productivity sectors (Chakraborty & Chatterjee, 2021; Raj &Sen, 2016). Further, women entrepreneurs have lower asset ownership, limited mobility, weaker business networks and face disproportionate unpaid care burdens and discriminatory lending practices (Klapper & Parker, 2011; Field et al., 2010). These constraints lead to lower capital accumulation and an inability to hire labour, further resulting in an inability to cross the critical "employer threshold" that defines the transition from OAE to NDE and DE categories. The aggregate "missing middle" debate (Hsieh & Olken, 2014) focuses on the shape of the overall firm-size distribution. This ceiling for women operates in the lower tail of the firm-size distribution. Even with improvements in broader structural conditions, female-owned firms may remain at a subsistence-level level of self-employment, probably due to persistent social and institutional constraints. Understanding whether access to digital technologies can reduce these gendered barriers is therefore crucial to explaining firm transition dynamics in the informal sector.

The paper is arranged as follows: Section 2 provides a brief overview of the background of the informal sector and digitalisation; Section 3 discusses the data and variables; Section 4 presents the methodology; and Section 5 reports the results. Section 6 has the robustness checks with the pseudo panel, and Section 7 concludes the paper.

2.Background

The overcrowding of firms in the smallest category, as own-account enterprises, reflects a lack of firm transition in India's informal sector. We observe the size distribution of small firms from the period of 2010-2023. We look at the number of firms in each category, the sales in each category, and the number of workers. The accumulation of firms in the lower end of the tail is a sustained characteristic in the economy. Figure 1 shows year-wise samples from the surveys

across the three size classifications. Throughout the period, there is overcrowding in the smallest firm category.

Fig 1: Size of enterprises for Male and Female Enterprises across Years

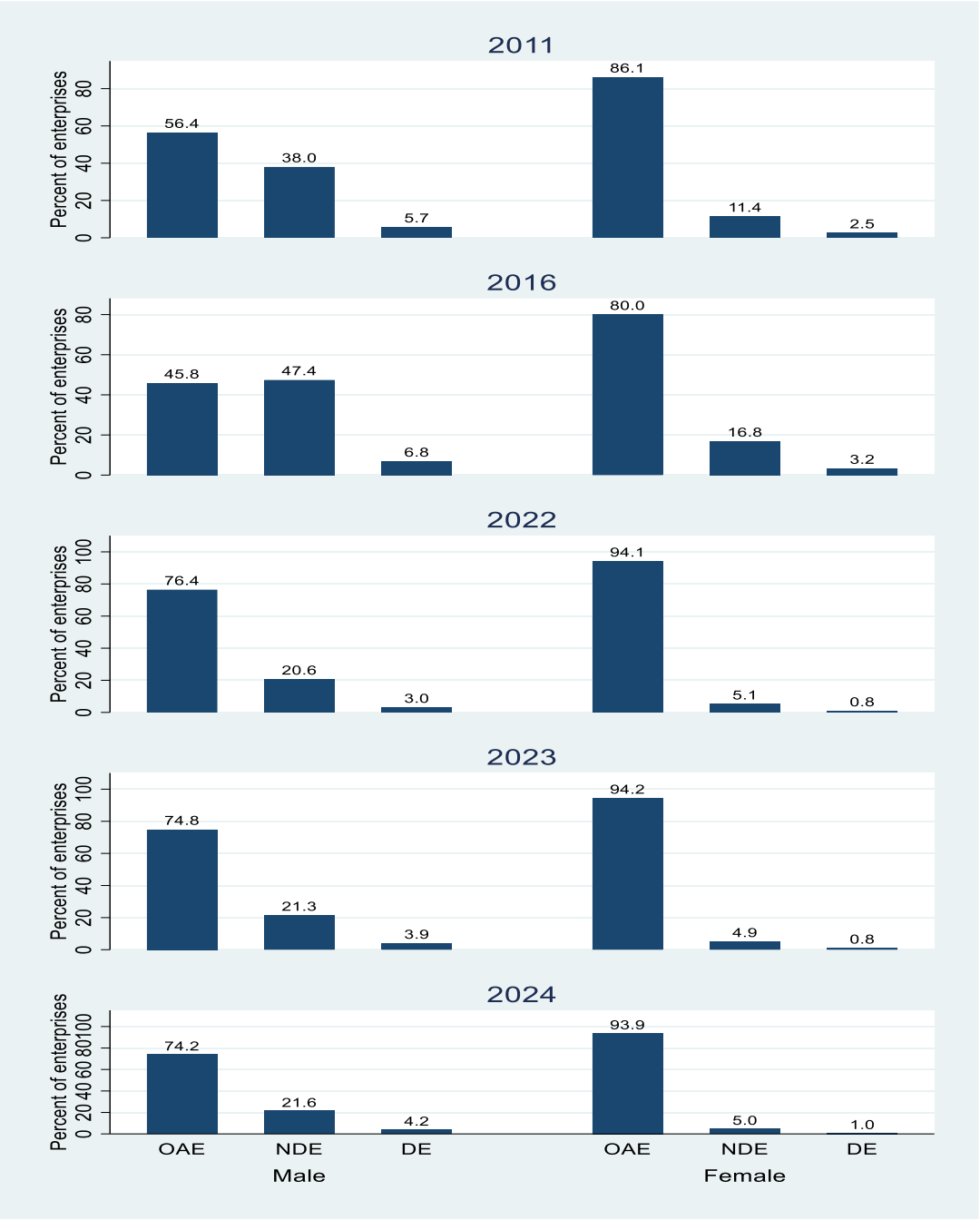
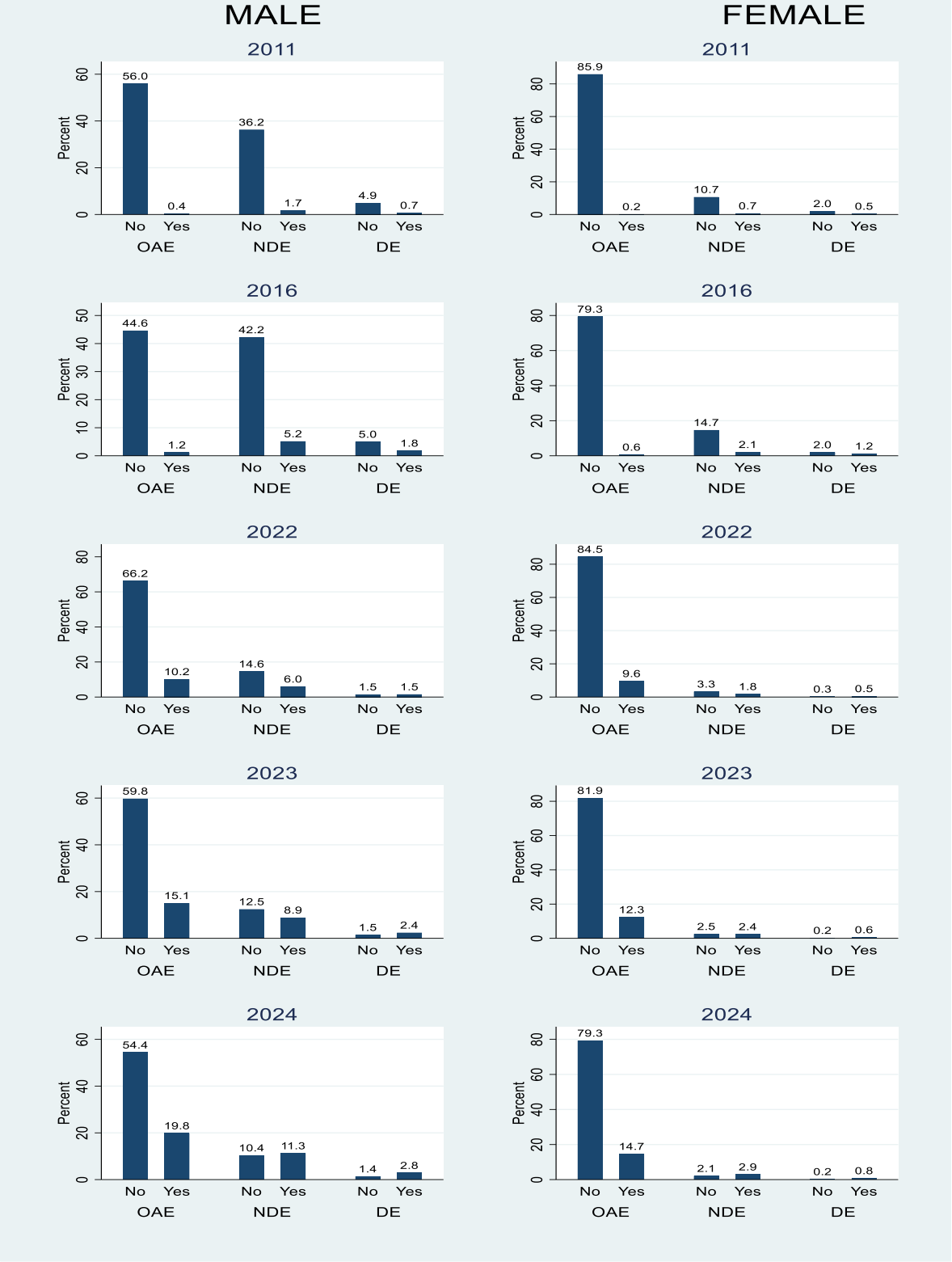


Fig 2: Internet Users vs Non-Users For Male And Female Enterprises Across Years



In this context, the rapid diffusion and adoption of digital technologies offer a way to ease growth constraints on small firms. Digital adoption has the potential to reduce information asymmetries, lower search and transaction costs, and expand market access beyond local networks (Goldfarb & Tucker, 2019). For own-account firms, the adoption could lower fixed

labour costs through digital payments and enhance transaction histories and thus ease access to formal credit (Alok et al., 2025). For female entrepreneurs, internet access may function as a "ceiling-breaker" by partially bypassing mobility restrictions, male-dominated trade networks, and localized discrimination. Online platforms have the potential to expand customer reach and reduce reliance on traditional intermediaries, thereby weakening gender-based constraints on scale. Fig 2 shows the increase in internet adoption over time across the three categories and genders.

However, the productivity effects of digital adoption are unlikely to be uniform across industries. In technologically intensive manufacturing sectors, digital tools may complement production processes, inventory management, and supply chain coordination, thereby amplifying returns to scale. Following Amin and Islam (2023), we hypothesise that complementarity between digital adoption and technological intensity raises the returns to expansion. Thus, internet adoption can increase the likelihood that firms cross the hiring threshold despite traditional size and gender constraints. By examining the interaction between gender and internet access, this paper identifies the conditions under which digitalisation enables OAEs and NDEs to cross the hiring threshold and escape the subsistence trap.

Despite their economic significance, these firms remain trapped at a subsistence scale, a stagnation often linked to constrained access to credit, limited adoption of technology, and institutional frictions.

3.Data and Variables

This study utilises five rounds of nationally representative enterprise surveys from India. We combine data from the National Sample Survey Office (NSSO), specifically the 67th and 73rd rounds, with the Annual Survey of Unincorporated Sector Enterprises (ASUSE) for the years 2021-22, 2022-23, and 2023-24. These surveys provide comprehensive firm-level data on ownership, employment, assets, and technology adoption for unincorporated non-agricultural enterprises not registered under the Factories Act (1948).

To capture the dynamics of upward mobility in the informal sector, the sample is limited to enterprises with 0-20 workers. Firms with more than 20 workers represent a negligible fraction of the unincorporated sector and are excluded to prevent outlier bias. To ensure clear identification of the owner's gender, we restrict our sample to proprietary enterprises and exclude partnership and Self-Help Group (SHG) firms. Although researchers have shown that female participation may be purely nominal or secondary to male partners in partnership firms,

we address this by considering only single-owner firms (Ghani et al., 2013; Sabarwal & Terrell, 2008). Following Bardasi et al. (2011) and Campos et al. (2015), we focus only on proprietary firms, as the owner's gender directly links to the firm's strategic choices and resource allocation.

We restrict the sample to enterprises less than 100 years old and exclude observations with missing or negative total employment values for data quality. Employment is defined strictly as full-time workers (including the proprietor) to maintain the hiring threshold. We focus on 17 major states, treating Andhra Pradesh and Telangana as a single entity to maintain longitudinal consistency following the 2014 bifurcation. While the cross-sectional nature of the data precludes tracking individual firm trajectories, the pooled sample allows us to identify the structural determinants of enterprise scaling across the different size categories.

The final estimate has 1425683 observations across the 5 survey rounds for 451 districts and 17 states. We dropped the districts with incomplete census matches and unbalanced district coverage across instruments.

This study draws on repeated cross-sectional data on unincorporated enterprises in India to explore how internet adoption relates to changes in the firm-size transition across the sample period. A key limitation of this approach is that the same enterprises cannot be tracked over time, since the data does not follow specific enterprises from one period to the next. As a result, when this paper refers to "transition", it does not imply that individual enterprises moved from a smaller size to a bigger size. Rather, it describes a broader change in the sample's overall structure. A pattern of true transition could be observed only if individual enterprises grew over time. However, this cannot be confirmed with the available data, or, more plausibly, if smaller enterprises shrank or exited while larger ones expanded or entered.

We categorise firm size according to the NSSO definition of own-account enterprises with no hired workers, and the second category includes those with total workers between 2 and 5, with at least one hired worker.[3]. The third category comprises firms with 6 to 20 workers. The first category includes household workers, but we do not include them in the main dependent variable. The 2010 NSSO survey does not differentiate between household workers and other workers, making it difficult to distinguish paid helpers or apprentices from unpaid household workers. Although it is defined as separate categories in the later rounds of the survey. So, we

[3] Hired worker is calculated adding formal hired worker and informal hired worker

take own-account enterprises as those run only by the owner; the next two categories consider total labour, which includes the owner, formal and informal hired workers, helpers, apprentices, and unpaid household workers[4] for classification[5]. Our key independent variable is a binary indicator of whether the enterprise has an internet connection, with 1 for Yes and 0 for No. Our other key dependent variable is the gender of the owner, which takes the value for male owners and 1 for female owners. We also examine the interaction between the internet and the female owner, as indicated in equation 1.

The other firm-level controls include social group, categorised into three groups, taking values 0, 1, and 2 for general, OBC, and SC/ST, respectively. The type of premise of the enterprise, meaning whether it is within the household (takes value 0) or outside the household (takes value 1). We also include the firm's age, registration status, and financial obstacle status. The registration status takes the value 0 for unregistered firms and 1 for registered firms. The financial obstacle status takes the value 0 if they do not face any financial obstacles and 1 if they do. The district-level controls are drawn from the Census of India Village Directory (VD) and Town Directory (TD) for 2001 and 2011, and include terrain ruggedness (Terrain Ruggedness Index, TRI), log banking penetration (total bank count across village and town directories), log population, log number of schools, and log domestic electricity. To ensure the instrument is not affected by reverse causality, we use the census substantially preceding each sub-sample period. The 2001 census for the 2010-11 and 2015-16 NSSO rounds, and the 2011 census for the 2021-23 ASUSE rounds. Since the distance of a district from its state headquarters is fixed by administrative geography and was recorded prior to the firm survey periods (even if it changes within those periods), it cannot be influenced by firms' subsequent internet adoption decisions, size outcomes, or connections.

Table 1: Descriptive statistics

Variables	*N*	*Mean*	*Standard deviation*	*Min*	*Max*
Dependent Variable					
Type of enterprises	1423925			0	2

[4] Unpaid household workers are not mentioned separately in the survey of 2010; they include them with other workers, but that might also include helpers who might get paid, so we do not include them in our primary independent variable.

.

OAE	1038060	0.729			
NDE(2-5)	331479	0.233			
DE (6-20)	54,386	0.382			
Independent variables					
Internet	1423925			0	1
Female	1423925	0.195		0	1
Social group of the owner	1423925	0.839		0	2
General/Others	438398	0.308			
OBC	775782	0.545			
SC and ST	209745	0.147			
Proportion of Urban	1423925	0.471		0	1
Premise of the enterprise	1423925			0	3
Within household	485614	0.341			
Outside household fixed	676040	0.475			
Outside household not fixed	262271	0.184			
Registration	1423925	0.381	0.486	0	1
Financial obstacles	1423925	0.066	0.248	0	1
Age	1423925	11.365	8.538	2	100

Apart from the district-level controls, we employ time (survey year) and state fixed effects to control for unobserved year- and region-specific technology adaptability. Time effects would capture the macro shocks that could affect the firm size. The state dummies would capture the state-level nuances of technology adoption. Industry effects control for industry-specific

technology or internet adoption rates, as the type of industry determines whether it adopts the internet and how that affects firm size.

4.Methodology

We estimate an ordered probit model by pooling the five rounds of cross-sectional survey data to examine the effect of internet adoption on firm transition.

$$P\left(Internet_{ijdt}\right) = \phi(\alpha_0 + \alpha_1 Internet_{ijdt} + \alpha_2(distance_d \times Female_i) + \gamma X_{ijdt} + \sigma Z_{dt} + D_d + J_j + T_t + \epsilon_{idt} \quad (1)$$

X_{ijdt} is a vector of firm-specific controls that includes the owner's social group, premises, log age, registration, and financial obstacles. Z_{dt} is the vector of district-specific controls from the census that include the district-level differences like the number of schools, domestic power usage, number of banks and the ruggedness of the district. We also capture state-level differences through state-specific effects; the time effect captures survey-year effects, and industry variations are accounted for with industry-specific effects.

Equation 1 estimates the role of internet adoption in firm transition within the informal sector of the economy. However, a naïve association between technology adoption and firm transition would likely yield biased estimates due to reverse causality and omitted variable bias. Firms that adopt the internet may differ systematically from non-adopters in unobserved variables, the skills of workers and managers, access to networks, capital availability, and other factors, which would have an obvious effect on firm size. This endogeneity may further complicate the selection bias across males, females, or social groups, explaining why some groups select the internet and why others do not. As market conditions and societal restrictions may affect females and marginalised castes, they may opt for digital adoption, whereas others may run a physical business. They are driven by different constraints and opportunities, generating gender- and social-group-differentiated biases in naive estimates. We address this using an instrumental-variables strategy implemented with the Two-Stage Residual Inclusion (2SRI) estimator.

To address endogeneity, we use the instrument of the distance between firms' districts and their state headquarters, drawn from the Census of Indian Towns and Cities. This variable captures the administrative and infrastructural readiness of a district, implying that districts farther from state headquarters would historically have received limited telephone and communication

infrastructure and would likely have been delayed as well. Since the rollout of communication lines tends to spread from the administrative and commercial centers, where the backbone infrastructure is first established. The instrument satisfies two conditions. First, the instrument is relevant, as distance to the state headquarters is a strong predictor of internet adoption at the firm because of the availability of local telecommunications infrastructure. Second, since our equation already controls for state-level and other district-level variations, distance to state headquarters would plausibly affect firm size through its effect on internet adoption, rather than through any direct channel. Infrastructural readiness could affect firm outcomes, raising the question of the exclusion restriction, but these effects are captured by our district-level controls. Infrastructure readiness, or a district's remoteness, is determined by the historical presence of banks, schools, and electricity access, as well as the district's population and terrain ruggedness, which capture the main dimensions of district development and heterogeneity that could affect it. One thing to note is that the instrument is from the town directory, but our sample includes both urban and rural areas. However, the instrument is at the district level and thus can capture the effect. The state-specific effects capture all the remaining time-invariant differences between states in development paths, policy rollouts, and infrastructure quality. The instrument is predetermined relative to outcomes, using the Census 2001 for the 2010-11 and 2015-16 survey rounds, and 2011 for the 2021-23 rounds. We also formally estimate the reduced-form equation to assess whether the instrument affects firm size, controlling for all other variables. The instrument is statistically insignificant, and the interaction with female ownership is also insignificant, thus providing evidence consistent with the exclusion restriction.

While we do not find any papers that use this particular instrument for internet adoption, the distance between firms and the closest telephone exchange site has been used as an instrument for broadband access (De Stefano et. al., 2018, 2023). Our identification strategy follows a larger body of literature that utilises geographic distance from pre-existing, predetermined cable network infrastructure nodes as an instrument for technology adoption (Czernich et al., 2011; Falck et al., 2014; DeStefano et al., 2018, 2023). We use the distance from districts to state headquarters as a proxy for the penetration rate and timing of telecommunications infrastructure rollout (as used in previous literature), implying that districts farther from state administrative centres received connectivity infrastructure later. The backbone networks in India radiate outward from state capitals through the State Wide Area Network (SWAN) architecture established under the National e-Governance Plan. This generates plausibly

exogenous variation in internet infrastructure availability across districts that is determined by administrative geography rather than local economic potential.

We also interact the instrument with female ownership (statehq_dist × female) to allow the instrument to predict internet adoption for female-owned firms, as our objective is to observe gender-heterogeneous treatment effects and to include a gender-interacted endogeneity correction in the second stage.

We employ the Two-Stage Residual Inclusion (2SRI) estimator (Terza et. al, 2008; Wooldridge, 2010; Terza, 2017), which is the appropriate control function approach when the endogenous variable is binary, and the outcome model is nonlinear. The instrument and all covariates are included to explain the endogenous regressor, the internet, and the interaction term with gender. The outcome model, including the predicted first-stage residuals, the endogenous regressor, and all other covariates, is estimated in the second stage. Unlike two-stage least squares, which is inconsistent in nonlinear models, 2SRI achieves consistency by including first-stage residuals directly in the nonlinear second-stage model, effectively controlling for the endogenous component of internet adoption. We first estimate the first-stage specification, where we regress the binary internet adoption variable on all covariates and the excluded instrument, which is the distance of the district from its state headquarters, and interacted with female ownership to allow the instrument to predict internet adoption by gender differentially. The regression function estimated in the first stage takes the following form:

$$P\left(Internet_{ijdt} = 1\right) = \phi(\alpha_0 + \alpha_1 Distance_{dt} + \alpha_2(Distance_d \times Female_i) + \gamma X_{ijdt} + \sigma Z_{dt} + D_d + J_j + T_t + \epsilon_{idt} \quad (2)$$

where i indicates firms, j indicates industry, d indicates districts, and t indicates time of the survey year. $\Phi(\cdot)$denotes the standard normal cumulative distribution function. $distance_{dt}$ denotes the distance of district d from its state headquarters, our excluded instrument. The interaction $Distance_d \times Female_i$ allows the instrument to differentially predict internet adoption for female-owned firms. $X_{ijdt}$is a vector of firm-level controls comprising gender of ownership, social group, premise type, log firm age, registration status, and financial obstacles. $Z_d$is a vector of district-level pre-period controls including terrain ruggedness, log banking penetration, log population, log schools, and log domestic electricity connections. D_d, J_j, and $T_t$denote state, industry, and year effects respectively.

From the first stage, we obtain predicted probabilities $\hat{p}_{id}$and compute two residual terms:

$$\hat{v}_{ijdt} = \text{internet}_{ijdt} - \hat{p}_{ijdt}$$
$$\hat{v}^{f}_{ijdt} = \hat{v}_{ijdt} \times \text{female}_{ijdt}$$

The first residual $\hat{v}_{id}$captures the overall endogenous component of internet adoption. The second residual $\hat{v}^{f}_{id}$captures the gender-specific endogenous component, we assume theta the endogeneity correction happens differently across gender.

Second Stage

In the second stage, we estimate an ordered probit model for firm size, augmenting the structural equation with both residual terms as control function corrections:

$$P(size_{ijdt} \leq s) = \Phi\Big(\kappa_s - \beta_1 internet_{ijdt} - \beta_2 female_{ijdt} - \beta_3 (internet_{ijdt} \times female_{ijdt} -$$
$$\mathrm{X}'_i\gamma - \mathrm{Z}'_d\delta - \lambda_s - \mu_k - \tau_t - \rho_1\hat{v}_{ijdt} - \rho_2\hat{v}^{f}_{ijdt}\Big) \quad (3)$$

for outcome thresholds $s = 0,1,2$ where κ_s are estimated cut points. $\beta_1$captures the causal effect of internet adoption on firm size for male-owned firms. $\beta_3$captures the differential effect for female-owned firms relative to men, with the total internet effect for women given by $\beta_1 + \beta_3$. $\rho_1$and $\rho_2$are endogeneity correction parameters, and their signs reveal the direction of selection bias separately for male and female owners.

The second stage includes the interaction $\text{internet}_{ijdt} \times \text{female}_{ijdt}$to allow the effect of internet adoption to vary by ownership gender. The owner of the firm, female_{ijdt} is exogenous. Under exogeneity of female_{ijdt}, the endogeneity of the interaction $\text{internet}_{ijdt} \times \text{female}_{ijdt}$is entirely from internet_{ijdt} alone, and the gender-interacted residual $\hat{v}^{f}_{ijdt} = \hat{v}_{ijdt} \times \text{female}_{ijdt}$ provides a complete control function correction for both the main endogenous variable and its interaction with the exogenous moderator (Wooldridge, 2015). This follows from the control function theory for models with endogenous regressors interacted with exogenous variables. When the moderating variable is exogenous, no additional instrument or first-stage equation is required for the interaction term, since the interaction inherits its exogenous variation directly from the moderator (Wooldridge, 2015; Terza et al., 2008).

Standard errors are clustered at the district level in all the models, to control for within-district correlation because of the district-level instrument. Since the delta-method of standard errors

from the second stage does not account for first-stage estimation uncertainty, inference in the second stage is based on bootstrapped standard errors with 200 replications, resampling at the district level to preserve the clustering structure. These marginal results for the internet and females are reported in the table.

Gender-Differentiated Selection and Endogeneity

We hypothesise in the identification equation that the endogeneity in internet adoption operates differently by gender. So, we utilise the gender-interacted residual $\hat{v}_{id}^{f}$ to correct for the possibility of endogeneity. If $\rho_1$and ρ_2 are oppositely signed, it implies that unobserved characteristics that affect internet adoption are positively associated with firm size for one gender and negatively associated for the other. This would imply that OLS estimates would be biased in opposite directions for male and female-owned firms simultaneously (Wooldridge, 2015). We test the joint significance of $\rho_1$and ρ_2, and separately test the net endogeneity for women ($\rho_1 + \rho_2$) through a linear combination test (lincom). Thus, we formally characterise the nature and direction of gender-based selection bias. We discuss the results of these tests in the next section.

5. Results

5.1 Baseline results

The baseline ordered probit results are provided in Table 2. We estimate the effect of internet connection on the size variable, followed by an interaction with the owner's gender, and then, in column 3. Our results suggest that internet connectivity is positively associated with firm size transitions. The coefficients are positive and significant, indicating that with an internet connection, a firm is more likely to transition to a larger firm. This finding is consistent across developing and developed country contexts. (Rabbani& Raj, 2024). The results also indicate that females are less likely to transition to larger firms. The interaction, though, shows a different story. We observe that female-owned firms that adopt the internet are more likely to transition.

Table 2: Regression results BASELINE

	(1)	(2)	(3)
	Type of ent =0	Type of ent =1	Type of ent =2
Internet	0.727***	0.642***	0.473***
	(0.021)	(0.019)	(0.016)

Female		-0.978***	-0.685***
		(0.026)	(0.034)
Female×Internet		0.415***	0.385***
		(0.032)	(0.037)
OBC			-0.218***
			(0.012)
SC and ST			-0.378***
			(0.020)
Urban			0.350***
			(0.013)
Fixed outside household			0.717***
			(0.024)
Not fixed outside household			-0.254***
			(0.022)
Log_age			0.024***
			(0.006)
Registered			0.734***
			(0.020)
Financial obstacle			-0.079***
			(0.017)
Standard deviation of TRI			0.004
			(0.003)
Log_banks			0.012
			(0.019)
Log_population			0.049***
			(0.012)
Log_schools			-0.005
			(0.020)
Log_domelec			-0.011
			(0.015)
State effects			Y
Time effects			Y

Industry effects			Y
/cut1	1.181***	1.027***	1.727***
	(0.022)	(0.021)	(0.318)
/cut2	2.282***	2.160***	3.217***
	(0.023)	(0.023)	(0.321)
Obs.	1423925	1423925	1423925
Pseudo R^2	0.040	0.079	0.296
Clustered Standard errors are in parenthesis *** $p<0.01$, ** $p<0.05$, * $p<0.1$			

The coefficients of firm-level characteristics are consistent with our hypothesis. Marginalized social groups are less likely to transition. Firms located outside the household with fixed premises are more likely to transition compared to within-household enterprises. While firms located outside the household with a non-fixed structure are less likely to transition. We also observe that both registered and older firms are more likely to transition to larger size categories. The negative coefficient on financial obstacles indicates that firm transition increases with fewer financial obstacles (Rabbani & Raj, 2024).

5.2 Endogeneity correction:2SRI results

Next, the results of the 2SRI are presented in Table 3, along with the margins across the three outcomes in Table 4. We employ our instrument at a distance from state headquarters, which is at the district level, and collect data from the 2001 and 2011 Census data of town directories. The main drawback of this instrument is that it does not vary across districts and is measured at the district level rather than the firm level. However, we use this because no other firm- or district-level exogenous instrument is available, and the literature justifies the use of a predetermined instrument at the aggregate level. We also check the exogeneity of the instrument by regressing the instrument and its interaction with female on the dependent variable; both have insignificant coefficients, satisfying the exogeneity condition. The first-stage probit results reveal that distance from state headquarters enters negatively and significantly, indicating that firms in districts farther from the state capital are less likely to adopt the internet, consistent with the rollout of the State Wide Area Network architecture. The instrument interaction with female ownership is also significant. The joint test of instrument relevance rejects the null.

Table 3: Margins of 2SRI estimation

	(1)	(2)	(3)
	Outcome 0	Outcome 1	Outcome 2
Internet	-0.093***	0.076***	0.017***
	(0.016)	(0.012)	(0.003)
Female	0.066***	-0.059***	-0.006***
	(0.003)	(0.002)	(0.001)
Internet effect on Female	-0.149***	0.133***	0.016***
	(0.019)	(0.016)	(0.003)
OBC	0.037***	-0.029***	-0.008***
	(0.002)	(0.002)	(0.0004)
SC and ST	0.060***	-0.048***	-0.012***
	(0.003)	(0.003)	(0.001)
Urban	-0.057***	0.046***	0.011***
	(0.002)	(0.002)	(0.0004)
Fixed outside household	-0.127***	0.107***	0.020***
	(0.004)	(0.004)	(0.001)
Not fixed outside household	0.029***	-0.026***	-0.003***
	(0.002)	(0.002)	(0.0003)
Log_age	-0.004***	0.003***	0.001***
	(0.001)	(0.001)	(0.0002)
Registered	-0.132***	0.109***	0.023***
	(0.004)	(0.004)	(0.001)
Financial obstacle	0.012***	-0.009***	-0.002***
	(0.003)	(0.002)	(0.001)
Standard deviation of TRI	-0.001	0.001	0.000
	(0.000)	(0.000)	(0.0001)
Log_banks	-0.002	0.002	0.000

	(0.003)	(0.002)	(0.001)
Log_population	-0.008***	0.006***	0.002***
	(0.002)	(0.001)	(0.0004)
Log_schools	0.001	-0.001	-0.000
	(0.003)	(0.002)	(0.001)
Log_domelec	0.002	-0.001	-0.000
	(0.002)	(0.002)	(0.0005)
Observations	1423925	1423925	1423925

*Clustered standard errors clustered at the district level in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1*

The margins reflect the quantitative interpretations of the full model. We observe that firms with internet adoption are less likely to be in the lowest size category. But firms with internet adoption are more likely to be in the second and third categories. We observe that female-owned firms have a positive coefficient and have a higher probability of being an own-account enterprise. Firms with OBC, SC, and ST owners are also more likely to be in the low categories as they have a positive coefficient. Urban firms registered are less likely to be an own-account enterprise. Firms located outside the household are more likely to be in a larger category than to be own-account enterprises. Whereas firms located outside the household without a fixed enterprise are more likely to be own-account enterprises. Older firms are also less likely to be own-account enterprises. In the outcome categories of firms with 2-5 workers and 6-20 workers, we observe similar trends in both. Females have a negative coefficient, indicating that female owners are less likely to be in larger firms. The marginalised social group owners also follow a similar trend, indicating that most run as own-account enterprises. We observe registered firms in the higher enterprise category, indicating formalization acting as a tool for transition. The margins for the 2SRI are presented in Table 4, and the results follow from our baseline model. The first stage regression coefficients are provided in the appendix. When we compare the probit estimates in Table 3 with the 2SRI marginal estimates in Tables 4 and 5, we see that endogeneity correction reveals significant importance. In the baseline specification, the coefficient on internet adoption is positive and significant, and the female×internet interaction is positive, suggesting that female-owned firms benefit from digital adoption. The 2SRI-corrected internet effect for female-owned firms is substantially larger than the baseline estimate; the true effect of internet adoption on female firm size transition is larger than the baseline estimates suggest, because women who adopt the internet tend to do so, probably

adopt internet from positions of constraint rather than advantage. These findings reveal the importance of accounting for gender-differentiated selection when evaluating the returns to digital technology adoption in the informal sector.

Beyond the internet adoption effect, our results consistently reveal a substantial baseline gender gap in firm size across all specifications. Female-owned firms are significantly more likely to be own-account enterprises and significantly less likely to employ hired workers, after controlling for industry, location, registration status, and district-level infrastructure. The marginal effect of female ownership on the probability of being in the smallest firm category is positive and quite high in the bootstrapped 2SRI specification, a large and robust disadvantage that reflects the structural barriers facing female informal entrepreneurs in India documented extensively in the literature (Klapper & Parker, 2011; Field et al., 2010; Ghosh, 2018). Though we observe that internet adoption substantially reduces this gap, female internet-adopting firms are less likely to be in the smallest category than male internet-adopting firms. It does not fully eliminate the gender disadvantage at the top of the firm size distribution, where the marginal effects for men and women converge. This suggests that digital adoption is a powerful but only partial equalizer in removing some constraints. Even with internet adoption, other significant binding constraints on female firm growth persist, including access to formal capital, physical scale limitations, and socially embedded norms.

The bootstrap margins are appropriate for the 2SRI model with 200 replications[6] and the coefficients for internet, female, and the interaction between internet and female are given in Table 4. The bootstrap standard errors are used because the residual inclusion term is generated from the first-stage estimation. The estimation uncertainty from both stages is propagated via bootstrapping, yielding a valid inference for the 2SRI estimator. The bootstrap regression outcomes are depicted in Appendix Table A2.

Table 4: Margins bootstrap

	0	1	2
Internet	-0.112***	0.085***	0.027***
	(0.017)	(0.012)	(0.005)

[6] The bootstrap accounts for the fact that the control-function residuals are estimated rather than observed. We use 200 replications to obtain consistent standard errors while keeping the computation feasible given the large sample size (over 1.4 million observations)

Female	0.108***	-0.090***	-0.018***
	(0.004)	(0.003)	(0.001)
Female×Internet	-0.185***	0.163***	0.022***
	(0.026)	(0.022)	(0.004)

*Clustered standard errors clustered at the district level in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1*

5.3 Channels of internet adoption effects

To investigate the mechanisms through which internet adoption affects firm size transition, we decompose internet use into three functional channels using binary indicators of specific internet activities reported in the surveys. The market access index aggregates the use of the internet for receiving orders, placing orders, and delivery services, ranging from 0 to 3. The financial access index aggregates use for banking and financial services and ranges from 0 to 2. The communication and information index aggregates for email, telephone calls, and information gathering. For each index, a higher value indicates more intensive internet use in that channel. We interact each index with female ownership to test whether the channel-specific effects differ by gender.

Table 6 presents ordered probit marginal effects for the three internet use channels among internet-adopting firms. All three channels significantly increase the probability of crossing the hiring threshold, with the communication and information channel exhibiting the largest effect. A one-unit increase in communication use intensity reduces the probability of own-account status by 6.2 per cent, almost twice as large as the market access (3.5 per cent) and financial access (3.3 per cent) channels. This pattern is consistent with information asymmetry reduction being the primary mechanism through which the internet enables firm growth among informal enterprises (Goldfarb & Tucker, 2019)

The gender interaction results reveal that the communication channel specifically drives the gender-differentiated internet effect documented in Table 5. Female-owned firms that use the internet for communication, email, and information gathering benefit significantly more from this channel than male-owned firms. This interaction reduces the probability of own-account status beyond the male baseline. By contrast, the market-access and financial-access gender interactions are insignificant, indicating that these channels provide equal returns regardless of gender.

Table 6: Marginal outcomes of the Channel of benefits analysis

	(1)	(2)	(3)
	Outcome 0	Outcome 1	Outcome 2
Market access Index	-0.035***	0.020***	0.015***
	(0.003)	(0.002)	(0.001)
Financial access index	-0.033***	0.019***	0.014***
	(0.005)	(0.003)	(0.002)
Communication access index	-0.062***	0.036***	0.026***
	(0.003)	(0.002)	(0.002)
Female × Market access Index	0.000	-0.000	-0.000
	(0.006)	(0.003)	(0.003)
Female × Financial access index	-0.007	0.004	0.003
	(0.005)	(0.003)	(0.002)
Female × Communication access index	-0.020***	0.011***	0.008***
	(0.005)	(0.003)	(0.002)
Observations	268623	268623	268623

Standard errors in parentheses

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

These findings are consistent with the hypothesis that internet adoption benefits female entrepreneurs by relaxing information and network constraints. Female entrepreneurs in India have historically had weaker access to business networks, market information, and peer learning, especially in the informal sector. This weak access to networks is due to mobility restrictions and social norms. These constraints could be addressed through technology adoption, which directly mitigates them and reduces information asymmetries. Market access and financial service channels, by contrast, face gender-neutral constraints that internet adoption cannot selectively alleviate for women.

6. Robustness

6.1 Pseudo panel

In this section, we create an additional robustness exercise to validate our main firm-level findings and to address the cross-sectional nature of the pooled data. We construct a cohort-level pseudo-panel following Kesar (2023) and Posti and Maiti (2023). This approach

aggregates firm-level observations into cohort-round cells defined by state × industry × age band × gender, and estimates a within-cohort panel regression with cohort and survey-round fixed effects following Deaton (1985). This approach does not impose stationarity assumptions on the joint distribution of outcomes across rounds, given the structural disruptions in the economy over the sample period (2015-16- 2023-24), which include demonetization (2016), GST implementation (2017), and the COVID-19 pandemic (2020-21). Given the nature of the data, we define cohorts as groups of firms that share time-invariant characteristics consistently identified across survey rounds. We define cohorts along four dimensions: state, NIC code (two-digit), gender of the owner and firm age band (0-5 years, 6-10 years, 11-20 years and above 20 years). Unlike previous studies on Indian informal firms, which rely on OLS estimation, we additionally use cohort-level internet adoption data from Bharatnet Gram panchayat data, thereby providing evidence of the internet effect on firm transition within the pseudo-panel framework.

For each cohort-round cell, we compute weighted shares for the hired-worker firms and the explanatory variables. The dependent variable is the share of firms in the cohort with hired workers, capturing the proportion of firms that have hired workers (at least one). The main independent variable is the share of internet adopters in the cohort. The other firm characteristics are constructed similarly as weighted shares of the corresponding firm-level characteristics within each cohort-round cell.

We estimate linear panel regression of the form:

$$share_hired_{crt} = \alpha + \beta\, share_internet_{crt} + X'_{crt}\gamma + W'_{st}\sigma + \mu_c + \tau_t + \epsilon_{crt}$$

Where c indexes cohorts, r indexes the gender dimension within cohorts (male or female), t indexes survey rounds, μ_c are cohort fixed effects absorbing all time-invariant cohort-level characteristics, and τ_t are round fixed effects capturing common time trends. X_{crt} is a vector of cohort-level weighted shares of firm characteristics, including premise type, registration status, location, owner's social group, financial obstacles and log of mean age. W_{st} is a vector of state-year level controls, including per capita power availability, human development index, crime rate, and social sector expenditure. Standard errors are clustered at the cohort level. Since the cohort is within the state our controls are also at the state level captured by W_{st} .

One thing to note is that in the pseudo-panel framework, we use a state-level instrument (the Gram Panchayats covered under the BharatNet programme). The BharatNet programme was

India's rural broadband initiative that connected gram panchayats to an optical fibre network in 2015 onward. We use data from the month-year preceding the survey start date[7].

The Table shows the pseudo-panel analysis. Column 1 shows the OLS estimates and confirms a positive within-cohort association between internet adoption and the share of firms that hire workers. Column 2 shows that the pooled IV estimate is larger and confirms that the OLS estimates underestimate the effect. The instrument is strong and meets the instrument-relevance tests.

Columns 3 and 4 reveal a finding that extends our main results. For male cohorts, the instrument is strong, and the IV estimate is significant, confirming the internet effect for male-owned firms. For female cohorts, the BharatNet gram panchayat coverage does not predict internet adoption among female-owned cohorts. This reveals that internet adoption is driven by the supply-side infrastructure of connectivity expansion. The instrument variable, BharatNet expansion, identifies a Local Average Treatment Effect for compliers, i.e., the firms that adopt the internet because their gram panchayat (in that state) received connectivity. Female entrepreneurs do not follow this pattern. They face demand barriers to adoption, digital literacy gaps, device ownership constraints, and social norms restricting independent technology use that physical connectivity alone cannot overcome. The firm-level 2SRI utilises the distance to state headquarters as an instrument to identify the effect for firms whose internet adoption is driven by variation in local telecommunications infrastructure quality.

Table 7 : Synthetic Panel Internet Effect on Firm Size by Gender

[7] So for the survey of 2015-16 we take the data for March, 2015; for the survey of 2021-22 we take the data available on March, 2021 and so on.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	ols_cohort	iv_cohort	ols_female	iv_female	ols_male	iv_male
Internet	0.112***	0.601***	0.110***	0.862	0.110***	0.593**
	(0.0136)	(0.226)	(0.0275)	(0.978)	(0.0157)	(0.239)
OBC owner	-0.0314**	-0.0255*	-0.0283	-0.00421	-0.0336**	-0.0321*
	(0.0127)	(0.0143)	(0.0200)	(0.0401)	(0.0161)	(0.0175)
SC & ST owner	-0.0693***	-0.0509**	-0.0511**	-0.0273	-0.0805***	-0.0584**
	(0.0170)	(0.0205)	(0.0246)	(0.0433)	(0.0231)	(0.0274)
Urban	0.0474***	0.0295*	0.0514***	0.0258	0.0461***	0.0270
	(0.0115)	(0.0151)	(0.0177)	(0.0382)	(0.0149)	(0.0192)
Premise	0.235***	0.190***	0.229***	0.167*	0.233***	0.189***
	(0.0142)	(0.0257)	(0.0265)	(0.0867)	(0.0170)	(0.0282)
Registration	0.196***	0.110**	0.215***	0.0598	0.191***	0.114***
	(0.0133)	(0.0429)	(0.0270)	(0.206)	(0.0153)	(0.0415)
Financial obstacle	0.00228	-0.00593	-0.0212	-0.0147	0.0138	0.00162
	(0.0212)	(0.0233)	(0.0360)	(0.0463)	(0.0258)	(0.0281)
Log age	0.0145	0.0331	0.0183	0.0451	0.0124	0.0322
	(0.0255)	(0.0293)	(0.0418)	(0.0626)	(0.0319)	(0.0361)
HDI	-1.149***	-1.023***	-2.127***	-1.271	-0.895**	-0.874**
	(0.352)	(0.381)	(0.784)	(1.451)	(0.392)	(0.418)
Per capita power	-0.0139***	-0.0253***	-0.00508	-0.0161	-0.0174***	-0.0297***
	(0.00296)	(0.00629)	(0.00560)	(0.0160)	(0.00348)	(0.00725)
Crime rate	0.00222**	-0.000201	0.00568***	0.00287	0.000656	-0.00192
	(0.000964)	(0.00149)	(0.00190)	(0.00418)	(0.00110)	(0.00170)
Social sector expenditure	-0.0444	-0.105	1.677	-1.853	-0.746	-0.267
	(1.173)	(1.278)	(2.250)	(5.374)	(1.367)	(1.487)
Constant	1.048***		1.563***		0.946***	
	(0.263)		(0.569)		(0.299)	
Observations	16,663	16,663	4,650	4,650	12,013	12,013

R-squared	0.703	-0.021	0.695	-0.233	0.682	-0.023
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
Round FE	Yes	Yes	Yes	Yes	Yes	Yes
Cluster	Cohort	Cohort	Cohort	Cohort	Cohort	Cohort
Instrument		BharatNet GP coverage		BharatNet GP coverage		BharatNet GP coverage

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The findings complement the main results by reflecting the requirement of specific interventions for female entrepreneurs to reach the adoption margins. Supply-side intervention generates male adoption whereas female adoption requires additional demand-side interventions. The policy implication is that the BharatNet programme, although it increases overall internet return but could narrow the gender gap if accompanied by digital literacy programmes, subsidized access to devices, and interventions addressing social norms around technology use. The negative coefficients on per capita power availability and HDI reflect a well-documented pattern in India's structural transformation. The informal sector firms in developing countries tend to concentrate at the subsistence own-account scale as growing firms transition to formal status, reducing the share of hired-worker firms in the informal sector as states develop over time. These coefficients capture the compositional effect of development on the informal firm size distribution. This compositional effect does not reflect a direct negative effect of infrastructure or human development on firm hiring decisions.

7. Conclusion

This paper examines whether internet adoption facilitates firm-size transitions among informal manufacturing enterprises in India and whether these gains differ by gender. Using five rounds of NSSO and ASUSE enterprise surveys, we find that internet adoption is significantly associated with the increase in the probability of firms crossing the hiring threshold from own-account enterprises to employer firms. These gains are substantially larger for female-owned firms, suggesting that digital technologies are associated with gender-specific barriers to firm growth. We also find evidence of gender-differentiated selection into internet adoption, implying that conventional estimates overstate the returns for male-owned firms while understating them for female-owned firms. Mechanism analysis further shows that communication and information use are the primary channels through which internet adoption promotes firm growth, especially for women.

Our cohort-level pseudo-panel analysis provides additional support to the firm-level findings. The results confirm a positive association between internet adoption and the share of worker enterprises within cohorts. However, while BharatNet expansion induces internet adoption among male-owned cohorts, it does not do so for female-owned cohorts, suggesting that broadband infrastructure alone is insufficient to encourage adoption among women entrepreneurs. This highlights the importance of additional demand-side interventions, such as digital literacy, affordable access to devices, and measures to address social barriers to technology use.

This study suffers from a few limitations. Firstly, the cross-sectional nature of the pooled data precludes tracking individual firm trajectories over time. As explained in the introduction, when this paper refers to firm transition, it describes changes in the structural composition of the firm size distribution over the sample period rather than confirmed transitions of individual enterprises from smaller to larger size categories. The compositional changes could also arise from firm growth, differential entry and exit across size categories, or changes in the survey sample composition across rounds. While our fixed effects and controls mitigate some of these concerns, the absence of a true panel dimension remains an inherent limitation of the data. The instrument captures the distance of a district from its state headquarters, which varies across districts rather than firms. The identification, therefore, relies on between-district variation within states rather than firm-level exogenous variation, and standard errors are clustered at the district level to account for this. It is quite common in the literature to use geographic instruments, which implies that our estimates capture the average treatment effect among firms in districts with varying levels of telecommunications infrastructure rather than a firm-specific causal effect. Third, the mechanism analysis was estimated on the subsample of internet-adopting firms and treats the type of internet use as exogenous. Since the channel through which a firm uses the internet is itself potentially endogenous, firms that use the internet for communication may differ systematically from those that use it for market access in ways that also affect firm size. The channel decomposition should be interpreted as suggestive evidence of a mechanism rather than causal identification of channel-specific effects.

Overall, the findings suggest that digital infrastructure investments can promote structural upgrading in India's informal sector. But the extent of this potential depends on addressing the gender-specific constraints that continue to limit women's digital adoption. Future research using longitudinal enterprise data could examine whether these gains persist as firms expand

beyond the initial hiring threshold and whether complementary interventions further strengthen the growth effects of digital adoption.

References

Alok, S., Ghosh, P., Kulkarni, N., & Puri, M. (2024). Open Banking and Digital Payments: Implications for Credit Access. *NBER Working Paper 33259*. National Bureau of Economic Research. https://doi.org/10.3386/w33259

Banerjee, A. V., & Duflo, E. (2014). Do Firms Want to Borrow More? Testing Credit Constraints Using a Directed Lending Program. *The Review of Economic Studies*, *81*(2), 572–607. https://doi.org/10.1093/restud/rdt046

Bruhn, M. (2009). Female-Owned Firms in Latin America Characteristics, Performance, and Obstacles to Growth. *World Bank Policy Research Working Paper No. 5122*. The World Bank. https://documents1.worldbank.org/curated/en/491141468045040504/pdf/WPS5122.pdf

Campos, F., & Gassier, M. (2024). Gender and Enterprise Development in Sub-Saharan Africa: A Review of Constraints and Effective Interventions. *Policy Research Working Paper;No. 8239.* . World Bank. https://openknowledge.worldbank.org/entities/publication/ad839540-cd59-5452-80db-5aea8c2593b3

Chakraborty, S., & Chatterjee, P. (2021). Women entrepreneurs in India: Where do they stand? *The Indian Journal of Labour Economics*, *64*(4), 1069–1092. https://doi.org/10.1007/s41027-021-00344-8

Czernich, N., Falck, O., Kretschmer, T., & Woessmann, L. (2011). Broadband infrastructure and economic growth. *The Economic Journal*, *121*(552), 505–532. https://doi.org/10.1111/j.1468-0297.2011.02420.x

Deaton, A. (2002). Panel data from time series of cross-sections. *Journal of Econometrics*, *30*(1–2), 109–126. https://doi.org/10.1016/0304-4076(85)90134-4

DeStefano, T., Kneller, R., & Timmis, J. (2018). Broadband Infrastructure, ICT Use and Firm Performance: Evidence for UK Firms. *Journal of Economic Behavior & Organization*, *155*, 110–139. https://doi.org/10.1016/j.jebo.2018.08.020

DeStefano, T., Kneller, R., & Timmis, J. (2022). The (Fuzzy) Digital Divide: The Effect of Universal Broadband on Firm Performance*. *Journal of Economic Geography*, *23*(1), 139–177. https://academic.oup.com/joeg/article/23/1/139/6573380

Falck, O., Gold, R., & Heblich, S. (2014). E-lections: Voting Behavior and the Internet. *American Economic Review*, *104*(7), 2238–2265. https://doi.org/10.1257/aer.104.7.2238

Field, E., Jayachandran, S., & Pande, R. (2010). Do traditional institutions constrain female entrepreneurship? A field experiment on business training in India. *American Economic Review*, *100*(2), 125–129. https://doi.org/10.1257/aer.100.2.125

Ghani, E., Kerr, W. R., & O'Connell, S. (2013). Spatial Determinants of Entrepreneurship in India. *Regional Studies*, *48*(6), 1071–1089. https://doi.org/10.1080/00343404.2013.839869

Goldfarb, A., & Tucker, C. (2019). Digital Economics. *Journal of Economic Literature*, *57*(1), 3–43. https://doi.org/10.1257/jel.20171452

Hsieh, C.-T., & Olken, B. A. (2014). The missing "missing middle." *Journal of Economic Perspectives*, *28*(3), 89–108. https://doi.org/10.1257/jep.28.3.89

International Labour Organization. (2018). Women and Men in the Informal Economy: A Statistical Picture. Third Edition. In *www.ilo.org*. International Labour Organization. https://www.ilo.org/publications/women-and-men-informal-economy-statistical-picture-third-edition

Islam, A. M., & Amin, M. (2023). The Gender Labor Productivity Gap Across Informal Firms. *World Development*, *167*(106229). https://doi.org/10.1016/j.worlddev.2023.106229

Kesar, S. (2024). Subcontracting linkages in India's informal economy. *Development and Change*, *55*(1), 38–75. https://doi.org/10.1111/dech.12817

Klapper, L. F., & Parker, S. C. (2010). Gender and the Business Environment for New Firm Creation. *The World Bank Research Observer*, *26*(2), 237–257. https://doi.org/10.1093/wbro/lkp032

Lokesh Posti, & Abhradeep Maiti. (2023). Firm size and productivity in the informal sector: Evidence from India. *The Indian Economic Journal*, *73*(3). https://doi.org/10.1177/00194662231202088

Mehrotra, S., & Giri, T. (2023). Enterprise informality in India: The Blind Spots in Public Policy. *The Indian Journal of Labour Economics*, *66*, 687–710 . https://doi.org/10.1007/s41027-023-00450-9

Rabbani, G., & Raj S. N., R. (2023). Small Firms in Indian Manufacturing: Some Stylized Facts. *Journal of the International Council for Small Business*, *4*(4), 374–386. https://doi.org/10.1080/26437015.2023.2190048

Rabbani, G., & Rajesh, N. (2024). The role of financial constraints in firm transition—Evidence from Indian manufacturing. *Journal of International Development*, *36*(4), 2193–2228. https://doi.org/10.1002/jid.3898

Raj S.N, R., & Sen, K. (2016). Out of the Shadows?: The informal sector in post-reform India. In *OUP Academic*. Oxford University Press. https://academic.oup.com/book/25518

Sabarwal, & K. Terrell. (2008). Does Gender Matter for Firm Performance? Evidence from Eastern Europe and Central Asia. *IZA Discussion Papers No. 3758*, Article 3758. Institute for the Study of Labor (IZA), Bonn. https://nbn-resolving.de/urn:nbn:de:101:1-20081126115

Terza, J. V. (2017). Two-stage Residual Inclusion Estimation: A Practitioners Guide to Stata Implementation. *The Stata Journal: Promoting Communications on Statistics and Stata*, *17*(4), 916–938. https://doi.org/10.1177/1536867x1801700409

Terza, J. V., Basu, A., & Rathouz, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, *27*(3), 531–543. https://doi.org/10.1016/j.jhealeco.2007.09.009

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data. In *MIT Press*. https://mitpress.mit.edu/9780262232586/econometric-analysis-of-cross-section-and-panel-data/

Appendix

Table A1: First stage Probit estimation

	(1) internet
Female	-0.322***
	(0.042)
OBC	-0.144***
	(0.021)
SC/ST	-0.191***
	(0.02)
Urban	0.255***
	(0.027)
Outside household fixed	0.291***
	(0.019)
Outside household not fixed	-0.213***
	(0.024)
Log_age	-0.12***
	(0.008)
Registration	0.593***
	(0.024)
Financial obstacle	0.171***
	(0.039)
Tri_std	-0.007
	(0.008)
Log_banks	0.094
	(0.072)
Log_population	-0.004
	(0.035)
Log_schools	-0.017
	(0.04)
Log_domelec	-0.033
	(0.05)
State_HQ distance	-0.001***
	(0.00014)
Female#State_HQ distance	0.000273*
	(0.00)

Constant	-3.446***
	(0.604)
Observations	1500623
Pseudo R^2	0.351

Standard errors are in parentheses

**** p<.01, ** p<.05, * p<.1*

Table A2 : 2SRI Ordered Probit Bootstrap SE (200 reps)

VARIABLES	(1) type_ent_cat
Internet	0.359***
	(0.0599)
Female	-0.738***
	(0.0258)
1.internet#1.female	0.773***
	(0.0786)
OBC	-0.225***
	(0.0110)
SC and ST	-0.429***
	(0.0165)
Urban	0.0660***
	(0.0121)
Fixed outside hh	0.739***
	(0.0156)
Not fixed outside hh	-0.280***
	(0.0217)
Log age	0.0302***
	(0.00480)
Registered	0.751***
	(0.0160)
Financial obstacle	-

	0.0666***
	(0.0126)
Standard deviation of TRI in polygon	0.000837
	(0.00229)
Log banks	0.00959
	(0.0193)
Log population	0.0230*
	(0.0139)
Log schools	0.0158
	(0.0231)
Log domestic electricity	-0.0155
	(0.0155)
nu_hat_b	0.188***
	(0.0583)
nu_hat_female_b	-0.640***
	(0.0765)
/cut1	0.806***
	(0.284)
/cut2	2.460***
	(0.286)
Observations	1,423,925
State FE	Yes
Industry FE	Yes
Year FE	Yes
District Controls	Yes

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1