ONLINE APPENDIX

Finance, gender, and entrepreneurship

India's informal sector firms

Ira N. Gang,[1] Rajesh Raj Natarajan,[2] and Kunal Sen[3]

November 2020

WIDER Working Paper 2020/144

Appendix to Section 3.2: Variables and descriptive statistics

Table A1: Summary statistics

Variable	Observations	Mean	SD	Min.	Max.
Entrepreneurial firm (E)	584,055	0.38631	0.48690	0	1
FIN1	584,055	0.06914	0.25369	0	1
FIN2	64,007	0.46959	0.48991	0	1
FIN3 DUM1	584,055	0.05354	0.22511	0	1
FIN3 DUM2	584,055	0.06156	0.24035	0	1
FIN3 DUM3	584,055	0.00468	0.06828	0	1
Female	584,055	0.12908	0.33529	0	1
Location	584,055	0.51206	0.49986	0	1
ST	584,055	0.05574	0.22943	0	1
SC	584,055	0.10004	0.30005	0	1
OBC	584,055	0.46240	0.49858	0	1
Age3–9	584,055	0.44885	0.49738	0	1
Age>9	584,055	0.42633	0.49454	0	1
Asst	584,055	0.01196	0.10870	0	1
Regis	584,055	0.40060	0.49002	0	1
lnLP	584,055	10.01993	0.98134	2.10843	15.9002

Source: authors' construction based on own estimates.

Firm characteristics as control variables (other variables are described in the main text)

Urban: This variable stands for the location of the firm, assuming the value 1 for urban firms and 0 for rural firms. Urban firms generally grow faster than their rural counterparts do, as their market is larger, and they have better access to infrastructure and inputs.

[1] Economics Department, Rutgers University, New Brunswick, USA; [2] Economics Department, Sikkim University, Sikkim, India, corresponding author: rajeshraj.natarajan@gmail.com; [3] UNU-WIDER, Helsinki, Finland, and University of Manchester, Manchester, UK.

Social group: Under the possibility that the socioeconomic group to which owners belong will affect firms' ability to transition, we capture the firm owner's social group using three dummy variables, for Other Backward Caste (OBC), Scheduled Caste (SC), and Scheduled Tribe (ST) membership. 'Others' is the omitted group.

Age: The age of the firm is measured as the number of years elapsed since its establishment. We classify firms into three age categories, those from zero to two years old (Age0–2), those from three to nine years old (Age3–9), and those active for ten years or more (Age>9). We introduce two dummy variables, for Age3–9 and Age>9, with Age0–2 as the reference category.

Assistance: This variable takes the value 1 for firms that have received government assistance towards training and marketing. Assistance may affect firm hiring within the informal sector.

Registration: This variable takes the value 1 if firms have registered under any Act and 0 if they have not. Registration itself may form a type of collateral, providing the firm with access to otherwise unavailable financial and non-financial resources (Levenson and Maloney 1998; Sharma 2014). The unorganized sector firms we are considering are not registered under the Factories Act of 1948 (see footnote 3 in the main text). However, they may be registered with state and municipal governments and co-operative authorities.

Labour productivity: Labour productivity is measured as the ratio of gross value added to employment. We follow the existing literature in including labour productivity as an additional control, as more productive firms are more likely to be entrepreneurial firms (Raj and Sen 2016). We use the log of labour productivity in regression estimations.

References in above section

Levenson, A.R., and W.F. Maloney (1998). *The Informal Sector, Firm Dynamics and Institutional Participation*. Washington, DC: The World Bank. https://doi.org/10.1596/1813-9450-1988

Sharma, S. (2014). 'Benefits of a Registration Policy for Microenterprise Performance in India'. *Small Business Economics*, 42: 153–64. https:/doi.org/10.1007/s11187-013-9475-y

Raj, R.S., and K. Sen (2016). 'Moving Out of the Bottom of the Economy? Constraints to Firm Transition in the Indian Informal Manufacturing Sector'. *IZA Journal of Labor & Development*, 5(1). https://doi.org/10.1186/s40175-016-0056-8

Appendix to Section 5: Robustness check using PSM-DID

For our DID strategy, we take advantage of the rule that regulators employed to select the set of under-banked districts under the 2005 reform of bank branch licensing in India. This rule compares the average number of persons per branch in a district against a statistic termed the 'national average' of population per branch for India (RBI 2009; Young 2019). This national average acts as a threshold, and the districts whose populations per branch exceeds this threshold receive treatment while others do not. The reform led to additional branch expansion over 2010–15 in some districts that were initially under-banked in 2010/11, so that they became banked by 2015/16. Our empirical strategy lies in examining whether informal firms show a greater propensity to become entrepreneurial in the districts that received the treatment, that is, the districts that changed status from under-banked to banked during the period of our analysis, relative to untreated districts, which remain under-banked throughout the period of our analysis.[1] In the period 2010–15, 22 districts changed from under-banked to banked, while 333 districts remained under-banked. Our unit of analysis is the firm, and we compare the entrepreneurship status of firms in the 22 treated districts relative to the 333 control districts.

A limitation of our DID strategy is that we do not have data on firm entrepreneurship status at the district level prior to 2010. Because of this we cannot explicitly test for parallel trends, which is a crucial assumption behind the validity of a DID estimation strategy. In other words, we cannot test for the assumption that the untreated districts provide the appropriate counterfactual of the trend that treated districts would have followed if they had not been treated. For example, control districts may differ significantly from treated districts in many important characteristics that are themselves correlated with why a particular district was treated. In order to guard against the possibility of the violation of the parallel trend assumption, we use PSM to construct a set of control districts that can be matched with treated districts in observable characteristics.

We therefore combine PSM and DID to estimate the causal impact of financial access on entrepreneurship. In the first stage, we employ PSM to construct matched control and treated districts, as the baseline. In the second stage, we apply the DID method in the matched data to estimate the impact of financial access on entrepreneurship. We follow the steps outlined in Unnikrishnan and Imai (2020).

First step: PSM

The PSM method matches the treated group of enterprises with the control group of enterprises based on observable characteristics. The intuition behind this method is to arrive at a control group of enterprises that were not exposed to the treatment whose observable characteristics are similar to those of the treated group. We match the enterprises based on the binary variable on the status of the districts as under-banked or not in the baseline period. In other words, we match the units of observation based on whether they are located in banked or under-banked districts. We then construct the propensity score based on the covariates that determine the treatment and also simultaneously affect the outcome (in our case, entrepreneurship).

Following Imbens and Rubin (2015), we estimate the propensity score by employing a logit regression. To do this, we first construct a binary variable for treatment, which takes only the

[1] As a further robustness check, we also compare the outcome changes in districts that received the treatment prior to the study period, that is, those districts that remained 'banked', and districts that received the treatment during the study period.

values 0 (for control group) or 1 (for treatment group). We then estimate the propensity scores as the fitted values that are derived from a logit estimation, with the binary treatment variable as the dependent variable and the covariates that are supposed to ensure balance between control and treatment groups as regressors.[2] As covariates, we include district-level variables that are likely to explain why a district is under-banked at the baseline and also to be correlated with entrepreneurship. These variables include proportion of villages that are on a bus route out of total inhabited villages, proportion of villages with electricity out of total inhabited villages, proportion of villages with a post and telegraph office out of total inhabited villages, proportion of villages with paved approach road out of total inhabited villages, proportion of villages with a primary school out of total inhabited villages, proportion of Scheduled Caste households out of total households, and proportion of Scheduled Tribe households out of total households.

The propensity scores are then used to match the control and treatment group enterprises. The key objective of the matching exercise is to find appropriate control group enterprises for treatment group enterprises. For matching, we use the kernel-matching algorithm, which employs weighted averages of all firms in the control group to build the counterfactual group to pair treatment with control firms.[3] As mentioned earlier, the matching is performed for two subsamples of firms: one subsample that includes firms in treated districts and firms in the districts that are under-banked throughout the study period, and another subsample that includes firms in treated districts and firms in the districts that were banked prior to the study period. Once the matching is done, we move to the second stage to disentangle the effect of bank branch penetration on entrepreneurship.

Second step: DID

We apply a version of the DID model to understand the effect of this policy change on entrepreneurship. We compare the firms in districts which remained under-banked (henceforth 'untreated') and those in districts which benefitted from the policy change over the period 2010/11–2015/16 (henceforth 'treated'). In the second stage, our estimation is confined to enterprises in the matched districts. Unlike in the typical DID settings, we lack the baseline data with untreated firms, as there were already both treated and untreated firms in our baseline year, 2010/11. To reduce any sample selection bias and attrition bias influencing our core findings, we follow the strategy employed by Unnikrishnan and Imai (2020). We use propensity scores (PS) as weights in regressions so that the regressions reflect the probability of firms being treated in 2010/11 and 2015/16 as different firms exhibit different probabilities of getting treated. This is certainly not a perfect strategy to eliminate selection bias given that the PS depends on the specification and the results of the probit model. However, we believe that the PS-weighted DID should yield a robust estimate given the data constraints. The generic model we estimate takes the following form:

$$E_{idt} = \beta_0 + \beta_1 U_{idt} + \beta_2 T_{idt} + \beta_3 U * T_{idt} + \beta_4 X_{idt} + \varepsilon_{idt} \quad (5)$$

where E is entrepreneurship and X is a vector of individual-level controls. The subscripts i, d, and t stand respectively for enterprise, district, and time. In our study, t equals 0 for pre-treatment and 1 for post-treatment. U is a dummy variable taking the value 0 for under-banked districts and 1 for other districts. T is a dummy variable that takes the value 1 if t equals 1 and 0 otherwise. The

[2] These fitted values would lie between 0 and 1.

[3] We tried this with different kernel-matching algorithms with different bandwidths and trimming levels to arrive at an ideal estimation model for this study.

interaction term of U and T identifies the effect of policy change on E_{idt}. The coefficient of the interaction term, β_3, therefore, yields a DID estimate that captures the effect of the programme change on the outcome variables.

We also confirm the robustness of our results by combining PSM with a DID strategy. The version of the DID model we use in this study helps us to address the bias resulting from self-selection and confounding. The first step of this method is to employ PSM to match the firms in the comparison group to similar firms in the treatment group. As discussed in the methodology, we used a kernel-matching algorithm to match each firm in the treatment group with firms in the control group. The estimates of DID are likely to be biased if outcomes in already banked districts or already under-banked districts are trending differently from outcomes in districts that witnessed a change in their status from under-banked to banked during the period under study.

The district-level variables that we use in the PSM are shown in Table 2 and include pre-intervention measures of infrastructure and human capital variables. There are eight such variables, namely SHSCPOP, SHSTPOP, MIDGRADEDU, ROADVILLG, ELECVILLG, POSTVILLG, BUSVILLG, and PRIMSCHVILLG. SHSCPOP and SHSTPOP represent the proportion of Scheduled Castes and of Scheduled Tribes in total population, respectively. MIDGRADEDU stands for the proportion of individuals educated to the secondary level and above and ROADVILLG represents the share of villages with paved approach roads in total villages. ELECVILLG, POSTVILLG, BUSVILLG, and PRIMSCHVILLG represent proportion of electrified villages, proportion of villages with post and telegraph offices, proportion of villages situated on a bus route, and proportion of villages with at least a primary school, respectively.

We then apply PS-weighted DID to the matched sample. Following Imbens (2000) and Hirano and Imbens (2001), we use the inverse of propensity scores as weights in the estimations.[4] The results of our DID estimations are presented in Table A3.[5] As mentioned earlier, we applied this method to two matched subsamples of firms. Our results clearly point to the positive effect of policy change on entrepreneurship. To be specific, we find that the probability of household firms becoming non-household firms is greater in districts that received the treatment during the period under study.

[4] Our results are also robust to alternative weighting schemes. For instance, we follow the weighting procedure proposed by Nichols (2007), where we weight the untreated subjects by $pi/(1 - pi)$ and treated ones by 1 (Table A4). We also carry out the estimation without weights, where we just use the matched districts from the first round for the second round (Table A5).

[5] Means of the covariates used in the first-stage PSM estimation for the treated and untreated before and after matching are presented in Table A2.

Table A2: Means of the covariates for the treated and untreated before and after matching

Variable			Before			After
	Treated	Untreated	StdDif	Treated	Untreated	StdDif
SHSCPOP	0.175616	0.168682	0.095617	0.16124	0.170671	−0.13005
SHSTPOP	0.088978	0.121001	−0.1945	0.143007	0.11887	0.146608
MIDGRADEDU	0.181736	0.116333	1.238938	0.126688	0.113365	0.252375
ROADVILLG	0.748775	0.553677	0.841523	0.606586	0.54467	0.267062
ELECVILLG	0.931218	0.755992	0.809186	0.874246	0.747923	0.583353
POSTVILLG	0.648715	0.427013	0.966474	0.505132	0.417673	0.381264
BUSVILLG	0.655319	0.387542	0.8585	0.529793	0.374197	0.498845
PRIMSCHVILLG	0.892691	0.830285	0.451128	0.850107	0.826545	0.170326

Note: StdDif stands for standardized difference between the treated and the untreated.

Source: authors' construction based on own estimates.

Table A3: Branch expansion and entrepreneurship: PSM-DID using PS weights

Variable	Group 1: unbanked versus unbanked to banked			Group 2: banked versus unbanked to banked		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Banked	−0.040	−0.126***	−0.159	−0.138***	−0.050	0.967***
	(0.040)	(0.047)	(0.132)	(0.030)	(0.035)	(0.133)
Time	0.444***	0.328***	0.354***	0.433***	0.219***	0.221***
	(0.008)	(0.014)	(0.014)	(0.010)	(0.013)	(0.013)
Banked*Time	0.257***	0.141**	0.009	0.098**	0.166***	0.175***
	(0.067)	(0.074)	(0.067)	(0.043)	(0.049)	(0.050)
Controls	N	Y	Y	N	Y	Y
District FE	N	N	Y	N	N	Y
Observations	290,642	278,456	278,456	195,743	183,278	183,278

Note: controls include gender, location, dummies for social group (ST, SC, and OBC), age categories (Age3to9 and Age>9), assistance, registration status, and labour productivity; we use the inverse of propensity scores as weights in the DID estimations.

Source: authors' construction based on own estimates.

Table A4: Branch expansion and entrepreneurship: PSM-DID using PS weights

Variable	Group 1: unbanked versus unbanked to banked			Group 2: banked versus unbanked to banked		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Banked	0.036	0.010	−0.226*	−0.068**	−0.082**	1.005***
	(0.029)	(0.034)	(0.121)	(0.029)	(0.034)	(0.135)
Time	0.392***	0.199***	0.290***	0.395***	0.164***	0.173***
	(0.015)	(0.018)	(0.019)	(0.015)	(0.019)	(0.020)
Banked*Time	0.124***	0.119***	0.075	0.121***	0.159***	0.196***
	(0.040)	(0.047)	(0.049)	(0.04)	(0.047)	(0.049)
Controls	N	Y	Y	N	Y	Y
District FE	N	N	Y	N	N	Y
Observations	290,642	278,456	278,456	195,743	183,278	183,278

Note: controls include gender, location, dummies for social group (ST, SC, and OBC), age categories (Age3to9 and Age>9), assistance, registration status, and labour productivity; we use the inverse of propensity scores as weights for untreated observations ($pi/(1 - pi)$) and 1 for treated observations.

Source: authors' construction based on own estimates.

Table A5: Branch expansion and entrepreneurship: PSM-DID

Variable	Group 1: unbanked versus unbanked to banked			Group 2: banked versus unbanked to banked		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Banked	0.181***	0.128***	−0.992***	−0.162***	−0.107***	0.689***
	(0.027)	(0.030)	(0.113)	(0.027)	(0.030)	(0.126)
Time	0.446***	0.287***	0.295***	0.435***	0.181***	0.201***
	(0.008)	(0.009)	(0.009)	(0.009)	(0.011)	(0.011)
Banked*Time	0.070*	0.085**	0.131***	0.080**	0.170***	0.215***
	(0.038)	(0.042)	(0.044)	(0.039)	(0.042)	(0.044)
Controls	N	Y	Y	N	Y	Y
District FE	N	N	Y	N	N	Y
Observations	290,642	278,456	278,456	195,743	183,278	183,278

Note: controls include gender, location, dummies for social group (ST, SC, and OBC), age categories (Age3to9 and Age>9), assistance, registration status, and labour productivity; we use the matched districts obtained from the first year for the second year.

Source: authors' construction based on own estimates.

Reference in the above section

Hirano, K., and G. Imbens (2001). 'Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization'. *Health Services and Outcomes Research Methodology*, 2(3): 259–78. https://doi.org/10.1023/A:1020371312283

Imbens, G. (2000). 'The Role of the Propensity Score in Estimating Dose-Response Functions'. *Biometrika* 87(3): 706–10. https://doi.org/10.1093/biomet/87.3.706

Imbens, G., and D. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Nichols, A. (2007). 'Causal Inference with Observational Data'. *Stata Journal*, 7: 507–41. https://doi.org/10.1177/1536867X0800700403

RBI (Reserve Bank of India) (2009). *Report of the Group to Review Branch Authorisation Policy*. Mumbai: RBI.

Unnikrishnan, V., and K.S. Imai (2020). 'Does the Old-Age Pension Scheme Improve Household Welfare? Evidence from India'. *World Development*, 134: 105017. https://doi.org/10.1016/j.worlddev.2020.105017

Young, N. (2019). 'Banking and Growth: Evidence from a Regression Discontinuity Analysis: Online Appendix'. Available at: https://nateyoungecon.com/wp-content/uploads/2019/01/NYoung_Banking_and_Growth.pdf (accessed 13 December 2019).