INTERNET APPENDIX

Road to Stock Market Participation

Sumit Agarwal, Meghana Ayyagari, Yuxi Cheng, and Pulak Ghosh

This document provides additional data descriptions and robustness tests. Figure A1 plots the time trends of road impacts across different development levels. Table A1 presents variable definition and sources of the main variables used in the paper. Table A2 presents additional summary statistics. Tables A3-A9 present additional robustness tests as described in the main manuscript. In Table A3, we show that roads improve market access regardless of literacy levels; Table A4 shows that the Connect dummy is associated with both the probability of a new state bank branch opening and number of new state bank branches opened; Table A5 shows the risk sharing channel; Table A6 presents placebo regressions; Table A7 presents a staggered DiD regression using the Callaway and Sant'Anna [2021] estimator and Figure A2 plots the associted dynamic effects; Table A8 shows that the baseline findings remain robust under Poisson regression; and Table A9 shows that our results are not driven by marginal investors experiencing a positive wealth shock, as proxied by rainfall shocks.

Figure A1: **Roads and Stock Market Participation- Time Trends by Regional Economic Development**

This figure illustrates the dynamic effects across different regions, estimated using the following equation::

$$Log(Y_{i,t}) = \sum_{k \in (-5,5)} \gamma_k \cdot D_k + \xi_i + \kappa_t + \varepsilon_{i,t} \tag{10}$$

where D_k are a series of indicator variables for each year relative to road construction. Following Borusyak et al. [2024] and Adukia et al. [2020], we omit two coefficients ($k = 1$ and $k = 5$) as reference groups since all pincodes in our sample eventually receive treatment. y-axis is the estimated coefficient, while x-axis represents the year before and after road construction with 0 being the year of road completion.

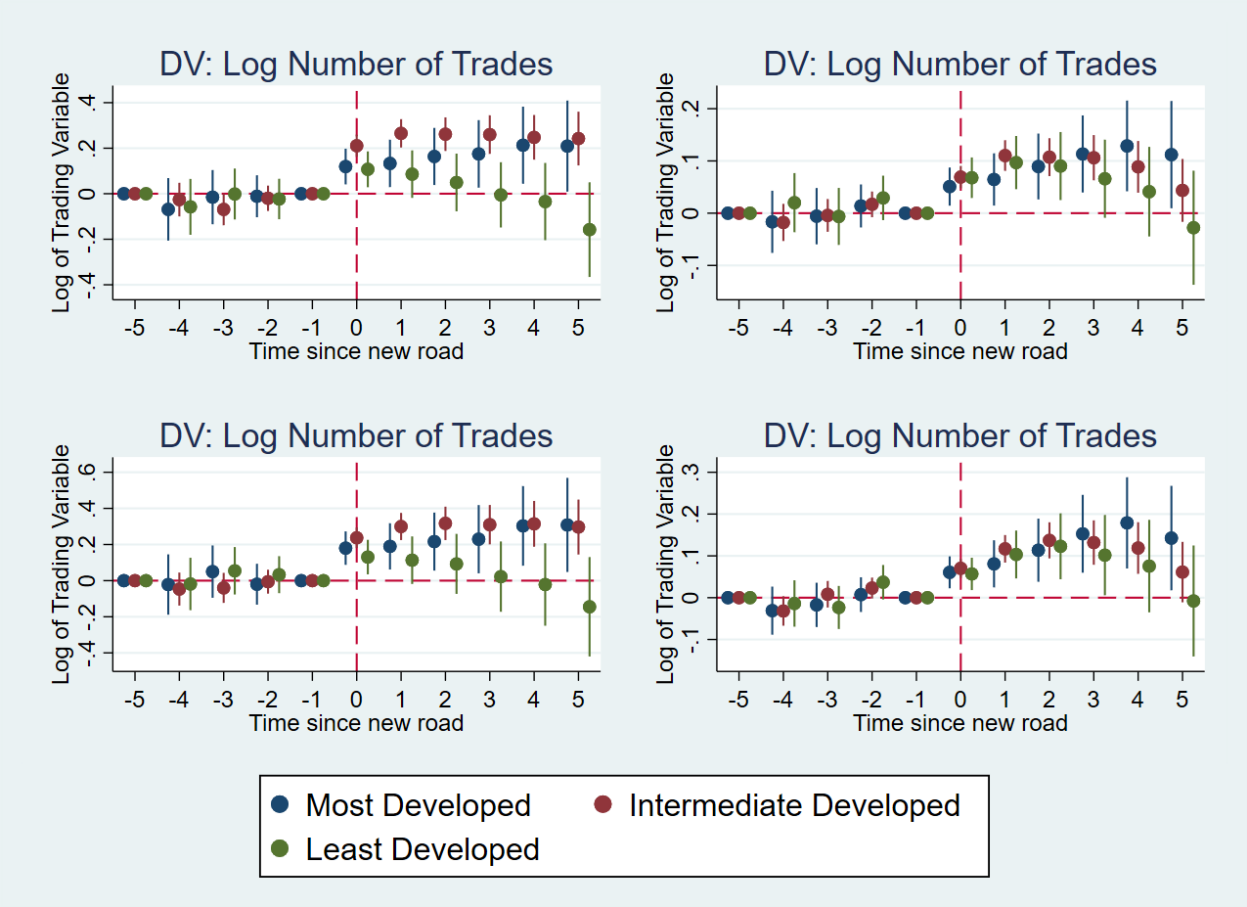

Figure A2: **Roads and Stock Market Participation: Time Trends**
These figures plot the dynamic effects of road on stock market participation across all type of investors using the staggered DiD setting following the approach in Callaway and Sant'Anna [2021] and estimated using the Stata package *csdid* where the y-axis represent the time-dynamic effects of staggered DiD estimators while x-axis represents the month since road construction.

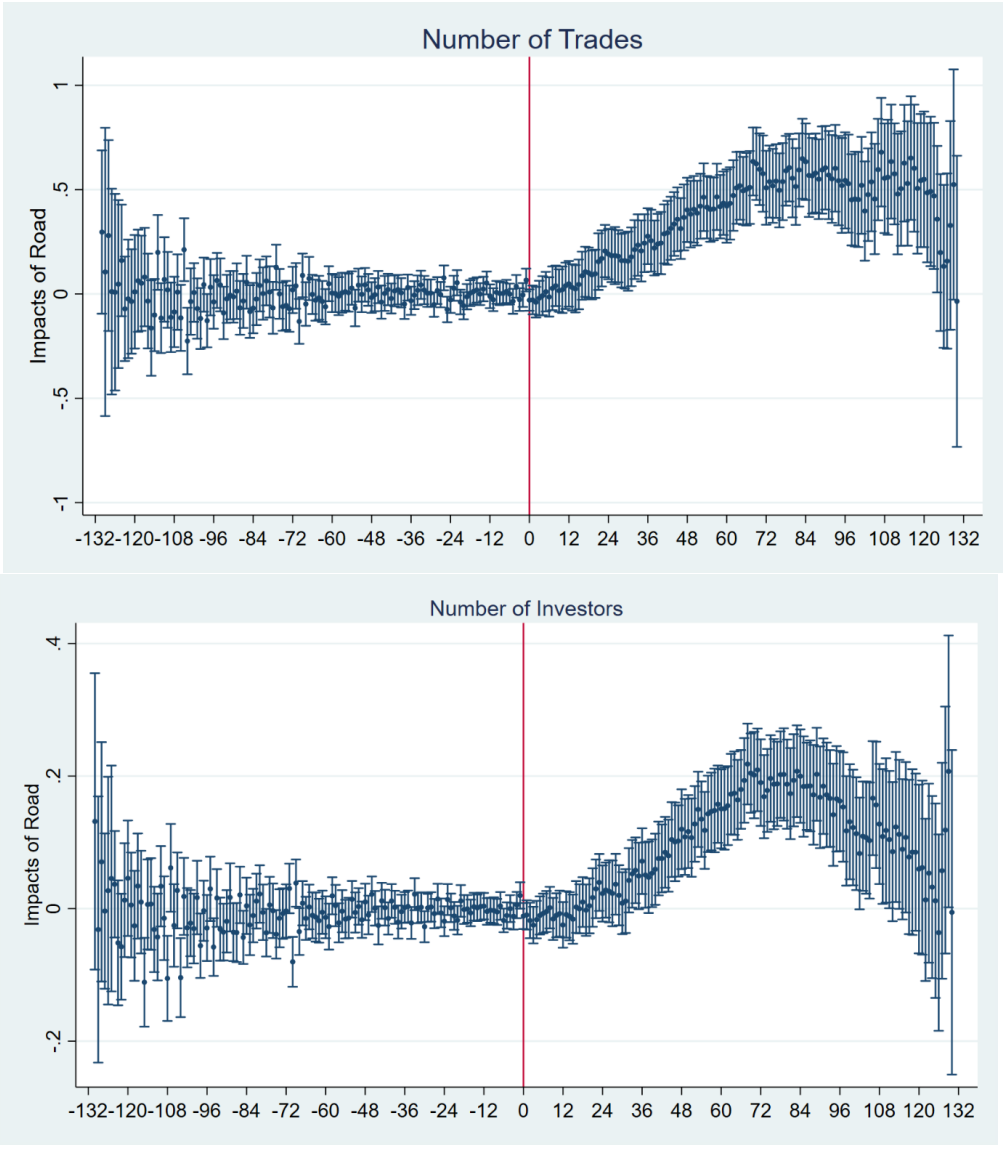

Figure A2: **Roads and Stock Market Participation: Time Trends (Continued..)**
These figures plot the dynamic effects of road on stock market participation across new investors using the staggered DiD setting following the approach in Callaway and Sant'Anna [2021] and estimated using the Stata package *csdid* where the y-axis represent the time-dynamic effects of staggered DiD estimators while x-axis represents the month since road construction.

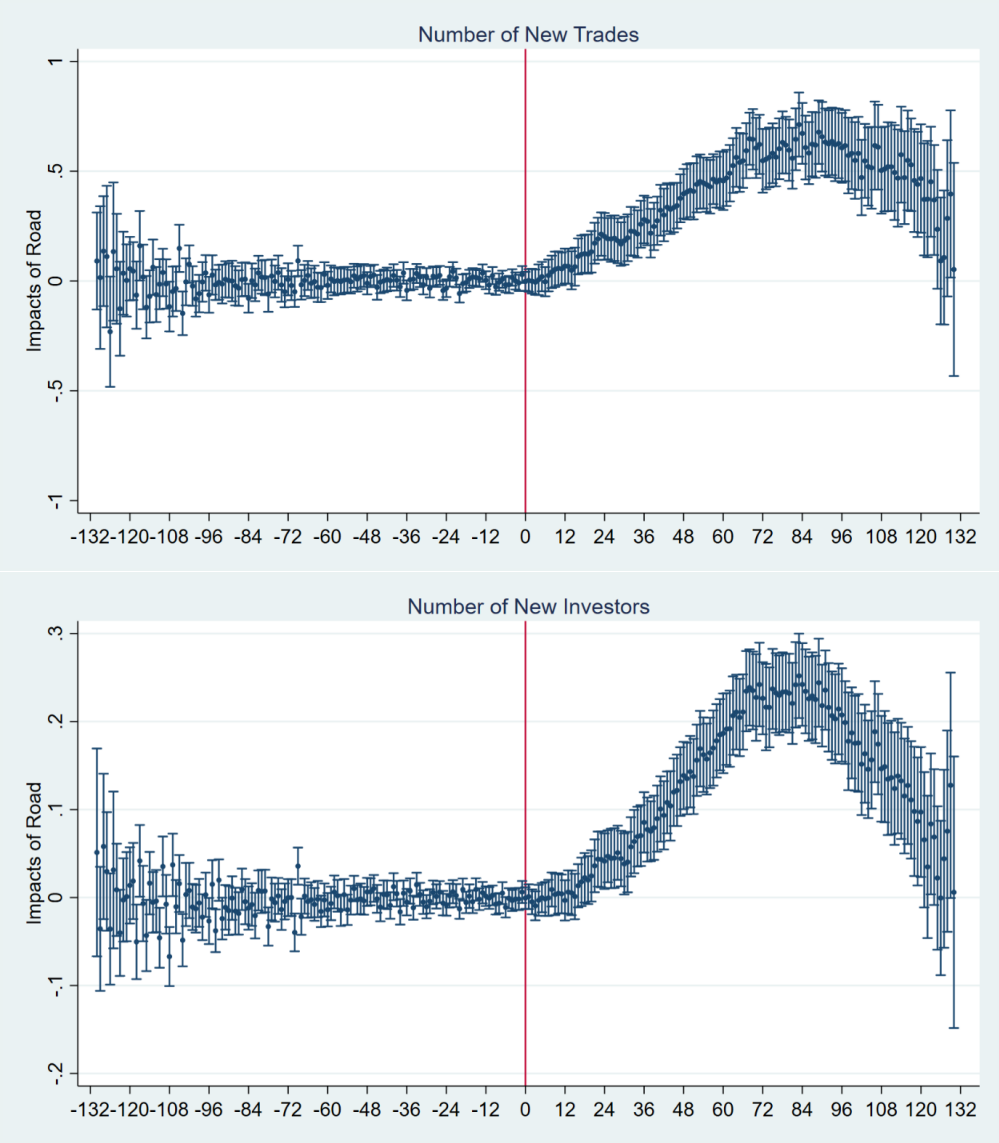

Table A1: **Variable Appendix**
This table reports the definition and source of the main variables used in this paper.

Variable	Definition Source	Source
Number of Trades	Total number of stock trades in pincode i in month t.	NSE
Number of Investors	Total number of investors in pincode i in month t.	NSE
Number of Tickers	Total number of unique tickers being traded in pincode i in month t.	NSE
New Investor	Investors whose trading account opening date is ≤ 3 years old	NSE
Experienced Investor	Investors who have had a trading account for more than 3 years	NSE
Young Investor	Investor whose age is between 18-30	NSE
Middle-Age Investor	Investor whose age is between 30-55	NSE
Mature Investor	Investor whose age is above 55	NSE
Monetary Profits	Average buy-and-hold profits in Indian Rupees (INR) calculated over different holding periods (1, 10, 25, and 140 trading days) for stocks traded in a pincode-month.	NSE
Excess BHR	Average buy-and-hold returns after subtracting the market return (using the NIFTY 50 index as benchmark) over different holding periods (1, 10, 25, and 140 trading days). This measures investors' performance relative to the overall market, indicating whether traders in a pincode outperform or underperform the market average	NSE
Connect	Indicator variable which equals 1 in the year-month (and thereafter) when a pincode is connected by a paved road under the PMGSY program and 0 otherwise	PMGSY
Rural	Post offices in India are classified into five categories by the Department of Posts: DO (Divisional office), GPO (General Post Office), HO (Head office), SO (Sub Office) and BO (Branch office). We classify a pincode as *Rural* if its post office type is BO.	Indian Post Office
Developed	Indicator variable which equals to 1 if the consumption per capita value of the pincode is above median value across all the pincodes in the sample.	SHRUG
Intermediate Development	Indicator variable which equals to 1 if the consumption per capita value of the pincode fall between the 50th and 10th percentile value across all the pincodes in the sample.	SHRUG
Least Developed	Indicator variable which equals to 1 if the consumption per capita (poverty rate) of the pincode is below the 10th percentile value across all the pincodes in the sample.	SHRUG
New Branch dummy	Indicator variable which equals to 1 if a new bank branch opened within a 3-year window after the completion of PMGSY road in pincode i.	RBI
Number of New Branches	Total number of new bank branches that opened within a 3-year window after the completion of PMGSY road in pincode i.	RBI
Distance	Distance between pincode location of investor and nearest city in 10km units. Cities are classified into Tier-1 towns (population exceeding 100,000 people) and Metros (population exceeding 1 million people) as defined by 2001 Population Census.	RBI
Consumption	Monthly total consumption of the household.	CMIE
Income	Monthly total income of the household.	CMIE
Deposit	Year amount of deposit in bank branch.	RBI-BSR
Literacy	Literate population divided by the total population in a given pincode month.	2001 Population Census
Financial Literacy	Index constructed as the average of three normalized household financial behavior measures: ownership of bank deposit accounts, credit/debit cards, and e-wallets that are sourced from the NSS 77th round of the All India Debt & Investment Survey.	Dash and Ranjan [2023]
Weather Risk	Standard deviation of monthly rainfall in pincode i over sample period.	CEDA
Night Light Index	Measure of human-generated light visible from space that serves as a proxy for economic activity and development. The index is derived from satellite observations of artificial lighting during nighttime hours, which is processed to remove natural light sources, atmospheric effects, and other confounding factors. The resulting values represent the intensity of light emissions in a given geographic area, typically measured at the pixel level with higher values indicating greater light intensity. The index is widely used as a reliable proxy for economic activity, urbanization, and living standards, as areas with greater economic development tend to produce more artificial light at night.	Li et al. [2020]
Market Persistence	Average duration (number of days) between an investor's first and last trade date in sample period across all activate investors in a given pincode-month.	NSE
Trading Intensity	Average number of days between consecutive trades within each pincode-month.	NSE
High Volatility Stock	Stock whose monthly price volatility is above the sample median for all stocks in the same month.	NSE

Table A2: **State-level distribution of investors**
This table reports the distribution of investors across Indian states who traded on the National Stock Exchange (NSE) from 2004 to 2015, showing both absolute numbers and percentages.

States	**Number of Investors**	**%**	**States**	**Number of Investors**	**%**
Maharashtra	2685160	19.875%	Assam	82744	0.612%
Gujarat	1954062	14.463%	Uttarakhand	78702	0.583%
Tamil Nadu	1046034	7.742%	Jammu and Kashmir	43869	0.325%
West Bengal	995263	7.367%	Himachal Pradesh	42182	0.312%
Karnataka	907781	6.719%	Goa	38684	0.286%
Uttar Pradesh	893399	6.613%	Pondicherry	10734	0.079%
Delhi	835927	6.187%	Tripura	10383	0.077%
Rajasthan	551151	4.079%	Dadra and Nagar Hav.	5176	0.038%
Telangana	532050	3.938%	Megalaya	4729	0.035%
Kerala	490173	3.628%	Chandigarh	4574	0.034%
Andhra Pradesh	432272	3.200%	Daman and Diu	3619	0.027%
Haryana	406657	3.010%	Manipur	2740	0.020%
Madhya Pradesh	398663	2.951%	Sikkim	2674	0.020%
Punjab	328561	2.432%	Nagaland	1998	0.015%
Bihar	228644	1.692%	Andaman and Nico.In.	1526	0.011%
Jharkhand	200192	1.482%	Arunachal Pradesh	1355	0.010%
Odisha	191229	1.415%	Mizoram	604	0.004%
Chattisgarh	96898	0.717%	Lakshadweep	64	0.000%
Total			**13510473**		

Table A3: **Heterogeneity by Literacy Levels**
This table estimates the following regression:

$$Log(Y_{i,t}) = \beta_1 \cdot Connect_{i,t} + \beta_2 \cdot Connect_{i,t} \times X_i + \xi_i + \kappa_t + \theta_{d,y} + \varepsilon_{i,t}$$

where Y is either *Number of Trades* or *Number of Investors* in pincode i in year-month t and is defined for the sample of New Investors (investors who opened trading accounts within the last three years). X_i is $Financial\ Literacy_i$ in cols. 1 and 3 and $Literacy_s$ in cols. 2 and 4. $Literacy_i$ is a pincode-level measure of literacy measured by the ratio of literate population to total population in each pincode (time-invariant). $Financial\ Literacy_s$ is a state-level measure of financial literacy capturing proportion of households in a state that have deposit accounts in banks, credit/debit cards, and e-wallets. $Connect_{i,t}$ is an indicator variable which equals 1 in the year-month (and thereafter) when a pincode is connected by a paved road under the PMGSY program and 0 otherwise. All regressions are estimated using pincode, year-month, and district-year fixed effects. Standard errors clustered by pincode are reported in parentheses. All variables are defined in the Variable Appendix. (***), (**), (*) denote statistical significance at 1%, 5%, and 10% levels respectively.

	1	2	3	4
	Number of Trades (New Investors)		Number of Investors (New Investors)	
Connect	0.645***	0.614***	0.263***	0.249***
	(0.063)	(0.064)	(0.030)	(0.029)
Connect X Financial Literacy	-0.000		0.001	
	(0.002)		(0.001)	
Connect X Literacy		0.063		0.080
		(0.130)		(0.059)
Pincode FE	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y
District-Year FE	Y	Y	Y	Y
N	592195	587983	592204	587995
Adj. R-sq	0.802	0.802	0.899	0.899

Table A4: **Roads and Bank Openings**
This table reports estimates from the following regression:

$$Y_{i,t} = \beta \cdot Connect_{i,t} + \xi_i + \kappa_t + \theta_{d,y} + \varepsilon_{i,t}$$

where Y is a dummy for New bank branch opening or the Number of new bank branches. Columns 1 and 4 show results for all banks, columns 2 and 5 show results for state-owned bank branches, and columns 3 and 6 show results for private banks branches. $Connect_{i,t}$ is an indicator variable which equals 1 in the year-month (and thereafter) when a pincode is connected by a paved road under the PMGSY program and 0 otherwise. All regressions are estimated using pincode, year-month, and district-year fixed effects. Standard errors clustered by pincode are reported in parentheses. All variables are defined in the Variable Appendix. (***), (**), (*) denote statistical significance at 1%, 5%, and 10% levels respectively.

	1	2	3	4	5	6
DV	New Branch Dummy	New Branch Dummy	New Branch Dummy	Number of New Branches	Number of New Branches	Number of New Branches
	All Banks	State Banks	Private Banks	All Banks	State Banks	Private Banks
Connect	0.002***	0.003***	-0.000	0.003**	0.003***	-0.001
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Pincode FE	Y	Y	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y	Y	Y
District-Year FE	Y	Y	Y	Y	Y	Y
N	598176	598176	598176	598176	598176	598176
Adj. R-sq	0.079	0.059	0.060	0.075	0.059	0.064

Table A5: **Financial Inclusion Channel-Robustness using Number of Branches**
Panel A reports estimates from the following regression:

$$\log(Y_{i,t}) = \beta_1 \text{Connect}_{i,t} + \beta_2 \text{Connect}_{i,t} \times \text{Number of New Branches}_i + \beta_3 \text{Night Light Index}_{i,t} + \beta_4 \text{Connect}_{i,t} \times \text{Night Light Index}_{i,t} + \xi_i + \kappa_t + \theta_{d,y} + \varepsilon_{i,t}.$$

Panel B reports estimates from:

$$\log(Y_{i,t}) = \beta_1 \text{Connect}_{i,t} + \beta_2 \text{Connect}_{i,t} \times \text{Number of New State Branches}_i + \xi_i + \kappa_t + \theta_{d,y} + \varepsilon_{i,t}.$$

where Y is either *Number of Trades* or *Number of Investors* in pincode i in year-month t and is defined for the sample of New Investors (investors who opened trading accounts within the last three years). $Connect_{i,t}$ is an indicator variable which equals 1 in the year-month (and thereafter) when a pincode is connected by a paved road under the PMGSY program and 0 otherwise. *Number of New Branches* are the number of new bank branches opened within a 3-year window after the PMGSY road completion in pincode i and *Number of New State Branches* is the number of new state bank branches opened within a 3-year window after the PMGSY road completion in pincode i. All regressions are estimated using pincode, year-month, and district-year fixed effects. Standard errors clustered by pincode are reported in parentheses. All variables are defined in the Variable Appendix. (***), (**), (*) denote statistical significance at 1%, 5%, and 10% levels respectively.

Panel A: New Bank Branch

	1	2	3	4
	Number of Trades		Number of Investors	
	(New Investors)		(New Investors)	
Connect	0.618***	0.564***	0.278***	0.298***
	(0.023)	(0.025)	(0.011)	(0.012)
Connect x Number of New Branches	0.024***	0.045***	0.008**	0.015***
	(0.007)	(0.008)	(0.003)	(0.004)
Night Light Index		0.006		0.009***
		(0.005)		(0.003)
Connect x Night Light Index		-0.010***		-0.004***
		(0.002)		(0.001)
Pincode FE	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y
District-Year FE	Y	Y	Y	Y
N	592195	592195	592204	592204
Adj. R-sq	0.802	0.802	0.899	0.9

Panel B: New Bank Branch - State vs. Private

	1	2	3	4
	Number of Trades	Number of Investors	Number of Trades	Number of Investors
	(New Investors)	(New Investors)	(New Investors)	(New Investors)
Reference Group	Pincodes with either no branch opening or branch openings of private banks		Pincodes with branch openings of private banks	
Connect	0.592***	0.277***	0.511***	0.252***
	(0.022)	(0.010)	(0.035)	(0.018)
Connect x Number of New State Branches	0.088***	0.019***	0.117***	-0.011
	(0.012)	(0.007)	(0.013)	(0.007)
Pincode FE	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y
District-Year FE	Y	Y	Y	Y
N	592195	592204	252358	252366
Adj. R-sq	0.802	0.899	0.854	0.931

Table A6: **Risk Sharing Channel**

This table reports estimates from the following regression:

$$Log(Y_{i,t}) = \beta_1 \cdot Connect_{i,t} + \beta_2 \cdot Connect_{i,t} \times WeatherRisk_i + \xi_i + \kappa_t + \theta_{d,y} + \varepsilon_{i,t}$$

where Y is either *Number of Trades* or *Number of Investors* in pincode i in year-month t and is defined for the sample of New Investors (investors who opened trading accounts within the last three years). $Connect_{i,t}$ is an indicator variable which equals 1 in the year-month (and thereafter) when a pincode is connected by a paved road under the PMGSY program and 0 otherwise. *Weather Risk* is measured by the standard deviation of rain fall in a pincode i over the entire sample period. In columns 3 and 4, we create dummies for Low (omitted category), Intermediate and High Weather Risk based on terciles of Weather Risk. Standard errors clustered by pincode are reported in parentheses. All variables are defined in the Variable Appendix. (***), (**), (*) denote statistical significance at 1%, 5%, and 10% levels respectively.

	1	2	3	4
	Number of Trades (New Investors)		Number of Investors (New Investors)	
Connect	0.677***	0.650***	0.281***	0.272***
	(0.031)	(0.026)	(0.015)	(0.012)
Connect X Weather Risk	-0.004		0.001	
	(0.003)		(0.001)	
Connect X Intermediate Weather Risk		-0.027		0.019
		(0.049)		(0.023)
Connect X High Weather Risk		0.000		0.048**
		(0.052)		(0.024)
Pincode FE	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y
District-Year FE	Y	Y	Y	Y
N	592193	592193	592202	592202
Adj. R-sq	0.802	0.802	0.899	0.899

Table A7: **Placebo Test: Roads and Stock Market Participation**
This table reports the placebo test of the baseline specification. In panel A, we randomize the road completion date for all pincodes across the entire sample period. In panel B, we randomize the road completion date for all pincodes within the same month. For each regression specification, we run 500 samples and average the regression coefficients and standard errors across these 500 regressions.

Panel A: Randomization for Entire Sample

	Number of Trades	Number of Investors	Number of Trades	Number of Investors	Number of Trades	Number of Investors
	All Investors		New Investors		Experienced Investors	
Connect	0.000102	-0.000004	0.000296	0.000029	-0.000043	0.000019
	(0.0027857)	(0.001339)	(0.003719)	(0.001549)	(0.003441)	(0.001518)
Pincode FE	Y	Y	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y	Y	Y

Panel B: Randomization within Each Month

	Number of Trades	Number of Investors	Number of Trades	Number of Investors	Number of Trades	Number of Investors
	All Investors		New Investors		Experienced Investors	
Connect	0.000042	-0.000066	0.000221	0.000012	0.000032	0.000062
	(0.002727)	(0.001344)	(0.003694)	(0.001615)	(0.003306)	(0.001405)
Pincode FE	Y	Y	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y	Y	Y

Table A8: **Roads and Stock Market Participation: Robustness using Callaway and Sant'Anna [2021] specification**
This table reports results from the staggered DiD setting following the approach in Callaway and Sant'Anna [2021] and estimated using the Stata package *csdid*. All regressions are estimated using pincode, year-month, and district-year fixed effects. Standard errors clustered by pincode are reported in parentheses. All variables are defined in the Variable Appendix. (***), (**), (*) denote statistical significance at 1%, 5%, and 10% levels respectively.

	1 Number of Trades	2 Number of Investors	3 Number of Trades	4 Number of Investors
	All Investors		New Investors	
Connect	0.481***	0.154***	0.538***	0.214***
	(0.103)	(0.045)	(0.152)	(0.012)
Pincode FE	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y

Table A9: **Baseline Effects: Poisson Regression**

This table replicates the results from the main baseline specification using Poisson regressions. Specifically, we use the count values of *Number of Trades* and *Number of Investors* as the dependent variables. All regressions are estimated using pincode, year-month, and district-year fixed effects. Standard errors clustered by pincode are reported in parentheses. All variables are defined in the Variable Appendix. (***), (**), (*) denote statistical significance at 1%, 5%, and 10% levels respectively.

	Number of Trades	Number of Investors	Number of Trades	Number of Investors	Number of Trades	Number of Investors
	All Investors		New Investors		Experienced Investors	
Connect	0.315***	0.092***	0.514***	0.134***	0.238***	0.063***
	(0.013)	(0.008)	(0.025)	(0.016)	(0.015)	(0.013)
Pincode FE	Y	Y	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y	Y	Y
District-Year FE	Y	Y	Y	Y	Y	Y
N	592195	592204	592195	592204	592195	592204
Adj-R2	0.972	0.969	0.941	0.940	0.970	0.963

Table A10: **Roads and Local Income Shocks**
This table reports estimates from the following regression:

$$Log(Y_{i,t}) = \beta_1 \cdot Connect_{i,t} + \beta_2 \cdot Rainfall\ Shock_{i,t} + \beta_3 \cdot Connect_{i,t} \times Rainfall\ Shock_{i,t} + \xi_i + \kappa_t + \theta_{d,y} + \varepsilon_{i,t}$$

where Y are the different measures of stock market participation defined in pincode i in year-month t: *Number of Trades* and *Number of Investors*. All estimates are presented for sample of New Investors, defined as investors whose trading account opening date is ≤ 3 years old. $Connect_{i,t}$ is an indicator variable which equals 1 in the year-month (and thereafter) when a pincode is connected by a paved road under the PMGSY program and 0 otherwise. $Rainfall\ Shock_{i,t}$ takes the value 1 if the rainfall in a pincode in a month is above 80^{th} percentile for the same pincode-month over the entire sample period; -1 if rainfall is below 20^{th} percentile for the same pincode-month over the entire sample period; and 0 otherwise. All regressions are estimated using pincode, year-month, and district-year fixed effects. Standard errors clustered by pincode are reported in parentheses. All variables are defined in the Variable Appendix. (***), (**), (*) denote statistical significance at 1%, 5%, and 10% levels respectively.

	1	2	3	4
	Number of Trades (New Investors)		Number of Investors (New Investors)	
Timing of rainfall	Concurrent	One month-lag	Concurrent	One month-lag
Connect	0.645***	0.641***	0.287***	0.284***
	(0.021)	(0.021)	(0.010)	(0.010)
Rainfall Shock	-0.003	-0.009**	-0.000	-0.000
	(0.004)	(0.004)	(0.002)	(0.002)
Connect X Rainfall Shock	-0.003	0.009	-0.004	-0.001
	(0.006)	(0.006)	(0.002)	(0.002)
Pincode FE	Y	Y	Y	Y
Year-Month FE	Y	Y	Y	Y
District-Year FE	Y	Y	Y	Y
N	592195	588044	592204	588051
Adj. R-sq	0.802	0.802	0.899	0.900