Roads and informal firms in India: Pre-Analysis Plan

Fiona Burlig*

May 31, 2016

1 Introduction

Infrastructure spending dominates budgets in the developing world. In 2014, the World Bank spent $24 billion on infrastructure (World Bank (2014)). There are good reasons for this level of investment: recent evidence has shown that this public goods funding leads to economic development, including Donaldson (2010), Asher and Novosad (2014c), and Banerjee, Duflo, and Qian (2012).[1] There remains limited evidence, however, on the effect of infrastructure on firms in developing countries. I hope to address a key gap in the literature: to the best of my knowledge, none of the papers that study infrastructural investments consider informal firms. Since these firms make up half of India's $1.85 trillion economy and employ over 80% of her non-farm workers, understanding what happens to them in the wake of large infrastructure investments is crucial (Barman (2013)). In this work, I aim to estimate the effects of several major Indian road construction projects on formal and informal firms in a variety of sectors, motivated by a theoretical model of trade with heterogeneous firms.

The remainder of this document describes my plan for assessing the effects of road construction on Indian firms. Section 2 describes firm dynamics in India, as well as the road construction projects I will be analyzing. Section 3 describes my theoretical framework and resulting testable

*University of California, Berkeley: Department of Agricultural and Resource Economics and Energy Institute at Haas. Email: fiona.burlig@berkeley.edu. I am generously supported by the National Science Foundation's Graduate Research Fellowship Program. Max Auffhammer, Kendon Bell, Lucas Davis, James Gillan, Erin Kelley, Jeremy Magruder, Edward Miguel, Louis Preonas, Andrés Rodríguez-Clare, Elisabeth Sadoulet, Andrew Stevens, Matt Woerman, and Catherine Wolfram provided useful feedback. All remaining errors are my own.

[1]Other important papers in this literature include: Faber (2014), Casaburi, Glennerster, and Suri (2013), and Duflo and Pande (2007)

hypotheses. Section 4 describes the datasets I will use, and explains how I can credibly pre-register an observational study using pre-existing data. Section 5 details my empirical strategies, intended robustness checks, and statistical inference.

2 Empirical context: modern India

2.1 Informal firms

In many sectors in India, firms are required to register formally with the government once they reach a certain size. This means that, unlike in many other developing country contexts, firms can legally be informal when they are below this size threshold. Informal (or "unregistered") firms do not pay taxes, and are not subject to a variety of labor regulations. By contrast, formal ("registered") firms are required to pay taxes, comply with workplace safety laws, and are subject to increased administrative burdens. A particularly important example seems to be the Factories Act of 1948 binds when firms reach 10 or 20 workers (firms using electric power must register at 10 workers, though there are a variety of labor regulations in India that begin to bind at different thresholds.

There is an existing literature on the effect of labor regulations on firm productivity in India. Besley and Burgess (2004), for example, demonstrates that pro-worker regulations can hurt aggregate firm productivity and increase the incentives for becoming informal. In a more recent example, Colmer (2015) shows that labor regulations can cause a misallocation of labor between sectors. Hsieh and Klenow (2009) argues that labor is substantially misallocated across firms in India, leading to losses in GDP of between 40 and 60 percent relative to marginal factor product equalization similar to that seen in the United States. Hasan and Jandoc (2012) find that states with more flexible labor laws have larger firms. Taken together, this evidence suggests that labor laws have substantially distorted firm size and productivity.

2.2 Road construction

In order to measure the effects of transportation infrastructure on firms, I will study India's recent major highway improvement projects.[2] In 2001, the National Highways Authority of India (NHAI) began work on the Golden Quadrilateral (GQ) project as the first phase of the National Highways Development Project (NHDP). The GQ project built nearly 5,900 kilometers of modern four and six lane highways connecting Delhi, Mumbai, Kolkata, and Chennai. It is the largest highway project in India to date, and the fifth longest in the world. 80% of the work on the GQ project was completed by 2004, and by 2006, 95% of the work had been finished. The project was declared finished in 2012, and cost the Indian government over $9 billion.

Figure 1: Indian Highway Upgrades

Notes: This figure shows India's largest highway upgrades to date. The left panel shows the Golden Quadrilateral highways, and the right panel the North-South East-West Corridor highways.

The second phase of the NHDP is the North-South East-West Corridor (NS-EW). Like the GQ project, the NS-EW Corridor connects major Indian regions using four and six lane highways. The NS-EW Corridor connects Srinagar, Kanyakumari, Porbandar, and Silchar. This project is larger

[2]If the 2012 Economic Census is released before this project is completed, I will study both the Golden Quadrilateral project and the North-South East-West Corridor project. If not, I will study only the Golden Quadrilateral upgrades. I describe both here.

than the GQ, and is slated to build over 7,000 kilometers of roadway by the time it is completed. By April 2011, 77% of the infrastructure improvements had been completed. The final stages of the project are still being finished. $12.3 billion were committed to the project. Figure 1 shows maps of these upgrades.

3 Theory

I am interested in four questions:
What are the effects of lowered trade costs on:

1. between-industry allocation of labor and firms?
2. within-industry allocation of labor and firms?
3. the allocation of labor and firms between the informal and the formal sector?
4. overall welfare?

I derive predictions from two sets of trade models: Ricardian, and "New New Trade" models.

Ricardian models assume perfect competition, which means that they do allow for useful predictions about heterogeneous firms. They are, however, useful for thinking about comparative advantage. Classic Ricardian models will predict that, as trade costs fall:

- Districts will specialize in goods (sectors) of comparative advantage.
- Welfare will increase.

In order to make predictions about firms, I turn to an extension of Melitz (2003), first derived by Aleman-Castilla (2006). I describe this model fully in Appendix A, and provide a brief overview here. Note that the predictions from this model address tradable goods only. The key features of this model are that consumers have CES love-of-variety preferences; firms are monopolistically competitive with free entry; firm heterogeneity comes from productivity draws φ; and firm size is monotonically increasing in φ. The model includes three types of firms: informal, formal, and exporting.

Informal firms pay no taxes, and pay a fixed cost f_I to enter. Their profits are:

$$\pi_I(\varphi) = \left(\frac{\varphi}{w}\right)^{\sigma-1} B - wf_I$$

where w is the wage, σ is the elasticity of substitution, and B is a price normalization.

Formal firms pay a tax rate α, and enjoy a productivity boost β, and pay fixed cost $f_F > f_I$. Their profits are:

$$\pi_F(\varphi) = \left[\frac{(1+\beta)\varphi}{(1+\alpha w}\right]^{\sigma-1} B - (1+\alpha)wf_F$$

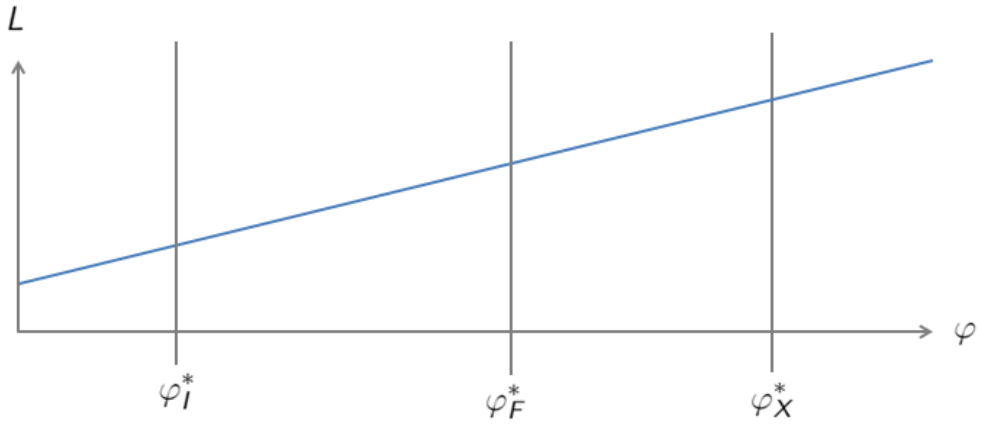
Exporters are formal firmss (with α and β as before, and pay fixed cost $f_X \geq f_F$. They can sell on the home market and receive profits $\pi_F(\varphi)$. Their additional profits from exporting are:

$$\pi_X(\varphi) = \left[\frac{(1+\beta)\varphi}{\tau(1+\alpha w}\right]^{\sigma-1} B - (1+\alpha)wf_X$$

where τ are iceberg trade costs.

In order to enter, a firm pays the fixed cost f. It then receives its productivity draw, φ, from a distribution with CDF $G(\varphi)$. If profits are negative, it will exit immediately. This leads to three entry cutoffs. Informal firms will enter if they make weakly positive profits, if and only if $\varphi \geq \varphi_I^*$. A firm will enter formally if and only if it makes higher profits than it would informally: $\varphi \geq \varphi_F^*$. A firm will export if and only if it makes higher profits than it would as a non-exporting formal firm: $\varphi \geq \varphi_X^*$. This implies a cutoff ordering: $\varphi_X^* > \varphi_F^* > \varphi_I^*$. Graphically:

Figure 2: Graphical setup

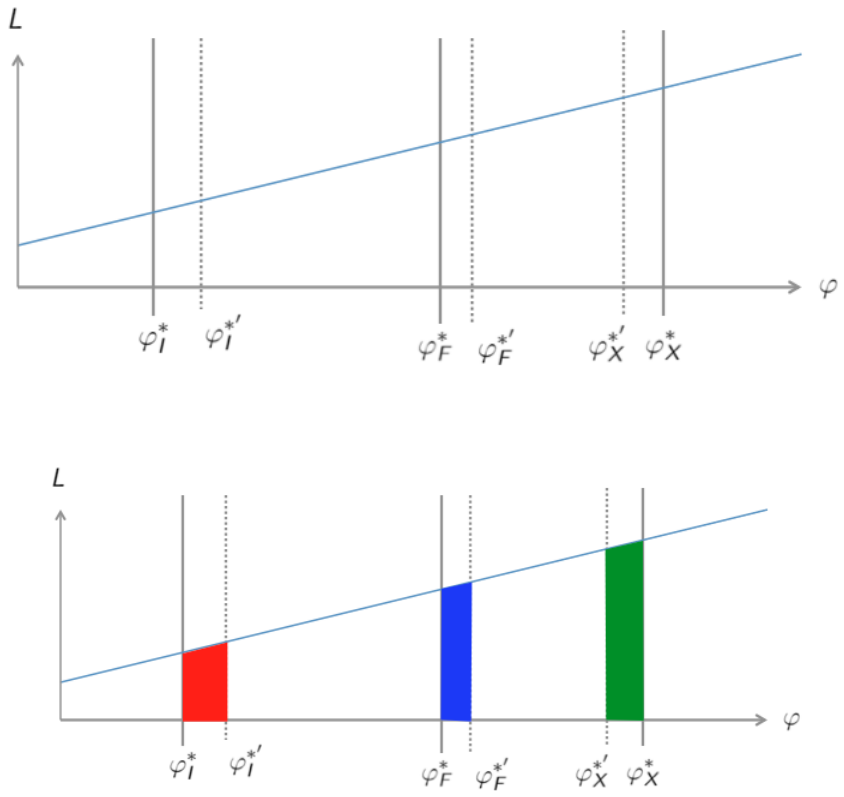


Notes: This figure shows the ordering of firm entry cutoffs on the X axis, and firm size L on the Y axis. Firm size is increasing in productivity, φ, and the informal entry cutoff is lower than the formal cutoff, which is lower than the exporting cutoff.

Falling trade costs will affect these cutoffs through GE effects. As τ falls, $\pi_X(\varphi)$ increases. This leads to a fall in φ_X^*, so firms enter the exporting sector, and their labor share increases. This in turn drives up wages. At the same time, exporting firms from foreign districts are entering the home market. These forces both increase the real wage: labor demand increases the wage, and the entry of foreign exporters decreases the price index. These effects drive up φ_F^* and φ_I^*. The number of formal firms (including exporters) falls, but the remaining formal firms (including exporters) employ more workers. The effect on the number of informal firms is ambiguous, because the lowest-productivity firms are forced to exit, but some of the firms that would have entered as low-productivity formal firms now enter informally instead. This causes a shift in the informal size distribution away from small firms and towards larger firms, and the informal sector overall employs more workers.

Graphically:

Figure 3: Graphical intuition



Notes: This figure shows the effects of falling trade costs on firm entry cutoffs. The upper panel shows that, as trade costs fall, φ_X^* falls to $\varphi_X^{*'}$; φ_F^* increases to $\varphi_F^{*'}$; and φ_I^* increases to $\varphi_I^{*'}$. The lower panel shows the changes in number of firms and labor shares: firms with $\varphi \in (\varphi_I^*, \varphi_I^{*'})$ now do not enter; firms with $\varphi \in (\varphi_F^*, \varphi_F^{*'})$ now enter informally rather than formally; and firms with $\varphi \in (\varphi_X^*, \varphi_X^{*'})$ now enter as exporters.

3.1 Predictions

This generates a number of predictions. As trade costs fall:

1. φ_X^* decreases, so firms enter the exporting sector, and their labor share increases
2. Wages increase
3. **Formal firms:** φ_F^* increases
   - The number of formal firms (including exporters) falls
   - The remaining formal firms (including exporters) employ more workers
   - The labor share in the formal sector is ambiguous

4. **Informal firms:** φ_I^* increases
   - The number of informal firms is ambiguous (the lowest-producing exit; but some low-productivity formal firms now become informal)
   - The informal firms employ more workers
   - The labor share in the informal sector is ambiguous
5. The total number of firms declines, and with it, the number of varieties produced at home declines.
6. Overall welfare increases.

The Ricardian models yield one additional prediction (and reinforces prediction 6):

7. Districts will specialize in sectors in which they have comparative advantage.

I also test several predictions that are not based on a trade model:

8. Women and marginalized groups will own more firms.
9. Employment of women will increase.
10. Employment of children will decrease.

In order to test these predictions, I combine a novel dataset with a quasi-experimental research design.

4 Data

I will use several datasets in this analysis. These are: the Economic Census of India (1990, 1998, 2005, and if possible, 2012 waves); detailed village and district boundary shapefiles; road construction data from NHAI and Ghani, Goswami, and Kerr (2015); and district-wise GDP data from the Indian Government.[3]

I already have access to several of these datasets. In particular, I have the geospatial data, GDP data, and the road construction data. In all of the analysis specified below, I will use district boundaries as defined in 2001 for consistency.

[3]Data here: `https://data.gov.in/catalog/district-wise-gdp-and-growth-rate-current-price1999-2000`.

4.1 Data access

The Economic Census of India data are available for purchase from the Ministry of Statistics and Programme Implementation.[4] I have recently secured funding through CEGA to purchase the 1990, 1998, and 2005 waves of these data (see attached email with grant confirmations). Once the 2012 wave becomes available, I hope to purchase these data as well.

Importantly, prior to registering this pre-analysis plan with the Open Science Framework today, I have not had access to these data. I can verify this as follows. First, the data are expensive (at least for a graduate student). As an Appendix to this PAP, I will upload my grant receipt for the purchase of these data (dated before this pre-analysis plan was submitted). I will also attach the receipt for the actual purchase of these data, along with a signed declaration from the Berkeley ARE department coordinator, documenting the date on which she released the data to me.[5] In addition, I know of only two economics researchers who have used these data before: Sam Asher and Paul Novosad. They have not shared these data with me (nor, as far as I know, any other researchers), nor have they made these data available online. Furthermore, the final wave of the dataset I hope to use has not yet been released by the Indian government. All of these factors allow me to credibly claim that I have not had access to the relevant data, meaning that my pre-analysis plan will be an effective check on the degrees of freedom with which I will conduct this analysis. Because I do not yet have the data, this PAP will include some conditional sections. Unlike the experimental setting, I am not collecting my own data, and as such, there is some uncertainty as to what information exactly is contained in the dataset, how well populated various variables are, etc.

4.2 Economic Census description

Here, I describe the dataset that I will use for my analysis, but currently do not have access to.

[4]The main website for the Economic Census is here: `http://mospi.nic.in/Mospi_New/site/inner.aspx?status=3&menu_id=87`; the description of the purchase process can be found here: `http://mospi.nic.in/Mospi_New/upload/Rate_List_EC.pdf`.

[5]The original plan was not to purchase the data until the PAP had been filed. Due to logistical constraints, this was impossible. Instead, the dataset was purchased, and shipped to Berkeley, where the department coordinator has it under lock and key until after this plan is submitted to a registry.

The Ministry of Statistics and Programme Implementation (MoSPI) conducted the 3rd, 4th, and 5th waves of the Economic Census in 1990, 1998, and 2005, respectively. They are in the process of finalizing the 6th wave, having collected these data in 2012. This Census is intended to be a complete list of all economic establishments in the country, with the exception of those firms planting, harvesting, and producing crops. Unlike other data on firms from India, firms of all sizes are included. Notably, the Census includes both formal and informal firms. According to the International Household Survey Network, MoSPI's data repository, the 5th wave of the Economic Census includes data on the location (village or town ward) of each firm; whether a firm has premises; which NIC code the firm's major activity falls under; whether the firm is publicly or privately owned; whether the owner is male or female, and whether the owner belongs to a Scheduled Caste, Scheduled Tribe, or Other Backwards Caste; what type of fuel is used in production; whether the firm is formal or not (and if so, under which category the firm is registered); the number of men, women, and children employed by the firm; the number of non-hired men, women, and child employees of the firm; and the firm's source of financing.[6] There does not seem to be a similar meta-data archive for the 3rd or 4th waves of the Census (nor for the currently-nonexistent 6th wave). In order to ensure that the variables I will use in my analysis are available for the other waves of the survey, I turn to the data descriptions in Asher and Novosad's papers which use the Economic Census.[7]

According to Asher and Novosad (2012, 2014a, 2014b, 2014c, and 2015), the 1990, 1998, and 2005 waves of the Economic Census include firm location, product category, public or private ownership, number of employees, registration status (formal or informal), source of energy and financing, and caste and gender of the owner and employees. Asher and Novosad note that location directories are required to decode the village/ward codes that accompany firm entries in the Economic Census. Using these directories in conjunction with fuzzy matching algorithms, it is possible to create a village/ward-level panel using the Economic Census data. In the majority of what follows, I will use districts, rather than villages, as my unit of analysis. The assumptions of my main model are more suitable to districts than to villages, and this also means I can perform my analyses even in the absence of Asher and Novosad's location codes. I discuss the possibility of village-level analysis

[6]These details can be found here: `http://catalog.ihsn.org/index.php/catalog/3384/datafile/F2`.

[7]Of course, Asher and Novosad also do not have access to the 6th wave. Any pre-specified analysis here that employs these data will be conditional on my ability to acquire the 6th wave, and conditional on the 6th wave including the appropriate variables. This is one of the challenges of using observational data in conjunction with pre-registration, but I believe that the costs outweigh the benefits.

below.

5 Empirical strategy

Because roads are not randomly assigned, and importantly, their placement is likely to be endogenous to firm outcomes, I require a quasi-experimental approach to recover an unbiased estimate of the causal effect of road construction on firms. Below, I outline my research design and detail how I will test the hypotheses described in Section 3. I then describe how I will look for heterogeneous effects; perform statistical inference; and outline some robustness checks that I believe will serve as useful tests for my main estimation strategy. Note that all of the below is written under the assumption that I will only have access to the 2005 Economic Census for the time being, since the 2012 wave has not yet been released. Section 5.5 describes changes I will make if I get access to the 2012 data.

5.1 Research design

In this section, I describe my main empirical strategy.[8]

Previous work (Khanna (2014); Ghani, Goswami, and Kerr (2015)) posits the following: the Golden Quadrilateral roads were designed to connect several "nodal cities" in India. These cities are Delhi, Mumbai, Kolkata, and Chennai[9], are several of the largest and most economically important in India, and their locations predate the planning of the Golden Quadrilateral highway by hundreds of years. A straight line between these nodal cities, which approximates the least-cost path between them, should be exogenous to district economic outcomes, and is plausibly a valid instrument. The two conditions for using being on the straight line between two nodal cities as an instrumental variable are 1) that being on this line has predictive power over being on the Golden Quadrilateral

[8]See Appendix B for alternative empirical approaches that have been used by previous researchers, and that I will employ as well.

[9]Khanna (2014) uses only Delhi, Mumbai, Kolkata, and Chennai as nodal cities, while Ghani, Goswami, and Kerr (2015) show results with and without Bangalore. In my main specification, I will include Bangalore. I will also include Gandhinagar, since this city also clearly forms one of the corners of the Quadrilateral, and Guntur, which prevents the predicted GQ from traveling across water. This will increase both the the power of my first stage and the external validity of my approach, since the North-West branch of the Quadrilateral will provide meaningful variation for the IV.

road network; and 2) that being on the straight line affects district-level economic outcomes only through the road network.

Graphically, this instrument is as follows:

Figure 4: Instrumental Variables Approach

Notes: This figure displays the actual Golden Quadrilateral highway network (in black) and the straight lines between the nodal cities of Delhi, Mumbai, Kolkata, Bangalore, and Gandhinagar, in blue.

My actual empirical approach combines the ability to use the panel structure of the data with the identification benefits of the instrumental variables approach. I will estimate the following system

of equations using two-stage least squares:

$$Y_{dt} = \beta \widehat{\text{Road} \times \text{Post}}_{dt} + \delta_t + \gamma_d + \varepsilon_{dt}$$

and

$$\text{Road} \times \text{Post}_{dt} = \beta \text{Line} \times \text{Post}_{dt} + \delta_t + \gamma_d + \nu_{dt}$$

where Road $\times$ Post$_{dt}$ is equal to one if any part of the GQ road lies within the district and the year is 2005, and zero otherwise. Similarly, Line $\times$ Post$_{dt}$ is equal to one if any part of the straight line between the centroids of two nodal cities lies within the district and the year is 2005, and zero otherwise. Y_{dt} is an outcome of interest in district d in year t; δ_t are year fixed effects, γ_d are district fixed effects, and ε_{dt} and ν_{dt} are idiosyncratic error terms.

The identifying assumption of this approach is that the instrument, being on the straight line between two nodal districts in 2005, is correlated with being on a GQ road in 2005, and affects firm outcome Y only through the presence of the road. Formally, the identifying assumptions are that $Cov(Line \times Post_{dt}, Road \times Post_{dt}) \neq 0$ and $Cov(Line \times Post_{dt}, \varepsilon_{dt}) = 0$. Using this estimation strategy allows for the quasi-random variation from the instrumental variables approach, but also allows for full use of the panel dataset, including district fixed effects which will help with statistical power. This is my preferred specification.

5.1.1 Heterogeneity

We care not only about the district that lie directly next to the Golden Quadrilateral highways, but also about the rest of India's districts. In order to understand the net effects of the Golden Quadrilateral upgrades on firms, I employ two models that are similar to my preferred specification. The first assumes that there is a linear effect of distance. Using 2SLS, I will estimate:

$$Y_{dt} = \beta \widehat{\text{Distance to Road}} \times \text{Post}_{dt} + \delta_t + \gamma_v + \varepsilon_{dt}$$

and

$$\text{Distance to Road} \times \text{Post}_{dt} = \beta \text{Distance to Line} \times \text{Post}_{dt} + \delta_t + \gamma_v + \varepsilon_{dt}$$

This model has the advantage of being parsimonious, and is less likely to suffer from power concerns, but it is possible that the linearity assumption is incorrect.

In order to less parametrically estimate the effects at different distances to the road, I employ a spatial distributed lag model. Again, using two-stage least squares, I estimate:

$$Y_{dt} = \sum_{k=1}^{\kappa} \beta^k \text{Road } \widehat{\text{Bin}^k} \times \text{Post}_{dt} + \delta_t + \gamma_d + \varepsilon_{dt}$$

and κ versions of the below equation, one for each bin:

$$\text{Road Bin}^k \times \text{Post}_{dt} = \beta \text{Line Bin}^k \times \text{Post}_{dt} + \delta_t + \gamma_d + \varepsilon_{dt}$$

where $\text{Road Bin}^k \times \text{Post}_{dt}$ is equal to one if the centroid of district d falls within $[(k-1) \times 50, k \times 50)$ kilometers of a GQ road, and the year is 2005. Similarly, $\text{Line Bin}^k \times \text{Post}_{dt}$ is equal to one if the centroid of district d falls within $[(k-1) \times 50, k \times 50)$ kilometers of the straight line between nodal cities, and the year is 2005. (Notice that I do not include the un-interacted bins in the regression specification, since they will be absorbed by the district fixed effects.) I estimate the results for all bins jointly. Notice also that I can calculate the cumulative effects for bins 1 to κ by calculating $\sum_{k=1}^{\kappa} \beta^k$; I will do this as well. The choice of 50 kilometer bins is somewhat arbitrary; I will also show results with bins with 25 and 100 km widths. I will omit the closest bin (0 to 50, 25, or 100 km), and my farthest bin will be 500 km and up in all cases. I will limit the width of my largest bin such that no two bins are overlapping.

5.1.2 Outcome variables

In Section 3 I outlined the hypotheses I will test. Here, I describe in detail how I will test each of them. Unless otherwise noted, I will use my preferred specification outlined above for each hypothesis. I lay out the outcome variables (and where necessary, alternative specifications) for each test below. For these tests, unless otherwise noted, I restrict my sample to tradable (manufacturing) goods, defined as 2004 NIC Divisions 15-37.

1. φ_X^* decreases, so firms enter the exporting sector, and their labor share increases
   Outcome variable: I cannot test this prediction directly, since I do not observe which firms export.

2. Wages increase
   Outcome variable: I cannot test this prediction directly using data from the Economic Census since I do not observe wages.[10]

3. **Formal firms:** φ_F^* increases
   - The number of formal firms (including exporters) falls
     Outcome variable: Number of formal firms.
   - The remaining formal firms (including exporters) employ more workers
     Outcome variable: Workers per firm (formal sector)
   - The labor share in the formal sector is ambiguous
     Outcome variable: Formal labor / total labor

4. **Informal firms:** φ_I^* increases
   - The number of informal firms is ambiguous (the lowest-producing exit; but some low-productivity formal firms now become informal)
     Outcome variable: Number of informal firms
   - The informal firms employ more workers
     Outcome variable: Workers per firm (informal sector)

[10] Data on agricultural wages are available from other sources. I have yet to find a dataset on manufacturing wages.

– The labor share in the informal sector is ambiguous

Outcome variable: Informal labor / total labor

5. The total number of firms declines, and with it, the number of varieties produced at home declines.

Outcome variable: Number of firms

Outcome variable: Number of (2,3,4-digit) NIC codes produced

6. Overall welfare increases.

Outcome variable: GDP/worker, GDP/capita

7. Districts will specialize in sectors in which they have comparative advantage (Note that figuring out which sectors have comparative advantage ex ante is non-trivial: theoretically, this is done using autarky prices, which I do not observe. Instead, these tests are ex post measures of specialization.)

Outcome variable: Total number of sectors with active firms

Outcome variable: $\max_i \left\{ \frac{Workers_i}{\sum_i Workers_i} \right\}$ where i is a manufacturing industry (2004 NIC Divisions 15-37).[11] As a robustness check, I also test this hypothesis using a Hirfendahl index over labor shares: $H = \sum_i \left(\frac{Workers_i}{\sum_i Workers_i} \right)^2$. I perform robustness checks using the three and four digit NIC codes as well.

8. Women and marginalized groups will own more firms.

Outcome variable: Number of women owning firms

Outcome variable: Number of Scheduled Castes, Scheduled Tribes, or Other Backwards Caste firm owners.

9. Employment of women will increase.

Outcome variable: Number of women employed (overall, and separately by formal/informal).

10. Employment of children will decrease.

Outcome variable: Number of children employed (overall, and separately by formal/informal).

[11] A Division is a two-digit NIC code.

5.2 Inference

In this section, I discuss the statistical inference procedures I will apply when estimating my causal effects. First, I discuss clustering of standard errors. Next, I lay out a plan for multiple testing corrections.

5.2.1 Standard errors

Because this is a two-period difference-in-difference model, these data are essentially already collapsed to "before" and "after" periods, so OLS standard errors should produce an appropriate rejection rate (see Bertrand, Duflo, and Mullainathan (2004) for details).[12] We might still be concerned about over-rejection due to spatial correlation. I will show two sets of robustness checks for my standard errors: first, I will cluster by state.[13] Second, I will use a Conley (2008) spatial HAC method, where I use bandwidths ranging from 10 to 1,000 kilometers from the district centroid. I will do this using both a uniform kernel and a triangular kernel. Clustering at the district level, however, remains my preferred standard error adjustment, in large part because there are enough districts that the asymptotic assumptions necessary for the clustering estimator are plausibly satisfied.

5.2.2 Multiple testing

Because I am testing multiple hypotheses that are likely not independent, I need to correct my recovered p-values to account for multiple testing. I do this in two ways.

First, I follow Kling, Liebman, and Katz (2007) in creating hypothesis-based indices. For each hypothesis category above (3, 4, 5, 7, and 8-9-10 as one category: 1 and 2 are not tested; 6 has only one test), I will create an index Y that is defined as the equally-weighted average of z-scores of each component, where an increase in economic activity or towards development is coded in the same direction each time. I calculate this z-score by taking a district's outcome measure, subtracting

[12] Indeed, clustering at the district level would be inappropriate: in a two-period model, district fixed effects induce negative correlation between the residuals of the two periods, leading clustered standard errors to understate the true uncertainty in the model.

[13] The reason that I do not make this my main standard error correction method is that clustered standard errors rely heavily on asymptotic assumptions. There are only 35 states in India, which is a small enough number that we might be concerned about biased standard error estimates. Nevertheless, I present standard errors clustered at the state level as an additional test.

the control group mean outcome, and dividing by the control group standard deviation (where the control group is a district that did not get a road.). These indices demonstrate whether roads had an effect on each of my three sets of related outcomes using a mean effects approach, which reduces the number of hypotheses that I test. Pre-specification here of my hypothesis categories is critical to the later credibility of this index-based approach.

Second, I also perform a multiple inference correction procedure to my p-values. For each hypothesis, I will present the standard p-value without any corrections. I will also present two sets of corrected p-values, using the family-wise error rate (FWER) and the false discovery rate (FDR) procedures outlined in Anderson (2008). In both of these corrections, I use the three numbered hypothesis categories above as "families." The FWER correction adjusts p-values for the chance that any test nested within a family has a false positive. The FDR procedure corrects p-values for the expected share of false positives within a family. I prefer the FDR, because the FWER can be seen as overly conservative, but for completeness, I will present both.[14]

5.3 Robustness checks

In this section, I pre-specify several robustness checks in order to engender confidence in my results. In particular, I run a robustness check assigning treatment at the district level; and I compare my results to placebo results for districts along the East-West and North-South Corridor Highways, which at the time of the 2005 Economic Census, were not yet constructed. I have also already laid out several robustness checks in the sections above, where I try to be clear about what is my main specification, and what alternative specifications are.

5.3.1 NS/EW Highways

During the time period that the Golden Quadrilateral project was planned, NHAI also planned an additional highway system, the North-South and East-West Corridor projects. At the end of 2004, only 4% of these systems were completed, and this figure *includes* the segments that these

[14] As a robustness check, I follow Casey, Glennerster, and Miguel (2012) and also present FWER and FDR corrections based on the indices described above. Because I have only three hypothesis categories, with ten individual tests, this will likely add little to the multiple inference corrections performed on each individual hypothesis, so I leave this as a robustness check rather than making it my main specification.

projects share with the Golden Quadrilateral. Essentially, then, none of these additional projects were completed by the 2005 Economic Census. As a result, I can run my main specification using these highways rather than the Golden Quadrilateral project as my treatment.[15] Ideally, I will see no effect of these highways on firm activity.

5.4 Districts vs. villages

Because this is a pre-analysis plan for data whose collection I did not design, it is important that I build in some conditionality. It may be possible to match villages in the Economic Census to my village shapefiles, if the location directory acquired by Asher and Novosad is included in my dataset.[16] If I am able to create a village-level panel, I will also conduct my analyses at the village level. To do this, I estimate the following system of equations using two-stage least squares:

$$Y_{vt} = \alpha + \beta \widehat{\text{Road} \times \text{Post}}_{vt} + \delta_t + \gamma_v + \varepsilon_{vt}$$

and

$$\text{Road} \times \text{Post}_{vt} = \alpha + \beta \text{Line} \times \text{Post}_{vt} + \delta_t + \gamma_v + \varepsilon_{vt}$$

where $\text{Road} \times \text{Post}_{vt}$ is equal to one if a village lies within a 25 kilometer buffer of the GQ road and the year is 2005, and zero otherwise.[17] Similarly, $\text{Line} \times \text{Post}_{vt}$ is equal to one if the village if a village is within a 25 kilometer buffer of the straight line between the centroids of two nodal cities and the year is 2005, and zero otherwise. I use these buffers to absorb measurement error in the road shapefile; to capture the fact that villages that are very close to the road likely experience similar treatment effects to those directly next to the road; and to make my results more comparable to the district level effects estimated earlier.[18]

Note, that the assumptions in my model are less likely to be satisfied at the village level. Neverthe-

[15] In doing this, I use Porbandar, Lucknow, Guwahati, and Silchar as nodal cities for the North-South section, and Srinagar, Delhi, Bangalore, and Thiruvananthapuram for the East-West section of the Corridor.

[16] It was requested upon the date of purchase, but is not a standard dataset from MoSPI.

[17] As above, I define a village to be within the buffer if the village's centroid lies within the buffer.

[18] Note that this is another advantage of the pre-analysis plan: I can guarantee that this choice of buffer was not driven by p-hacking.

less, testing whether specialization and industry concentration occur at both short and large spatial scales when opening to trade has policy relevance.

5.5 2012 Economic Census

The above plan has been written for use with the 1990, 1998, and 2005 waves of the Economic Census only. When the 2012 Economic Census is released, I will apply for additional funding to purchase it.[19] The addition of this wave will lead to several adjustments to the above analysis, because the majority of the North-South East-West Corridor was completed by the time of data collection for the 2012 Economic Census.[20] I detail these changes below.

5.5.1 Main analysis

I will use the same outcome variables and specifications outlined above with the 2012 Economic Census. With this additional wave, however, I will consider not only Golden Quadrilateral highway upgrades, but also North-South East-West Corridor highway upgrades. I will run the same main specification and binned specification, with the adjustment that the treatment indicator will turn on in 2005 for Golden Quadrilateral districts (and remain on in 2012), and will turn on in 2012 for North-South East-West Corridor districts. The nodal cities for my North-South East-West Corridor straight lines will be: Porbandar, Lucknow, Guwahati, and Silchar as nodal cities for the North-South section, and Srinagar, Delhi, Bangalore, and Thiruvananthapuram for the East-West section of the Corridor. In order to avoid concerns about endogeneity of treatment timing, I will consider districts "treated" by the North-South East-West Corridor in 2012 whether they had gotten their upgrade by 2012 or not.[21] In general, my main analysis will remain largely unchanged - I will simply update it to incorporate the North-South East-West Corridor upgrades as well.

[19] I assume here that the 2012 Census is released before my paper is completed. If this is not the case, I will use up through the 2005 wave only.

[20] 77% of the NS-EW Corridor was completed by April 30, 2011. See the image here: `http://www.nhai.org/images/April11/NHDP%20PH.I,%20II,%20III%20%20.jpg`.

[21] This assigns the same treatment timing to every districts along the North-South East-West Corridor. The downside to doing this is that it introduces some measurement error in the treatment variable; in all likelihood, however, considering villages "treated" when they have not yet in fact been treated should attenuate any treatment effect estimates.

5.5.2 Robustness checks and heterogeneity

In the event that I get the 2012 Economic Census data, I will adjust my robustness checks and heterogeneous effects analyses to incorporate the North-South East-West Corridor upgrades. Below, I describe how I will carry out the placebo test specified above; and I discuss heterogeneous effects for the two sets of upgrades.

Because the 2012 Economic Census data will allow me to look at villages who have experienced North-South East-West Corridor upgrades, I can no longer use the planned North-South East-West Corridor as a placebo treatment for my Golden Quadrilateral roads in the same way. In order to preserve this robustness check, I will run my main specification for the Golden Quadrilateral roads using data up to 2005 only. Using this more limited dataset, I will run the placebo analysis specified above, just as though I did not have access to the 2012 dataset. I unfortunately do not have a similar placebo check for the North-South East-West Corridor upgrades.

We might think that the treatment effects of roads on firms changes over time, such that the effect of a road built one or two years ago is different from the effect of a road built seven or eight years ago. In order to investigate this possibility, I will add the following specification:

$$Y_{dt} = \beta_1 \widehat{\text{Road}^{GQ} \times 2005_{dt}} + \beta_2 \widehat{\text{Road}^{GQ} \times 2012_{dt}} + \beta_3 \widehat{\text{Road}^{NSEW} \times 2012_{dt}} + \delta_t + \gamma_d + \varepsilon_{dt}$$

and

$$\begin{aligned}
\text{Road}^{GQ} \times 2005_{dt} &= \beta \text{Line}^{GQ} \times 2005_{dt} + \delta_t + \gamma_d + \varepsilon_{dt} \\
\text{Road}^{GQ} \times 2012_{dt} &= \beta \text{Line}^{GQ} \times 2012_{dt} + \delta_t + \gamma_d + \varepsilon_{dt} \\
\text{Road}^{NSEW} \times 2012_{dt} &= \beta \text{Line}^{NSEW} \times 2012_{dt} + \delta_t + \gamma_d + \varepsilon_{dt}
\end{aligned}$$

Again, I estimate these equations jointly. In this specification, Road $(\text{Line})^{GQ} \times 2005_{vt}$ is equal to one if a GQ road (straight line) lies anywhere within a district and the year is 2005, and zero otherwise (this is identical to the treatment variable above); Road $(\text{Line})^{GQ} \times 2012_{dt}$ is equal to one

if a GQ road (straight line) lies anywhere within a district and the year is 2012, and zero otherwise; and Road (Line)$^{NSEW} \times 2012_{dt}$ is equal to one if a NS-EW (straight line) road lies anywhere within a district and the year is 2012, and zero otherwise. Using this specification, I can test whether having had the road for a longer period of time is different than having had the road for one or two years by comparing β_1 to $\beta_1 + \beta_2$; I can also compare the effect of the Golden Quadrilateral road to the effect of the North-South East-West corridor road, by comparing β_1 to β_3 (or $\beta_1 + \beta_2$ to β_3).

References

Aleman-Castilla, Benjamin (2006). "The Effect of Trade Liberalization on Informality and Wages: Evidence from Mexico". http://cep.lse.ac.uk/pubs/download/dp0763.pdf.

Anderson, Michael L. (2008). "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". In: *Journal of the American Statistical Association* 103.484, pp. 1481–1495. ISSN: 0162-1459. DOI: 10.1198/016214508000000841.

Asher, Sam and Paul Novosad (2012). *Seed Capital: The Impact of Agricultural Output on Local Economic Activity in India*. Working Paper.

– (2014a). *Digging for Development: Mining Booms and Local Economic Development in India*. Working Paper.

– (2014b). *Dirty Politics: Natural Resource Wealth and Politics in India*. Working Paper.

– (2014c). *The Employment Effects of Road Construction in Rural India*. Working Paper.

– (2015). *Politics and Local Economic Growth: Evidence from India*. Working Paper.

Banerjee, Abhijit, Esther Duflo, and Nancy Qian (2012). *On the Road: Access to Transportation Infrastructure and Economic Growth in China*. Working Paper 17897. National Bureau of Economic Research.

Barman, Abheek (2013). *Informal workers, making up 90% of workforce, won't get a good deal till netas notice them*. The Economic Times. (Visited on 08/25/2015).

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). "How Much Should We Trust Differences-In-Differences Estimates?" In: *Quarterly Journal of Economics* 119.1.

Besley, Timothy and Robin J. Burgess (2004). "Can Labor Regulation Hinder Economic Performance? Evidence from India". In: *The Quarterly Journal of Economics* 119.

Casaburi, Lorenzo, Rachel Glennerster, and Tavneet Suri (2013). *Rural Roads and Intermediated Trade: Regression Discontinuity Evidence from Sierra Leone*. SSRN Scholarly Paper ID 2161643. Rochester, NY: Social Science Research Network.

Casey, Katherine, Rachel Glennerster, and Edward Miguel (2012). "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan*". In: *The Quarterly Journal of Economics* 127.4, pp. 1755–1812. ISSN: 0033-5533, 1531-4650. DOI: 10.1093/qje/qje027.

Colmer, Jonathan (2015). *The Productivity Effects of Labour Reallocation: Evidence from India*. Working Paper.

Conley, Timothy G. (2008). "Spatial Econometrics". In: *The New Palgrave Dictionary of Economics*. Ed. by Steven N. Durlauf and Lawrence E. Blume. 2nd ed. Basingstoke: Nature Publishing Group, pp. 741–747. ISBN: 978-0-333-78676-5.

Donaldson, Dave (2010). *Railroads of the Raj: Estimating the Impact of Transportation Infrastructure*. Working Paper 16487. National Bureau of Economic Research.

Duflo, Esther and Rohini Pande (2007). "Dams". In: *The Quarterly Journal of Economics* 122.2, pp. 601–646. ISSN: 0033-5533, 1531-4650. DOI: 10.1162/qjec.122.2.601.

Faber, Benjamin (2014). "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System*". In: *The Review of Economic Studies*. ISSN: 0034-6527, 1467-937X. DOI: 10.1093/restud/rdu010.

Ghani, Ejaz, Arti Grover Goswami, and William R. Kerr (2015). "Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing". In: *The Economic Journal*, n/a–n/a. ISSN: 1468-0297. DOI: 10.1111/ecoj.12207.

Hasan, Rana and Karl Robert L. Jandoc (2012). *Labor Regulations and the Firm Size Distribution in Indian Manufacturing*. Working Paper.

Hsieh, Chang-Tai and Peter J. Klenow (2009). "Misallocation and Manufacturing TFP in China and India". In: *The Quarterly Journal of Economics* 124.4, pp. 1403–1448. ISSN: 0033-5533, 1531-4650. DOI: 10.1162/qjec.2009.124.4.1403.

Khanna, Gaurav (2014). *The Road Oft Taken: The Route to Spatial Development*. SSRN Scholarly Paper ID 2426835. Rochester, NY: Social Science Research Network.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz (2007). "Experimental Analysis of Neighborhood Effects". In: *Econometrica* 75.1, pp. 83–119. ISSN: 1468-0262. DOI: 10.1111/j.1468-0262.2007.00733.x.

Melitz, Marc J. (2003). "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity". In: *Econometrica* 71.6, pp. 1695–1725. ISSN: 1468-0262. DOI: 10.1111/1468-0262.00467.

World Bank (2014). *World Bank Group's Infrastructure Spending Increases to US $24 Billion*. The World Bank: News.

Pre-Analysis Plan Appendix

Appendix A Theoretical Appendix

In this section, I describe the theoretical underpinnings of the predictions discussed in Section 3, and outline a theoretical model of informal firms and trade. I draw on two canonical classes of trade models to shape my theoretical predictions: the Ricardian model of comparative advantage, popularized by Dornbusch, Fischer, and Samuelson (1977); and the "new new trade theory" monopolistic competition model of Melitz (2003).

I first discuss the predictions of the Ricardian models.

Ricardian predictions

These models are driven by comparative advantage, and yield the prediction that in autarky, districts will produce goods in a wide range of industries. As trade costs decrease, districts will specialize (completely or incompletely) in industries (in particular, districts will specialize in the industries in which they have comparative advantage). As a result, the number of industries whose goods a village produces decreases as trade costs decline. The share of firms and labor increases in the "specialization" industries, and declines in the non-specialization industries. Overall welfare increases.

New new trade predictions

These models are driven by consumers with love-of-variety preferences and monopolistically competitive firms with increasing returns to scale technologies. Critical to the set up of the model is that firms, upon deciding to enter the market, get a productivity draw from a distribution. Based on this draw, each firm immediately decides whether to exit and not produce, enter as an informal firm, or enter as a formal firm. Informal firms do not pay taxes. Formal firms do pay taxes, and also benefit from a productivity shifter. Zero profit and free entry conditions pin down two productivity cutoffs: one below which no firms will enter, and another below which firms will enter informally. This allows for a mass of firms that are informal to coexist with a mass of formal firms. In the

open economy version of the model, a subset of formal firms have productivity draws high enough that allow them to export. Again, I derive a new productivity cutoff above which formal firms will decide to export.

As trade costs fall, it becomes easier for firms to enter the exporting sector, lowering the export productivity cutoff. These high-productivity, high labor-demand exporting firms are able to achieve high profits. As trade costs fall, labor share in this sector increases. These firms will bid up wages. This makes it more challenging for lower-productivity firms to enter the market: the cutoffs for informal firm entry and for formal (non-exporter) firm entry will increase with a decline in trade costs. However, the informal and formal firms that are able to remain will have higher productivity and therefore higher labor demand than before. These two competing effects make the overall labor demand effect in the non-exporting sector ambiguous. The number of firms, and therefore varieties, produced at home falls as trade costs decline, but consumers have access to more varieties than before, which, when combined with love-of-variety preferences, leads to direct welfare gains.

More formally, I borrow the following extension to the Melitz model from Aleman-Castilla (2006) in order to discuss the effects of opening to trade on informal and formal firms.

THE CLOSED ECONOMY:

Consumption:

Suppose that consumers have Dixit-Stiglitz love of variety preferences over a continuum of differentiated goods, ω:

$$U = \left[\int_{\omega\in\Omega} x(\omega)^{(\sigma-1)/\sigma} d\omega\right]^{\sigma/(\sigma-1)}$$

It is standard to show that:

$$x(\omega) = Q\left[\frac{p(\omega)}{P}\right]^{-\sigma}$$

and

$$r(\omega) = R\left[\frac{p(\omega)}{P}\right]^{1-\sigma}$$

where $p(\omega)$ is the price of variety ω, $P \equiv \left[\int_{\omega\in\Omega} p(\omega)^{1-\sigma} d\omega\right]^{1/(1-\sigma)}$, $R = PQ = \int_{\omega\in\Omega} r(\omega) d\omega$ is aggregate expenditure, and σ is the elasticity of substitution.

Production:

Each firm will choose to produce a unique horizontally-differentiated variety. Firms are monopolistically competitive, and have heterogeneous labor productivity φ. All firms with the same φ will behave identically, so I continue by indexing firms only by φ. Production requires a fixed cost of production, f, and a constant marginal cost that is inversely proportional to a firm's productivity, all paid in units of labor, so the total amount of labor required to produce x units of a given variety is:

$$l = f + \frac{x}{\varphi}$$

Firm profits can be written:

$$\pi(\varphi) = px - w\left(f + \frac{q}{\varphi}\right)$$

It is straightforward to show that firms will charge:

$$p = \frac{\sigma}{\sigma - 1}\frac{w}{\varphi}$$

Using the solution to the consumer's problem, the profit-maximizing level of profits is:

$$\pi(\varphi) = \varphi^{\sigma-1} B - wf$$

Where

$$B \equiv \frac{E\left(\frac{\sigma-1}{\sigma}\right)^{\sigma-1} w^{1-\sigma}}{\sigma P^{1-\sigma}}$$

For notational convenience, define:

$$\rho = \frac{\sigma - 1}{\sigma}$$

The timing of the firm is as follows: firms pay a fixed cost of entry, f. They then receive a productivity draw from a distribution with CDF $G(\varphi)$. Upon observing this draw, if profits are negative, they will exit immediately. Free entry will drive profits to zero, such that:

$$\begin{aligned} \pi(\varphi^*) &= 0 \\ \pi(\varphi^*) &= \varphi^{\sigma-1} B - wf = 0 \\ \varphi^{\sigma-1} B &= wf \\ \varphi^* &= \left(\frac{wf}{B}\right)^{\sigma-1} \end{aligned}$$

φ^* is the cutoff below which firms will instantly exit the market.

Informality:

Now suppose that formal and informal firms behave somewhat differently. In particular, informal firms do not pay taxes (in my empirical context, they do so legally). Formal firms, in contrast, do pay taxes $0 < \alpha < 1$ on every unit of labor. Formal firms also benefit from a productivity boost, β, relative to the informal sector. Suppose further that $\beta > \alpha$. Firms have to pay a fixed cost to enter the market. Informal firms pay f_I, whereas formal firms pay f_F, where $f_F > f_I$.

We begin with a closed economy. We can write firm profits as:

$$\pi_I(\varphi) = B\left(\frac{\varphi}{w}\right)^{\rho/(1-\rho)} - wf_I$$

Note that this is the same expression for profits as before, only with a new subscript on fixed costs. This implies that:

$$\varphi_I^* = \left(\frac{wf_I}{B}\right)^{\sigma-1}$$

Notice that $\varphi_I^* \equiv \varphi^*$, the overall industry cutoff. We now turn to formal firms:

$$\pi_F(\varphi) = B\left(\frac{(1+\beta)\varphi}{(1+\alpha)w}\right)^{\rho/(1-\rho)} - (1+\alpha)wf_F$$

A firm will choose to become formal if and only if its expected profits are greater in the formal sector than the informal sector:

$$\left[B\left(\frac{(1+\beta)\varphi}{(1+\alpha)w}\right)^{\rho/(1-\rho)} - (1+\alpha)wf_F\right] \geq B\left(\frac{\varphi}{w}\right)^{\sigma-1} - wf_I$$

We can write the formal firm cutoff productivity as:

$$\varphi_F^* = \left[\frac{w^{1/(1-\rho)}[(1+\alpha)f_F - f_I]}{B\left[\left(\frac{1+\beta}{1+\alpha}\right)^{\rho/(1-\rho)}\right]}\right]^{(1-\rho)/\rho}$$

In equilibrium, φ_F^* will pin down the share of firms in the formal and informal sectors.

Aggregation:

Let M be the total mass of active firms in the district, where this is a fraction of an unobserved potential mass of entrants $M_{ENTRANTS}$. We can define the probability that a firm will be able to successfully enter the market as $p = 1 - G(\varphi_I^*)$. Therefore, $M = pM_{ENTRANTS}$.

Note that a fraction of the active firms will produce formally; this fraction is determined by the proportion of firms that draw productivities above the formal cutoff: $p_F = 1 - G(\varphi_F^*)$. This allows us to write: $M_F = p_F M$ and $M_I = (1 - p_F)M$, so, finally:

$$M = M_I + M_F$$

Because any firm drawing $\varphi < \varphi^*$ will exit immediately, I can define the distribution of productivities

conditional on successful entry as:

$$\mu(\varphi) = \begin{cases} \frac{g(\varphi)}{1-G(\varphi^*)} & \text{if } \varphi \geq \varphi^* \\ 0 & \text{otherwise} \end{cases}$$

Notice that we can compute the aggregate productivity level among the informal and formal firms as a function of their respective cutoffs:

$$\tilde{\varphi}_I(\varphi^*) = \left[\frac{1}{1 - G(\varphi_I^*)} \int_{\varphi_I^*}^{\varphi_F^*} \varphi^{\sigma-1} g(\varphi) d\varphi \right]^{1/(\sigma-1)}$$

and

$$\tilde{\varphi}_F(\varphi^*) = \left[\frac{1}{1 - G(\varphi_F^*)} \int_{\varphi_F^*}^{\infty} \varphi^{\sigma-1} g(\varphi) d\varphi \right]^{1/(\sigma-1)}$$

This allows us to write down an expression for the weighted total productivity level in the economy, which takes into account the productivity shifters present among formal firms:

$$\tilde{\varphi}_{TOTAL} = \left[\frac{1}{M} \left(M_I {\tilde{\varphi}_I}^{\sigma-1} + M_F \left[\frac{(1+\alpha)}{(1+\beta)} \tilde{\varphi}_F \right]^{\sigma-1} \right) \right]^{1/(\sigma-1)}$$

The open economy:

Trade:

We have thus far only considered a closed economy. Were there to be trade, I assume that only formal firms are able to export. In order to export, firms pay two trade-related costs: a fixed cost of exporting, $f_X \geq f_F$, and an iceberg trade cost $\tau \geq 1$. In order to partition the formal sector between firms producing only domestically and traders, I assume (as is standard) that $\tau^{\sigma-1} f_X > f_F$. We can write the exporting firm's profit as:

$$\pi_X(\varphi) = B \left(\frac{(1+\beta)\varphi}{\tau(1+\alpha)w} \right)^{\rho/(1-\rho)} - (1+\alpha) w f_X$$

A firm will choose to export if:

$$
\begin{aligned}
B\left(\frac{(1+\beta)\varphi}{\tau(1+\alpha)w}\right)^{\rho/(1-\rho)} &\geq (1+\alpha)wf_X \\
\left(\frac{(1+\beta)\varphi}{\tau(1+\alpha)w}\right)^{\rho/(1-\rho)} &\geq \frac{(1+\alpha)wf_X}{B} \\
\frac{(1+\beta)\varphi}{\tau(1+\alpha)w} &\geq \left(\frac{(1+\alpha)wf_X}{B}\right)^{(1-\rho)/\rho} \\
\varphi_X^* &= \left(\frac{(1+\alpha)wf_X}{B}\right)^{(1-\rho)/\rho}\frac{\tau(1+\alpha)w}{(1+\beta)}
\end{aligned}
$$

Notice that this cutoff is increasing in the wage, w, and in trade costs, τ.

In the model with trade, we can write down the new mass of firms in the economy as:

$$M = M_I + M_F + M_X$$

where M_X is now the mass of firms that are exporting.

$$
\begin{aligned}
\tilde{\varphi}_I(\varphi_I^*) &= \left[\frac{1}{1-G(\varphi_I^*)}\int_{\varphi_I^*}^{\varphi_F^*}\varphi^{\sigma-1}g(\varphi)d\varphi\right]^{1/(\sigma-1)} \\
\tilde{\varphi}_F(\varphi_F^*) &= \left[\frac{1}{1-G(\varphi_F^*)}\int_{\varphi_F^*}^{\varphi_X^*}\varphi^{\sigma-1}g(\varphi)d\varphi\right]^{1/(\sigma-1)} \\
\tilde{\varphi}_X(\varphi_X^*) &= \left[\frac{1}{1-G(\varphi_X^*)}\int_{\varphi_X^*}^{\infty}\varphi^{\sigma-1}g(\varphi)d\varphi\right]^{1/(\sigma-1)} \\
\tilde{\varphi}_{TOTAL} &= \left[\frac{1}{M}\left(M_I{\tilde{\varphi}_I}^{\sigma-1} + M_F\left[\frac{(1+\alpha)}{(1+\beta)}\tilde{\varphi}_F\right]^{\sigma-1} + M_X\left[\frac{(1+\alpha)}{\tau(1+\beta)}\tilde{\varphi}_X\right]^{\sigma-1}\right)\right]^{1/(\sigma-1)}
\end{aligned}
$$

As in the standard Melitz model, as τ decreases, φ^* increases. The intuition for this result is that opening to trade leads to a higher average profit per firm, driven by the increased profits of the exporters. The zero profit condition then shifts up, since exporters gain profits due to increased export opportunities. But since foreign exporters are entering the home market, non-exporting firms experience a declining price index, and so φ^* increases while φ_X^* declines.

Notice also that welfare unambiguously increases, though the number of varieties produced domestically will fall (which is in this model equivalent to saying that the number of firms will fall).

The new piece of this is informality. Because we have monopolistic competition and CES preferences, trade does not lead to increased product market competition. Instead, a decrease in trade costs will increase the labor demanded by the more productive firms. New entrants into this sector will bid up wages, forcing the least productive firms to exit. Earlier, we defined:

$$\varphi_F^* = \left[\frac{w^{1/(1-\rho)}[(1+\alpha)f_F - f_I]}{B\left[\left(\frac{1+\beta}{1+\alpha}\right)^{\rho/(1-\rho)}\right]} \right]^{(1-\rho)/\rho}$$

Differentiating this with respect to τ yields:

$$\frac{\partial \varphi_F^*}{\partial \tau} = \frac{1}{\rho} \left[\frac{w^{1/(1-\rho)}\left[(1+\alpha)f_F - f_I\right]}{B\left[\left(\frac{1+\beta}{1+\alpha}\right)^{\rho/(1-\rho)}\right]} \right]^{(1-2\rho)/\rho} w^{\rho/(1-\rho)} \frac{\partial w}{\partial \tau} < 0$$

That is, a decrease in τ will lead to an *increase* in φ_F^*, since $\frac{\partial w}{\partial \tau} < 0$.

So: a decline in trade costs leads to an increase in φ_I^* and φ_F^*, and a decrease in φ_X^*. But what does it mean for labor?

Because both φ_I^* and φ_F^* increase, the effect on informal sector labor is ambiguous. Because φ_X^* declines with a decline in τ, there will be an increase in the labor force in the exporting formal

sector. This is because prices and outputs in each sector are defined as follows:

$$
\begin{aligned}
p_I &= \frac{w}{\rho\varphi} \\
q_I &= Q\left[\frac{w}{P\rho\phi}\right]^{-\sigma} \\
p_F &= \frac{(1+\alpha)w}{\rho(1+\beta)\varphi} \\
q_F &= Q\left[\frac{(1+\alpha)w}{P\rho(1+\beta)\varphi}\right]^{-\sigma} \\
p_X &= \frac{\tau(1+\alpha)w}{\rho(1+\beta)\varphi} \\
q_X &= Q\left[\frac{\tau(1+\alpha)w}{P\rho(1+\beta)\varphi}\right]^{-\sigma}
\end{aligned}
$$

This, along with technology, pins down labor demands:

$$
\begin{aligned}
l_I &= f_I + Q\left[\frac{P\rho}{w}\right]^{\sigma}\varphi^{\sigma-1} \\
l_F &= f_F + Q\left[\frac{P\rho(1+\beta)}{w(1+\alpha)}\right]^{\sigma}\varphi^{\sigma-1} \\
l_X &= f_X + Q\left[\frac{P\rho(1+\beta)}{\tau w(1+\alpha)}\right]^{\sigma}\varphi^{\sigma-1}
\end{aligned}
$$

Total employment, then, is:

$$
L = L_e + \int_{\varphi_I^*}^{\varphi_F^*} l_I(\varphi)\mu(\varphi)d\varphi + \int_{\varphi_F^*}^{\varphi_X^*} l_F(\varphi)\mu(\varphi)d\varphi + \int_{\varphi_X^*}^{\infty} l_X(\varphi)\mu(\varphi)d\varphi
$$

where L_e is the labor used to pay for the fixed entry costs, and $\mu(\varphi)$ is the conditional productivity distribution defined above.

Market clearing requires that:

$$
L_e = M_e f_e
$$

where M_e is the mass of new entrants. Equilibrium also requires that some firms exit. In order to do this, each successfully-entered firm has a probability δ that they will draw a negative shock and exit. This probability is exactly calibrated to match the probability of new entrants, such that:

$$(1 - G(\varphi_I^*))M_e = \delta M$$

We can write the free entry condition as:

$$\bar{\pi} = \frac{\delta f_e}{1 - G(\varphi_I^*)}$$

This pins down the labor employed by new entrants:

$$L_e = M_e f_e = \frac{\delta M}{1 - G(\varphi_I^*)} f_e = M\bar{\pi}$$

From here, notice that because a decline in τ increases φ_I^*, it will also increase $\bar{\pi}$, and in turn, L_e.

It is also easy to show that a decline in τ will increase the labor demand in the formal exporting sector:

$$\int_{\varphi_X^*}^{\infty} l_X(\varphi)\mu(\varphi)d\varphi = \frac{1}{1 - G(\varphi_I^*)} \int_{\varphi_X^*}^{\infty} \left[f_X + Q \left(\frac{P\rho(1+\beta)}{\tau(1+\alpha)w} \right)^{\sigma} \varphi^{\sigma-1} \right] g(\varphi)d(\varphi)$$

Clearly, as τ decreases, the term in brackets increases, φ_I^* increases, so that $\frac{1}{1-G(\varphi_I^*)}$ increases, and φ_X^* decreases, so that we are now integrating over a larger range. Therefore, the employment share in the export sector is increasing as τ declines.

Finally, we can consider the impact on labor in the non-exporting formal sector:

$$\begin{aligned} \int_{\varphi_F^*}^{\varphi_X^*} l_F(\varphi)\mu(\varphi)d\varphi &= \frac{1}{1 - G(\varphi_I^*)} \int_{\varphi_F^*}^{\varphi_X^*} \left[f_F + Q \left(\frac{P\rho(1+\beta)}{(1+\alpha)w} \right)^{\sigma} \varphi^{\sigma-1} \right] g(\varphi)d(\varphi) \\ &= \frac{1 - G(\varphi_F^*)}{1 - G(\varphi_I^*)} \left[f_F + Q \left(\frac{P\rho(1+\beta)}{(1+\alpha)w} \right)^{\sigma} \tilde{\varphi}_F(\varphi_F^*)^{\sigma-1} \right] \end{aligned}$$

This is ambiguous: $\frac{\partial \tilde{\varphi}_F(\varphi_F^*)}{\partial \tau} < 0$, since a decline in τ leads to an increase in φ_F^*, and $\tilde{\varphi}_F$ is increasing in φ_F^*. But without further distributional assumptions, we cannot sign $\frac{1-G(\varphi_F^*)}{1-G(\varphi_I^*)}$.

To summarize, these results suggest that as trade costs fall:

1. φ_X^* will decrease, so more firms will enter the exporting sector, and their labor share will increase
2. This drives up wages.
3. φ_F^* will increase, making it more difficult to enter the formal sector, but the remaining formal firms employ more workers, so the overall labor effect is ambiguous.
4. φ_I^* will increase, making the informal firms with the lowest φ draws exit, but the remaining informal firms employ more workers, so the overall labor effect is ambiguous.
5. Overall welfare increases due to an increase in available varieties.

Note that the model above was defined for goods within one sector. It is trivial to extend the model to multiple sectors: simply give consumers two-tiered preferences, with a homothetic (usually Cobb-Douglas) upper tier over goods, and a CES lower tier over varieties. The above results should hold for each sector.

Appendix B Empirical Appendix

The simplest model that has been used in this literature is a difference-in-differences strategy. I will follow Datta (2012) in estimating an equation of the following form:

$$Y_{dt} = \beta \text{Road} \times \text{Post}_{dt} + \delta_t + \gamma_d + \varepsilon_{dt}$$

where Y_{dt} is an outcome for district d in time t; $\text{Road} \times \text{Post}_{vt}$ is an indicator equal to one if a Golden Quadrilateral road project passes through district d and the year is 2005; δ_t is a time fixed effect; γ_d is a district fixed effect; and ε_{dt} is an idiosyncratic error term. The coefficient of interest in this model is β. I will estimate this model for the sake of consistency with the existing literature.

We might worry, however, that because road placement is not random, that β recovers a biased estimate of the causal effect of road construction on firm outcomes.

In order to assuage these concerns, Ghani, Goswami, and Kerr (2015) and Khanna (2014) use an instrumental variables approach, similar to that used by Banerjee, Duflo, and Qian (2012). These authors posit the following: the Golden Quadrilateral roads were designed to connect several "nodal cities" in India. These cities, Delhi, Mumbai, Kolkata, and Chennai[22], are several of the largest and most economically important in India, and their locations predate the planning of the Golden Quadrilateral highway by hundreds of years. A straight line between these nodal cities, which approximates the least-cost path between them, should be exogenous to district economic outcomes, and is plausibly a valid instrument. The two conditions for using being on the straight line between two nodal cities as an instrumental variable are 1) that being on this line has predictive power over being on the Golden Quadrilateral road network; and 2) that being on the straight line affects district-level economic outcomes only through the road network. These assumptions seem likely to be satisfied in this context. Following the previous literature, I estimate:

$$Y_d = \beta \widehat{\text{Road}}_d + \varepsilon_d$$

and

$$\text{Road}_d = \pi \text{Line}_d + \eta_d$$

using the post-period (2005) data only. I estimate the above equations using two-stage least squares, where Y_d is an outcome in district d, Road_d is equal to one if a Golden Quadrilateral road passes through the district, and ε_d is an idiosyncratic error term. In the first stage equation, Line_d is equal to one if district d is on the straight line between two nodal cities, and η_d is an idiosyncratic error term. I also run an instrumental variables specification following Khanna (2014), where rather than using a binary indicator for being on a GQ road, I use the as-the-crow-flies distance from the Golden Quadrilateral road as my right-hand-side variable.[23] In this specification, I instrument for distance

[22]Khanna (2014) uses only Delhi, Mumbai, Kolkata, and Chennai as nodal cities, while Ghani, Goswami, and Kerr (2015) show results with and without Bangalore. My main specification will include Bangalore, following their visual argument, but I will show robustness checks without it.

[23]I will define distance to the road/line as perpendicular distance from a village's *centroid* to the road/line.

to road with distance to the straight line between nodal districts, so the estimation becomes:

$$Y_d = \beta \widehat{\text{Distance to}}\ \text{Road}_d + \varepsilon_d$$

and

$$\text{Distance to Road}_d = \pi \text{Distance to Line}_d + \eta_d$$

This is my preferred instrumental variables specification, I am interested in the overall effects of falling trade costs; it is easy to conceive of trade costs falling for districts that are adjacent to the highway projects. Note that here and everywhere that follows, I exclude the nodal districts themselves from the analysis for fear of endogeneity.

Appendix References

Banerjee, Abhijit, Esther Duflo, and Nancy Qian (2012). *On the Road: Access to Transportation Infrastructure and Economic Growth in China*. Working Paper 17897. National Bureau of Economic Research.

Datta, Saugato (2012). "The impact of improved highways on Indian firms". In: *Journal of Development Economics* 99.1, pp. 46–57. ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2011.08.005.

Ghani, Ejaz, Arti Grover Goswami, and William R. Kerr (2015). "Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing". In: *The Economic Journal*, n/a–n/a. ISSN: 1468-0297. DOI: 10.1111/ecoj.12207.

Khanna, Gaurav (2014). *The Road Oft Taken: The Route to Spatial Development*. SSRN Scholarly Paper ID 2426835. Rochester, NY: Social Science Research Network.