Labor Regulations and the Cost of Corruption: Evidence from the Indian Firm Size Distribution*

Amrit Amirapu† Michael Gechter‡

September 21, 2018

Abstract

In this paper we estimate the costs associated with an important suite of labor regulations in India by taking advantage of the fact that these regulations only apply to firms above a size threshold. Using distortions in the firm size distribution together with a structural model of firm size choice, we estimate that the regulations increase firms' unit labor costs by 35%. This estimate is robust to potential misreporting on the part of firms and enumerators. We also document a robust *positive* association between regulatory costs and exposure to corruption, which may explain why regulations appear to be so costly in developing countries.

*First version: November 22, 2014. We wish to thank Dilip Mookherjee, Arvind Subramanian, Kevin Lang, Hiroaki Kaido, Kehinde Ajayi, Sam Bazzi and Marc Rysman for input, guidance and encouragement. We are grateful to the editor, Rohini Pande, and three anonymous referees for comments that substantially improved the paper. We also thank Sam Asher, Laveesh Bhandari, Aditya Bhattacharjea, Sharon Buteau, Areendam Chanda, Urmila Chatterjee, Nancy Chau, Bivas Chaudhuri, Bibek Debroy, Andy Foster, Francois Gourio, Jayinthi Harinath, Rana Hasan, Ravi Kanbur, Saibal Kar, Dan Keniston, Anirban Mitra, Devashish Mitra, Sandip Mitra, Paul Novosad, G. Sajeevan, Johannes Schmieder, Pronab Sen, Anushree Sinha, Sandip Sukhtankar, Shyam Sundar, Anil Suraj, John Van Reenen and numerous seminar and conference participants for many useful comments. Representatives in the state governments of BR, KL, MH, TN and WB provided invaluable assistance in understanding the data collection process for the Economic Census. We gratefully acknowledge financial support from the Weiss Family Program Fund for Research in Development Economics. Liu Li, Stefan Winata and Haiqing Zhao provided expert research assistance, supported by the BU MA-RA mentor program.

†School of Economics, University of Kent. Email: amrit.amirapu@gmail.com.

‡Department of Economics, The Pennsylvania State University. Email: mdg5396@psu.edu.

1 Introduction

Restrictive labor regulations have been blamed for some of the most significant problems faced by developing countries, including low labor force participation rates and low levels of employment in the formal sector (Besley and Burgess (2004); Botero, Djankov, La Porta, Lopez-De-Silanes, and Shleifer (2004)). It has even been suggested that regulations may distort the allocation of labor across firms, thus contributing to the substantially lower levels of aggregate productivity seen in developing countries (Hsieh and Klenow (2009)). What is not clear is why labor regulations should be so much costlier in a developing country setting, particularly since enforcement agencies there are typically characterized by severe resource constraints, low compliance and widespread corruption (Svensson (2005), Chatterjee and Kanbur (2013); Kanbur and Ronconi (2015)). Moreover, previous work on the subject in developing countries has focused on a small subset of labor regulations: namely, laws related to employment protection (e.g. firing restrictions) and minimum wages.[1] In actuality, labor regulations are multifaceted, encompassing many different types of employment-related laws, such as workplace safety requirements and the provision of mandated benefits (including health insurance, social security legislation, payment of gratuities, etc.).

In this paper, we address both of these gaps and make several further contributions to the growing literature on labor regulations in developing countries. In particular, we estimate the costs associated with a suite of labor regulations in India whose components include workplace safety regulations, social security taxes and business registration requirements.[2] What the regulations have in common is that they only apply to firms with 10 or more workers, a feature we exploit to identify the magnitude of the costs they impose on firms.

[1] See Djankov and Ramalho (2009); Freeman (2010); Nataraj, Perez-Arce, Kumar, and Srinivasan (2014) for excellent reviews of the literature which reveal this focus, and Dougherty, Frisancho, and Krishna (2014) for a notable exception.

[2] Business registration requirements are generally considered separately from labor regulations. However, in our context labor regulations intended to apply to all firms are much more likely to be enforced once enforcement agencies have records of a firm's existence obtained through registration. This view is consistent with recent research experimentally defraying the costs of registration (de Mel, Mckenzie, and Woodruff (2013); de Andrade, Bruhn, and McKenzie (2014)), which finds that informal firms behave as if registration imposes costs on them over and above the costs of registration alone.

Because our methodology takes advantage of this objective feature of the laws, we do not need to rely for identification on inherently subjective assessments of differences in the text of the laws across regions - a criticism which has dogged some of the best known work in the literature (see Besley and Burgess (2004), Bhattacharjea (2009) and Fagernas (2010)).

Instead, our methodology translates observed firm behavior in response to the 10-worker threshold into estimates of the increase in unit labor costs associated with these regulations. Because our estimates are derived from firm behavior in response to actual enforcement rather than from the text of the laws, we refer to our estimated labor cost increase as representing "*de facto* regulatory costs" in what follows. We find that these regulations effectively increase firms' unit labor costs by 35%, substantially distorting economic decisions relative to a counterfactual regime without these regulations. We also apply our method to India's most stringent, controversial piece of employment protection legislation, Chapter VB of the Industrial Disputes Act (IDA), which stipulates that any industrial establishment with more than 100 workers (in most states) must obtain prior permission from the state government before laying off workers or closing the establishment. In contrast to the substantial costs we uncover at the 10-worker threshold, we find only a small and statistically insignificant impact on unit labor costs from operating at or above the 100-worker threshold.

The next contribution of the paper is to provide suggestive evidence that the distortionary effect of regulations is associated with the quality of governance through the extent of corruption present in regulatory enforcement. We show that *de facto* costs are lower in states that reformed rules to constrain the power of inspectors and higher in states with greater levels of corruption. If extortionary corruption is a significant determinant of regulatory costs (as our results suggest), this may explain why regulations appear to be more costly in developing countries than in developed countries: it is not the regulations themselves that are particularly problematic, but the way in which they are enforced.

We develop our argument as follows. We begin by exhibiting the Indian establishment size distribution using data from the Economic Census of India (EC). The EC aims to be a complete enumeration of all non-farm establishments[3] in India and, unlike all other Indian establishment-level datasets, it is not censored by size or restricted to include only the formal or informal sector. It is thus the only Indian dataset that permits estimation of the complete establishment size distribution - across all sizes and types of establishments. We find that a power law distribution fits the data well, except for a discontinuous and proportional decrease in the density of establishments with 10 or more workers (see Figures 1 and 2). We take this distortion of the establishment size distribution as qualitative evidence that the regulations which become binding there do affect firms' hiring decisions.

To understand and quantify the effect of the regulations on firm cost structure, we develop a simple model in which managers are endowed with heterogenous productivities and must choose their optimal employment levels. Firms that report hiring more than a threshold number of workers face higher unit labor costs due to the presence of regulations, and are thus smaller than they would be otherwise. Garicano, Lelarge, and Van Reenen (2016) (henceforth GLV) show that the magnitude of the increase in costs can be identified from characteristics of the distribution including, most importantly, the size of the downshift in the density above the threshold. Our model augments GLV to allow for the possibility of strategic misreporting. That is, managers may choose to deliberately misreport their employment levels at some cost, with the goal of avoiding some or all of the additional labor costs that apply to firms above the threshold size.[4] Fitting the model's predicted size distribution to the one observed from the EC data, we generate an estimate of the additional labor costs that apply to firms above the 10 worker threshold that is robust to the possibility of strategic misreporting.

Although large on average, we show that there is substantial heterogeneity in the mag-

[3] The EC refers to these as "entrepreneurial units" and defines them as any unit "engaged in the production or distribution of goods or services other than for the sole purpose of own consumption." As is common in the literature, we occasionally refer to them as "firms" even though the unit of observation in the data is actually a factory or an establishment, rather than a firm (i.e. multiple establishments may belong to the same firm). We do this primarily for expositional purposes, but also based on the observation that only a minute proportion of establishments belong to multi-establishment firms.

[4] The importance of allowing for strategic misreporting is explained in greater detail in Section 4.1. Note that "strategic misreporting" is distinct from the issue of corruption in the enforcement of labor regulations.

nitude of *de facto* regulatory costs along several dimensions including state, industry and ownership type. In the first place, we find that privately-owned establishments face the highest *de facto* regulatory costs while government-owned establishments show no significant cost increase when employing 10 or more workers. This supports our interpretation that the downshift in the distribution starting at 10 workers is indeed due to the regulations, since many regulations either do not apply to government-owned establishments or are less binding for such establishments in practice. We also document higher effective costs for businesses run by members of disadvantaged social groups (Scheduled Castes, Scheduled Tribes and women), suggesting that regulatory enforcement is unequal and may be linked to the bargaining power of firm owners. Using variation across states and industries, we find a strong and robust positive correlation between our estimated regulatory costs and several different measures of corruption. The link between high regulatory costs and corruption may appear surprising if one thinks of corruption as collusive, "greasing the wheels" in a highly-regulated economy by allowing firms to reduce their effective regulatory burden by bribing inspectors (e.g., Huntington (1968)). Our results are instead consistent with the concern that corrupt inspectors may overreport violations relative to honest inspectors in order to extract greater bribes - which we deem extortionary corruption.[5]

Our finding that the most contentious component of India's employment protection legislation, Chapter VB of the IDA, does *not* have a substantial effect on unit labor costs differs from much of the earlier academic work on the subject and belies the attention the IDA has received from academics (e.g Besley and Burgess (2004); Hasan, Mitra, and Ramaswamy (2007); Aghion, Burgess, Redding, and Zilibotti (2008); Adhvaryu, Chari, and Sharma (2013); Chaurey (2015)) and the business press (e.g. Bajaj (2011); Ghosh (2016)) alike. We attribute this difference, first, to the fact that our methodology for estimating the impact of the legislation is very different from the strategies employed in previous papers.

[5]See Banerjee (1994); Mookherjee (1997); Hindriks, Keen, and Muthoo (1999); Polinsky and Shavell (2001); Mishra and Mookherjee (2013) for theoretical treatments. Empirically, Sequeira and Djankov (2014) and Asher and Novosad (2016) also provide evidence for the importance of extortionary corruption.

Our identification is based on the presence and degree of distortions in the distribution of establishments larger than the size at which Chapter VB's firing restrictions become binding. Most previous work identifies the effect of India's employment protection legislation based on differences in the growth of mean outcomes across states which have been coded as initiating pro-worker or pro-employer reforms to the full IDA. The coding of states into these three groups (pro-employer, pro-worker, or neutral) has been the subject of controversy in the subsequent literature (see Fagernas (2010); Bhattacharjea (2006, 2009)), and an advantage of our identification strategy is that we can side-step this controversy. Hsieh and Olken (2014), the only other paper in the literature to focus on establishment size distributions, report finding no visually striking change in the size distribution of Indian establishments at the 100-worker threshold, in accordance with our quantitative results. Another source of difference is our focus on Chapter VB, rather than the full IDA. This is partially a question of the feasibility of applying our approach (Chapter VB is size-based), but we also see focusing on Chapter VB as providing a proof of concept: if Chapter VB's complete restriction on firing is not very distortionary, it would be surprising if the law's more mild provisions are.

In addition to our contributions to the empirical literature on labor regulations, our extension of the GLV model to allow firms to strategically misreport their sizes should find applications in many other settings. Robustness to strategic misreporting in response to a size threshold is particularly crucial when working with survey datasets from developing countries, because such data are typically self-reported. By contrast, in the high-income country administrative data used by studies such as GLV, it may be harder for establishment managers to misreport information when desired. We show that in the presence of strategic misreporting, a naive approach to estimating GLV's model can dramatically overestimate the increase in labor costs associated with a size-based regulation. In explicitly modeling the decision to misreport, we also show - under a range of different modeling assumptions - that while misreporting can be extensive near the threshold, the reported firm size distribution approaches the true distribution at large firm sizes, so one can minimize bias in the estimate of regulatory costs by focusing the estimation on large firm sizes and discarding the observations close to the threshold. In our case, if one fails to account for the possibility of misreporting, the estimated increase in per-worker costs rises from 35% to 101%.[6]

We identify firms' real responses using a reasonable theoretical restriction: that the cost of misreporting be strictly convex in the degree of misreporting. Under this assumption, misreporting can be extensive near the threshold, but becomes increasingly costly for large values. In fact, we show that the reported firm size distribution becomes arbitrarily close to the true distribution at large firm sizes, so one can minimize any bias in the estimate of regulatory costs by focusing the estimation on large firm sizes and discarding the observations close to the threshold. In our case, if one fails to account for the possibility of misreporting, the estimated increase in per-worker costs rises from 35% to 101%.

Before closing the introduction, it is worth noting three significant limitations of our methodology. The first is that we cannot separately identify the costs of individual regulations. Our cost estimates refer to the costs associated with *all* of the regulations that become binding at the 10 worker threshold, and are likely to also include effects of regulations at the 20 worker threshold. The second limitation is that, due to our misreporting framework, the only kinds of costs that we can capture are unit labor costs, because these are the only costs that result in a downshift of the log firm size distribution. Fixed costs affect the firm size distribution in a way that is similar to strategic misreporting: they both lead to higher reported mass just below the threshold and lower mass just above the threshold. For this reason, their effects cannot be separately identified, so that if the regulations we study have significant fixed cost components our methodology will not detect these. Finally, because our methodology involves comparing firms below the size threshold with those far above the threshold, we lose the ability to compare firms of similar sizes (9 versus 11), which are more likely to face similar production and demand conditions. For this reason, and the fact that

[6] The fact that misreporting is sufficiently large in magnitude to produce an almost threefold distortion in estimated costs speaks to low state capacity and serves as a cautionary tale for users of government statistics in such environments. We thank an anonymous referee for pointing this out.

firm size is a choice variable, our methodology is quite unlike regression discontinuity studies, and must instead rely on assumptions regarding the distribution of firms and economic theory regarding firm behavior.

The rest of the paper is organized as follows. In the next section (Section 2), we provide an overview of the relevant institutional details regarding Indian labor and industrial regulations. Section 3 introduces the data and provides qualitative evidence on the importance of size-based regulations using the size distribution of establishments in India. In Section 4 we describe the theoretical model and empirical strategy. Section 5 provides the main results. In Section 6, we interpret the findings and investigate the connection between our estimated costs and corruption. Section 7 concludes.

2 Labor Regulations in India

Many labor regulations in India only apply to establishments that are larger than a certain threshold, where size is most often measured in terms of the number of workers in the establishment. There are several thresholds at which different labor regulations start to apply, but the two most prominent thresholds occur once an establishment employs at least 10 and at least 100 workers.[7] In most states in India, establishments that employ more than 100 permanent workers (excluding contract workers) must abide by India's most controversial piece of employment protection legislation: Chapter VB of the IDA.[8] Under this regulation, establishments over the threshold must be granted government permission before closing the establishment or laying off workers. It is the IDA - of which Chapter VB is a part - that has been the subject of most academic papers on labor regulations in India.

In contrast, the 10 worker threshold has received far less attention from academics, even

[7] There are other thresholds, such as at 20 workers (at which point establishments must contribute to the "Employees' Provident Fund Organisation," which operates a pension scheme for formal sector workers) and at 50 workers (at which point severance payment obligations increase under Chapter VA of the Industrial Disputes Act), but we do not separately analyze these thresholds because they are less contentious and do not appear to substantially distort the establishment size distribution.

[8] In 2005, the year to which our analysis applies, this threshold was 100 workers for all states except West Bengal, where the threshold was 50 workers.

though it is extremely important due to the large number of varied regulations that start to become binding at that threshold as well as the fact that this threshold is most commonly associated with the formal/informal divide. The major regulations that start to apply once an establishment employs 10 or more workers include the following: establishments must register with the government, meet various workplace safety requirements (under the Factories Act for manufacturing establishments that use power and The Building and Other Construction Workers' Act for construction-related establishments, for example), pay insurance/social security taxes (under the Employees' State Insurance Act), distribute gratuities (under the Payment of Gratuity Act) and they must bear a greater administrative burden (under, for example, the Labor Laws Act).

In Appendix A we provide a table which includes a comprehensive list of all central (i.e. federal) size-based labor regulations in India. For each law, we briefly describe the regulation as well as the nature of the size-based threshold. The table documents variation in the regulatory burden across industries and ownership type. Many regulations cover specific industries (especially manufacturing), although many others are explicitly universal in scope. We note that government establishments are explicitly included in some laws and explicitly excluded from others. Some important size-based laws (e.g. the Payment of Gratuity Act and the Payment of Bonus Act) which may apply to government establishments on paper are not relevant in practice, because gratuities and bonuses for government workers' in establishments of all sizes are set by pay commissions, and are far in excess of those required in these laws.

Other regulations are indirectly - thought not explicitly - size-based, because they reference laws with size-based aspects. For example, the Maternity Benefits Act only applies to establishments designated as "factories" under the Factories Act, which means it only applies to establishments with more than 10 workers. Furthermore, there appears to be a salience effect associated with the 10 worker threshold: for example, in interviews with small business owners in Chennai, we discovered that several of them appeared to believe that certain regulations (such as the Provident Fund Act) apply once you have 10 workers, when in fact they did not.

In addition to - or in lieu of - the explicit costs associated with complying with the regulations, establishments with 10 or more workers may be subject to implicit costs associated with increased interaction with labor inspectors, who often have the power to extract bribes and tighten (or ease) the administrative burden firms face. Indeed, inspectors in India have a large amount of discretion regarding the enforcement of administrative law.[9]

It has been argued that the ability to extract bribes is exacerbated by the antiquated and/or arbitrary nature of certain components of the laws (Debroy (2013)). TeamLease Services (2006) provides some telling examples: "[r]ules under the Factories Act, framed in 1948, provide for white washing of factories. Distemper won't do. Earthen pots filled with water are required. Water coolers won't suffice. Red-painted buckets filled with sand are required. Fire extinguishers won't do... And so on." The result of such rules is that almost all firms can be found guilty of some violation or another under the letter of the law - even if they are in compliance with the spirit of the law. Firm owners who choose not to comply with such regulations face costs (i.e. fines and possible prison sentences) if discovered and convicted.

This kind of behavior has been referred to as "harassment bribery" (Basu (2011)). Anecdotal evidence of inspectors using the complexity, arbitrariness and sheer amount of paperwork as a way to extract bribes is easy to come by. For example, we have included a selection of citizen reports from "ipaidabribe.com" in Appendix H, which demonstrate just this kind of behavior.[10] Interestingly, some of the reports suggest that the size of the bribe paid is a direct linear function of the number of employees - which will be relevant for interpretation of our results in Section 6.

[9] For example, in some cases, the definition of what constitutes a "day" is at the discretion of the inspector, and it is a commonly held view that "[w]hile grave violations are ignored, minor errors become a scope for harassment" (TeamLease Services (2006)).

[10] We thank Andrew Foster for this suggestion.

3 Data and the Size Distribution in India

3.1 Data

We will use the Economic Census of India (EC) as our main data source to investigate the costs associated with the regulations described in the previous section. The EC is meant to be a complete enumeration of all (formal and informal) non-farm business establishments in India at a given time. As such, it contains a very large number of units: the 2005 wave, which we will principally use, has almost 42 million observations. It is the only Indian dataset that represents the unconditional distribution of establishment size, which is essential for our analysis. Other datasets, such as the CMIE's Prowess Database, the Annual Survey of Industries (ASI) or the National Sample Survey's (NSS) Unorganized Manufacturing Surveys cover only certain parts of the distribution and are thus unsuitable for our analysis. The ASI, for example, only covers establishments in the manufacturing sector that have registered with the government under the Factories Act. However, registration under this Act is only required for establishments with 10 or more workers if the unit uses power (20 or more workers if the establishment uses no power). Therefore, the selection into the ASI varies discontinuously at precisely one of our points of interest.

The price to pay for uniform coverage and large sample size is that the EC does not contain very detailed information on each observation. For each establishment in the data, there is only information on a handful of variables including the total number of workers usually working, the number of non-hired workers (such as family members working alongside the owner), the registration status, the 4-digit NIC industry code, the type of ownership (private, government, etc) and the source of funds for the establishment. There is no information on capital, output or profits, and the data are cross-sectional.

We supplement our analysis with data from a variety of other sources. We get data on state and industry level corruption from a) Transparency International's "India Corruption Study 2005", b) the RBI, and c) the World Bank Enterprise Survey for India (2005). Data on state-level regulatory enforcement come from the Indian Labour Year Book.[11] Other measures of state-level regulations come from Aghion et al. (2008) and Dougherty (2009).

3.2 The Size Distribution of Establishments in India

Figure 1 shows the distribution of establishments by the number of total workers (hired and non-hired - typically family - workers) in 2005 on a log scale. Four things are striking about figure 1. First, the distribution is extraordinarily right-skewed. Indeed, about half of all establishments are single person establishments. Second, the natural log of the density is a linear function of the natural log of the number of total workers. This implies that the unlogged distribution follows a power law in the number of total workers. This pattern will be important for the analysis that follows but it is not very surprising in and of itself: power law distributions in firm sizes have been documented in many countries (Axtell (2001) and Hernández-Pérez, Angulo-Brown, and Tun (2006)). The third and more unique feature of the distribution is that there appears to be a level shift downward in the log frequency for establishment sizes greater than or equal to 10. To the best of our knowledge, ours is the first paper to document this phenomenon in India. Finally, we do not see any discernible change in the distribution at 100 workers, the relevant threshold for employment protection legislation. We will confirm this fact in our formal analysis.

Also apparent from the figures above is that there is a significant amount of non-classical measurement error due to rounding of establishment sizes to multiples of 5. The existence of rounding is not surprising given that the data are self-reported and that respondents are asked to give the "number of persons usually working [over the last year]". Our estimation procedure, described in the next section, accommodates this measurement error pattern.

[11] We would like to thank Anushree Sinha and Avantika Prabhakar for their considerable and generous help in obtaining these data.

4 Model and Empirical Strategy

4.1 Modeling Size Based Regulations with Strategic Misreporting

To interpret the downward shift from Figure 1 in economic terms, we develop a model which is based on the framework from GLV, but augmented to allow for the possibility that managers of plants may strategically misreport their size to government officials - including labor inspectors and/or EC enumerators. For example, if plant managers are aware of the increased regulatory burden that is associated with employing 10 or more workers, and if they believe the EC enumerators will relay information to government regulatory bodies,[12] they may wish to hide the fact that their actual employment exceeds the threshold or more generally under-report their actual employment.

In the GLV framework size-based regulations increase the unit labor costs of firms that exceed the size threshold, which results in a parallel downward shift in part of the theoretical logged firm size distribution. From the magnitude of the downshift observed in the empirical distribution, one can back out the additional labor costs imposed by the regulations. If firms are allowed to misreport their size, however, the reported firm size distribution may differ from the true distribution. In what follows we will show how a naive estimation procedure - which does not take misreporting into account - may result in biased estimates of the labor costs. We also present our solution, which minimizes the bias from misreporting.

The primitive object in our framework - following GLV as well as Lucas (1978), on which both our model and that of GLV is based - is the distribution of managerial ability ($\alpha \sim \phi : [\underline{\alpha}, \alpha_{max}] \rightarrow \mathbb{R}$). Firms whose managers have higher ability (α) are more productive and can profitably employ more workers. Homogenous workers are allocated to firms through a competitive labor market with a single, market clearing wage (w). As is common in the literature, we assume that the distribution of managerial ability follows a power law ($\phi(\alpha) =$

[12] In point of fact firms' answers to Economic Census enumerators have no impact on their regulatory burden, but it is quite possible that firms believe otherwise, and that is what is relevant. If firms believe that that reporting has no effect on their regulatory burden, then there should be no incentive to misreport, and no need to adjust our estimation so that it is robust to misreporting.

$c_\alpha \alpha^{-\beta_\alpha}$), which will then generate a power law in the theoretical firm size distribution. Our model differs from the basic GLV framework by allowing firms to choose not only their true employment (n), but also their *reported* employment (l).[13] Both are relevant when calculating expected costs due to the size-based regulations. In particular, a firm with productivity α faces the following profit-maximization problem:

$$\pi(\alpha) = \max_{n,l} \alpha f(n) - wn - \tau wl * \mathbb{1}\{l > N\} - M(n,l) \tag{1}$$

where n is the number of workers a firm actually employs, l is the number of workers the firm reports to government officials (inspectors and enumerators alike), $f(n)$ is a production function (with $f'(n) > 0$ and $f''(n) < 0$), w is a constant wage paid to all workers, $\tau \geq 0$ is a proportional tax on labor that firms pay on their *reported* employment *if* their reported employment exceeds the regulatory threshold, and $M(n,l)$ is an expression which captures the expected costs of misreporting.[14]

The term capturing regulatory costs ($\tau wl * \mathbb{1}\{l > N\}$) creates an incentive for firms to misreport their employment in a downward direction (i.e. to set $l < n$). Counteracting this incentive is that misreporting firms may be caught by the authorities and made to pay a fine. We think of the expected misreporting costs as being the product of three distinct terms: $M(n,l) = q(n) * p(n,l) * F(n,l)$. Firms are inspected with probability $q(n)$; conditional on being inspected, they are caught misreporting with probability $p(n,l)$; and if caught they are made subject to a fine, $F(n,l)$. As written above, the probability of being inspected, the probability of being caught and the magnitude of the fine may in general depend on n or l in an arbitrary way. Going forward we will make the following assumptions regarding the general structure of $M(n,l)$, which enable us to identify τ in the presence

[13]A summary of the basic GLV framework (i.e. without misreporting) is provided in Appendix B.1. For a detailed derivation of the model, see Garicano et al. (2016).

[14]It may be worth pointing out that labor demand will be lower in a regime with these regulations than without them. Therefore, the regulations will have a general equilibrium effect on employment and output through the wage, w. However, this will not affect our estimation of τ, our object of interest, because τ measures the increase in unit labor costs for larger firms as a proportion of the wage, which is common to all firms.

of misreporting under minimal additional parametric assumptions. As we discuss further below, these assumptions are sufficient but not necessary to identify τ.

Assumption 1. *Let $u \equiv n - l \geq 0$ denote the degree of misreporting and $M(u)$ denote the expected costs of misreporting. We assume that $M(0) = 0$ and that $M(u)$ is a continuous, increasing and strictly convex function of u alone. Conditional on u, $M(u)$ is thus independent of firm size, n.*

Under this assumption, we show that for large enough values of firm size, x, the difference between the log of the reported density, $\psi(x)$, and the log of the true density, $\chi(x)$, becomes vanishingly small.

Proposition 1. *Suppose a firm's profit maximization problem takes the form of Equation 1 and Assumption 1 holds. Then*

$$\lim_{x \to \infty} log\chi(x) - log\psi(x) = 0.$$

Proof. See Appendix B.3. □

Proposition 1 implies that an estimation based on large enough firm sizes will be minimally biased because the reported distribution becomes arbitrarily close to the true distribution at large sizes. We will demonstrate how this is accomplished below with a specific example, but some discussion of the assumptions is necessary at this point.

The assumption that misreporting costs should be strictly convex in the degree of misreporting is both standard in the literature (e.g. Almunia and Lopez-Rodriguez (2015); Kumler, Verhoogen, and Frias (2015)) and intuitive given our understanding of the context in which Indian businesses make such decisions.[15] One important implication of Assumption

[15] The intuition is that hiding larger and larger numbers of employees from enumerators or inspectors should get increasingly difficult until at some point it is impossible. This intuition is supported by our understanding of the context in which firms make such decisions, which has been informed by our interviews with small firms. Among such enterprises it is common to hear accounts of business owners ushering employees out the back door of the establishment whenever labor inspectors arrive, but this type of behavior is clearly only possible for relatively small numbers of employees. We would like to thank Sharon Buteau and Balasekhar Sudalaimani from IFMR for helping to set up these interviews.

1 - that the extent of misreporting should be relatively lower for larger firms - finds empirical support in recent literature (e.g. Goyette and Kouame (2016)). The other substantive assumption imposed above (i.e. that $M(u)$ is independent of firm size, n) is restrictive, but in fact neither it nor the convexity assumption is necessary; they are both primarily useful in illustrating how identification proceeds without making parametric assumptions on the exact functional form of misreporting.

In Appendix B.4, we consider a range of alternative specifications for the functional form of misreporting, including several which depart from both of the assumptions above. What this exploration reveals is that most reasonable specifications either yield the same conclusions as Proposition 1 or else they are incompatible with the observed data. The only specifications that are both consistent with the data and for which it is impossible to correctly identify τ are those which cause all firms to misreport a constant fraction of their true employment.

We now proceed by informally characterizing the solution to the firm's problem (from equation 1) under Assumption 1, for firms at every level of productivity. The lowest productivity firms (those with α below some threshold, α_1) will be effectively unconstrained, in the sense that they choose to hire at most N workers ($n \leq N$) and thus do not fall under the purview of the size-based labor regulations. For this reason, there is no incentive for them to misreport, so they report truthfully ($l = n$). There will be a second set of firms with higher productivity ($\alpha \in [\alpha_1, \alpha_2]$) that will find it optimal to exceed the regulatory threshold in practice (choosing $n > N$), but will misreport their employment to avoid the higher regulatory costs (setting $l = N$). These firms will only *appear* to be bunched up at N, but will in fact have higher employment.

The last category of firms are those with $\alpha > \alpha_2$, which are productive enough to warrant hiring work forces so large that they cannot completely avoid the regulation without being detected/fined with sufficiently high probability/severity and thus report $l > N$. Even these firms, however, with both $n > N$ and $l > N$ do not find it profit-maximizing to report truthfully. They can save on their unit labor costs by shading their reported employment, and will choose $l = n - M'^{-1}(\tau w)$. Note that the degree of misreporting is by a constant amount, which is a direct implication of Assumption 1, as spelled out in Appendix B.3. Importantly, this last set of firms face higher unit labor costs than they would in the absence of the regulations and therefore employ fewer workers by a constant proportion - resulting in a "downshift" in the logged firm size distribution. The fact that the degree of misreporting is by a constant amount implies that the difference between the true and reported distributions goes to zero with size, and thus the downshift in the reported distribution will match that of the true distribution at large sizes.

To derive a closed form solution for the true and reported firm size distributions, it is necessary to make some functional form assumptions. Doing so will clarify how we estimate τ and explore the implications of Proposition 1 for our estimation procedure. The first parametric assumption we make, following GLV, is that firm output is a power function of labor: $f(n) = n^\theta$. The second is to impose a specific functional form for misreporting that satisfies Assumption 1: $M(n,l) = \frac{F}{n_{max}}(n-l)^2$.[16] One way to generate this function is to suppose that the probability of inspection is proportional to firm size (e.g. $q(n) = \frac{n}{n_{max}}$, where the denominator is chosen to guarantee that $q(n) \leq 1$ on the observed support), that the probability of being caught - conditional on being inspected - is proportional to the *fraction* of employees that are misreported ($p(n,l) = \frac{n-l}{n}$) and that the fine if caught is proportional to the level of misreporting (i.e. $F(n,l) = F * (n-l)$). With these two substitutions, the firm's profit maximization problem from Equation 1 becomes:

$$\pi(\alpha) = \max_{n,l} \alpha n^\theta - wn - \tau wl * \mathbb{1}\{l > N\} - \frac{F}{n_{max}}(n-l)^2$$

As in our informal characterization above, the optimal choices of n and l will depend on the productivity of the firm. A full mapping between productivity α and the true firm size

[16] Again, we point out that this assumption is not necessary for identifying τ. The assumption that production is a power function, however, does have implications for identification, in the sense that our estimate of τ will depend on our estimate of θ.

n, as well as between α and reported firm size l, is given by the following equations:

$$n^*(\alpha) = \begin{cases} (\frac{\theta}{w})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} \leq N & if \quad \alpha \in [\underline{\alpha}, \alpha_1] \\ n_2^*(\alpha) & if \quad \alpha \in (\alpha_1, \alpha_2] \\ (\frac{\theta}{w})^{\frac{1}{1-\theta}}(1+\tau)^{-\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} > N & if \quad \alpha > \alpha_2 \end{cases}$$

$$l^*(\alpha) = \begin{cases} (\frac{\theta}{w})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} \leq N & if \quad \alpha \in [\underline{\alpha}, \alpha_1] \\ N & if \quad \alpha \in (\alpha_1, \alpha_2] \\ (\frac{\theta}{w})^{\frac{1}{1-\theta}}(1+\tau)^{-\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} - \frac{n_{max}}{2F}w\tau > N & if \quad \alpha > \alpha_2. \end{cases}$$

Because there is a strictly monotonic relationship between α and n as well as α and l (except for the bunching), one can obtain expressions for the distributions of true and reported firm size, $\chi(n)$ and $\psi(l)$, as transformations of the distribution of managerial ability, $\phi(\alpha)$. Simplifying terms, one can write the log of the density of firms with true employment n as

$$log\chi(n) = \begin{cases} logA - \beta log(n) & if \quad n \in [n_{\min}, N) \\ log[\xi(n)] & if \quad n \in [N, n_m(\alpha_2)] \\ - & if \quad n \in (n_m(\alpha_2), n_t(\alpha_2)) \\ logA - \frac{\beta-1}{1-\theta}log(1+\tau) - \beta log(n) & if \quad n \geq n_t(\alpha_2) \end{cases} \tag{2}$$

and the log of the density of firms with reported employment l as

$$log\psi(l) = \begin{cases} logA - \beta log(l) & if \quad l \in [l_{\min}, N) \\ log(\delta_l) & if \quad l = N \\ - & if \quad l \in (N, l_t(\alpha_2)) \\ logA - \frac{\beta-1}{1-\theta}log(1+\tau) - \beta log(l + \frac{n_{max}}{2F}w\tau) & if \quad l \geq l_t(\alpha_2) \end{cases}$$

where A is a function of constants and terms have been simplified and collected. Appendix B.2 provides a derivation of this result, along with all missing steps.

Comparing the expressions for the reported and true size distributions above, there are several points worth noting. 1) For the range $l < N$, the true distribution coincides with the reported/observed distribution. 2) There appears to be bunching at N in the reported distribution, but some of these firms in fact have greater than N workers. 3) Compared to the distribution for $n < N$, both the true distribution *and the reported distribution* for $n \gg N$ are downshifted, and by exactly the same function of τ as in GLV's model without misreporting (the intercepts for both distributions are $logA - \frac{\beta-1}{1-\theta} log(1+\tau)$ for larger firms versus $logA$ for smaller firms).[17] 4) As stated in Proposition 1, the difference between the log of the reported distribution and the log of the true distribution converges to 0 for large firms. The intuition is straightforward: the only difference in the two expressions is the constant amount $\frac{n_{max}}{2F} w\tau$, the contribution of which becomes negligible at large sizes.

Together, these observations allow us to back out an estimate of τ - the extra unit labor costs faced by firms above the size threshold. In particular, 1) and 4) tell us that if we focus on firms below the size threshold and those well above it, the reported density will be arbitrarily close to the true density. Observation 3) - that a function of the tax enters additively in the log density for all firms above the threshold - tells us that τ can be determined from the size of the downshift observed in the log firm size distribution between large and small firm sizes. Given these observations, our identification strategy is quite simple. For very small and very large firm sizes (i.e. $n < N$ or $n \gg N$), one can express the log of the density according to the following equation:

$$\log(\chi(n)) = \log\left[\left(\frac{1-\theta}{\theta}\right)^{1-\beta}(\beta-1)\right] - \beta\log(n) + \log((1+\tau)^{-\frac{\beta-1}{1-\theta}})\mathbb{1}\{n > N\}, \quad (3)$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. To see how τ is identified from $\chi(n)$, rewrite Equation

[17] Appendix B.1 includes the firm size distribution from GLV for comparison.

3 as

$$\log(\chi(n)) = \alpha - \beta \log(n) + \delta \mathbb{1}\{n > N\}. \tag{4}$$

α, β and δ can be identified by applying Equation 4 to the observed size distribution. θ is a function of α and β and is thus also identified. τ is given by

$$\tau = \exp(\delta)^{-\frac{1-\theta}{\beta-1}} - 1,$$

which is identified as long as θ and β are identified.

In principle, by choosing some threshold n_L satisfying $n_L \gg N$, one should be able to produce a value for τ by using ordinary least squares to estimate the specification

$$\log(\chi(n)) = \alpha - \beta \log(n) + \delta \mathbb{1}\{n > N\} + \epsilon(n), \tag{5}$$

where $\epsilon(n)$ represents any deviation of the observed firm size distribution from the model coming from an idiosyncratic tendency for firms to cluster to or away from a particular size.

In practice, however, Proposition 1 is problematic for estimating the parameters of the model from raw data using Equation 5, because Equation 5 must be estimated using data from relatively small establishments with sizes outside the range $[N, n_L]$. This is due to the fact that the empirical probability of observing an establishment of a given size is truncated at $\frac{1}{\#\ of\ observations}$ (this is visually apparent in figure 1). Truncation makes the relationship between the log of the empirical probability and the log of the total number of workers nonlinear. To preserve the linear relationship, a researcher would have to omit establishment sizes large enough that truncation is not an issue.[18] However, Proposition 1 tells us that the misreported distribution is close to the true distribution only at large sizes and that the misreported distribution may be biased downward at establishment sizes close to a

[18]The suggestion to focus on relatively smaller establishments appears in Appendix B of Garicano, Lelarge, and Van Reenen (2013).

regulatory threshold. This leads to downward bias in δ (the downshift in the log density) and consequently upward bias in τ. Instead, we will develop an empirical approach that deals with the truncation problem and allows us to focus on large firms, where the the difference between the log of the reported distribution and the log of the true distribution is close to zero.

Since our approach involves estimating parameters of the firm's problem by fitting features of the theoretical density to the observed empirical density, it is worth noting the following discrepancy between the model and the data. The log density of reported employment that is generated by the model is undefined for $l \in (N, l_t(\alpha_2))$ because the density of reported employment contains a hole in this region (see Equation 2). Reporting $l \in (N, l_t(\alpha_2))$ is dominated by choosing either $l = N$ or $l \geq l_t(\alpha_2)$. However, Figure 1 clearly shows that there are firms who report employing 11 workers.[19] Based on interviews with firms and accountants, our understanding of the discrepancy is that small firms tend to be inattentive to the regulatory threshold, while large firms tend to be attentive. Attentive firms are aware of the regulations as well as the expected costs and benefits of misreporting, while inattentive firms are simply not aware of the relevant regulations - and hence do not bother to misreport their firm size. In Appendix B.6 we present a version of our model from the previous section that combines inattention and misreporting. In particular, we show there that if the fraction of inattentive firms is large at small firm sizes and small at large firm sizes[20] the model's

[19]This type of discrepancy - in which many agents are observed to make strictly dominated choices - is common in analyses of behavior in response to "notches" or "kinks" (e.g. Kleven and Waseem (2013)).

[20]This assumption can be motivated theoretically if one imagines that managers must pay a fixed cost (which varies idiosyncratically across firms) in order to learn regulatory details - including the location of the thresholds. In practice this would involve hiring an accountant or attorney who is knowledgable about the text of labor regulations. Under the plausible assumption that the distribution of fixed costs does not vary with firm size, the fact that the benefits of adjusting employment in response to the threshold rise with size implies that all large firms will adjust while only some small firms will.

The assumption is also consistent with our firm interviews which revealed that small firms are commonly unaware of regulatory details (including the size thresholds themselves) while large firms expend considerable time and money in ensuring that they have correct information regarding the regulations (often by hiring accounting, legal and human resource departments). This difference may be due to the fact that large firms are more likely to be audited and inspected by labor regulators (Almeida and Ronconi (2016)) or that small firms tend to be managed by those with lower levels of human capital (La Porta and Shleifer (2014)) and are hence less likely to be knowledgable about the law.

predicted density will closely resemble the observed density, and τ remains identified using the method described here (in particular, by focusing primarily on firms with employment levels far above N).

Before proceeding, it is worth noting a second possible source of misreporting: Economic Census enumerators themselves. EC enumerators were required to fill out an extra form containing the address of any establishment that reported 10 or more workers. It is conceivable that enumerators might have found it preferable to under-report the number of workers for establishments with 10 or more workers in order to avoid the extra burden of filling in the "Address Slip". However, as we show in Appendix B.5, this type of misreporting - like the previous one - only generates bias in the reported distribution for establishment sizes close to N.[21] In particular, such enumerator-driven misreporting is likely to contribute to the "bunching" at 10 and the "valley" just after 10, but it cannot lead to a downshift in the firm size distribution at large firm sizes, which is how we identify τ. Moreover, it is easy to show that any estimation technique that is robust to the possibility of manager-driven misreporting will also be robust to the possibility of enumerator-driven misreporting.

4.2 An Empirical Approach Robust to Strategic Misreporting

In this subsection, we develop a way of estimating Equation 5 using establishments that are least affected by the misreporting. These include establishments that are below the bunching point as well as those that are far above the size threshold. As we remarked in the previous subsection, we cannot estimate Equation 5 directly on large establishments because of truncation in the empirical probability of observing an establishment of a given size. Furthermore, as discussed in Section 3, the empirical size distribution is characterized by substantial rounding to multiples of 5 workers, especially at larger sizes. Setting aside the truncation problem, OLS estimation of Equation 5 will produce downward bias in δ

[21]The two types of misreporting can be modeled similarly. The main difference is that the marginal benefit of misreporting in the manager-driven misreporting model is replaced with a fixed benefit of misreporting in the enumerator-driven misreporting model.

because sizes that are multiples of 5 are treated as single observations. Instead, their excess establishments should be distributed to nearby sizes.

To address both issues, we non-parametrically estimate the density associated with larger sizes using the method described in Markovitch and Krieger (2000) (MK). MK propose a nonparametric density estimator for heavy-tailed distributions that achieves L_1 consistency. L_1 consistency fails for any distribution with heavier tails than an exponential for the standard Parzen-Rosenblatt kernel density estimator,

$$\hat{f}(l) = \frac{1}{Eh} \sum_{i=1}^{E} K\left(\frac{l - L_i}{h}\right), \tag{6}$$

where L_i, for our purposes, is establishment i's total number of workers, l is a number of workers for which we would like to know the density, E is the total number of establishments in the 2005 EC, $K(\cdot)$ is a kernel function, and h a smoothing parameter or "bandwidth." L_1 consistency is known to hold for distributions with compact support, so MK suggest the simple approach of estimating the density of a transformation of L_i which has compact support, then inverting back for an estimate of the density of the original l.

Specifically, we first apply the transformation recommended by MK, $T(l) = \frac{2}{\pi}\arctan(l)$, to each establishment's number of workers. Our estimate of the density associated with a specific number of workers, l, is given by

$$\hat{\psi}(l) = \hat{f}(T(l))T'(l)$$

where $\hat{f}(T(l))$ applies Equation 6 to the transformed data, $T(L_i)$, and evaluates at the transformed number of workers of interest $T(l)$. $T'(l)$ is the derivative of the transformation evaluated at l. We use the Epanechnikov kernel function. An advantage of this approach from our perspective is that a constant bandwidth[22] applied to the transformed data expands

[22] Note that in this case, the bandwidth must be chosen. We cannot use cross-validation to choose the optimal bandwidth because it will recover the rounding pattern found in the data.

asymmetrically with respect to the original data. As we move to the right in the distribution, where data are more scarce, our kernel begins to put positive weight on observations farther away. This accords with our observation that rounding in the reported distribution becomes more severe at larger sizes. We use the empirical probability for small sizes, where the establishment size distribution is better represented as a discrete variable.

We apply a modified version of Equation 5 to the log of the estimated density $\hat{\psi}(l)$ for all observed sizes. For example, when analyzing the effect of regulations that apply to firms with 10 or more workers, we remove the effect of misreporting close to the threshold by adding dummy variables for size 8 and 9 and for sizes 10 - 20. The choice of 20 as the largest size for which we include a dummy is unimportant and we show in Table 10 in the appendix that our results are robust to alternative choices of dummy variables. Since Equation 5 treats each establishment size as one observation, and since the range of establishment sizes in the 2005 EC runs from 1 to 22,901, the model is primarily estimated using data far from the 10-worker cutoff.[23] Finally, we include dummies for having 1 or 2 workers because own account and 2-worker establishments are likely to be household enterprises and may therefore differ fundamentally in character from their larger counterparts.[24]

Figure 2 depicts the strategy. The dark grey dots show the raw data. The light grey circles represent the result of the first step: nonparametric density estimates associated with each establishment size. The line shows the fit of the model in Equation 5, augmented by

[23] We note here that our misreporting-robust strategy for estimating δ from Equation 5 bears a resemblance to the estimator for the average treatment effect in "donut" regression discontinuity designs (RDDs, e.g. Almond and Doyle (2011), Barreca, Lindo, and Waddell (2016), Eggers, Freier, Grembi, and Nannicini (2018)). Donut RDDs omit data where the assumption of quasi-random assignment of the running variable is considered dubious. The average treatment effect at the threshold is obtained by taking difference in regression functions estimated on both sides of the donut and extrapolated to the threshold. Our estimated value of δ is similarly obtained from the difference in heights of a linear fit to the log density of firm size above and below our misreporting "donut."

However, the fact that our model estimates the log of the density associated with a firm size, rather than a regression function, generates a key difference in interpretation. In donut designs, the values of extrapolated regression functions at the threshold represent the counterfactual treated and untreated expected outcomes at that point. In our case, log density prediction generated by extrapolating the log density from the left of our donut to the right does not represent the counterfactual density with no regulation at the 10-worker threshold. Because of the requirement that the observed and counterfactual densities integrate to one, we expect the density to the left of our donut to be reduced without regulations.

[24] Table 10 shows that our results are robust excluding the 2-worker dummy.

the dummy variables, to the nonparametric density estimates. Figure 2 also provides some evidence for the model described in section 4.1. The observed establishment size distribution appears to converge back to a power law with the same slope as for establishments with fewer than 10 workers, but deviates slightly from that slope at sizes just above the 10-worker cutoff. In the next section we report the results of the estimation.

5 Results

5.1 Regulations Applying to Firms Employing 10+ Workers

Table 1 reports estimates for the increase in per-worker costs associated with the increased regulatory burden of crossing this threshold, τ, at the all-India level and for a selection of states, industries and ownership types. Estimates for all states, industries and ownership types are reported in Appendix C. Standard errors, displayed beside the point estimates in parentheses, are obtained from a clustered bootstrap procedure with 200 replications. Following GLV, we cluster by industry at the 4 digit NIC code level. This allows for the possibility that differences in production technology - which could affect the firm size distribution and therefore our estimates - may be correlated by industry.[25] The top panel of Table 1 gives the all-India estimate of τ using our methodology. The point estimate is .35 and is significant at the $< 1\%$ level. This means that, on average, establishments in India that employ more than 9 workers act as though they must pay additional labor costs of 35% of the wage per additional worker.

By contrast, estimating the model without accounting for misreporting in any way yields much larger estimates. In particular, estimating Equation 5 on the size distribution omitting

[25] For robustness we have also tried alternative procedures, including a wild bootstrap and non-parametric bootstrap - both clustered at the firm size level. Clustering in this fashion allows for the possibility that reporting errors may be correlated by size and provide the most conservative approach to inference (especially the latter procedure). Under these procedures, statistical significance of our All-India estimates and estimates for most large states and industries survive (although some are only significant at the 10% level under the latter procedure).

sizes larger than 99 workers and including the same dummy variables as in our own specification would lead us to conclude that exceeding the 10-worker threshold increases per-worker costs by 101%. This is due to a combination of rounding and the fact that the density associated with establishment sizes 21 - 99 converges only slowly back to the downshifted power law it follows at larger sizes, as predicted in our misreporting model. In other words, a "naive" estimation puts undue weight on firm sizes whose densities are biased downward by misreporting. In what follows we will focus our discussion on our misreporting-robust estimates of τ.

The lower panels of Table 1 show substantial variation in the magnitude of our misreporting-robust estimates of the per-worker tax by State, Industry and Ownership Type. For example, the point estimate on τ for the State of Kerala is .14 and is not statistically significant, while the estimate for Bihar, on the other hand, is .70 and is statistically significant at the 5% level, implying that establishments in Bihar act as though they must pay a tax of 70% of the wage for each additional worker they employ past 9 workers.

For industries, we see that *de facto* regulatory costs are high for establishments in manufacturing, construction and retail and wholesale trade. Some industries have very noisy estimates, at times producing negative point estimates for τ. This is also true of some of the smaller states and ownership categories (as one can see in Appendix C), and is explained by the fact that the power law relationship can break down when there are a small number of observations in a category - as is the case for electricity, gas and water. In a few cases, negative point estimates reflect the fact that the production and market characteristics of these industries can vary greatly from our model so that it provides a poor fit of the data.[26]

When looking at the differences by ownership type, we find that the estimates for τ are highest for private firms (particularly unincorporated proprietorships, which form by far the largest category of private firms) and insignificant for government-owned firms. This is to

[26] Andhra Pradesh, the largest state to show a negative point estimate for τ, has a size distribution distorted in ways that are different from all other states, and which produces a poor fit (figure available upon request). We have concluded that this is the result of errors in data collection or recording rather than deliberate misreporting.

be expected, since a) the regulatory burden does not vary as much across the 10 worker threshold for government establishments (see Section 2), and b) inspectors are less likely to engage in extortionary corruption with government establishments - which, we will argue in the following section, is a primary determinant of the high effective regulatory costs.

For unincorporated proprietorships we can observe information about the gender and social group of the owner. The results in Table 3 show that the effective regulatory costs, τ, appear to be much higher for disadvantaged social groups (i.e. members of "Scheduled Tribe" and "Scheduled Caste" communities) who may lack bargaining power over government officials. The estimate of τ is also higher for female-owned establishments than for male-owned ones, although this difference is not statistically significant.[27] Since there is no difference in the substance of the law across gender or caste, the results imply that much of the variation in τ is driven by differences in how it is enforced. We explore this idea further in the next section.

The results above derive from the 2005 Economic Census, but we have also used data from the 1998 Economic Census to test whether there is inter-temporal variation in regulatory costs. Using the same empirical methodology described in Section 4, we estimate τ at the All-India level to be equal to .48 (.12) in the earlier data. Although somewhat larger in magnitude, it lies within the confidence interval of our 2005 estimate.[28] Interestingly, the downshift in the 1998 firm size distribution is not as visually striking as that observed in the 2005 data, which may reflect the fact that incentives related to misreporting were different in the two time periods, for example due to the Address Slip reporting requirement added in 2005. Let us reiterate, however, that while misreporting may be responsible for visible distortions around the threshold, it is not likely to affect our estimates of τ. This is especially

[27] Note that while the estimated regulatory costs for establishments owned by members of disadvantaged communities are high, their contribution to the overall costs is relatively low, since there are few owners from these backgrounds (4.05% of proprietorships are members of Scheduled Tribes, 9.82% are members of Scheduled Castes, and 9.48% are female).

[28] Relatedly, estimates of τ based on an alternative dataset comprised of the 2005/6 ASI and the 2005/6 NSSO Unorganized Manufacturing Enterprises Survey are also similar to our estimate of τ using the 2005 EC.

true of enumerator misreporting (see Appendix B.5 for details).

5.2 Employment Protection Legislation

In this subsection we report the results obtained by using our empirical strategy to test for an increase in per-worker costs for establishments that hire more than 100 workers and thus fall under the ambit of Chapter VB of the IDA, the most stringent component of India's employment protection legislation. As before, we run the test on the 2005 Economic Census and report the standard error in parentheses. One difference in the estimation procedure is that we use the number of "hired workers" of the firm, as opposed to the "total workers", since the IDA excludes non-hired workers.[29] Another difference is that we now include dummy variables for firm sizes 1 to 20, so we are effectively comparing the distribution from 21 to 99 with that from 100 onwards. We include the dummies from 1 to 9 because we do not want to conflate the effect of the 100 worker threshold with that of the 10 worker threshold, and we include the dummies from 10 to 20 because those values will be most contaminated by misreporting - as implied by our model. Finally, we exclude West Bengal in this analysis because its VB IDA threshold is different. The results are shown in the table and largely conform to what the figures in Section 3 informally suggest: there is little evidence of a downshift. The implied τ is only .03 and is not statistically significant.[30] Chapter VB of the IDA does *not* therefore appear to have an adverse effect on *the unit labor costs* of firms.[31]

[29]In fact, the threshold for Chapter VB of the IDA is meant to include only permanent workers, but the number of hired workers is the best proxy we have.

[30]We obtain similar results when testing for a downshift at the 50 worker threshold (Chapter VA of the IDA): τ is -.069 (.050). However, these results must be interpreted with some caution given that our model implies that firm sizes larger than 10 can be affected by the regulations at 10 (and 20).

[31]Note that our procedure is only capable of capturing distortions in the unit labor costs of firms, as those are the only ones that would show up as a downshift in the log firm size distribution. If the IDA imposes fixed costs, our procedure will not detect them. GLV identify fixed costs from bunching at N, but this is not possible for us because reported bunching may not reflect actual bunching, as discussed in Section 4.1.

6 Discussion and Investigation of Mechanisms

In the previous section we documented considerable variation - across states, industries and ownership types - in our estimates of the costs of regulations (τ) applying to firms that employ 10 or more workers. In this section we explore the determinants of this variation and show that differences in regulatory enforcement across states (particularly inspector bargaining power and levels of corruption) help explain the variation in τ. Before getting to the results, we note that the analyses we run in this section are necessarily somewhat speculative, since we do not claim to have isolated as-good-as-random variation in regulatory enforcement. Note also that all relevant variables in the following analysis have been rescaled to have mean zero and standard deviation one, with the goal of allowing comparability between regression coefficients in different specifications.

6.1 τ vs Measures of Regulation and Corruption

We begin by regressing our state-level estimates of τ against other established measures of the regulatory environment.[32] These include the "Besley Burgess" (BB) measure of labor regulations from Aghion et al. (2008) as well as several measures of regulatory reform from Dougherty (2009). The former is a measure of the number of amendments that a state government has made to the IDA in either a "pro-worker" or "pro-employer" direction, as interpreted by Aghion et al. (2008), who update the measure to include amendments up to 1997.[33] Positive values of the BB measure indicate more "pro-worker" amendments, which are assumed to imply a more restrictive environment for firms operating in those states. Dougherty (2009) provides state level reform indicators that reflect "the extent

[32] Note that the estimates of τ we use in all the analysis below were generated using the procedure in Section 4.2. For this reason, we weight observations using analytic weights inversely proportional to the variance of our estimate of τ in all regressions which include τ as the regressand. In general, using a dependent variable that is generated with error leads to standard errors that are biased upward. Weighted least squares is a standard approach for improving precision by weighting more heavily those observations that are estimated more precisely (see Allcott (2015) for another example). Nevertheless, our conclusion does not depend on this procedure as we obtain qualitatively similar results when using unweighted regressions.

[33] Since there were no state-level amendments to the IDA between 1997 and 2005, this measure is appropriate for use with 2005 data.

to which procedural or administrative changes have reduced transaction costs in relation to labor issues" by "limiting the scope of regulations, providing greater clarity in their application, or simplifying compliance procedures".[34] Higher values therefore indicate an improved environment for firms. Dougherty's measures are unique in that they cover a wide range of labor-related issues - not just the IDA.[35] In the analysis below, we will focus on an overall measure of reforms from Dougherty (2009) as well as a measure of reforms regarding the role of inspectors, which aims to capture the extent to which states have reformed rules to constrain the influence of inspectors and includes such actions as limiting the number of inspector visits to one per year and requiring authorization for specific complaints.

Table 4 reports the results of regressing τ against the two measures from Dougherty (2009) and the BB measure. The main finding is a robust correlation between the Dougherty (2009) measures and τ: states that saw more "transaction cost reducing" reforms - particularly if they constrained the power of inspectors - have significantly lower τs.[36] This result is to be expected because Dougherty's measures includes reforms that change how firms are impacted by laws that vary across the 10 worker threshold. For example, reforms that affect the powers of inspectors certainly have a differential impact on firms above and below the threshold since firms above the threshold fall under the legal ambit of many more inspectors than firms below the threshold. By contrast, we find no strong correlation between τ and the BB measure. This is perhaps unsurprising, as the BB measure captures variation only due to state amendments to the IDA, which does not vary over the ten person threshold. On the other hand, many studies use the BB measure to proxy for the general regulatory environment (e.g. Adhvaryu et al. (2013)) so we might expect it to correlate with our own

[34]These measures are the result of surveying "a labour expert designated by the AIOE [All-India Association of Employers] or Federation of Indian Chambers of Commerce and Industry (FICCI) affiliate in the state capital" of each state, and adjusting the answers "through discussions with local union leaders, independent labour experts, employers and state labour commissioners" (Dougherty (2009)).

[35]Other than the IDA, the specific areas covered in Dougherty's index include the "Factories Act, State Shops and Commercial Establishments Acts, Contract Labour Act, the role of inspectors, the maintenance of registers, the filing of returns and union representation" (Dougherty (2009)).

[36]τ is not significantly correlated with most of the 7 other subcomponent measures from Dougherty (2009). One notable exception is reforms related to the use of contract workers (not depicted here).

measure of regulatory costs.

In Table 11 of the appendix we report the results of regressing τ against other measures of the labor environment: in particular, per capita measures of strikes, man-days lost to strikes, lockouts, man-days lost to lockouts and the percentage of registered factories that have been inspected. The only measure that is significantly correlated with τ, echoing the results of Table 4, is the percentage of registered factories inspected.

If imposing reforms that constrain the powers of inspectors is correlated with lower effective regulatory costs for firms, this might be because constraining inspectors allows firms to avoid the *de jure* costs associated with following the rules, or it might be because constraining inspectors makes it harder for them to extort firms for bribes. If the latter, we should expect a strong link between τ and the corruption level of the environment.

6.2 τ and Corruption

Indeed, the results of Table 5 show a large and robust *positive* association between τ and two measures of corruption. The first three columns of Table 5 report the results of regressing τ against state level corruption as measured in a 2005 Transparency International (TI) Survey.[37] One might be concerned, however, that the TI measure may be flawed as it is partly the result of individuals' perceptions. Therefore, Columns 4 - 6 of Table 5 report the results of τ regressed against the (normalized) percent of a state's available electricity that was lost in transmission and distribution in 2005. This variable has been used by other researchers as a proxy for corruption and poor state capacity, and has the virtue of being a concrete and objective measure that does not depend on perceptions (Kochhar, Kumar, Rajan, Subramanian, and Tokatlidis (2006)).[38]

[37] The TI corruption measure is based on a survey of the perceptions and experiences regarding corruption in the public sector among 14,405 respondents (Indian households) in 20 Indian states.

[38] Because the samples vary significantly across specifications, we provide results in the Appendix (Tables 12 and 13) that restrict the analysis to include only the 18 largest states by Net State Domestic Product (NSDP), for which data are most consistently available, and because our estimates of τ are most accurately measured for the biggest states. We also provide partial residual plots associated with Columns 3 and 6 of Table 5 to demonstrate that the results are not driven by outliers (Figures 3 and 4 in Appendix E).

Although the state-level correlations between τ and corruption are robust, the regressions are subject to the concern that our measures of corruption may be correlated with omitted variables that also influence τ. To partially address this concern, we provide analysis in Appendix E.1 that corroborates our results using a conceptually different source of variation by taking advantage of within-state, industry level heterogeneity in the exposure to corruption.[39]

The implication that corruption may *increase* regulatory costs appears counter-intuitive given that much of the literature on regulations and corruption (e.g. Khan, Khwaja, and Olken (2015)) has emphasized the role corruption may play in *reducing* regulatory burden. However, if one allows for the possibility that corrupt inspectors can extort firms by threatening to impose large fines for technical violations of the letter of the law while honest inspectors merely require firms to obey the spirit of the law (i.e. a more "reasonable" interpretation of the law that is less costly to abide by), the relationship between regulatory burden and corruption becomes theoretically ambiguous and can easily be positive. We sketch the basic points of such a framework in Appendix G, and in Section 2 we explain why it is likely that the Indian setting would provide a fertile ground for extortionary corruption.

7 Conclusion

This paper makes several contributions to the literature on labor regulations in developing countries. We provide estimates of the unit labor costs associated with a suite of regulations whose components have hitherto received little attention. These regulations include mandatory benefits, workplace safety provisions, and reporting requirements where the lit-

[39] In particular, we hypothesize that if regulations are especially costly *due to* corruption in their enforcement, then we would expect costs to be highest for those businesses that are engaged in regulation-heavy industries *when* they are located in states with high corruption. Using data from the World Bank's 2005 Firm Analysis and Competitiveness Survey of India (FACS) to create an industry level measure of "regulatory intensity," we test this hypothesis in Table 14 of Appendix E and find that the interaction between industry level "regulatory intensity" and state level corruption is indeed associated with higher effective regulatory costs (τ). In this analysis, bias will only arise if omitted variables exist that are correlated with the interaction between state level corruption and industry level "regulatory intensity," which is harder to argue.

erature has previously emphasized employment protection legislation and minimum wage laws. In the Indian context, we find that the costs associated with this suite of regulations are much larger than those associated with the most stringent portion of the country's employment protection legislation. Our results suggest that these types of regulations deserve more attention than they have received to this point.

Our results also suggest a mechanism that may explain why these regulations are so costly in a developing country context: high *de facto* regulatory costs appear to be driven by extortionary corruption on the part of inspectors. Specifically, we show that Indian states that have reformed their inspector-related regulations in a positive way face lower regulatory costs and states with the highest levels of corruption also have the highest levels of regulatory costs. This analysis points to the size of regulatory costs' having more to do with the way regulations are implemented than with the content of the specific laws themselves.

In addition to the above, our paper also makes a methodological contribution. We extend GLV's theoretical model to allow firms to strategically misreport their sizes and simultaneously develop an empirical strategy to estimate costs from a firm size distribution under the assumptions of our model. We show that ignoring the problem of misreporting can lead to vastly over-estimating the actual costs of the regulations. We believe this contribution will find applications in other developing country settings, where the costs of strategic misreporting are typically low.

We close by noting that our analysis reveals the net costs of regulations borne by firms, but does not speak directly to the possible benefits to workers. Our results do suggest that the current regulations make it easy for inspectors to penalize firms for technical violations rather than violations of grave consequence. To the extent that this is so, it is unlikely that workers derive as much protective benefit from the regulations as they might otherwise. It is difficult to arrive at more concrete conclusions, as data do not exist that would allow one to measure how workers would benefit if their employers were made to follow the spirit rather than the letter of the law. However, the results hint at an intriguing possibility: by simplifying regulations identified as costly - or by clarifying compliance and enforcement - it may be possible to reduce the costs borne by firms without diminishing effective protection for workers.

References

Adhvaryu, A., A. Chari, and S. Sharma (2013). Firing Costs and Flexibility: Evidence from Firms' Labor Adjustments to Shocks in India. *Review of Economics and Statistics 95*(3).

Aghion, P., R. Burgess, S. J. Redding, and F. Zilibotti (2008). The Unequal Effects of Liberalization: Evidence from Dismantling the License Raj in India. *American Economic Review 98*(4), 1397–1412.

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics 130*, 1117–1165.

Almeida, R. K. and L. Ronconi (2016). The Enforcement of Labor Law in the Developing World: Some Stylized Facts from Labor Inspections. *Industrial Relations 55*(3), p468–489.

Almond, D. and J. J. Doyle (2011). After Midnight: A Regression Discontinuity Design in Length of Postpartum Hospital A Regression Discontinuity Design in Length of Postpartum Hospital Stays*. *American Economic Journal: Economic Policy 3*(3), 1–34.

Almunia, M. and D. Lopez-Rodriguez (2015). Under the Radar: The Effects of Monitoring Firms on Tax Compliance. *Working Paper*.

Asher, S. and P. Novosad (2016). Politics and Local Economic Growth: Evidence from India. *American Economic Journal: Applied Economics, forthcoming*.

Axtell, R. (2001). Zipf Distribution of U.S. Firm Sizes. *Science 293*(5536), 1818–1820.

Bajaj, V. (2011, November 30). Outsourcing Giant Finds It Must Be Client, Too. *The New York Times*.

Banerjee, A. (1994). Eliminating Corruption. In M. Quibria (Ed.), *Proceedings of the Third Annual Conference on Development Economics*. Asian Development Bank.

Barreca, A. I., J. M. Lindo, and G. R. Waddell (2016). Heaping-Induced Bias in Regression-Discontinuity Designs. *Economic Inquiry 54*(1), 268–293.

Basu, K. (2011). Why, for a Class of Bribes, the Act of Giving a Bribe should be Treated as Legal. *Working Paper*.

Besley, T. and R. Burgess (2004). Can Labor Regulation Hinder Economic Performance? Evidence from India. *The Quarterly Journal of Economics 119*(1), 91–134.

Bhattacharjea, A. (2006). Labour Market Regulation and Industrial Performance in India: A Critical Review of the Empirical Evidence. *Indian Journal of Labour Economics 49*(2).

Bhattacharjea, A. (2009). The Effects of Employment Protection Legislation on Indian Manufacturing. *Economic and Political Weekly XLIV*(22), 55–62.

Botero, J. C., S. Djankov, R. La Porta, F. Lopez-De-Silanes, and A. Shleifer (2004). The Regulation of Labor. *The Quarterly Journal of Economics 119*(4), 1339–1382.

Chatterjee, U. and R. Kanbur (2013). Regulation and Non-Compliance: Magnitudes and Patterns for India's Factories Act. *Working Paper*.

Chaurey, R. (2015). Labor Regulations and Contract Labor Use: Evidence from Indian Firms. *Journal of Development Economics 114*, 224–232.

de Andrade, G. H., M. Bruhn, and D. McKenzie (2014). A Helping Hand or the Long Arm of the Law? Experimental Evidence on What Governments Can Do to Formalize Firms. *The World Bank Economic Review* (May), lhu008.

de Mel, S., D. Mckenzie, and C. Woodruff (2013). The Demand for, and Consequences of, Formalization among Informal Firms in Sri Lanka. *American Economic Journal: Applied Economics 5*(2), 122–150.

Debroy, B. (2013). India's Segmented Labour Markets, Inter-State Differences, and the Scope for Labour Reforms. In B. Debroy, L. Bhandari, S. Aiyar, and A. Gulati (Eds.), *Economic Freedom of the States of India*, Chapter 4. Academic Foundation.

Djankov, S. and R. Ramalho (2009). Employment laws in developing countries. *Journal of Comparative Economics 37*(1), 3–13.

Dougherty, S., V. Frisancho, and K. Krishna (2014). State-Level Labor Reform and Firm-Level Productivity in India. *India Policy Forum 10*.

Dougherty, S. M. (2009). Labour Regulation and Employment Dynamics at the State Level in India. *Review of Market Integration 1*(3), 295–337.

Eggers, A. C., R. Freier, V. Grembi, and T. Nannicini (2018). Regression Discontinuity Designs Based on Population Thresholds. *American Journal of Political Science 62*(1), 210–229.

Fagernas, S. (2010). Labor Law, Judicial Efficiency, and Informal Employment in India. *Journal of Empirical Legal Studies 7*(2), 282–321.

Freeman, R. B. (2010). *Labor regulations, unions, and social protection in developing countries: Market distortions or efficient institutions?* (1 ed.), Volume 5. Elsevier BV.

Garicano, L., C. Lelarge, and J. Van Reenen (2013). Firm Size Distortions and the Productivity Distribution: Evidence from France. *NBER Working Paper 18841*.

Garicano, L., C. Lelarge, and J. Van Reenen (2016). Firm Size Distortions and the Productivity Distribution: Evidence from France. *American Economic Review 106*(11), 3439–3479.

Ghosh, A. (2016, February 28). Why Make in India is stumbling over our labour laws (Comment: Special to IANS). *Business Standard*.

Goyette, J. and W. Kouame (2016). Tax Evasion, Social Infrastructures and Misallocation. *Working Paper*.

Hasan, R., D. Mitra, and K. V. Ramaswamy (2007). Trade reforms, labor regulations, and labor-demand elasticities: Empirical evidence from India. *The Review of Economics and Statistics 89*(August), 466–481.

Hernández-Pérez, R., F. Angulo-Brown, and D. Tun (2006). Company Size Distribution for Developing Countries. *Physica A: Statistical Mechanics and its Applications 359*, 607–618.

Hindriks, J., M. Keen, and A. Muthoo (1999). Corruption, Extortion and Evasion. *Journal of Public Economics 74*(3), 395–430.

Hsieh, C. T. and P. J. Klenow (2009). Misallocation and Manufacturing TFP in China and India. *The Quarterly Journal of Economics 124*(4), 1403–1448.

Hsieh, C.-T. and B. A. Olken (2014). The Missing "Missing Middle". *Journal of Economic Perspectives 28*(3), 89–108.

Huntington, S. P. (1968). *Political Order in Changing Societies.* New Haven, CT: Yale University Press.

Kanbur, R. and L. Ronconi (2015). Enforcement Matters: The Effective Regulation of Labor. *CEPR Discussion Paper No. 11098*.

Khan, A., A. Khwaja, and B. Olken (2015). Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors. *The Quarterly Journal of Economics, forthcoming*.

Kleven, H. J. and M. Waseem (2013). Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan. *Quarterly Journal of Economics 128*(2), 669–723.

Kochhar, K., U. Kumar, R. Rajan, A. Subramanian, and I. Tokatlidis (2006). India's Pattern of Development: What Happened, What Follows? *Journal of Monetary Economics 53*(5), 981–1019.

Kumler, T., E. Verhoogen, and J. Frias (2015). Enlisting Employees in Improving Payroll-Tax Compliance: Evidence from Mexico. *Working Paper*.

La Porta, R. and A. Shleifer (2014). Informality and development. *28*(Working Paper 20205), 109–126.

Lucas, R. E. J. (1978). On the size distribution of business firms. *The Bell Journal of Economics 9*(2).

Markovitch, N. and U. Krieger (2000). Nonparametric Estimation of Long-Tailed Density Functions and its Application to the Analysis of World Wide Web Traffic. *Performance Evaluation 42*(2-3), 205–222.

Mishra, A. and D. Mookherjee (2013). Controlling Collusion and Extortion: The Twin Faces of Corruption. *Working Paper*.

Mookherjee, D. (1997). Incentive Reforms in Developing Country Bureaucracies. In *Annual World Bank Conference on Development Economics*. World Bank.

Nataraj, S., F. Perez-Arce, K. B. Kumar, and S. V. Srinivasan (2014). The impact of labor market regulation on employment in low-income countries: A meta-analysis. *Journal of Economic Surveys 28*(3), 551–572.

Polinsky, A. M. and S. Shavell (2001). Corruption and Optimal Law Enforcement. *Journal of Public Economics 81*(1), 1–24.

Sequeira, S. and S. Djankov (2014). Corruption and firm behavior: Evidence from African Ports. *Journal of International Economics 94*, 277–294.

Svensson, J. (2005). Eight Questions about Corruption. *Journal of Economic Perspectives 19*(3), 19–42.

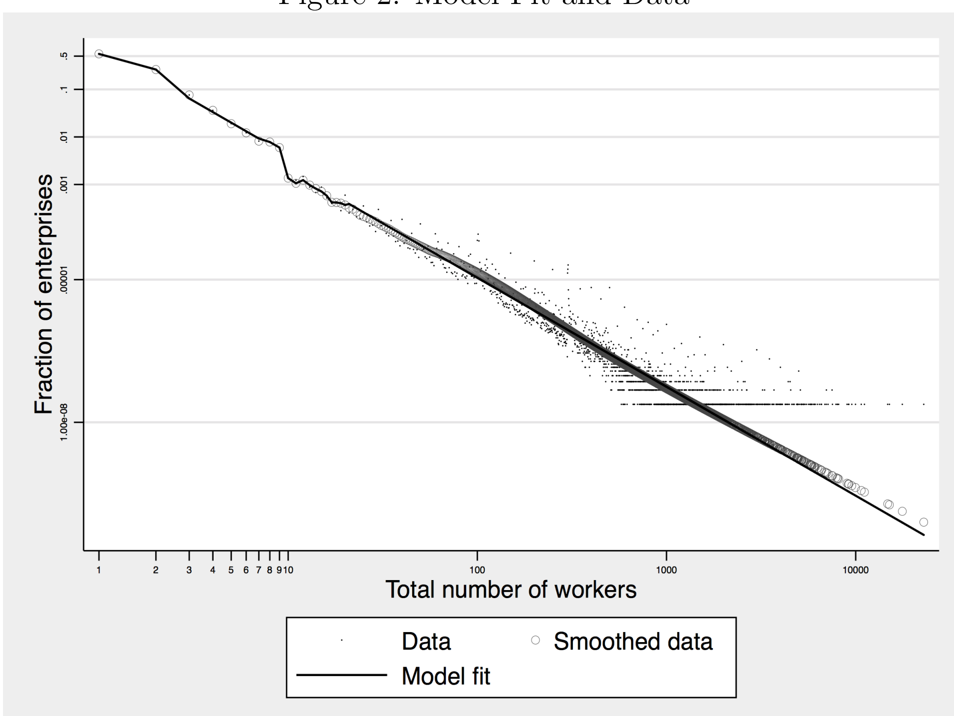
TeamLease Services (2006). India Labour Report 2006: A Ranking of Indian States by their Labour Ecosystem. Technical report.

Figures and Tables

Figure 1: 2005 Log-Log Distribution of Establishment Size

Note: Both axes are on a log scale. Total number of workers is the number of workers usually working daily in an establishment. Source: 2005 Economic Census of India.

Figure 2: Model Fit and Data



Note: This figure shows the fit of the model described in Section 4.2 (the black line) to the data (dark grey points). Model estimation involves non-parametric smoothing using the method described in Markovitch and Krieger (2000) with a bandwidth of 0.005 as a first step. The smoothed density estimates are shown in light grey. The second step is to fit the black line to the light grey points. Both axes in log scale. Source: 2005 Economic Census of India.

Table 1: Estimates of τ at 10 worker threshold

Level	τ	s.e.
All-India		
	0.347	(0.081)
By State		
Bihar	0.693	(0.302)
Gujarat	0.165	(0.151)
Kerala	0.138	(0.196)
Karnataka	0.520	(0.156)
Uttar Pradesh	0.502	(0.254)
By Industry		
Wholesale & retail trade	0.637	(0.094)
Manufacturing	0.268	(0.085)
Construction	0.478	(0.549)
Electricity, gas and water	-0.367	(0.145)
By Ownership Type		
Government and PSU	-0.092	(0.128)
Unincorporated Proprietary	0.430	(0.059)

Note: This table presents estimates of regulatory costs faced by establishments with 10 or more workers, using the methodology described in Section 4 with a bandwidth of 0.005. Standard errors generated using a clustered bootstrap procedure with 200 replications are presented in parentheses. Clustering is done at the 4 digit (NIC code) industry level, following Garicano et al. (2016). Estimates are presented for a subset of states, industries and ownership types, as well as at the All-India level. Results for all states, industries and ownership types are available in Appendix C. Source: 2005 Economic Census of India.

Table 2: Estimate of τ at 100 (hired) worker threshold

All-India	τ	s.e.
	0.0298	(0.0295)

Note: This table presents an estimate of regulatory costs faced by establishments that hire 100 or more workers, using the methodology described in Section 4. The standard error generated using a clustered bootstrap procedure with 200 replications is presented in parentheses. Clustering is done at the 4 digit (NIC code) industry level, following Garicano et al. (2016). The estimate is presented for the All-India level using "hired workers" only. Source: 2005 Economic Census of India.

Table 3: Estimates of τ at 10 worker threshold by social group of owner

Level	τ	s.e.
By Gender of Owner		

Male	0.424	(0.060)
Female	0.525	(0.207)
By Social Group of Owner		
Scheduled Tribe	1.016	(0.335)
Scheduled Caste	0.890	(0.233)
Other Backward Caste	0.425	(0.086)
Other	0.326	(0.059)

Note: This table presents estimates of regulatory costs faced by establishments that employ 10 or more workers, using the methodology described in Section 4 with a bandwidth of 0.005. Standard errors generated using a clustered bootstrap procedure with 200 replications are presented in parentheses. Clustering is done at the 4 digit (NIC code) industry level, following Garicano et al. (2016). Estimates are presented by gender and social group of owner using only privately owned establishments (for which these data are available). Source: 2005 Economic Census of India.

Table 4: Tau vs Other Measures of Regulations (all States and UTs)

	(1) tau	(2) tau	(3) tau	(4) tau	(5) tau	(6) tau
Dougherty measure	-0.360**	-0.394*				
(all reforms)	(0.169)	(0.199)				
Dougherty measure			-0.480***	-0.623***		
(inspector reforms)			(0.162)	(0.148)		
Besley-Burgess					0.223	0.235
measure (regs)					(0.178)	(0.177)
log of net state		-0.394*		-0.428**		-0.494*
domestic product pc		(0.192)		(0.181)		(0.237)
share of privately		1.411		8.538		-10.35
owned establishments		(6.022)		(5.655)		(8.065)
Constant	0.131	2.900	0.209	-2.952	-0.00266	14.20*
	(0.181)	(5.514)	(0.140)	(5.401)	(0.280)	(7.402)
Observations	21	21	21	21	16	16

Note: This table tests for correlations between our estimated regulatory costs (tau) and other established measures of the regulatory environment from the previous literature. Robust standard errors are reported in parentheses. Observations are weighted by the inverse variance of tau and include all Indian States and Union Territories for which data are available. Sources: Dougherty(2009); Besley and Burgess(2004); RBI.

Table 5: Tau vs State Level Measures of Corruption (all States and UTs)

	(1) tau	(2) tau	(3) tau	(4) tau	(5) tau	(6) tau
TI Corruption Score	0.617**	0.587*	0.685***			
	(0.286)	(0.321)	(0.127)			
electricity losses				0.268	0.254	0.593***
				(0.303)	(0.226)	(0.153)
log of net state domestic product pc		-0.156	-0.215		-0.176	-0.271**
		(0.306)	(0.277)		(0.221)	(0.107)
share of privately owned establishments		-1.037	6.274*		6.820**	7.842*
		(4.469)	(3.227)		(3.291)	(4.313)
Dougherty measure (inspection reforms)			-0.594***			-0.494***
			(0.0910)			(0.0917)
Electricity available (GWH)					0.139	
					(0.197)	
Constant	0.247	2.736	-2.864	-0.486*	-4.347	-3.698
	(0.233)	(4.438)	(3.497)	(0.251)	(3.570)	(3.937)
Observations	20	20	19	35	32	21
Measure of Corruption	TI	TI	TI	TDLs	TDLs	TDLs

Note: This table reports the results of our estimated regulatory costs (tau) regressed against two different measures of corruption. Robust standard errors are reported in parentheses. Observations are weighted by the inverse variance of tau and include all Indian States and Union Territories for which data are available. Sources: Transparency International (2005); RBI; Dougherty(2009).

Appendices for Online Publication Only

A List of Size-Based Regulations in India

Name of Act	Size-based threshold	Nature of size-based threshold	Definition of worker/employee	Notes
The Building and Other Construction Workers (Regulation of Employment and Conditions of Service) Act, 1996	10+ building workers	Regulates employment and conditions of service. Also includes registration requirements for all establishments with 10+ workers.	" 'building worker' means a person who is employed to do any ... work for hire or reward, whether the terms of employment be expressed or implied."	

Name of Act	Size-based threshold	Nature of size-based threshold	Definition of worker/employee	Notes
The Employees' State Insurance Act, 1948	10+ workers	Provides for health, maternity and employment injury insurance to industrial workers. Sets up medical benefit councils, requires enterprises to "furnish returns and maintain registers", etc.	An "employee" is a person employed for wages - whether directly or indirectly, permanent or temporary	applies to government establishments; originally applicable to all "factories" but since extended to shops, hotels, etc
The Factories Act, 1948	10+/20+: 10+ workers with power, 20+ workers without power	Regulates employment and conditions of service. Also includes registration requirements for all establishments with 10+/20+ workers.	" 'worker' means a person employed, directly or by or through any agency (including a contractor) with or without the knowledge of the principal employer, *whether for remuneration or not*"	applies to government establishments (but not armed forces)
The Industrial Disputes Act, 1947	10+, 50+, 100+ [permanent] workers	Sets out procedures to govern the investigation and settlement of industrial disputes between firms, workers and trade unions.	the thresholds at 50 and 100 are for permanent employees only (i.e. contract workers are excluded)	applies to government establishments
The Maternity Benefits Act, 1961	10+ employees	regulates maternity benefits		applies to government establishments
Payment of Gratuity Act, 1972	10+ employees	Requires and regulates the payment of gratuities to workers upon resignation or retirement (at a rate of 15 days salary per year of service).	any person employed for wages, whether the terms of such employment are express or implied	government establishments have their own version of gratuities as laid out in the pay commissions
Labour Laws (Exemption from Furnishing Returns and Maintaining Registers by Certain Establishments) Act, 1988	10+; 20+ employees	Allows "small" (10-19) and "very small" (1-9) establishments to keep significantly simpler records and registers.		

Name of Act	Size-based threshold	Nature of size-based threshold	Definition of worker/employee	Notes
Payment of Bonus Act, 1965	10+/20+ workers; applies to "every factory" AND "every other establishment" with 20+ workers	Requires bonuses to be paid out of profits based on productivity. The minimum bonus was 8.33% of salary in 2005.	"any person employed on salary not exceeding Rs. 10,000 per month in any industry to do any ... work"	government establishments have their own version of bonuses as laid out in the pay commissions
Payment of Wages Act, 1936	10+/20+: applies to persons employed in any factory or an assortment of other industries specified in the law.	"[R]egulate[s] the payment of wages of certain classes of employed persons"; This act requires employers to pay employees in a timely fashion and establishes a significant amount of bureaucracy to ensure this end.	includes workers hired through subcontractors	this law was passed during the colonial era to address the problem of non-payment or delayed payment of workers.
The Personal Injuries (Compensation Insurance) Act, 1963	10+/20+: applies to persons employed in any factory and a few other industries specified in the law (e.g. mines).	Imposes liability on employers to compensate workers for injury.		
The Contract Labour (Regulation and Abolition) Act, 1970	20+ contract workers	Regulates employment of contract labour.	contract workers (hired through a contractor)	
The Employees' Provident Fund and Miscellaneous Provisions Act, 1952	20+: Applies "to every establishment which is a factory engaged in any industry specified in Schedule I and in which twenty or more persons are employed"	Establishes "contributory provident funds" for workers in certain industries and establishments (currently 24% of wage - split equally between employer and employee).	" 'employee' means any person who is employed for wages in any kind of work"	does not apply to government establishments

Name of Act	Size-based threshold	Nature of size-based threshold	Definition of worker/employee	Notes
Inter-State Migrant Workmen (Regulation of Employment and Conditions of Service) Act,1979	5+; applies to every establishment or contractor that employs 5 or more inter state migrant workers	This act came about in response to perceived exploitation of migrant workers and thus aims "to regulate the employment of inter-State migrant workmen and to provide for their conditions of service..."		
The Motor Transport Workers Act, 1961	5+: "applies to every motor transport undertaking employing five or more motor transport workers"	Regulates employment and conditions of service (like Factories Act but specifically for Motor Transport Workers). Also includes registration requirements for all establishments with 5+ workers.	"a person who is employed in a motor transport undertaking directly or through an agency, whether for wages or not..."	
The Trade Unions Act, 1926	7+: Requires at least 7 members of a non-registered trade union to apply for the union's registration.	Defines the conditions under which it is possible to form a trade union.		
The Plantations Labour Act, 1951	15+ (and 5+ hectares of land): applies to plantations that grow tea, coffee, rubber or cardamom, and which have 15+ workers AND 5+ hectares of land	Regulates employment and conditions of service (like the Factories Act but specifically for Plantation Workers in Tea, Coffee, etc plantations). Also includes registration requirements for all establishments with 15+ workers and 5+ hectares of land.		
The Industrial Employment (Standing Orders) Act, 1946	100+	Requires establishments to define "the conditions of recruitment, discharge, disciplinary action, [etc]".		

Note: This table includes a comprehensive list of size-based labor regulations in India. Empty cells denote non-applicable or missing information. Source: All information including direct quotes taken from Universal Law Publishing Company (2013).

B Omitted Proofs Regarding the Theoretical Log Density of Firm Size With and Without Misreporting

B.1 Derivation of the theoretical log density without misreporting

The model we present in Section 4.1 of the paper augments the GLV framework by incorporating strategic misreporting. In this part of the appendix, we provide a summary of the basic GLV framework (i.e. without misreporting), including a derivation of the theoretical log density of firm size.[40]

As noted earlier, the primitive of the model is the distribution of managerial ability (α), which is assumed to follow a power law: $\phi(\alpha) = c_\alpha \alpha^{-\beta_\alpha}$, where $c_\alpha \equiv (\beta_\alpha - 1)\underline{\alpha}^{\beta_\alpha - 1}$ and $\underline{\alpha}$ is the minimum possible value of α[41].

A firm with productivity α and output given by a power function of labor ($f(n) = n^\theta$) faces the following profit-maximization problem:

$$\pi(\alpha) = \max_n \alpha n^\theta - w\bar{\tau}n \tag{7}$$

where again, n is the number of workers a firm employs, w is a constant wage paid to all workers, and $\bar{\tau}$ is a proportional tax on labor that takes the value 1 if $n \leq N$ and $1+\tau$ if $n > N$, where $\tau > 0$. The resulting first order condition suggests the following general relationship between employment and productivity: $n^*(\alpha) = (\frac{\theta}{w\bar{\tau}})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}}$. However, this relationship is discontinuous at N and only applies for interior solutions: some firms will find it profitable to choose the corner solution of $n(\alpha) = N$ rather than $n^*(\alpha)$.

In fact, firms can be sorted into three categories, according to their productivity, α.

1) Firms with the lowest values of $\alpha (\in [\underline{\alpha}, \alpha_1])$ are not affected by the regulation and choose their optimal employment in an unrestricted way. In particular they choose $n^*(\alpha) = (\frac{\theta}{w})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} \leq N$. α_1 is defined such that $\pi(n^*(\alpha_1)) = \pi(N)$, where $n^*(\alpha_1) = (\frac{\theta}{w})^{\frac{1}{1-\theta}}(\alpha_1)^{\frac{1}{1-\theta}}$.

2) Firms with productivity larger than α_1 but lower than another threshold ($\alpha \in (\alpha_1, \alpha_2]$), find it optimal to choose $n^*(\alpha) = N$, rather than exceed the threshold and expose themselves to the discontinuously higher costs associated with the size-based regulation.

3) The last category includes those firms with the highest productivities ($\alpha > \alpha_2$). These firms find it optimal to exceed the threshold even though it means paying higher unit labor costs: $n^*(\alpha) = (\frac{\theta}{w(1+\tau)})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}}$. α_2 is defined such that $\pi(n^*(\alpha_2)) = \pi(N)$, where $n^*(\alpha_2) = (\frac{\theta}{w(1+\tau)})^{\frac{1}{1-\theta}}(\alpha_2)^{\frac{1}{1-\theta}}$.

To summarize then, a full mapping between productivity α and firm size n is given by:

$$n(\alpha) = \begin{cases} (\frac{\theta}{w})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} \leq N & if \quad \alpha \in [\underline{\alpha}, \alpha_1] \\ N & if \quad \alpha \in (\alpha_1, \alpha_2] \\ (\frac{\theta}{w(1+\tau)})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} > N & if \quad \alpha > \alpha_2 \end{cases}$$

An exact expression for the distribution of firm size, $\chi(n)$, can now be recovered as a transformation of the distribution of managerial ability, $\phi(\alpha)$, since the first-order conditions

[40] The analysis here follows closely from that of GLV (see their Appendix B for their derivation).

[41] These last two assumptions are needed to satisfy $\int_{\underline{\alpha}}^{\infty} \phi(\alpha) = 1$.

on the firms' maximization problems imply the monotonic relationship between α and n described above. Specifically, we transform $\phi(\alpha)$ into $\chi(n)$ using the change of variables formula along with the complete expression for $n(\alpha)$ above:

$$\chi(n) = \begin{cases} \left(\frac{1-\theta}{\theta}\right)^{1-\beta}(\beta-1)n^{-\beta} & if \quad n \in [n_{\min}, N) \\ \left(\frac{1-\theta}{\theta}\right)^{1-\beta}\left(N^{1-\beta} - (1+\tau)^{-\frac{\beta-1}{1-\theta}} n_u^{1-\beta}\right) & if \quad n = N \\ 0 & if \quad n \in (N, n_u) \\ \left(\frac{1-\theta}{\theta}\right)^{1-\beta}(\beta-1)(1+\tau)^{-\frac{\beta-1}{1-\theta}} n^{-\beta} & if \quad n \geq n_u \end{cases}$$

From here one can complete the derivation for the case without misreporting by simply taking logs to obtain the theoretical log density of firm size:

$$log\chi(n) = \begin{cases} log[\left(\frac{1-\theta}{\theta}\right)^{1-\beta}(\beta-1)] - \beta log(n) & if \quad n \in [n_{\min}, N) \\ log[\left(\frac{1-\theta}{\theta}\right)^{1-\beta}\left(N^{1-\beta} - (1+\tau)^{-\frac{\beta-1}{1-\theta}} n_u^{1-\beta}\right)] & if \quad n = N \\ - & if \quad n \in (N, n_u) \\ log[\left(\frac{1-\theta}{\theta}\right)^{1-\beta}(\beta-1)] - \left(\frac{\beta-1}{1-\theta}\right)(1+\tau) - \beta log(n) & if \quad n \geq n_u \end{cases}$$

B.2 Full Derivation of the Theoretical Log Density With Misreporting

In this part of the appendix, we include the steps omitted in section 4.1 when deriving the theoretical log density of true and reported firm size in the presence of misreporting. We begin by restating the profit-maximization problem (in Equation 1) for a firm that is now allowed to choose both its true employment (n) and its *reported* employment (l):

$$\pi(\alpha) = \max_{n,l} \alpha f(n) - wn - \tau wl * \mathbb{1}\{l > 9\} - q(n) * p(n,l) * F(n,l)$$

where $\alpha, f(n), w$ and τ are all defined as they were previously. As noted, the problem is similar to the case without misreporting except that now firms pay the extra marginal cost, τw, only on their *reported* employment, and not on their true employment. Furthermore, they only pay this cost if their reported employment exceeds the regulatory threshold, $N = 9$. The other new elements are $q(n)$, which gives the probability of inspection as a function of firm size, $p(n,l)$, which is the probability that a misreporting firm is caught by the authorities conditional on being inspected and $F(n,l)$, which is the fine that a firm caught misreporting must pay. Therefore, firms must trade off the benefit of lower regulatory costs from under-reporting their employment against the expected cost of being caught and fined for under-reporting ($M = q * p * F$).

For now we assume a particular functional form for the misreporting costs, $M(n,l) = \frac{F}{n_{max}}(n-l)^2$, but we show in the next two subsections of this Appendix that 1) our main result obtains for any convex form of misreporting costs, and 2) non-convex misreporting costs either do not fit the data or generate similar conclusions. We also reintroduce our earlier assumption that firms' production functions are power ($f(n) = n^{\theta}$), so that the profit maximization problem for a firm with productivity α is:

$$\pi(\alpha) = \max_{n,l} \alpha n^{\theta} - wn - \tau wl * \mathbb{1}\{l > 9\} - F * \frac{(n-l)^2}{n_{max}}.$$

As before, the solution to this problem looks different depending on which of three different productivity categories the firm falls into:

1) Firms with the lowest values of $\alpha(\in [\underline{\alpha}, \alpha_1])$ are not affected by the regulation and choose their optimal employment in an unrestricted way. In particular they choose $n_1^*(\alpha) = (\frac{\theta}{w})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} \leq N$. Because their true employment is below the regulatory threshold, they have no incentive to misreport and hence choose $l_1^*(\alpha) = n_1^*(\alpha)$. α_1 is defined such that $\pi(n_1^*(\alpha_1)) = \pi(N)$, where $n_1^*(\alpha_1) = (\frac{\theta}{w})^{\frac{1}{1-\theta}}(\alpha_1)^{\frac{1}{1-\theta}}$.

2) Firms with productivity larger than α_1 but lower than another threshold ($\alpha \in (\alpha_1, \alpha_2]$), will choose $n > N$, exceeding the regulatory threshold, but will find it profitable to misreport their employment, setting $l_2^*(\alpha) = N$.[42] These firms will only *appear* to be "bunched" up at 9, but will in fact have higher employment. The employment function, $n_2^*(\alpha)$, for these firms is defined implicitly from the first order condition: $\alpha\theta n_2^*(\alpha)^{\theta-1} - w - \frac{2F}{n_{max}}[n_2^*(\alpha) - N] = 0$.

3) The last category includes those firms with the highest productivities ($\alpha > \alpha_2$). These firms are productive enough to warrant hiring work forces so large that they cannot choose $l = 9$ while simultaneously avoiding detection with reasonable probability and must report $l > 9$. Even these firms, however, with both $n > 9$ and $l > 9$ do not find it profit-maximizing to report truthfully. From the first order condition on l we get: $l_3^*(\alpha) = n_3^*(\alpha) - \frac{n_{max}}{2F} w\tau$. In other words, these firms can save on their unit labor costs by shading down their reported employment. Importantly, the degree of misreporting is by a constant amount that is independent of firm size ($\frac{n_{max}}{2F} w\tau$).[43] These firms set their true employment according to: $n_3^*(\alpha) = (\frac{\theta}{w(1+\tau)})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}}$. α_2 is defined such that $\pi(n_3^*(\alpha_2)) = \pi(n_2^*(\alpha_2))$, where $n_3^*(\alpha_2) = (\frac{\theta}{w(1+\tau)})^{\frac{1}{1-\theta}}(\alpha_2)^{\frac{1}{1-\theta}}$, and thus marks the productivity of a firm that is indifferent between reporting employment just below the threshold ($l(\alpha_2) = N$) (while choosing true employment $n_2^*(\alpha_2) > N$), and admitting that it is over the threshold ($l^*(\alpha_2) = n_3^*(\alpha_2) - \frac{n_{max}}{2F} w\tau$), which would allow it to choose a higher level of true employment ($n_3^*(\alpha_2) > n_2^*(\alpha_2)$). We will denote $n_2^*(\alpha_2)$ by $n_m(\alpha_2)$, $n_3^*(\alpha_2)$ by $n_t(\alpha_2)$ and, similarly, $l_3^*(\alpha_2)$ by $l_t(\alpha_2)$. The region between $n_m(\alpha_2)$ and $n_t(\alpha_2)$ is a strictly dominated region in which no firms should set their optimal employment level.

To summarize then, full mappings between productivity α and the true firm size n, as well as between productivity α and reported firm size l, are given by:

$$n^*(\alpha) = \begin{cases} (\frac{\theta}{w})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} \leq N & if \quad \alpha \in [\underline{\alpha}, \alpha_1] \\ n_2^*(\alpha) & if \quad \alpha \in (\alpha_1, \alpha_2] \\ (\frac{\theta}{w})^{\frac{1}{1-\theta}}(1+\tau)^{-\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} > N & if \quad \alpha > \alpha_2 \end{cases}$$

[42]Conditional on misreporting a positive amount, setting $l = N$ is the "optimal lie" for these firms since it yields the largest benefit (by reducing firms' regulatory burden to 0) while minimizing the misreporting costs.

[43]This outcome is a result of the convex cost assumption on the misreporting function.

$$l^*(\alpha) = \begin{cases} (\frac{\theta}{w})^{\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} \leq N & if \quad \alpha \in [\underline{\alpha}, \alpha_1] \\ N & if \quad \alpha \in (\alpha_1, \alpha_2] \\ (\frac{\theta}{w})^{\frac{1}{1-\theta}}(1+\tau)^{-\frac{1}{1-\theta}}(\alpha)^{\frac{1}{1-\theta}} - \frac{n_{max}}{2F}w\tau > N & if \quad \alpha > \alpha_2 \end{cases}$$

From these functions we can obtain expressions for the distributions of true and reported firm size, $\chi(n)$ and $\psi(l)$, as transformations of the distribution of managerial ability, $\phi(\alpha)$ (where $\phi(\alpha) = c_\alpha \alpha^{-\beta_\alpha}$), by the change of variables formula:

$$\chi(n) = \begin{cases} c_\alpha(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}} n^{-\beta} & if \quad n \in [n_{\min}, N) \\ \left|\frac{d\alpha_2^*(n)}{dn}\right| \phi(\alpha_2^*(n)) & if \quad n \in [N, n_m(\alpha_2)) \\ 0 & if \quad n \in [n_m(\alpha_2), n_t(\alpha_2)) \\ c_\alpha(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}}(1+\tau)^{-\frac{\beta-1}{1-\theta}} n^{-\beta} & if \quad n \geq n_t(\alpha_2) \end{cases}$$

$$\psi(l) = \begin{cases} c_\alpha(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}} l^{-\beta} & if \quad l \in [n_{\min}, N) \\ \int_{\alpha_1}^{\alpha_2} \phi(\alpha) d\alpha = \delta_l & if \quad l = N \\ 0 & if \quad l \in (N, l_t(\alpha_2)) \\ c_\alpha(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}}(1+\tau)^{-\frac{\beta-1}{1-\theta}}[l + \frac{n_{max}}{2F}w\tau]^{-\beta} & if \quad l \geq l_t(\alpha_2) \end{cases}$$

where $\beta = \theta + \beta_\alpha - \theta\beta_\alpha$ and $\alpha_2^*(n)$ is the inverse function of $n_2^*(\alpha)$, implicitly defined above. Taking the logarithm of each expression delivers the version of the distributions shown in the main text:

$$log\chi(n) = \begin{cases} logA - \beta log(n) & if \quad n \in [n_{\min}, N) \\ log[\xi(n)] & if \quad n \in [N, n_m(\alpha_2)] \\ - & if \quad n \in (n_m(\alpha_2), n_t(\alpha_2)) \\ logA - \frac{\beta-1}{1-\theta}log(1+\tau) - \beta log(n) & if \quad n \geq n_t(\alpha_2) \end{cases}$$

$$log\psi(l) = \begin{cases} logA - \beta log(l) & if \quad l \in [l_{\min}, N) \\ log(\delta_l) & if \quad l = N \\ - & if \quad n \in (N, l_t(\alpha_2)) \\ logA - \frac{\beta-1}{1-\theta}log(1+\tau) - \beta log(l + \frac{n_{max}}{2F}w\tau) & if \quad l \geq l_t(\alpha_2) \end{cases}$$

where terms have been simplified with two substitutions.[44] These are the densities described in the main text.

B.3 Proof that Convex Misreporting Costs Imply Convergence Between True and Reported Firm Size Distributions

An important implication of the misreporting model from section 4.1 is that the difference between the log density of reported firm size, $\psi(l)$, and the log density of true firm size,

[44] We substituted A for the expression $c_\alpha(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}}$ and $\xi(n)$ for $\left|\frac{d\alpha_2^*(n)}{dn}\right|\phi(\alpha_2^*(n))$.

$\chi(n)$, converges to 0 for large values of n, l. In this section of the appendix we show that this result does not hinge on a specific functional form for the misreporting costs, but instead holds whenever the expected costs of misreporting are increasing and strictly convex in the degree of misreporting *and* are independent of firm size.

We replace our former expression for the expected costs of misreporting, $q(n) * p(n,l) * F(n,l)$, with the simpler expression $M(n,l)$, since here we do not need to distinguish between the fine if caught and the probability of being caught. Furthermore, we impose Assumption 1, which allows us to write the problem of a firm with the option of misreporting as follows:

$$\pi(\alpha) = \max_{n,u} \alpha f(n) - wn - \tau w(n-u) * \mathbb{1}\{n-u > 9\} - M(u)$$

This is identical to Equation 1 except for the change in variables (u for $n-l$) and the more general expression for the expected costs of misreporting. Under Assumption 1, the first order condition on u for a large firm (i.e. one whose reported employment exceeds the threshold) is: $\tau w - M'(u) = 0$. The first term denotes the benefit of increasing u by one unit (in terms of regulatory costs avoided) while the second term captures the marginal cost of u. The first term is constant, while the second starts from 0 (for $u = 0$) and increases at an increasing rate. There exists therefore some value of misreporting that satisfies the first order condition, given by $u^* = M'^{-1}(\tau w)$. Note that the optimal value for misreporting, u^*, does not depend on α. This means that, for the largest set of firms, misreporting is by the same constant amount, regardless of firm size or productivity: $l(\alpha) = n(\alpha) - u^*$.

To see that this result is all that is required for the difference between the reported distribution and the true distribution to converge to 0, consider the analysis in the previous sub-appendix, but with the more general result that $u^* = M'^{-1}(\tau w)$. Then, it is straightforward to show that $log\chi(n) = logA - \frac{\beta-1}{1-\theta}log(1+\tau) - \beta log(n)$ and $log\psi(l) = logA - \frac{\beta-1}{1-\theta}log(1+\tau) - \beta log(l+u^*)$ for firms above the threshold. For large values (i.e. $l = n = x \to \infty$), the difference between these two density functions goes to 0:

$\lim_{x\to\infty} log\chi(x) - log\psi(x) = \beta log(x+u^*) - \beta log(x) = \beta log(x(1+\frac{u^*}{x})) - \beta log(x) = \beta log(1+\frac{u^*}{x}) = 0.$

B.4 Alternative Forms of Misreporting

In this part of the appendix, we consider the implications of alternative assumptions regarding the form of strategic misreporting by firms. It is difficult to characterize a general solution that does not impose any structure on the functional form of misreporting, so instead we proceed by considering several different specifications in turn. We conclude by demonstrating that there is only one potential case (or class of cases) under which our results would be biased.

Recall that n represents the number of workers actually employed within a firm, while l represents the number of workers that a firm reports. The level of misreporting is represented by u, where $u \equiv n - l$. We consider four different specific functional forms for the expected costs of misreporting, $M(n,l)$. The expected cost of misreporting is the product of, respectively, the probability of being inspected, the probability of being caught conditional on being inspected, and the fine levied if caught: $M(n,l) \equiv q(n) * p(n,l) * F(n,l)$. It is reasonable to suppose that the probability of being inspected, $q(n)$, is increasing with respect to firm size because it is easier for inspectors to find larger firms (this supposition receives empirical support from Almeida and Ronconi (2016)). Let us in fact assume that q is directly proportional to firm size: $q(n) = \frac{n}{n_{max}}$. The probability of being caught conditional on being inspected should also be an increasing function of either the level of misreporting (u) or the fraction of employees being misreported ($\frac{u}{n}$). We consider the following two specific forms, representing either case: $p(n,l) = \frac{u}{n_{max}}$ or $p(n,l) = \frac{u}{n}$. Last, we allow the fine that a firm must pay if caught to be either fixed or directly proportional to the level of misreporting:[45] $F(n,l) = F$ or $F(n,l) = F * u$.

This leaves us with 4 possible cases for the functional form of $M(n,l)$. Below we consider the implications for our estimation of τ under each case in turn.

B.4.1 Alternative Misreporting Case 1

The first case is that which corresponds to the one we focus on in the text (Section 4.1). In this case, the probability of inspection is directly proportional to firm size, the probability of being caught conditional on inspection is proportional to the fraction of employees being misreported, and the fine if caught is proportional to the level of misreporting: $q(n) = \frac{n}{n_{max}}$; $p(n,l) = \frac{u}{n}$; $F(n,l) = F * u$.

Then, $M(n,l) = q(n) * p(n,l) * F(n,l) = \frac{n}{n_{max}} * \frac{n-l}{n} * F * u = \frac{(n-l)^2}{n_{max}} * F$. This is precisely the case we consider in the text, so the implications are the same.

B.4.2 Alternative Misreporting Case 2

In the next case, we assume that the probability of inspection is directly proportional to firm size, the probability of being caught conditional on inspection is proportional to the fraction of employees being misreported, and the fine if caught is fixed (i.e. independent of the degree of misreporting): $q(n) = \frac{n}{n_{max}}$; $p(n,l) = \frac{u}{n}$; $F(n,l) = F$.

Then, $M(n,l) = q(n) * p(n,l) * F(n,l) = \frac{n}{n_{max}} * \frac{n-l}{n} * F = \frac{(n-l)}{n_{max}} * F$

Under this functional form for the expected cost of misreporting, the firm's problem is as follows:

$$\pi(\alpha) = \max_{n,l} \alpha n^{\theta} - wn - \tau wl * \mathbb{1}\{l > 9\} - F * \frac{(n-l)}{n_{max}}$$

For large firms (i.e. those that would report having more than 9 workers), the marginal cost of increasing l is τw, while the marginal benefit is $\frac{F}{n_{max}}$. There are thus only two possible cases. Either $\tau w > \frac{F}{n_{max}}$ or $\tau w > \frac{F}{n_{max}}$.[46] If $\tau w > \frac{F}{n_{max}}$, then the marginal cost of increasing l is always too high, so that all large firms will thus choose $l = 9$ and appear to be bunched up at that size. Clearly, this is not consistent with the observed firm size distribution. On the other hand, if $\tau w < \frac{F}{n_{max}}$, then the marginal benefit of increasing l always exceeds the

[45] In fact it is more intuitive to assume that the fine is proportional to the level of misreporting, but this is not certain from the text of the laws.

[46] There is of course a third possibility, $\tau w = \frac{F}{n_{max}}$, in which firms would be completely indifferent regarding their choice of l, but this is a knife-edge case.

marginal cost so that large firms will find it optimal to set $l = n$. In other words, there will be no misreporting among large firms, so our estimate of the regulatory costs based on the observed firm size distribution will be unbiased (although it is worth noting that in this case our estimate would actually be determined by F, not τ).

B.4.3 Alternative Misreporting Case 3

For case 3, we again require that the probability of inspection be directly proportional to firm size, and that the fine if caught is proportional to the level of misreporting, but we now allow the probability of being caught conditional on inspection to be directly proportional to the *level* of employees being misreported (u): $q(n) = \frac{n}{n_{max}}$; $p(n,l) = \frac{u}{n_{max}}$; $F(n,l) = F * u$.

Then, $M(n,l) = q(n) * p(n,l) * F(n,l) = \frac{n}{n_{max}} * \frac{n-l}{n_{max}} * F * u = \frac{n(n-l)^2}{n_{max}^2} * F$, and the firm's problem is as follows:

$$\pi(\alpha) = \max_{n,l} \alpha n^{\theta} - wn - \tau wl * \mathbb{1}\{l > 9\} - \frac{F}{n_{max}^2} * n(n-l)^2$$

In this case it is helpful to rewrite the problem in terms of u instead of l:

$$\pi(\alpha) = \max_{n,u} \alpha n^{\theta} - wn - \tau w(n-u) * \mathbb{1}\{n-u > 9\} - \frac{F}{n_{max}^2} * n * u^2$$

From the first order condition on u, we get that $u = \frac{\tau w}{2F'} * \frac{1}{n}$, where $F' = \frac{F}{n_{max}^2}$. Thus, the level of misreporting, u, is a decreasing function of firm size, n. Indeed, $\underset{n\to\infty}{u} = 0$. The first order condition on n yields the following expression for n: $n = \left[\frac{w(1+\tau)+F'u^2}{\theta}\right]^{\frac{1}{\theta-1}} \alpha^{\frac{1}{1-\theta}}$. Using the previous result this expression simplifies to the following for large firms: $n = \left[\frac{w(1+\tau)}{\theta}\right]^{\frac{1}{\theta-1}} \alpha^{\frac{1}{1-\theta}}$. This is exactly the same as the solution for n in the absence of misreporting (see Appendix Subsection B.1). Thus, although the analytical solution for this case is not very tractable, the implications are the following: First, the log distribution of true firm size, $log\chi(n)$, will be downshifted at large firm sizes by exactly the same amount as in the case without misreporting. Second, since the level of misreporting is a decreasing function of firm size, with $\underset{n\to\infty}{u} = 0$, the reported distribution will converge to the true distribution at large firm sizes. Therefore, if our estimate of τ is based primarily on large firm sizes, it will be unbiased, as in the primary specification reported in the main text.

B.4.4 Alternative Misreporting Case 4

In case 4, the probability of inspection is again directly proportional to firm size, the probability of being caught conditional on inspection is proportional to the *level* of employees being misreported as in Case 3, but now the fine if caught is fixed: $q(n) = \frac{n}{n_{max}}$; $p(n,l) = \frac{u}{n_{max}}$; $F(n,l) = F$.

Then, $M(n,l) = q(n) * p(n,l) * F(n,l) = \frac{n}{n_{max}} * \frac{n-l}{n_{max}} * F = \frac{n(n-l)}{n_{max}^2} * F$, and the firm's problem becomes:

$$\pi(\alpha) = \max_{n,l} \alpha n^{\theta} - wn - \tau wl * \mathbb{1}\{l > 9\} - \frac{F}{n_{max}^2} * n(n-l)$$

From the above it can be seen that for large firms (again, meaning those that would report having more than 9 workers) the marginal cost of increasing l is τw, while the marginal benefit is $\frac{F}{n_{max}^2} * n$. While the marginal cost is fixed, the marginal benefit of increasing l is increasing in n. Thus, for large enough firms, it will become less and less attractive to misreport their true number of employees. Indeed, *all* firms larger than a certain size will find it optimal to report $l = n$, so there will be no misreporting in the right tail, and our estimate of τ will be unbiased - as in the previous cases.

B.4.5 The problematic case

In the cases considered above, we see that, under a variety of potential specific functional forms for misreporting, our estimate of τ remains unbiased. However, it is possible to construct a case that is consistent with the data and still leads to a biased estimate. Below we will describe this case and explain its implications for our estimation strategy.

Suppose now that the probability of inspection is actually independent of firm size. For the sake of specificity, we fix $q(n) = 1$. Further assume that the probability of being caught conditional on inspection is proportional to the fraction of employees being misreported, and that the fine if caught is proportional to the level of misreporting: $p(n,l) = \frac{u}{n}$; $F(n,l) = F * u$.

Then, $M(n,l) = q(n) * p(n,l) * F(n,l) = 1 * \frac{n-l}{n} * F(n-l) = \frac{(n-l)^2}{n} * F$.[47] Under this functional form for the expected cost of misreporting, the firm's problem is as follows:

$$\pi(\alpha) = \max_{n,l} \alpha n^{\theta} - wn - \tau wl * \mathbb{1}\{l > 9\} - F * \frac{(n-l)^2}{n}.$$

From the first order condition on l for large firms, we are able to see that $l = f_1(\tau) * n$, where $f_1(\tau) = 1 - \frac{\tau w}{2F}$. Thus, as long as $0 \leq \tau \leq 2F$, then $0 \leq f_1(\tau) \leq 1$ and l will be a *constant fraction* of n. From the first order condition on n (again, for large firms), we can write $n = \left[\frac{w}{\theta} + f_2(\tau)\right]^{\frac{1}{\theta-1}} \alpha^{\frac{1}{1-\theta}}$, where $f_2(\tau) = \frac{\tau w}{\theta}(1 - \frac{\tau w}{4F})$ and $f_2(\tau) > 0$ as long as $0 \leq \tau \leq 2F$. This is reminiscent of the expression for n in the case without misreporting (it would be identical if $f_2(\tau) = \frac{\tau w}{\theta}$), but it is different, which will imply that the distribution will be downshifted by a different (and lesser) degree.

Using these expressions, the distribution of productivity, and the standard change of variables formula, one can show that the log distribution of true firm size at large sizes is given by: $log\chi(n) = logA_1 - \frac{\beta-1}{1-\theta} log\left[\frac{w}{\theta} + f_2(\tau)\right] - \beta log(n)$, where $A_1 \equiv c_{\alpha}(1-\theta)$ and $f_1(\tau)$, $f_2(\tau)$ are defined as above. The log distribution of reported firm size (at large sizes) is given by $log\psi(l) = logA_1 - \frac{\beta-1}{1-\theta} log\left[\frac{w}{\theta} + f_2(\tau)\right] + (\beta - 1) log f_1(\tau) - \beta log(l)$. Comparing these with each other as well as the log distribution at small firm sizes (i.e. $n, l < 10$) and the log distribution in the case without misreporting, one can reach the following conclusions about this case.

[47] We have specified particular funcitonal forms for $q(n)$, $p(n,l)$ and $F(n,l)$, but any combination of functional forms that results in the condition that $M(,n,l) \propto \frac{(n-l)^2}{n}$ will have the same implications.

First, the log distribution of true firm size, $log\chi(n)$, is downshifted at large firm sizes - but by less than in the case without misreporting. In other words, being able to misreport lessens the actual cost of the regulations, even for large firms. Second, firms misreport their true employment by a constant fraction, $(1 - \frac{\tau w}{2F})$, which leads to a downshift in the log distribution of reported firm size *over and above* the downshift in the log distribution of true firm size. This would cause our estimate of τ to be biased upwards. In short, what causes the bias in the case above is that large firms misreport their true employment by a constant *fraction*. Any functional form of misreporting costs that leads firms to behave in this way will have the same implications as above.

In closing we note that this case has a singular quality to it. Since the marginal cost of increasing l is constant ($\frac{\partial \pi(\alpha,n,l)}{\partial l} = \tau w$), the marginal benefit of increasing l, $\frac{\partial \pi(\alpha,n,l)}{\partial l}$, must be increasing with n in just such a way as to ensure that the optimal choice of l increases as a constant fraction of n. If the marginal benefit of increasing l, $\frac{\partial \pi(\alpha,n,l)}{\partial l}$, increases with n "too quickly", then larger firms will find it optimal to report an increasing (not constant) fraction of their true employment, leading to Proposition 1's convergence result. If $\frac{\partial \pi(\alpha,n,l)}{\partial l}$ increases with n "too slowly", larger firms will find that the marginal cost of increasing l outstrips the marginal beneft at ever smaller fractions, which will cause them to report a smaller and smaller fraction of their true employment. In this sense, the problematic case requires a very particular functional form for misreporting to obtain.

B.5 Misreporting by Enumerators

In the text we referred to a second potential source of misreporting: not only might firms lie to enumerators about their size, enumerators themselves might lie when recording the figures reported to them by firms. One reason this might happen is that Economic Census enumerators were required to fill out an extra form containing the address of any establishment that reported 10 or more workers. It is therefore conceivable that enumerators might have found it preferable to under-report the number of workers for establishments with 10 or more workers in order to avoid the extra burden of filling in the "Address Slip". To show that this other source of potential misreporting is unlikely to bias our results, we consider a very simple model of enumerator misreporting. The model demonstrates that, since the cost of filling in the address slip is a fixed cost, it is not likely to lead to a "downshift" in the distribution, which means it is therefore unlikely to bias our estimate of τ.

The model begins with firms facing the same problem specified in Equation 7, with the same resulting distribution of the true firm size, $\chi(n)$. Then, all firms are matched with an enumerator, who must decide how to report the size of the firm they meet. In general, the reported size, l, may or may not be equal to the true firm size, n. If an enumerator reports a size $l > 9$, they must face the burden of filling out an address slip, at cost $C > 0$. If they report $l \leq 9$, they pay no such cost. Importantly, the cost C is constant and does not depend on the size of the firm.[48] Furthermore, enumerators face expected costs of misreporting, $M(u)$, where $u \equiv n - l$.[49] $M(u)$ captures both the probability of being caught as well as the penalty faced if caught. The only assumptions we make on $M(u)$ are that it is

[48]This reflects the fact that it is no more costly (in terms of time or hassle) to fill in an address slip for a firm of size 20 than a firm of size 200.

[49]As before we will only consider nonnegative values of u, since there will be no incentive to over-report

strictly increasing in u and that $M(0) = 0$.[50] Then the utility maximization problem faced by an enumerator matched with a firm of size n is:

$$U(n) = \max_{u} \; -C\mathbb{1}\{n - u > 9\} - M(u)$$

For enumerators matched with firms of size $n < 9$, the optimal decision is to choose $l = n$, or $u = 0$, because there is no need to misreport. Then, their utility is maximized at 0. Enumerators matched with firms larger than 9 must decide whether to report the size truthfully and bear the address slip cost or lie in order to avoid the cost of filling out the address slip. Since $M(u)$ is increasing in u, misreporting costs will be lower than C for low values of u and higher than C for high enough values.[51] Therefore, enumerators matched with firms of an intermediate size (i.e. $n \in [9, \bar{n}]$) will find it optimal to lie by choosing $l = 9$ or $u = n - 9$, in order to avoid the fixed cost of filling out the address slip. $\bar{n}$ is defined such that $C = M(\bar{n} - 9)$, so that an enumerator matched with a firm of size $\bar{n}$ would be indifferent between misreporting ($l = 9$) and reporting truthfully ($l = \bar{n}$). For enumerators matched with firms of size $n > \bar{n}$, the cost of misreporting exceeds the cost of filling out the address slip, so they are better off bearing the address slip cost and reporting truthfully (i.e. setting $l = n$ or $u = 0$).

To summarize, we obtain the following relationship between reported and true employment:

$$l(n) = \begin{cases} n & if \quad n \in [n_{min}, 9) \\ 9 & if \quad n \in [9, \bar{n}] \\ n & if \quad n > \bar{n} \end{cases}$$

Given this mapping between reported and actual employment, the reported firm size density is given by:

$$\psi(l) = \begin{cases} \chi(n) & if \quad l \in [n_{\min}, 9) \\ \int_9^{\bar{n}} \chi(n)dn = \delta_e & if \quad l = 9 \\ 0 & if \quad l \in (9, \bar{n}] \\ \chi(n) & if \quad l \geq \bar{n} \end{cases}$$

In other words, the enumerator misreporting will cause the reported distribution to exhibit "bunching before the threshold as well as a "valley" after the threshold, even if neither phenomena exists in the true distribution. However, the reported distribution coincides with the true distribution before the threshold and for values far above the threshold. In other words, enumerator misreporting does not cause a downshift in the reported distribution in excess of the downshift in the true distribution and hence does not bias our estimate of τ.

firm size.

[50]The latter assumption is made only for simplicity of exposition.

[51]The other possibility is that $M(u) < C \;\; \forall\, u$, but this case is not very interesting and is clearly not borne out by the data (it would suggest that enumerators should *always* misreport firm size to be 9).

B.6 The Firm Size Distribution With Misreporting and Inattention

As we noted in Section 4.2, there is a significant discrepancy between the firm size distribution predicted by our model and that observed in the data. In the model, there should be many firms "bunched up" with sizes just below 10, and there should be a "valley" (i.e. *no* firms) reporting sizes just above 10. Yet, Figure 1 clearly shows that there are firms reporting sizes just above 10. In this section we account for this discrepancy.

Our explanation, related to those considered by Kleven and Waseem (2013), is that small firms tend to be inattentive to the regulatory threshold while large firms tend to be attentive. Attentive firms are aware of the regulations as well as the expected costs and benefits of misreporting, while inattentive firms are simply not aware of the relevant regulations - and hence do not bother to misreport their firm size. This explanation is based on the observation - itself the result of interviews conducted with a number of firms - that small firms are commonly unaware of regulatory details (including the size thresholds themselves) while large firms commonly expend considerable time and money in ensuring that they have correct information regarding the regulations (often by hiring accounting, legal and human resource departments within the firm). This difference is likely due to the fact that large firms cannot fly under the radar as easily and are much more likely to be audited and inspected by labor regulators (Almeida and Ronconi (2016)). It may also be related to the well documented fact that small firms tend to be managed by those with lower levels of human capital (La Porta and Shleifer (2014)) and are hence less likely to be knowledgable about the law.

The assumption that larger firms are more likely to be attentive can also be motivated theoretically. Imagine that managers must pay a fixed cost (which varies idiosyncratically across firms) in order to learn regulatory details - including the location of the thresholds. In practice this would involve hiring an accountant or attorney who is knowledgable about the text of labor regulations. Under the plausible assumption that the distribution of fixed costs does not vary with firm size, the fact that the benefits of adjusting employment in response to the threshold rise with size implies that all large firms will adjust while only some small firms will.

In what follows we will amend the theoretical model from Section 4.1 to include the assumption that a large fraction of firms are inattentive, but that this fraction goes to 0 at large firm sizes. More formally:

Assumption 2. $\exists$ *a given proportion* $p(\alpha)$ *of entrepreneurs that are attentive (i.e. they are aware of the size-based regulations, as well as the expected costs and benefits of misreporting) while* $(1 - p(\alpha))$ *of entrepreneurs are inattentive (i.e. they ignore or are unaware of the regulations* and - *therefore - they report their size truthfully). Furthermore,* $p'(\alpha) > 0$ *and* $\lim\limits_{\alpha \to \infty} p(\alpha) = 1$.

Under this assumption, one can show that the theoretical firm size distribution will match the main features of the empirical density while still yielding minimally biased estimates of τ under our estimation procedure. To demonstrate this result, we proceed below by solving the problem that faces each type of firm in turn, beginning with the attentive firms. We will derive the firm size distribution for each type, and then put them together to determine the firm size distribution for the entire population of firms.

B.6.1 The Attentive Firms' Problem

The problem of the attentive firm is exactly as modeled in the paper, and thus the associated firm size distribution for attentive firms is also the same (Section 4.1). For clarity, we repeat the main steps of the derivation with functional form assumptions. Let the expected costs of misreporting be given by $M(n,l) = F * \frac{(n-l)^2}{n_{max}}$. Then, the firm's problem is the following:

$$\pi(\alpha) = \max_{n,l} \ \alpha n^{\theta} - wn - \tau wl * \mathbb{1}\{l > 9\} - F * \frac{(n-l)^2}{n_{max}}.$$

By combining solutions for n and l from the first order conditions with the distribution of managerial ability ($\phi(\alpha) = c_{\alpha}\alpha^{-\beta_{\alpha}}$), one can determine the distributions of true and reported firm size, $\chi(n)$ and $\psi(l)$:

$$\chi(n) = \begin{cases} c_{\alpha}(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}} n^{-\beta} & if \quad n \in [n_{\min}, N) \\ \left|\frac{d\alpha_2^*(n)}{dn}\right| \phi(\alpha_2^*(n)) & if \quad n \in [N, n_2^*(\alpha_2)) \\ 0 & if \quad n \in [n_2^*(\alpha_2), n_3^*(\alpha_2)) \\ c_{\alpha}(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}}(1+\tau)^{-\frac{\beta-1}{1-\theta}} n^{-\beta} & if \quad n \geq n_3^*(\alpha_2) \end{cases}$$

$$\psi(l) = \begin{cases} c_{\alpha}(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}} l^{-\beta} & if \quad l \in [n_{\min}, N) \\ \int_{\alpha_1}^{\alpha_2} \phi(\alpha) d\alpha = \delta_l & if \quad l = N \\ 0 & if \quad l \in (N, l_3^*(\alpha_2)) \\ c_{\alpha}(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}}(1+\tau)^{-\frac{\beta-1}{1-\theta}}[l + \frac{n_{max}}{2F} w\tau]^{-\beta} & if \quad l \geq l_3^*(\alpha_2) \end{cases}$$

where $\beta \equiv \theta + \beta_{\alpha} - \theta\beta_{\alpha}$. Then, the logged distributions of true and reported firm size, $log\chi(n)$ and $log\psi(l)$, are given by:

$$log\chi(n) = \begin{cases} log(A) - \beta log(n) & if \quad n \in [n_{\min}, 9) \\ log[\xi(n)] & if \quad n \in [9, n_m(\alpha_2)] \\ - & if \quad n \in (n_m(\alpha_2), n_t(\alpha_2)) \\ log(A) - \frac{\beta-1}{1-\theta} log(1+\tau) - \beta log(n) & if \quad n \geq n_t(\alpha_2) \end{cases}$$

$$log\psi(l) = \begin{cases} log(A) - \beta log(l) & if \quad l \in [l_{\min}, 9) \\ log(\delta_l) & if \quad l = 9 \\ - & if \quad n \in (9, l_t(\alpha_2)) \\ log(A) - \frac{\beta-1}{1-\theta} log(1+\tau) - \beta log(l + \frac{n_{max}}{2F} w\tau) & if \quad l \geq l_t(\alpha_2) \end{cases}$$

where $A \equiv c_{\alpha}(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}}$ and $\xi(n) \equiv \left|\frac{d\alpha_2^*(n)}{dn}\right| \phi(\alpha_2^*(n))$. What these expressions show is that, for firms above the threshold, misreporting will be by a constant/fixed amount, and will thus be increasingly inconsequential in affecting the firm size distribution at larger and larger firm sizes. Formally, the difference between the reported firm size distribution, $\psi(x)$, and the true firm size distribution, $\chi(x)$, converges to 0 as $x \to \infty$.

B.6.2 The Inattentive Firms' Problem

Now let us consider the inattentive firm's problem. This problem is much simpler, because we model the inattentive firms as being entirely unaware of and unresponsive to the regulatory thresholds. In particular, because the firm is not attentive to the regulations, the issue of misreporting does not come up, so the reported density will be the same as the true density. Thus, the problem for such firms is as follows:

$$\pi(\alpha) = \max_{n} \alpha n^{\theta} - wn$$

The solution to this problem is straightforward: $n = \left[\frac{\theta}{w}\right]^{\frac{1}{1-\theta}} \alpha^{\frac{1}{1-\theta}}$, $\forall \alpha \in [\underline{\alpha}, \infty)$. Together with the distribution of managerial ability ($\phi(\alpha) = c_{\alpha}\alpha^{-\beta_{\alpha}}$), this implies that the firm size distribution of inattentive firms is given by the following expression:

$$\Gamma(n) = c_{\alpha}(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}} n^{-\beta}, \quad \forall n \in [n_{min}, \infty)$$

Because there is no misreporting, the reported density is equivalent to the true density: $\Gamma(l) = \Gamma(n), \quad \forall l = n$. Thus, the logged version of both the true and reported densities is given by:

$log\Gamma(n) = log(c_{\alpha}(1-\theta)) - \frac{\beta-1}{1-\theta} log(\frac{w}{\theta}) - \beta log(n), \quad \forall n \in [n_{min}, \infty)$, and

$log\Gamma(l) = log(c_{\alpha}(1-\theta)) - \frac{\beta-1}{1-\theta} log(\frac{w}{\theta}) - \beta log(l), \quad \forall n \in [n_{min}, \infty)$

In other words, the reported firm size density is a simple - unbroken - power law at all firm sizes.

B.6.3 Combining the Densities for Attentive and Inattentive Firms

Now that we have derived the true and reported distributions for attentive and inattentive firms, we can combine them to generate the *complete reported* distribution for all firms. We rewrite the density (true and misreported) for attentive firms after collecting and renaming terms:

$$\chi(n) = \begin{cases} An^{-\beta} & if \quad n \in [n_{\min}, N) \\ \xi(n) & if \quad n \in [N, n_2^*(\alpha_2)) \\ 0 & if \quad n \in [n_2^*(\alpha_2), n_3^*(\alpha_2)) \\ A(1+\tau)^{-\frac{\beta-1}{1-\theta}} n^{-\beta} & if \quad n \geq n_3^*(\alpha_2) \end{cases}$$

$$\psi(l) = \begin{cases} Al^{-\beta} & if \quad l \in [n_{\min}, N) \\ \delta_l & if \quad l = N \\ 0 & if \quad l \in (N, l_3^*(\alpha_2)) \\ A(1+\tau)^{-\frac{\beta-1}{1-\theta}} [l + \frac{n_{max}}{2F} w\tau]^{-\beta} & if \quad l \geq l_3^*(\alpha_2) \end{cases}$$

where $A \equiv c_{\alpha}(1-\theta)(\frac{\theta}{w})^{\frac{\beta-1}{1-\theta}}$ and $\xi(n) \equiv \left|\frac{d\alpha_2^*(n)}{dn}\right| \phi(\alpha_2^*(n))$. The simplified density for inattentive firms is just:

$$\Gamma(n) = An^{-\beta}, \quad \forall n \in [n_{min}, \infty)$$

Given that some fraction ($p(\alpha)$) of the firms are attentive while the rest ($1 - p(\alpha)$) are inattentive, the complete distributions for true and reported firm size are each convex combinations of the true and reported distributions (respectively) for attentive and inattentive firms, where the weight of the distribution at a given point is given by the proportion of entrepreneurs that are either attentive or inattentive. In particular, the true distribution *for all firms* is given by: $\Omega(n(\alpha)) = p(\alpha)\chi(n(\alpha)) + (1 - p(\alpha))\Gamma(n(\alpha))$, where n is written explicitly as a function of α. Substituting in our expressions for $\chi(n)$ and $\Gamma(n)$ from above, we get:

$$\Omega(n(\alpha)) = \begin{cases} An(\alpha)^{-\beta} & if \quad n(\alpha) \in [n_{\min}, N) \\ p(\alpha)\xi(n) + (1 - p(\alpha))An(\alpha)^{-\beta} & if \quad n(\alpha) \in [N, n_2^*(\alpha_2)) \\ (1 - p(\alpha))An(\alpha)^{-\beta} & if \quad n(\alpha) \in [n_2^*(\alpha_2), n_3^*(\alpha_2)) \\ A[p(\alpha)(1+\tau)^{-\frac{\beta-1}{1-\theta}} + (1 - p(\alpha))]n(\alpha)^{-\beta} & if \quad n(\alpha) \geq n_3^*(\alpha_2) \end{cases}$$

We can do the same to generate the *reported* distribution for all firms, $\Lambda(l(\alpha)) = p(\alpha)\psi(l(\alpha)) + (1 - p(\alpha))\Gamma(l(\alpha))$, or:

$$\Lambda(l(\alpha)) = \begin{cases} Al(\alpha)^{-\beta} & if \quad l(\alpha) \in [n_{\min}, N) \\ p\delta_l & if \quad l(\alpha) = N \\ (1 - p(\alpha))Al(\alpha)^{-\beta} & if \quad l(\alpha) \in (N, l_3^*(\alpha_2)) \\ p(\alpha)A(1+\tau)^{-\frac{\beta-1}{1-\theta}}[l(\alpha) + \frac{n_{max}}{2F}w\tau]^{-\beta} + (1 - p(\alpha))Al(\alpha)^{-\beta} & if \quad l(\alpha) \geq l_3^*(\alpha_2) \end{cases} \tag{8}$$

These expressions generate some different conclusions from the model without inattentive firms. One conclusion that is the same is that τ will again lead to a downshift in the true logged distribution at large firm sizes. However, the term which describes the downshift ($p(\alpha)(1+\tau)^{-\frac{\beta-1}{1-\theta}} + (1 - p(\alpha))$) is now a combination not only of τ, but also of the proportion of attentive ($p(\alpha)$) and inattentive firms ($1 - p(\alpha)$). However, an implication of Assumption 2 is that $\exists\ \alpha_L$ *such that* $p(\alpha) \approx 1 \quad \forall\ \alpha > \alpha_L$. Then, the tails of the true and reported distributions are well approximated by the following expressions:

$$\Omega(n(\alpha)) = A[(1+\tau)^{-\frac{\beta-1}{1-\theta}})]n(\alpha)^{-\beta}, \quad \forall\ n(\alpha) > n(\alpha_L)$$

$$\Lambda(l(\alpha)) = A(1+\tau)^{-\frac{\beta-1}{1-\theta}}[l(\alpha) + \frac{n_{max}}{2F}w\tau]^{-\beta}, \quad \forall\ l(\alpha) > l(\alpha_L)$$

However, these are just the same as in the version of the model without inattention. Moreover, the expressions for the reported and true firm size distributions *at small (i.e.* $n <$

10*) firm sizes* also coincide with those of the model without inattention (because attentive and inattentive small firms behave in the same way). Thus, under these assumptions, our current procedure for estimating τ using the firm size distribution at large sizes (i.e. leaving out the middle of the distribution which is likely to impacted by both misreporting and now also inattention) allows us to estimate τ with little bias.

One final point is worth making: if $p \approx 0$ at small firm sizes, that would explain why there is not more bunching nor a very significant "valley" in the firm size distribution around 10. Thus, under the assumption that small firms are mostly inattentive, while large firms are mostly attentive, we can make sense of the relatively small observed distortions in the distribution around 10, while still being able to estimate the cost of the regulations on large firms.

C Full Results by State, Industry, Ownership Type and Social Group for the 10-worker Threshold

Table 7: Estimates of τ by State

State	*Tau*	*Standard Error*
Bihar	.693	.302
Karnataka	.52	.156
Uttar Pradesh	.502	.254
Delhi	.427	.213
Tamil Nadu	.397	.154
Jharkhand	.388	.194
Madhya Pradesh	.379	.203
Maharashtra	.332	.107
Assam	.322	.345
Rajasthan	.32	.174
Orissa	.283	.139
Gujarat	.165	.151
West Bengal	.151	.071
Kerala	.138	.196
Punjab	.096	.158
Haryana	.007	.168
Andhra Pradesh	-.159	.053
Himachal Pradesh	-.165	.159

Note: This table presents estimates of regulatory costs faced by establishments that employ 10 or more workers, using the methodology described in Section 4. Standard errors generated using a clustered bootstrap procedure with 200 replications are presented in parentheses. Clustering is done at the 4 digit (NIC code) industry level, following GLV. Estimates are presented for the 18 largest states (by NSDP). Source: 2005 Economic Census of India.

Table 8: Estimates of τ by Industry

Industry	*Tau*	*Standard Error*
Wholesale and retail trade	.637	.094
Real estate, renting and business activities	.601	.158
Construction	.478	.549
Hotels and restaurants	.468	.222
Transport, storage and communications	.334	.209
Manufacturing	.268	.085
Other service activities	.264	.186
Health and social work	.076	.149
Mining and quarrying	-.042	.294
Financial intermediation	-.105	.074
Education	-.173	.15
Public administration and defence	-.311	.034
Electricity, gas and water supply	-.367	.145

Note: This table presents estimates of regulatory costs faced by establishments that employ 10 or more workers, using the methodology described in Section 4. Standard errors generated using a clustered bootstrap procedure with 200 replications are presented in parentheses. Clustering is done at the 4 digit (NIC code) industry level, following GLV. Estimates are presented for all major industry categories. Source: 2005 Economic Census of India.

Table 9: Estimates of τ by Ownership Type

Ownership Type	*Tau*	*Standard Error*
Unincorporated proprietary	.43	.059
Co-operative	-.007	.075
Non profit institution	-.04	.095
Unincorporated partnership	-.058	.053
Government and public sector undertaking	-.092	.128
Corporate financial	-.18	.055
Corporate non financial	-.197	.05

Note: This table presents estimates of regulatory costs faced by establishments that employ 10 or more workers, using the methodology described in Section 4. Standard errors generated using a clustered bootstrap procedure with 200 replications are presented in parentheses. Clustering is done at the 4 digit (NIC code) industry level, following GLV. Estimates are presented by ownership type of the establishment. Source: 2005 Economic Census of India.

D Robustness of τ estimates to choice of fixed effects

Table 10: Robustness of τ estimates to inclusion/exclusion of specific firm size fixed effects

Maximum size omitted	Include size 2 and 8	Omit size 8	Omit size 2	Omit size 2 and 8
18	0.390	0.361	0.386	0.349
	(0.077)	(0.081)	(0.089)	(0.081)
19	0.389	0.359	0.385	0.348
	(0.082)	(0.081)	(0.089)	(0.082)
20	0.388	0.358	0.384	0.347
	(0.082)	(0.075)	(0.086)	(0.084)
21	0.386	0.357	0.382	0.345
	(0.086)	(0.071)	(0.088)	(0.080)
22	0.385	0.356	0.381	0.344
	(0.083)	(0.080)	(0.088)	(0.075)
23	0.384	0.354	0.380	0.343
	(0.081)	(0.082)	(0.083)	(0.080)
24	0.382	0.352	0.378	0.341
	(0.077)	(0.079)	(0.082)	(0.088)
25	0.381	0.351	0.376	0.340
	(0.084)	(0.073)	(0.086)	(0.077)
26	0.379	0.350	0.375	0.338
	(0.077)	(0.076)	(0.084)	(0.075)
27	0.378	0.349	0.374	0.337
	(0.080)	(0.076)	(0.075)	(0.071)
28	0.377	0.348	0.373	0.336
	(0.076)	(0.070)	(0.079)	(0.076)
29	0.376	0.347	0.372	0.335
	(0.074)	(0.069)	(0.077)	(0.072)
30	0.375	0.346	0.371	0.334
	(0.077)	(0.070)	(0.074)	(0.077)
31	0.374	0.345	0.370	0.333
	(0.079)	(0.076)	(0.081)	(0.070)
32	0.373	0.344	0.369	0.332
	(0.079)	(0.074)	(0.073)	(0.074)
33	0.372	0.343	0.368	0.331
	(0.069)	(0.076)	(0.077)	(0.078)
34	0.371	0.342	0.367	0.331
	(0.072)	(0.070)	(0.075)	(0.074)
35	0.371	0.341	0.366	0.330
	(0.069)	(0.072)	(0.073)	(0.074)

Note: This table presents estimates of regulatory costs faced by establishments with 10 or more workers, using the methodology described in Section 4 with a bandwidth of 0.005. Standard errors generated using a clustered bootstrap procedure with 200 replications are presented in parentheses. Clustering is done at the 4 digit (NIC code) industry level, following Garicano et al. (2016). Estimates are presented for a variety of choices for the largest firm size to omit from estimation by including a fixed effect for that size. For each choice of the largest firm size to omit, we show robustness to the choice of omitting firms of size 2, 8, or both from estimation. Source: 2005 Economic Census of India.

E Further Results Related to Exploration of Mechanisms

In the main body of the paper we reported the results of regressing τ against state-level differences in the statutory, procedural and administrative aspects of the regulations, as well against state-level differences in measures of corruption. In this section we provide some robustness tests related to the former analyses (see Tables 12 and 13, as well as Figures 3 and 4). We also provide some further tests which report the results of regressing τ against other measures of the labor environment. In particular, Table 11 reports the results of τ regressed against per capita measures of strikes, man-days lost to strikes, lockouts, man-days lost to lockouts and the percentage of registered factories that have been inspected. One might imagine that strikes and lockouts capture relevant features of the regulatory and labor environment,[52] but we do not find them to be robustly correlated with τ. We do find, echoing the results of Tables 4 and 12, a robust correlation between τ and the percentage of registered factories inspected.

Table 11: Tau vs Strikes and Lockouts

	(1) tau	(2) tau	(3) tau	(4) tau	(5) tau
strikes per capita	-0.0584 (0.208)				
mandays lost due to strikes per capita		-0.00607 (0.0771)			
lockouts per capita			0.0367 (0.0789)		
mandays lost due to lockouts per capita				0.0289 (0.0750)	
percent of factories inspected					0.736** (0.290)
log of net state domestic product pc	-0.398 (0.286)	-0.387 (0.258)	-0.407 (0.264)	-0.407 (0.261)	0.547 (0.743)
Constant	4.085 (2.729)	3.946 (2.597)	4.132 (2.630)	4.138 (2.612)	-5.644 (7.681)
Observations	18	17	18	18	10

Note: This table tests for correlations between our estimated regulatory costs (tau) and other miscellaneous measures of the labor environment. Robust standard errors are reported in parentheses. Observations are weighted by the inverse variance of tau and include only the 18 largest Indian States, as measured by NSDP. Sources: Indian Labour Year Book (2005).

[52] For example, some industrial regulations explicitly undermine or support the rights of parties to engage in strikes or lockouts.

Table 12: Tau vs Other Measures of Regulations

	(1)	(2)	(3)	(4)	(5)	(6)
	tau	tau	tau	tau	tau	tau
Dougherty measure	-0.424*	-0.431*				
(all reforms)	(0.212)	(0.235)				
Dougherty measure			-0.549***	-0.652***		
(inspector reforms)			(0.170)	(0.150)		
Besley-Burgess					0.223	0.237
measure (regs)					(0.182)	(0.183)
log of net state		-0.463*		-0.527**		-0.518**
domestic product pc		(0.262)		(0.204)		(0.235)
share of privately		-0.0632		8.524		-12.80
owned establishments		(7.095)		(6.707)		(8.354)
Constant	0.203	4.945	0.288*	-1.923	0.00679	16.67*
	(0.222)	(6.800)	(0.149)	(6.529)	(0.292)	(7.798)
Observations	18	18	18	18	15	15

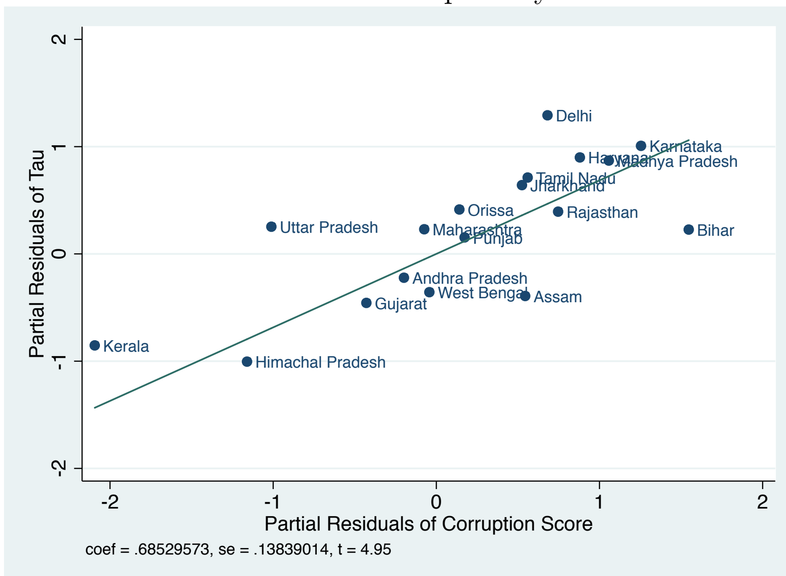
Note: This table tests for correlations between our estimated regulatory costs (tau) and other established measures of the regulatory environment from the previous literature. Robust standard errors are reported in parentheses. Observations are weighted by the inverse variance of tau and include only the 18 largest Indian States, as measured by NSDP. Sources: Dougherty(2009); Besley and Burgess(2004); RBI.

Table 13: Tau vs State Level Measures of Corruption

	(1)	(2)	(3)	(4)	(5)	(6)
	tau	tau	tau	tau	tau	tau
TI Corruption Score	0.812**	0.806**	0.685***			
	(0.290)	(0.342)	(0.129)			
electricity losses				0.925***	0.948**	0.575***
				(0.310)	(0.326)	(0.190)
log of net state		-0.0667	-0.213		-0.219	-0.350**
domestic product pc		(0.338)	(0.285)		(0.131)	(0.133)
share of privately		-2.323	6.335*		1.365	7.900
owned establishments		(4.493)	(3.376)		(4.940)	(5.153)
Dougherty measure			-0.594***			-0.500***
(inspection reforms)			(0.0932)			(0.116)
Electricity					0.109	
available (GWH)					(0.134)	
Constant	0.334	3.075	-2.935	0.476**	1.365	-2.956
	(0.213)	(4.682)	(3.836)	(0.169)	(4.634)	(4.728)
Observations	18	18	18	18	18	18
Measure of Corruption	TI	TI	TI	TDLs	TDLs	TDLs

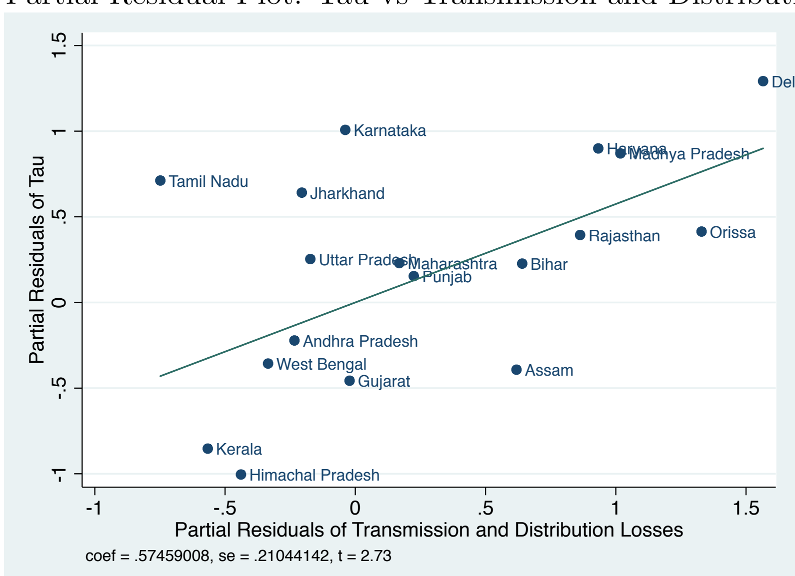
Note: This table reports the results of our estimated regulatory costs (tau) regressed against two different measures of corruption. Robust standard errors are reported in parentheses. Observations are weighted by the inverse variance of tau and include only the 18 largest Indian States, as measured by NSDP. Sources: Transparency International (2005); RBI; Dougherty(2009).

Figure 3: Partial Residual Plot: Tau vs Transparency International Corruption Score



Note: This figure is the graphical analogue of column 3 in Table 4. It depicts the relationship between the components of tau and the TI Corruption Score that are unexplained by the other covariates. Sources: Transparency International (2005); Dougherty(2009); RBI.

Figure 4: Partial Residual Plot: Tau vs Transmission and Distribution Losses



Note: This figure is the graphical analogue of column 6 in Table 4. It depicts the relationship between the components of tau and electricity transmission and distribution losses that are unexplained by the other covariates. Sources: Dougherty(2009); RBI.

E.1 τ and Corruption: State X Industry Analysis

In this portion of the Appendix, we explain our State X Industry analysis (described in Section 6.1) in more detail. The purpose of this analysis is to partially address concerns that the state-level correlations between τ and corruption lack exogenous variation and may be biased if our measures of corruption are correlated with omitted variables that also influence τ. To do so, we take advantage of State X Industry level heterogeneity as an additional source of variation. We use data from the World Bank's 2005 Firm Analysis and Competitiveness Survey of India (FACS) to create an industry level measure of the extent to which regulations are problematic, which we term "regulatory intensity". Specifically, Indian firms in the 2005 FACS were asked whether "regulations specific to [their] industry" were problematic for their "operation and growth". Averaging the firm-level responses by industry, we classify industries according to how likely businesses are to complain about *industry-specific* regulations. If regulations are especially costly due to corruption in their enforcement, then we would expect costs to be highest among those businesses in regulation-heavy industries *and* in states with high corruption. That is, we would expect the *interaction* between industry level "regulatory intensity" and state level corruption to be positive.

To test our hypothesis we generate our measures of τ at the State X Industry level[53] and regress those measures against interactions of state level corruption with industry level regulatory intensity. The results, shown in Table 14 with and without interaction terms, support our hypothesis. First, when excluding the interaction terms (columns 1 and 3), the main effects (state level corruption and industry level regulatory intensity) are significantly correlated with τ in the expected directions. When interaction terms are included (columns 2 and 4), their coefficients are also large and significant, suggesting that the presence of industry specific regulations is most costly when firms are located in a corrupt environment.

F Possible Consequences of τ

In Section 6 we argued that our estimated costs (τ) are mostly due, not only to the substance of the regulations themselves, but also to high levels of corruption. In this subsection we will indicate possible consequences of high values of τ. In what follows we use two distinct measures of τ: one which is created using all the establishments in a state, regardless of economic sector (τ) and another which is created using only the establishments engaged in manufacturing (τ_{manuf}).

Table 15 displays the results of employment growth in the manufacturing sector between 2010 and 2005 at the state level regressed against our two measures of labor market distortions (τ and τ_{manuf}) as well alternative measures (Dougherty and BB). For each of the four measures, we observe its performance as a predictor of future employment growth in *registered* manufacturing as well as *unregistered* manufacturing. Interestingly, in the regressions of employment growth in registered manufacturing against τ_{manuf}, the coefficient on τ_{manuf} is negative and significant at the 5% level, while the coefficient for employment growth in *unregistered* manufacturing is positive and significant. This result makes sense: we should expect higher costs to negatively affect the sectors to which the costs apply - in this case the

[53]Industries here are categorized according to their groupings in the World Bank Enterprise Surveys, which distinguishes 23 distinct industry categories. Examples include "auto components", "leather and leather products", and "food processing". We only generated τ for state X industry cells with a sufficient number of observations (in particular, for those with at least 40 observations in the size distribution), and were thus left with only 190 observations out of a possible 414 (23*18).

Table 14: Tau vs State Level Corruption Interacted with Industry Level "Regulatory Intensity"

	(1)	(2)	(3)	(4)
	tau	tau	tau	tau
log of net state domestic product pc	-0.348***	-0.312***	-0.186***	-0.142***
	(0.0298)	(0.0311)	(0.0288)	(0.0298)
TI Corruption Score	0.0592***	0.143***		
	(0.0140)	(0.0291)		
electricity TDLs			0.226***	0.224***
			(0.0215)	(0.0206)
Regulatory Intensity	0.259***	0.314***	0.0937**	0.103***
	(0.0288)	(0.0328)	(0.0284)	(0.0274)
TI Corruption Score X Regulatory Intensity		0.160**		
		(0.0492)		
electricity TDLs X Regulatory Intensity				0.176***
				(0.0435)
Constant	-0.424***	-0.404***	-0.534***	-0.576***
	(0.00849)	(0.0103)	(0.0121)	(0.0156)
Observations	190	190	190	190

Note: This table reports the results of our estimated regulatory costs (tau) regressed against against state level corruption, industry level regulatory intensity, and their interaction. Robust standard errors are reported in parentheses. Observations are now at the state X industry level but are still weighted by the inverse variance of tau and include only the 18 largest Indian States, as measured by NSDP. Sources: Transparency International (2005); RBI; World Bank Firm Analysis and Competitiveness Survey of India (2005).

registered sector, since that is under the ambit of labor regulations while the unregistered sector is not.[54] If these correlations reflect a causal chain, it would mean that high levels of regulatory costs and corruption (as measured by τ) are pushing employment from the registered to the unregistered sector.

Also included in Table 15 are the results of employment growth in manufacturing regressed against the BB and Dougherty measures. Neither regressor has a coefficient that is statistically significant or of a meaningful magnitude.[55]

[54] It also makes sense that the coefficients on τ are insignificant, since τ is measured across all sectors and will be less pertinent to manufacturing performance than τ_{manuf}.

[55] One might argue that it is not quite fair to regress growth between 2010 and 2005 on a regressor that uses data from 1997, as is the case for the BB measure. However, we have duplicated these results using growth from 1997 to 2002 and the results are the same. Furthermore, the Besley Burgess measure from Aghion et al. (2008) should be the same in 2005 due to the lack of state level reforms between 1997 and 2005.

Table 15: Manufacturing Employment Growth (2005 - 2010) vs Tau and Other Measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	reg manuf	unreg manuf	reg manuf	unreg manuf	reg manuf	unreg manuf	reg manuf	unreg manuf
tau	-0.0240	0.00197						
	(0.0176)	(0.0233)						
tau (manuf)			-0.0471**	0.0623**				
			(0.0217)	(0.0256)				
Besley-Burgess measure (regs)					-0.00525	0.00979		
					(0.00731)	(0.0142)		
Dougherty measure (all reforms)							0.0226	-0.0143
							(0.0130)	(0.0159)
log of net state domestic product pc	0.00312	0.0189	0.0107	0.0192	0.00413	0.0140	0.0212	0.0136
	(0.0178)	(0.0214)	(0.0145)	(0.0161)	(0.00863)	(0.0168)	(0.0154)	(0.0195)
share of employment in manufacturing	-0.393	0.00558	-0.708**	0.435	0.0194	-0.559	-0.515*	0.0525
	(0.258)	(0.329)	(0.258)	(0.325)	(0.186)	(0.362)	(0.245)	(0.323)
Constant	0.0969	-0.182	0.0372	-0.229	0.0209	-0.0675	-0.0861	-0.131
	(0.173)	(0.209)	(0.139)	(0.152)	(0.0825)	(0.160)	(0.147)	(0.182)
Observations	18	17	18	17	15	15	18	17

Note: This table reports the results of employment growth in the registered and unregistered manufacturing sectors against several measures of the regulatory environment, including our own estimated regulatory costs (tau). Robust standard errors are reported in parentheses. Observations are unweighted and include only the 18 largest Indian States, as measured by NSDP. Sources: Besley and Burgess (2004); Dougherty(2009); RBI.

G Collusionary vs Extortionary Corruption (i.e. Harassment Bribery)

The results of Section 6.1 documented a robust *positive* correlation between effective regulatory costs (τ) and corruption/poor governance. To explain this phenomenon, we distinguish between two types of corruption that could take place between corrupt inspectors and firms: collusion and extortion. In order to explain our conception of the difference between these two types of corruption, it is necessary to make a further distinction: between following the letter of the law and the spirit of the law. An honest inspector will require firms to follow the spirit of the law (that is, a reasonable interpretation of the law). A corrupt inspector will threaten firms with the maximum penalty possible if they have not followed the letter of the law (which may be a literal but "unreasonable" interpretation of the law), and then bargain over the surplus gained from not following through with the penalty to obtain a bribe.

If the costs of following the letter of the law - or the penalties that inspectors can threaten for *not* following the letter of the law - are similar or less than the cost of following the spirit of the law, then firms will bear lower costs when dealing with corrupt inspectors than with honest inspectors. If the cost of following the letter of the law - or the maximum penalty that can be threatened under the letter of the law - is much greater than the cost of following the spirit of the law, firms will bear higher costs when dealing with corrupt inspectors. We refer to the former case as "collusionary corruption" and the latter case as "extortionary corruption". Under an extortionary regime (that is, if the cost of following the letter of the law is relatively high), then greater levels of corruption (i.e. more corrupt inspectors) will be associated with greater effective regulatory costs (τ). In a collusionary regime, the opposite association should be observed. If we are willing to interpret the positive correlations in Section 6.1 as indicative of a *causal* relationship, then the results suggest we are under an extortionary regime.

This may not be difficult to believe, as the regulations which apply to firms with more than 10 workers appear so complex as to make it almost impossible (or prohibitively costly) for any firm to be fully in compliance with all aspects of the law as written. As mentioned in Section 2, many of the laws have components that are antiquated, arbitrary, contradictory and confusing. That the laws may be impossible to fully comply with is suggested by some of the anecdotes we provide in Appendix H and the descriptions of the excessive specificity of the regulations we provided in Section 2. In this case, a dishonest inspector can, at any time, choose to subject a firm under his jurisdiction to a penalty, which may include financial (e.g. fines) and/or non-financial elements (e.g. harassment, time needed to defend claims of violations, prison terms). One could think of the extent of the penalty as a function of state governance: properly functioning governments hire and motivate inspectors to pursue substantive violations rather than minor ones, while inspectors in corrupt or dysfunctional governments can get away with threatening to impose high penalties for even minor technical violations if a bribe is not paid (i.e. extortion). Then, firms located in states with more corrupt governments would be expected to pay higher effective regulatory costs, as suggested by the results of Section 6.1.

H Qualitative Evidence Regarding Harassment Bribery from "ipaidabribe.com"

"I am a small factory owner in Kirti Nagar Industrial Area. We follow almost all rules laid down by government for the welfare of workers. Now, even if we follow everything there is always somethings where we lack and which needs improvement. We have a factory inspector by the name of Mr. ——— (M: ———-). He comes to all the factories in our area, inspects them, find mistakes and then harass and blackmails us. According to him he can get our factories sealed. To avoid this, to save our time and to save the unnecessary paperwork we pay him every year. I have paid him twice in two years i.e. 10000 & 15000 and this is common with all factories. Please take a strict action against him so that he learns a lesson. I am sure he is not alone. All his colleagues are equally corrupt."

(Reported on August 11, 2014 from New Delhi, Delhi | Report #131791)

"During the routine labor verification process by the labor department at our office, we were advised by the consultant to pay the labor inspector a bribe to ensure that they don't keep calling us for needless paperwork."

(Reported on June 28, 2011 from Chennai, Tamil Nadu | Report #35064)

“The Labour Department requires a dozen odd registers to be maintained some of them which are totally outdated and pointless. E.g: Salary register, Attendance register, Leave register etc.

Our IT office has an electronic system that logs all entries/exits and leave taken. We have the records and offerred to provide it to them in a printout.

Salaries are paid electronically via bank transfer.

The officer declined and said it must be maintained in a manual register!

Finally an arrangement was made where we maintain a few records manually and the rest he would overlook.

Cost of arrangement Rs 1500 twice a year even if the officer shows up only once a year for the inspection!

He is supposed to inspect twice so expects to be paid even for the time he did not show up!” (Reported on October 13, 2010 from Chennai, Tamil Nadu | Report #44950)

“Well i had gone to renew my labour license and after all the running around in the bank and the department, the signing authority asked me to pay Rs.500 for signing. When asked why 500, i was told since there are 5 employees for Rs.100 each.” (Reported on December 31, 2010 from Hyderabad, Andhra Pradesh | Report #43509)

“... in my third visit i met one of office peon in Labour office he guided me for the bribe he also investigated and *advised me for bribe according to the number of Employees deployed on contract basis* and for this valueble suggestion he charged me Rs. 100. Again with full confidence i went to the ALCs desk and straight away i offered him the packet which was contains the amount of Bribe Rs. 3000/- ... He issued me the license after office hours...” (Reported on March 30, 2011 from Mumbai, Maharashtra | Report #39133)

“Applying for shop & establishment [registration] & procured all documents relating to the registration. Finally inspectors are asking Rs.1000 as a bribe. If any other notice received by the company for resolving that another Rs.2000 and above , it depends on the company” (Reported on March 28, 2014 from Bangalore, Karnataka | Report #99016)

“Officer name ———- . Mobile no. ——— He is asking for a bribe of 60,000 and is saying will issue a negative report under labour laws.” (Reported on January 24, 2014 from Gurgaon, Haryana | Report #83365)