Efficient Informal Trade: Theory and Field Experimental Evidence[*]

Niklas Bengtsson[†]

November 6, 2014

Abstract

Informal sectors in developing countries are often thought of as responses to rigid and cumbersome market regulations. In this paper I discuss informal trade as a first-best outcome. The model I propose has the curious aspect that "bad" regulations can be necessary to achieve efficiency even though they are always sidestepped in equilibrium. The key assumption is that the regulations define the trading parties' fall-back position in case the informal bargaining process breaks down. I set up a field experiment to test the model's mechanisms in the Cape Town market for metered taxis. Consistent with the model, I find that sidestepping the regulations increases ex post efficiency without significantly affecting the predictability of the terms of trade (i.e. ex ante efficiency). However, a subsample analysis indicates that when information asymmetries are severe, informal trade increases ex post efficiency at the expense of reduced ex ante efficiency.

Keywords: *Informal sector, Market regulations, Taxi Experiment, Incomplete Contracts, Transaction Costs, Institutions, Natural Field Experiment, Environmental Economics*

JEL: O17, L51, H10

[*] I would like to thank Johan Arntyr for providing extraordinary assistance with this research, and Tore Ellingsen, Per Engström, Bertil Holmlund, Steinar Holden, Espen Moen, Åsa Rosén, Mathias Sutter and numerous seminar participants for helpful comments. Financial assistance from the Swedish International Development and Cooperation Agency (Sida) is gratefully acknowledged.

[†] The Department of Economics, Uppsala University and UCLS. Contact: niklas.bengtsson@nek.uu.se

1 Introduction

Well-designed regulations can correct market imperfections driven by asymmetric information. However, in practice regulations are often imperfect themselves, and agents frequently sidestep them – sometimes informally – to make mutually advantageous agreements *ex post*. Such behavior is often described as *ex ante* inefficient, because they reduce the predictability of the terms of trade and give rise to transaction costs problems (Williamson 1979). Following the work of de Soto (1989), a growing number of papers have verified that heavier regulations are associated with more informal trade and lower output in formal sectors (Djankov et al. 2002; Besley and Burgess 2004).

In this paper, I point to an apparently overlooked aspect of informal trade in regulated markets, namely that the regulations can define the trading parties' fall-back position. As such, incomplete regulations can be interpreted as *option contracts*,[1] which will frame the informal bargaining process. This implies that even though incomplete regulations are not strictly followed, they can still prevent holdups through their role as threat points. The option contract perspective on regulations has a rather curious implication: "bad" regulations might be necessary for efficiency, even though they are not strictly followed. In such markets, informal trade is a first-best equilibrium; both deregulation and more strict enforcement would reduce efficiency (which marks an important difference to the policy advice in e.g. de Soto 1989).

I present a simple model with the above features and provide evidence of the mechanisms involved using a randomized field experiment in the Cape Town taxi industry. In many ways, the urban taxi market represents the textbook case of a market that might benefit from regulations. The reason is that cabdrivers and customers must search for each other before they can trade. Absent regulations that enforce commitment to posted prices, search costs are sunk at the time cabdrivers and customers meet, and the bilateral trade relationship that arises will not internalize these sunk costs (Diamond 1971; Williamson 1975; Grout 1984). This means that the trade surplus might be allocated in such a way that one of the parties would have preferred not to enter the market at all. In a seminal paper, Diamond (1971) found that even trivial search costs can result in no trade at all in these kinds of markets.

In search markets, governments can – and do – enforce regulatory policies that are meant to reduce the uncertainty of the terms of trade – such as requiring taxis to install sealed taximeters in their vehicles. The Cape Town taxi market is no exception. By law, Cape Town taxis must maintain sealed meters that govern the kilometer fare. However, compliance to the meter system is not complete. Specifically, taxi drivers and customers frequently agree on a fixed fare in advance, and when doing so, drivers disable the meter and mask the implied kilometer fare. The Cape Town market for metered taxis thus provides an opportunity to vary the degree of commitment to the regulations without affecting the natural context of the transaction.[2] The experimental design followed a 2:2 protocol. Half of the trips were randomly assigned to a treatment group who encountered a customer asking for a fixed, unregulated fare with and without counteroffers. The other half was randomly assigned to a comparison group compelled to utilize the regulated fare, either passively or through (mild) enforcement.

The experimental findings reveal that, on average, there are significant ex post gains from sidestepping the meter. If a fixed fare is agreed upon before the trip has begun, firms disable the meter, and the average distance traveled is decreased by almost ten percent. The cost incentives

[1] On option contracts, see Huberman and Kahn (1988); Chung (1991); Aghion et al. (1994); MacLeod and Malcomson (1993); Edlin and Reichelstein (1996); Nöldeke and Schmidt (1995).

[2] Using the taxonomy in Harrison and List (2004), the study would be labeled a "natural" field experiment

are quite intuitive. By disabling the meter, an additional kilometer no longer adds to the firm's profits, so there are no incentives to take longer routes to increase mileage. By contrast, when the meter is running, drivers tend to take detours that increase both the time spent in the cab and gasoline costs. Since complying to the regulated fare increases mileage, the regulations do not only reduce the trade surplus, but also inflict a negative externality on the non-trading population through pollution. Using information on distance traveled, the size of the vehicle and the speed of the trip, I estimate that strict obedience to the meter regulations increase carbon dioxide emissions by 8 percent in this market.

I next ask how the gains from trade are allocated between the customer and the firm. I find that the average price paid is quite insensitive to informal bargaining. Although counteroffers drives down the price, the effect is not statistically significant on the full sample, and the reduction in driving costs offset the bulk of the lost revenues. The insensitivity of the final price in the treatments involving bargaining suggests that the regulations serve an implicit role by framing the informal renegotiation in an efficient manner, consistent with the model of regulations as option contracts. Thus, although the empirical results suggest that the regulations do indeed increase both private and environmental costs, it does not follow that deregulation would unambiguously increase welfare in the Cape Town taxi market.

The empirical results are in line with a theory of "efficient informal trade": sidestepping the metered fare increases the size of the cake without affecting the allocation of the trade surplus. However, while I cannot reject the efficiency hypothesis on the full sample there are some anomalies suggesting not all informal transactions are efficient. Notably, many customers and firms appear to opt for the metered fare even though this behavior reduces the trade surplus on average. This side-result can be explained by parameter heterogeneity (i.e. different bargaining strengths across different customers and firms). Using quasi-random variation in moral hazard opportunities, I show empirically that when information asymmetries are severe, Pareto gains from informal trade are larger, but the predictability of the terms of trade is lower. This implies that trade is sometimes prevented even in the presence of Pareto gains from trade. Thus, under severe information asymmetries, the regulator does face a tradeoff between ex ante efficiency and ex post efficiency: he cannot achieve both without designing (unrealistically) detailed regulations.

This work is related to the long-term discussion of "efficient corruption" in overly regulated markets in developing countries. In a classic formulation, Huntington (1968) wrote that "in terms of economic growth, the only thing worse than a society with a rigid, over-centralized, dishonest bureaucracy is one with a rigid, over-centralized and honest bureaucracy". Myrdal (1968) emphasized the second best nature of "efficient corruption", noting that the rigid bureaucracies are themselves a function of corrupt officials. More recent accounts of corruption note typically emphasis the negative aspects of corruption (Rose-Ackerman 2003; Bardhan 1997; Svensson 2005). However, the present paper studies informality not among public officials but among private agents.

The present paper also joins an emerging literature using experimental methods to test microeconomic mechanisms in the field. Past experimental work on holdup and bargaining has been conducted in laboratories. These studies have typically found that sunk investments matter more than they should according to standard economic models (Hackett 1994; Ellingsen and Johannesson 2004), although perhaps not enough to prevent the Diamond paradox in search markets (Grether et al. 1988). The field experimental evidence on holdups and decentralized trade is still very limited in scope. A closely related paper is Keniston (2011), who use a structural and experimental approach to study how the utility costs of bargaining affects efficiency in the market for autorickshaws in Jaipur, India. The focus in Keniston (2011) is similar, as it compares fixed prices and bargained prices, but in contrast the experiment studied in this paper, Keniston (2011) assumes that rickshaw drivers cannot take detours and inflate the length of the journey. There is thus no moral hazard element in his study. Another study is Iyer and Schoar (2010), who vary the specificity of a wholesale order to test for holdups among pen manufacturers in India.

A related literature focus specifically on “credence goods” – markets in which the seller knows more about which service or good the consumer needs than the consumer himself. Darby and Karni (1973) introduce the term and mention repair services as the typical example, but taxi markets also fit well into this category (see the survey by Dulleck and Kerschbamer 2006). In this vein of research, there are a number of “taxi experiments” used to study related mechanisms. Most closely related to this work is a study by Balafoutas et al. (2011), which explores how taxi drivers in Athens, Greece respond to different displays of customer wealth and information. In addition, Habyarimana and Jack (2011) study how mini-bus drivers in Kenya respond to safety incentives and Castillo et al. (2012) study how the bargaining process is affected by the gender of the person giving counteroffers.

The paper is set-up as follows. Section 2 presents a model of efficient noncompliance, Section 3 presents the institutional details and the experiment, Section 4 describes the identification strategy and the main results, Section 5 discusses an clarifies the empirical relevance of parameter heterogeneity and Section 6 concludes.

2 Equilibrium framework

2.1 Set up

Although the model presented in this section is meant to illustrate the typical taxi market, it is quite general. The need for rigid regulations arises from fixed search costs on both sides of the market, which are sunk at the time customers and sellers meet. The regulations are, however, incomplete, and therefore give rise to ex post rents. These starting points seemingly imply that the regulator is facing a tradeoff, and must choose between ex ante efficiency or ex post efficiency. However, as we shall see, there is a combination of regulations and informal trade that achieves the first best (both ex ante and ex post efficiency).

The experiment will test two propositions relating to this efficiency results. The experiment is conduced on the match-level, which means that it does not test propositions regarding equilibrium level of output. For this reason, this conceptual framework focus on the main empirical propositions, and therefore sidestep equilibrium issues (what determines the market equilibrium and hence the social optimum).

In this section I will leave out Taxi drivers produce a service worth y to the customer, which is the value of traveling to the destination (instead of, say, walking). To produce the service, taxis use inputs captured by the term $c > c_L$, where c_L is the minimum costs required to produce the service. Trade cannot occur instantly, however. Both customers and firms must make ex ante investments before they can enter the market, which are sunk at the time customers and firms meet. In the taxi market, it is natural to think about these ex ante investments as the cost of search (or waiting time). The driver’s exogenous cost of entering the market is k, and the customer’s exogenous cost is b.

Firms and customers are matched according to a matching function $m = m(v, n)$, which is homogenous to degree one. v is the share of unmatched firms, and n is the share of unmatched customers. The probability that a firm finds a customer is equal to $q(\theta) = m(v,n)/v$, where $\theta = \frac{v}{n}$ is equal to market "tightness". Conversely, due to the homogeneity of the matching function, the probability that a customer finds a firm is equal to $\theta q(\theta) = m(v,n)/n$. All matched trade relationships are bilateral (one customer, one firm).

The customer will direct his or her search to one out of at least two submarkets and enter the market with the lowest expected costs. The market entry condition is

$$R \geq \theta q(\theta)(y-p) - b \tag{1}$$

where R is the expected cost of entering any other submarket. The number of submarkets are exogenously determined and the total population is fixed.[3] In equilibrium, this means that all submarkets will offer the same price, so (22) will hold with equality.

For the firm, the value of entering the market is equal to

$$V = q(\theta)(p-c) - k. \tag{2}$$

Free entry implies that $V = 0$ in equilibrium such that

$$q(\theta) = \frac{k}{(p-c)}. \tag{3}$$

The endogenous variables are share of unmatched firms v, costs c and the total price p.

2.2 Social efficiency

Social efficiency is characterized by the solution to the social planner's problem of maximizing social welfare in this economy. All customers in this market engage in search and are evenly spread across submarkets according to (22). The condition for social efficiency is thus found by asking how many available firms (v) the planner would prefer, given a fixed population (fixed n):

$$\begin{aligned}
\max_{v,c} \quad U &= Rn + Vv & (4)\\
&= n\theta q(\theta)(y-p) + vq(\theta)(p-c) - vk - nb & (5)\\
&= vq(\theta)(y-c) - vk - nb.
\end{aligned}$$

Notably, the planner puts the same utility weight on firms and customers and does not care about the division of the surplus (that is, the price). The solution to (25) is $c = c_L$ (costs trivially bind at the lowest possible level) and

$$q(\theta) = \frac{k}{(y-c_L)(1-\eta)} \tag{6}$$

where η is a positive constant equal to (the negative of) the elasticity of $q(\theta)$:

$$\eta = -\frac{\partial q(\theta)}{\partial \theta}\frac{\theta}{q(\theta)}. \tag{7}$$

[3] Only two firms need to compete for an efficient competitive (Bertrand) equilibrium (Acemoglu and Shimer 1999).

Equation (27) defines the socially efficient level of search frictions in this economy.

2.3 Unregulated bargaining

We now ask under what regulatory regimes the market will internalize the search externalities and minimize costs. Absent regulations, once the two parties have matched and formed a bilateral trade relationship, the price p is determined through bargaining. Firms are residual claimants on productivity so $c = c_L$. The price is given by the solution to the Nash bargaining problem, defined as:

$$\max_p (y-p)^{\beta}(p-c_L)^{1-\beta} \tag{8}$$

Notably, since the ex ante investments k and b are sunk, they will not be internalized in the price. The final price equals

$$p = \beta c_L + (1-\beta)y. \tag{9}$$

Since the final price will be independent of k and b, nothing guarantees that the trading parties will be compensated for their ex ante search investments. If customers have low bargaining power ($\beta = 0$) the consumer surplus (22) will be negative. Conversely, if customers have too high bargaining power ($\beta = 1$) the producer surplus (23) is negative. The problem is due to what Williamson (1985) calls the “fundamental transformation” of markets: the transition from competition to bilateral trade relationships. At the decentralized bargaining stage, the ex ante investments are sunk, which creates a holdup problem. Even though there are mutual benefits from trade, the market does not clear.

If the bargained price happens to lie on a point between the two parties' reservation prices, trade will occur but the search frictions in the market will in general not be efficient. Combining (30) and the free entry formula (24), the competitive level of tightness is given by

$$q(\theta) = \frac{k}{(y-c_L)(1-\beta)}. \tag{10}$$

Equation (31) says that an unregulated market will be efficient only under the so-called Hosios-condition (Hosios 1990), that is, when $\beta = \eta$. Both β and η are exogenous constants. Thus, even under competition, there are no market forces assuring this condition holds when the price setting is completely decentralized.

2.4 Rigid and flexible regulations

We will now study how governmental price regulations affects the holdup situation. A common regulation in taxi markets, and indeed in most markets, is that firms are required to post and commit to unit price contracts related to the costs of production (“one dollar per hour”, “one dollar per mile”, “one dollar per kilo”, etc.). In an urban taxi market, this is the familiar *metered fare.* Such a price policy appears promising because it forces firms to commit to the terms of trade ex ante. However, metered fares are similar to procurement contracts (Laffont and Tirole 1986), which are incomplete as they only specify the relationship between the costs and the price, c and

p, but omits cost minimization (i.e. that $c = c_L$) from the contract. As such, they might severely circumscribe the gains from trade by creating moral hazard incentives to inflate costs c.

To see this, consider the case where the regulations are strictly enforced, meaning that governments have some technology that allows them to oversee that the unit price is binding ex post (such as installing sealed taximeters in cabs). At the moment of the match, firms and customers take the advertised price p_c as given, so the only variable factor is c. Ex post profits (leaving aside the sunk cost k) is thus $p_c c - c$, and firms have no incentives to decrease costs c as long as $p_c > 1$. While the customer might observe c and c_L, and might infer that costs are not minimized, he cannot prove this in a court of law – being unproductive is not illegal, at least within bounds. Thus, using the jargon of the incomplete contract literature, the problem is that c_L is "observable but nonverifiable".

Suppose firms can increase costs up to a certain level $c = c_H$, after which the costs would be verified as "excessive" in a court of law. The final price will then depend on the structure of the ex ante market. Since firms are residual claimants on productivity, they cannot commit to cost minimization and the final price p will equal $p = pc_H$.

Although costs are unlikely to be minimized, the strictly regulated price will internalize search frictions because the regulations will bridge the competitive stage and the price-setting stage. Following the competitive search literature (Moen 1997; Shimer 1996), the equilibrium unit price p_c is found by maximizing (23) with respect to p_c at $c = c_H$ and $p = p_c c_H$, subject to the customer's entry condition (22). That is, the firms maximize ex ante profits, taking into account that a higher price will increase ex post profits but also attract less customers to their submarket. The solution is

$$p \quad = p_c c_H = \eta c_H + (1 - \eta) y. \tag{11}$$

Using (32) and (24) to solve for $q(\theta)$, frictions are given by

$$q(\theta) = \frac{k}{(y - c_H)(1 - \eta)}. \tag{12}$$

Comparing (33) with (27), strictly enforced price contracts will not maximize social welfare. The reason is that the costs will be too high $c_H > c_L$.

A special case occurs if customers cannot verify productivity at all. As $c_H \rightarrow y$, firms monopolize all consumer surplus from trade. However, customers anticipate this, which implies that firms must lower the per-unit price p_c in order to attract any customers to their submarket. The competitive unit price p_c will therefore tend to unity. As this happens, the surplus from trade approaches zero in equilibrium. The outcome is similar to the paradoxical result found in Diamond (1971) except that market failure is double-sided in this case. Even if the entry costs k and s are trivial, neither the firms nor the customers find it worthwhile to enter the market.

The above scenario can be seen as a case of regulatory failure. In an attempt to solve the problems associated with search frictions and holdups, the regulator reduces the size of the cake without assuring that the allocation of the cake will guarantee that trade occurs. The problem lies in the incomplete nature of the regulations. The (perhaps obvious) solution would be to force firms to post a complete price contract, relating p_c to y instead of c, but let us suppose this is not possible.[4] However, there is a regulatory environment that achieves both ex ante and ex post

[4] In the taxi market, this means that the taxis post a complete list of fixed departure and destination sites (like a bus shuttle company). More generally, exactly why incomplete price contracts exists has been an active area

efficiency, even when the regulator cannot dictate the ex post allocation. The key is to construct the regulations in such a way that the incomplete price regulation becomes an *option contract*, that stipulates the fallback positions in case the renegotiation breaks down.

Under the regulations-as-options perspective, the trading parties will sidestep the regulations and engage in informal bilateral bargaining when they meet (notice that the model is agnostic about whether doing so is illegal or not). The solution to the regulations-as-options case is found by backwards induction. First, we ask what the end price p is, given the posted price p_c. The solution to the bargaining problem is given by

$$\max_p (y - p - [y - p_c c_H])^{\beta} (p - c_L - [p_c c_H - c_H])^{1-\beta}, \tag{13}$$

where the threat point terms (in brackets) have been added to the bargaining problem in (29). If the bargaining is successful and a fixed price is agreed upon before the costs are realized, the regulated price p_c is no longer dictating the outcome. In this case, the firm have no incentive to maximize costs, so $c = c_L$. By contrast, in case bargaining fails, the cabdriver will turn on the meter and maximize costs ($c = c_H$).

The solution to (34) is

$$p = p_c c_H - \beta(c_H - c_L). \tag{14}$$

Both customers and firms correctly anticipate that (35) will determine the final price. Therefore, firms maximize profits (23) with respect to p_c, subject to both the demand constraint (22) *and* the bargaining outcome (35). The per unit price is thus given by

$$p_c = \frac{y(1-\eta) + \eta c_L + \beta(c_H - c_L)}{c_H} \tag{15}$$

Using expression (36) in (35), the end price equals

$$p \quad = p_c c_H = \eta c_L + (1-\eta) y. \tag{16}$$

Using (37) and (24) we see that the market equilibrium level of tightness $q(\theta)$ coincides with the condition for the socially optimal level of tightness (27). The regulations-as-options solution thus guarantees that trade occurs *and* that the costs are minimized ex post.

The option contract perspective offers a new perspective on informal trade and the act of sidestepping regulations. More specifically, it implies that "bad" regulations might be necessary for efficiency, even though they are not strictly followed in equilibrium. This finding has stark policy implications. Following the work of de Soto (1989), a number of papers have found that heavier regulations are associated with more informal trade (Djankov et al. 2002; Besley and Burgess 2004). At the face of it, these empirical findings appear to provide a strong argument for regulatory reform, perhaps in particular for deregulation: if the rules are sidestepped and renegotiated, what harm can be caused by removing them? In the model presented in this section, this conclusion is not

of contract theoretical research since Maskin and Tirole (1999) argued that uncertainty alone is not a satisfactory explanation as long as agents are forward-looking and rational. One possible explanation for incomplete regulations is that it is simply too costly to oversee and monitor more sophisticated regulations (i.e. a menu cost-problem). A more convincing argument, however, is that that the regulator might not *need* to provide more complete regulations. The theoretical point made in this section is that the resulting allocation of the trade surplus can be efficient anyway.

justified. Deregulation would just take us back to the holdup problem described in equation (29).

2.5 Match-specific bargaining strengths

The socially optimal result described in the previous sections critically hinges on the ability of a single posted price to internalize the preference and technology parameters. With parameter heterogeneity, every single transaction is not likely to be efficient. Suppose, for instance, that customers differ in bargaining strength β_i (using i as subscript to denote that the β is match-specific). Unable to post a menu of prices over different bargaining strengths, the firms must post a single unit price based on some representative measure of bargaining strength, say $\bar{\beta}$. Under regulations-as-options, the match-specific end price will then equal

$$p \quad = p_c c_H = \eta c_L + (1-\eta)y + (\bar{\beta} - \beta_i)(c_H - c_L). \tag{17}$$

Thus, in the case with heterogenous bargaining strength and informal bargaining, the end price can drive some firms and customers out of the market, and the regulator's dilemma is again whether the problems with holdups are worse than the problems of cost maximization. Since we shall deliberately vary the intensity of counteroffers, this variation of the model is more relevant when interpreting some results from the experiment. More specifically, equation (38) suggest a route for interaction variables, as the impact of making counteroffers – i.e. the instability of the predicted price – will be stronger if the firm's moral hazard opportunities $(c_H - c_L)$ are stronger. Notably, an efficient outcome, independent of an individual's bargaining strength, is approached as $c_H \to c_L$ (that is, when the possibility to inflate costs decreases).

2.6 Notes on the theoretical literature

While the final point about the efficiency of informal trade in regulated markets appears novel to the literature, this section draws on past literature in search theory and contract theory. Diamond (1971) first demonstrated that if firms are price setters and make no ex ante price commitments, even trivial search costs will result in no trade at all in equilibrium. That a condition enabling social efficiency exists in search markets is discussed in Mortensen (1982), Pissarides (1984) and Hosios (1990). Shimer (1996) and Moen (1997) show that this condition holds if firms can advertise complete terms of trade ex ante. In the literature on incomplete contracts, the first formalization of the holdup problem is Grout (1984), although the description of the holdup problem goes back to Williamson (1975), Klein et al. (1978), and Goldberg (1976). The tradeoff between institutional rigidity and flexibility in the presence of incomplete contracts and ex ante investments is discussed in Williamson (1985) and Williamson (1983). Subsequent studies in this line of research have studied what type of ex ante contracts can achieve ex post efficiency (Huberman and Kahn 1988; Chung 1991; Aghion et al. 1994). In particular MacLeod and Malcomson (1993), Edlin and Reichelstein (1996) and Nöldeke and Schmidt (1995) discuss how markets and legal institutions can reproduce the first best in these settings.

3 Experiment

3.1 Institutional background

Although the tradition of using taximeters predates transaction costs economics, modern rationalizations of the practice draws (implicitly) on the idea that market frictions increase uncertainty about the final terms of trade – as in the conceptual framework discussed in the previous section. For instance, in a policy paper for South Africa's Human Sciences Research Council, Lowitt (2006) argues in the following way:

> "In a cruising taxi market, customers typically do not know how frequently a taxi will pass. Moreover, if fares are unregulated and can vary between taxis, passengers will not know whether to accept or reject the first and subsequent fares offered, as there is a fundamental asymmetry of information. In this situation most customers will take the first taxi that stops, regardless of price. Under these circumstances, taxi drivers might not benefit from offering lower fares in hopes of increasing usage. [...] As a consequence, taxi drivers may either end up .charging so little that they fail to make an acceptable living or they may increase fares to protect total earnings and eventually potential cut off demand." (Lowitt 2006).

The metered taxi service in Cape Town is regulated by the National Land and Transport Transition Act (Act No. 22 of 2000), the Western Cape Regulations on Operating Licences 2002 and the City of Cape Town's bylaws. A metered taxi service is defined in the National Land and Transport Transition Act as a "a public transport service operated by means of a motor vehicle (...) to carry fewer than nine seated persons, including the driver, where that vehicle (a) is available for hire by hailing, by telephone or otherwise; (b) may stand for hire at a rank; and (c) is equipped with a sealed meter, in good working order, for the purpose of determining the fare payable." (City of Cape Town 2007, p. 185).

Prospective taxi firms apply to the Provincial Operating Licence Board (POLB). As point (b) indicates, the taxi operator is authorized to operate from a base area. This base area is typically a taxi rank, which marks the center of a smaller pick-up radius and a larger service radius. According to the Transport Department of Cape Town, the taxi ranks are more important in Cape Town than elsewhere in South Africa. The city owns and maintains official taxi ranks located in tourist-friendly places (e.g., hotels, malls or communication hubs). City bylaws require taxis to return to their ranks once the customer has been dropped off. Although hailing for taxis is not prohibited, taxis are not allowed to "roam" for customers outside of their designated areas.

The application to the POLB must state the fare and the area of operation. The fare typically consists of a fixed component (a "fall flag" price) and a variable, per-kilometer component. Additional fees for luggage and waiting time are also allowed, but there are no other requirements for the price structure. The fare for each license is posted in the National Gazette. In the February 2011 issue of the National Gazette (one month prior to the experiment), all of the posted prices had a fixed component of R2 and a variable component of R10 per kilometer. Past studies by the City of Cape Town (2007) report that a starting price of R2 is the most common fall flag fare, though they list R5 as the main alternative. Additionally, these studies report that the variable component varies between R7 and R12 per kilometer.

According to city bylaws, the meter must be fully visible in the vehicle. Drivers are prohibited from breaking the seal of the meter. More specifically, drivers are not allowed to "tamper or interfere with any tyre, mechanism or fitting of a taxi so as to cause the taximeter fitted thereto to register any fare or charge other than a fare or charge in accordance with the prescribed tariff." (City of Cape Town 2007, p. 236). The local bylaws further stipulate the following:

- Taxi drivers are prohibited from demanding a payment greater than the prescribed tariff.
- However, the driver and the customers are free to agree on a fixed price: "No agreement for the payment of a fare or charge exceeding that permitted by the prescribed tariff shall be binding in respect of any journey in a taxi and a passenger shall, notwithstanding any such agreement, be entitled to refuse to pay any amount demanded in excess of the fare and charges so permitted ..."
- The customer is entitled to a receipt showing the fare, the date and time during which the passenger entered the taxi, the place of departure and destination, the distance traveled and the taxi's registration number.
- The driver is not allowed to ask for a payment before the journey is completed.
- No customer shall "hire a taxi knowing that he will not be able to pay the fare" or "unlawfully refuse to pay the fare" that has been agreed upon.

Notably, both the firms and customers are free to renege on the regulated fare if there is mutual consent. Moreover, the law stipulates that the asking price should not exceed that "permitted by the prescribed tariff." However, the law does not specify at what distance the prescribed tariff should be evaluated, only that the meter should be turned on even though a fixed fare is agreed upon.

Before launching the experiment, we performed in-depth interviews with the local authorities (the Public Transport Regulations and Surveys) and local taxidrivers to get a more complete characterization of the reception and usage of the meter system. The authorities indicated that the meter system is widely used in practice, but admitted that compliance with the rules governing the meter system can be imperfect. The Public Transport Regulations and Surveys claimed to be aware of an (illegal) second-hand market for meters that had been tampered with and deregistered. In 2007, the City of Cape Town did a survey of taxicabs operating from the most popular ranks, and found a non-trivial share of illegal vehicles with meters installed in their cars.

One reason the meter system is so widely used is that it also serves as a monitoring device for the fleet-owner vis-a-vis the cabdrivers. The larger companies (Unicab, Marine Taxis, Sport, CabXpress, Intercab and CAB-NET) has installed their own meter system in their vehicles, which are used to compare distance travelled with the official taximeter. The drivers, in turn, indicate that using the meter is in their own self-interest, since the revenue registered in the meter guarantees a minimum amount of income. However, our interviewees indicated that there are ways to tamper with the system in order to cheat the car owner. For instance, for earlier versions of the meter, drivers could put the car on a jack and drive it in reverse – apparently this would reverse the meter's recording of the distance travelled. Most cabdrivers appears to have a kinked remuneration scheme, however, for example the first 200 ZAR in daily revenue is given to the car-owner, and then the rest becomes the driver's income.

3.2 Experimental set-up and empirical predictions

In sum, the city bylaws stipulate that the metered fare is binding for the seller at the moment of the match. Before the trip starts, the customer has the option of initiating a renegotiation of the fare by demanding a fixed price. The context is especially suited to the field experimental approach because it allows us to naturally manipulate the price contract in an experimental manner without introducing artefactual elements into the contractual arrangement. In this section, we discuss the experimental design and how our tests relate to the predictions laid out in Section 2.

The experiment was executed in 18 days (i.e., from February 26 to March 27 2011). A total number of 176 trips were conducted (a median of 10 trips per day). The trips were connected to the most trafficked taxi ranks within a 15-kilometer radius of Mount Nelson Hotel in Cape Town. The experimental design followed a 2:2 protocol, with two main treatment groups and two supplementary treatments (subtreatments) for each group. Half of the trips were randomly assigned to a treatment group compelled to utilize the taxi meter, and the other half was assigned to a fixed fare treatment. The experimental design is illustrated in Figure 1.

A research assistant undertook the trips while acting as a tourist with limited language abilities. Once an available taxi vehicle was identified at the predetermined taxi rank, the assistant approached the driver with a request to go to a predetermined destination. In the first treatment group (M1), the assistant was instructed to simply state his destination without further interacting with the driver. Thus, this treatment group received a minimum degree of interference. The second subtreatment (M2) introduced a mild enforcement mechanism by instructing the assistant to ask the driver to "run the taximeter". The last two subtreatments introduced new terms of trade: a fixed fare rather than a metered fare. In subtreatment F1, the assistant was instructed to ask for a fixed price before the trip and to accept the driver's first offer. The second renegotiation treatment (F2) introduced the concept of bargaining by instructing the assistant to ask for a fixed price, wait for a first offer, make a counteroffer equal to 75 % of the original offer, and then accept the driver's new offer immediately.

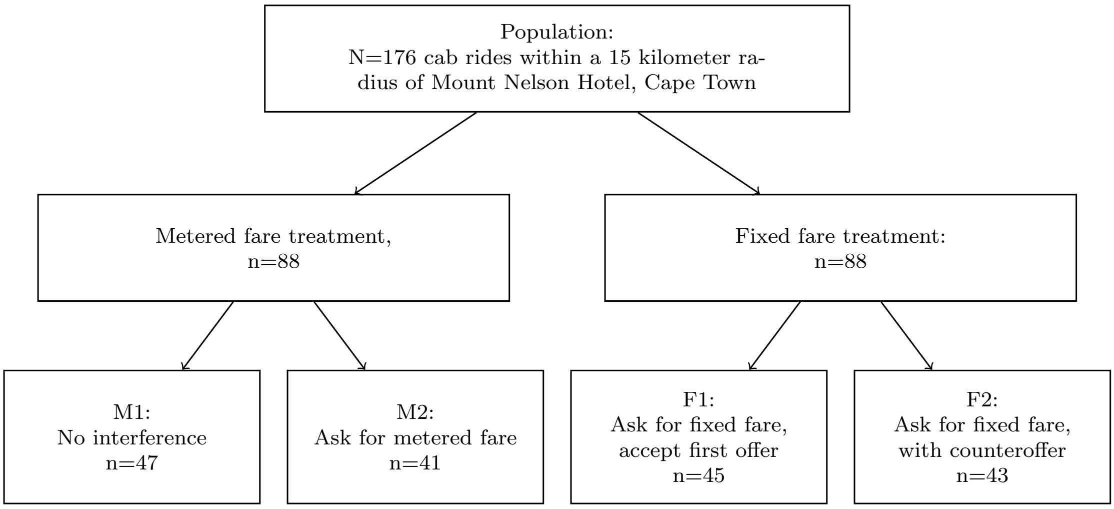

Figure 1: Experiment. A total of 176 cab rides in Cape Town was randomly assigned to one out of four subtreatments (M1, M2, F1, F2).

The main theoretical has two empirical implications: (1) renegotiation minimizes ex post costs, regardless of threat point and (2) if the metered fare serves as a threat point, the renegotiated price will be constrained, implying that no party is worse off from renegotiation. The empirical test of the first prediction is that drivers inflate mileage when the metered fare is turned on. Theory provides some guidance in constructing an efficient test of this prediction. Notably, the moral hazard incentives to inflate mileage should only differ across the main treatment groups. Therefore, when testing for excessive driving under the metered fare, the statistical power can be increased by grouping the subtreatments. The experiment was designed with this categorization in mind, and the randomization was stratified in order to assure that each route had at least one trip assigned to F1 or F2 and one trip assigned to M1 or M2.

The second prediction is that, in equilibrium, no party should be worse off from informal trade. This prediction stands in stark contrast to the standard bargaining solution, which predicts that a passive consumer ($\beta = 0$) would be held up in the bargaining process. Of course, interpreting the weights in the sharing rule and incorporating them into an experimental design is a challenge. A popular interpretation of the bargaining weights is that they represent the probability of making a counteroffer in a strategic bargaining game (Osborne and Rubinstein 1990). Therefore, one could interpret F1 as $\beta = 0$ and F2 as $\beta > 0$. A first issue with this interpretation is that the game theoretical underpinnings of Nash bargaining presume perfect foresight, which implies that there are no counteroffers in practice. In the empirical analysis, I make a heuristic interpretation of F1 as "low β" and F2 as "higher β". A second issue is that the equilibrium concept presumes that there is a representative, or "typical" β for the entire economy. With individual-specific β, the efficiency results no longer hold for each transaction (although it might hold on average). In Section 5, I will return to the issue of heterogenous bargaining strengths.

3.3 Implementation

The assistant (the "customer") was given strict instructions on how to execute the four treatments. Under M1, the assistant was instructed to simply state the destination ("Destination X please.") and travel to the planned destination without further interacting with the driver. In M2, the assistant was instructed to state the destination and also ask the driver to use the meter ("Destination X please, and please run the meter") before beginning the trip. The fixed fare treatments (F1 and F2) began with the assistant asking, "Can you give me a fixed price for [destination] please?" If the driver gave a negative response, the assistant would probe once more. Under F1, the assistant was instructed to immediately accept the first offer. Under F2, the assistant was instructed to make a counteroffer equal to approximately 75 percent of the opening offer: "Could you get me there for [counteroffer]?" At this point, the driver could accept, insist on the old offer, or make a new offer. The assistant was instructed to accept any of these outcomes. However, if the driver asked the assistant to come up with a new counteroffer, the assistant was instructed to end the negotiation by asking "What is your price?"

Once the trip commenced, further communication between the driver and the assistant was restricted. If the driver asked about the assistant's background (e.g., "Where are you from?"), the assistant would answer: "Sweden". If the driver asked about the assistant's business in Cape Town (e.g., "What are you doing here?"), the assistant would only answer: "I'm a tourist." In response to all other questions or inquiries, the assistant would pretend not to comprehend and shrug his shoulders. At the time of arrival, the assistant was instructed to pay the asking price, rounded up to the nearest multiple of ten, without haggling (on the round-up, see Section 5.1).

A selection of the busiest taxi ranks in Cape Town was used to construct 42 distinct routes. According to a survey by the Transport Department in 2005 (City of Cape Town 2007), the most trafficked official ranks were (in descending order): Waterfront Aquarium, Waterfront Victoria Wharf, Cable Way Station, Upper Adderley Street, Green Market Square, Plein Street, Camps Bay, Cullinan Hotel, Holiday Inn, Upper Long Street, the International Convention Centre, and the Seaport ranks. Out of these 12 ranks, nine were included in the experiment. In addition, a few larger unofficial ranks were used to avoid interacting too much with only a few ranks. The distribution of the departure and destination ranks used in the experiment is displayed in Table 1.

Table 1: Departure and destination ranks

	Number of trips	
Rank name	From	To
Cable Way Station	27	29
Café Neo	0	2
Camps Bay	33	33
Central Train Station	12	12
Clifton Beach	4	6
Four Seasons Hotel	15	5
Gardens Centre	5	14
Green Market Square	0	5
International Convention Centre	15	12
Kirstenbosch Botanical Gardens	2	2
Mount Nelson Hotel	9	10
Sea Point	18	17
Upper Adderley Street	8	1
Upper Long Street	8	6
Waterfront, Victoria Wharf	5	5
Waterfront Aquarium	12	15
Waterkant Mall	3	2
Sum	176	176

Note: The table displays the distribution of departure and destination ranks used in the experiment.

Each route is represented at least two times in the sample. The experimental design randomized the order of the trips within each route while balancing the randomization so that each route contained at least one observation assigned to the regulated (metered) fare (M1 or M2) and one observation assigned to the renegotiated fare (F1 or F2). In practice, the last trip on each route was assigned to the inverse of the second-to-last trip. For example, if the second-to-last trip was randomly assigned to subtreatment F2, the last trip of the route was assigned to subtreatment M2; if the second-to-last trip was randomly assigned to M1, the last trip on that route was assigned to F1. This so-called block randomization is more efficient than simple randomization, particularly if one is using route-level controls (block randomization ensures that the treatment varies within each route in practice and not only in expectation, see Fisher 1926). Block randomization increase efficiency even when it is "incomplete" (that is, uses an uneven number of observations per block), as in this case.

There is no evidence of experimenter bias. The treatments were randomly assigned to individual trips in the route schedule once the departure and destination ranks were established, and the assistant was not able to cherry pick routes based on the treatment assignments. However, hypothetically, the assistant could have (unconsciously) allowed the treatment assignment to influence the actual timing of the trip. As always, there is also the possibility that the assistant failed to follow protocol. Table 2 displays the results from orthogonality tests conducted across a set of predetermined variables. As expected, the statistical probability that a trip was assigned to a particular treatment does not depend on the available control variables, such as departure rank, destination rank, date, time of day, map distance or any determinant of vehicle quality.

Table 2: Test of randomization

	Predetermined variables (t-test)			
	Regulated fare treatment	Unregulated (fixed) fare treatment	Difference	p-value
Map distance	6.29	6.21	0.08	0.791
			(0.29)	
Advertised price (imputed)	64.89	64.13	0.76	0.791
			(2.86)	
Start hour	13.74	13.63	0.11	0.715
			(0.31)	
Radio in car	0.32	0.38	-0.06	0.431
			(0.07)	
Independent operator	0.16	0.22	-0.06	0.337
			(0.06)	
Visible meter in car	0.93	0.97	-0.03	0.307
			(0.03)	
Seat belt for passenger	0.69	0.73	-0.03	0.621
			(0.07)	
Car type: Hatchback	0.25	0.22	0.03	0.595
			(0.06)	
Car type: Sedan	0.51	0.48	0.03	0.653
			(0.08)	
Car type: Station Wagon	0.24	0.31	-0.07	0.313
			(0.07)	
Obs	87	87	176	

	Categorical variables (F-test)			
	Nominator d.f.	Denominator d.f.	F	p-value
Departure rank	14	162	.378	.979
Destination rank	16	160	.658	.831
Date of trip	17	159	.88	.598
Route (block unit)	46	130	.35	.99995

Notes: First panel display uncorrected, raw means and t-tests of the difference across treatment group. Standard errors in parenthesis. Second panel display F-tests of orthogonality between treatment and categorical variables. The fixed fare treatment equals F1 or F2; the metered fare treatment equals (M1 or M2). The advertised fare is imputed based on equation (21).

3.4 Intention-to-treat and driver response

The experiment deliberately abstained from imposing artefactual incentives to induce the treatments (e.g., odd rewards or threats of legal action). Rather, the slight variation in the inducement instruments (subtreatments) captures naturally occurring circumstances – different “type” of customers that taxi drivers regularly serve in practice. In the experiment, taxi drivers always had the final word on the terms of trade. The assistant never insisted on a particular treatment once the request had been made, but accepted the driver’s preferred terms of trade in the end.

Table 3 displays the extent to which a fixed fare was renegotiated under the different subtreatments and across the two main treatments. 85 % of the taxi drivers used the meter under the passive subtreatment (M1). When the drivers were asked to use the meter, the compliance rate was 97.5 % (only one driver refused to comply). Under the fixed price treatments, renegotiation occurred in 95.6 % and 97.7 % of the cases, respectively. According to the assistant's logbook, the reasons for refusing to use the metered fare treatment were "No taximeter in car" (4 obs) and "Driver did not start taximeter" (4 obs). The reasons for refusing to offer a fixed fare under the fixed fare treatment were "Driver refused to give a fixed price" (2 obs) and failure to ordinate the correct treatment (1 obs). All trips were conducted and recorded regardless of driver response.

Table 3: Response to treatment, across intention-to-treat group

Treatment	Driver response	
Regulated fare	Used meter	Offered fixed price
M1: No interference	40 (85.1)	7 (14.9)
M2: Ask for meter	40 (97.6)	1 (2.4)
Sum control	80 (90.9)	8 (9.1)
Unregulated (fixed) fare	Used meter	Offered fixed price
F1: Accept first offer	2 (4.4)	43 (95.6)
F2: With counteroffer	1 (2.3)	42 (97.7)
Sum treatment	3 (3.4)	85 (96.6)
Overall no. of compliers	165 (93.75)	

Notes: Total number of observations: 176. Percent in parenthesis.

Cape Town taxi drivers are not obliged to comply with a a request to offer a fixed fare, nor are they restrained from initiating a renegotiation over a fixed fare themselves. Relating compliance to theory, it is notable that most drivers opt for the metered fare under the passive subtreatment (M1). However, few drivers (virtually no one) refused to offer a fixed fare when the customer initiated a renegotiation. The experimental design thus turned out to be quite sharp despite the rather mild inducement mechanisms. Given these compliance rates, we opt to estimate the intention-to-treat (ITT) effects.[5]

[5] The appendix contains the results from using alternative estimation methods, including instrumental variables estimation.

4 Results

4.1 Outcomes and estimation

The main outcomes are the distance of the trip and the asking price. A GPS tracker (in a smart phone) was used to record the distance of the trip. The assistant started the GPS at the beginning of the trip, casually giving the appearance that he was using the smart phone to send text messages.

Consider the following regression model for trip i and route j:

$$\text{observed distance}_{ij} = \delta_0 + \delta_1 T_{ij} + m_j + \varepsilon_{ij}. \tag{18}$$

where m_j is a route-specific effect, and ε_{ij} is an error term. T_{ij} is a vector of assigned treatment dummies. By design, T_{ij} is orthogonal to both m_j and ε_{ij}, so putting m_j in the error term and using OLS gives us consistent estimates. However, maximum precision is achieved by estimating random effects at the route-level with generalized least squares. Alternatively, one could use fixed route effects.

An efficient estimation is achieved by noting that most of the explanatory power of the traveled route stems from the actual (map) distance between the departure and destination ranks. Therefore, we consider the following regression equation:

$$\text{observed distance}_i - \text{map distance}_j = \delta_0 + \delta_1 T_{ij} + \varepsilon_{ij}. \tag{19}$$

In (19), the dependent variable is the difference between the actual distance traveled (as measured by the GPS) and an exogenous reference distance. Henceforth, we shall refer to this difference as the "excessive driving". Google Maps was used to obtain a reference estimate of the map distance between each departure and destination rank. As it turns out, the map distance explains more than two-thirds of the variation in the observed distance. Thus, the regression estimates using the difference specification will have considerably less unexplained variation than the level estimates. Notably, because the map distance is observed and not estimated, specification (19) will not punish the estimation by reducing the degrees of freedom, as opposed to using dummies to estimate the fixed route effects.

Analogously, we estimate a model with the price as the dependent variable while accounting for the variation in metered fares across routes:

$$\text{asking price}_i - \text{imputed price}_j = \gamma_0 + \gamma_1 T_{ij} + \varepsilon_{ij}, \tag{20}$$

where

$$\text{imputed price}_j = R2 + R10 \times \text{map distance}_j. \tag{21}$$

The dependent variable in (21) can be interpreted as the asking price relative to the most commonly advertised price. This variable is a measure of price excessiveness.

A challenge to our attempts to measure the price was the tipping factor. Typically, Cape Town cab drivers expect to be tipped at the end of the ride. Had this expectation been equal across the treatment and control groups, we could have safely ignored the tipping factor in the empirical analysis. The crux is that the metered fare more often ended on an odd figure, whereas the fixed fare almost always resulted in an even multiple of R10. During the pilot, it became clear that the drivers did not expect to return any change for the ride within a multiple of R10. The assistant was instructed to round up the fare to the nearest R10 (regardless of treatment) before paying.

It is not obvious that the empirical analysis should be adjusted by the tipping amount. Despite its prevalence, tipping is difficult to rationalize in conventional strategic bargaining games (tipping is not consistent with a subgame perfect equilibrium in one-shot games). Moreover, although the custom of rounding up the fare is arguably common, it is not certain that the rule used in the experiment represents the tipping convention in Cape Town. Therefore, the empirical analysis will use the "asking price", which does not include any tips. The "end price", equal to the price that the customer actually paid in the end, will inform the main analysis in supplementary regressions.

4.2 The costs of regulations

The raw, unadjusted averages across the treatment groups are presented in Table 4. The trips are approximately half a kilometer longer under the regulated fare treatments, which represent approximately ten percent relative to the map reference. Although the raw estimates of distance are imprecisely estimated, the efficiency gained by subtracting the map distance from the outcome is considerable. The difference specifications are therefore precisely estimated.

Figure 2 displays graphical evidence of the treatment effect. The figure displays the estimated probability density functions of excessive driving across the main treatment groups. Nonparametric tests (Wilcoxon rank-sum) show that the two distributions are significantly different from each other, and the shape of the distributions suggests that the treatment effect is more pronounced around the means.

Table 4: Main results. Experimental evidence of excessive driving under regulated and unregulated fares. Raw averages and t-tests.

	Means			
	Unregulated fare (fixed price)	Regulated fare (metered fare)	Difference	p-value
Observed distance	6.21	6.83	-0.62* (0.33)	0.061
Observed distance minus map distance	-0.01	0.56	-0.56*** (0.19)	0.003
Asking price	74.89	78.30	-3.41 (4.26)	0.425
Asking price minus advertised price	10.75	13.40	-2.65 (2.94)	0.369
Observations	88	88	176	

Notes: Estimates display uncorrected, raw means and t-tests of the difference across treatment group. Standard errors in parenthesis. Metered fare is treatment group M1 or M2; fixed fare is treatment group F1 or F2. Two observations on distance are missing due to faulty GPS-recording. Prices in Rands

The graphical evidence does not seem to call for specifications other than the average effects. However, for brevity, Table 5 contains additional specifications. When a linear term for map distance is included in the regression, efficiency is further improved. Using random effects at the route level also produces more precise estimates, but fixed effects at the route level reduces precision due to the reduction in the degrees of freedom (although, notably, the estimates are somewhat smaller with the fixed route effects). Additional controls for the starting hour of the trip, the visibility of the meter in the car, the car type and the operator do not enhance precision once map distance has been controlled for. Overall, the different models do not yield statistically different estimates, and we do not expect them to. Given the weak assumptions underlying the causal interpretation of the treatment effects in randomized controlled trials, there are no reasons to put more emphasis on the regression evidence over the raw, nonparametric estimates displayed in Table 4 and Figure 2.

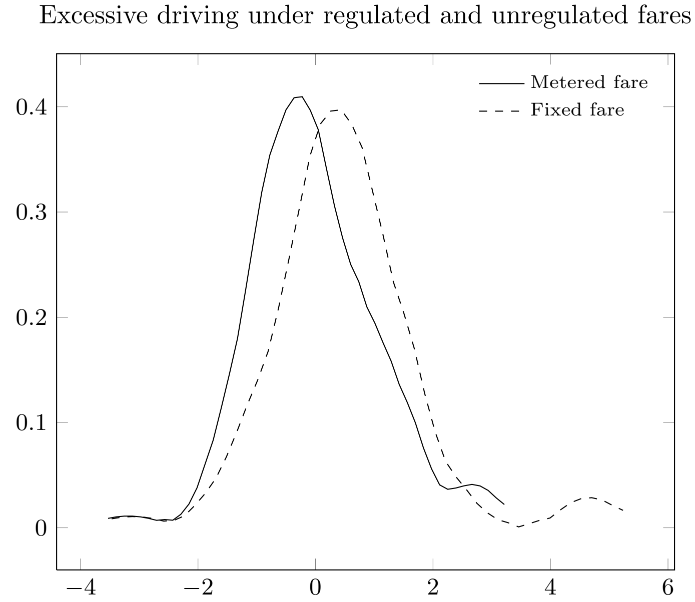

Figure 2: Figure displays the PDFs of the observed trip distance minus map distance in kilometers. The solid line represents the trips assigned to metered fares (Control groups M1 or M2, n=87); the dashed line represents the trips assigned to fixed fares (Treatment groups F1 or F2, n=87). The PDFs are estimated using kernel (Epanechnikov) density regressions. Wilcoxon rank-sum test of equal distributions: z=-3.160, p=0.0016.

The average distance under the fixed fare treatment is 6.21 kilometers, whereas the average distance under the metered fare treatment is 6.83 kilometers; a ten percent difference. However, given that the results presented thus far are ITT estimates, the true behavioral response to being forced to use the meter is likely to be somewhat higher. As a complement to the preferred estimates, the appendix presents results from per-protocol, on-treatment and instrumental variables estimations (Table B.2). The IV estimates are obtained by two stage least squares, with the intention to treat as an instrument for the probability of using the meter. Given the high compliance rate, the IV effects are only marginally larger than the intention-to-treat effects. The differences between the ITT, on-treatment, per-protocol and IV estimates are statistically insignificant.

The empirical findings are easily summarized. If the regulated fare is enforced, the drivers increase mileage. The main effects are both economically and statistically significant as well as consistent with the moral hazard incentives discussed in Section 2. Exactly how the drivers increase mileage is not revealed by the statistical figures alone. However, a visual inspection of the GPS recordings reveals that the drivers take detours when the metered is running. As an indicative example, Figure 3 depicts two recorded routes between Cableway Station and Waterfront Mall. When the meter is running, the drivers chose the sea route instead of going straight to the destination, as seen by the dark, northwestern GPS-recording. This particular comparison represents one of the more glaring examples in the data: the difference between the distances traveled is more than 4 kilometers.

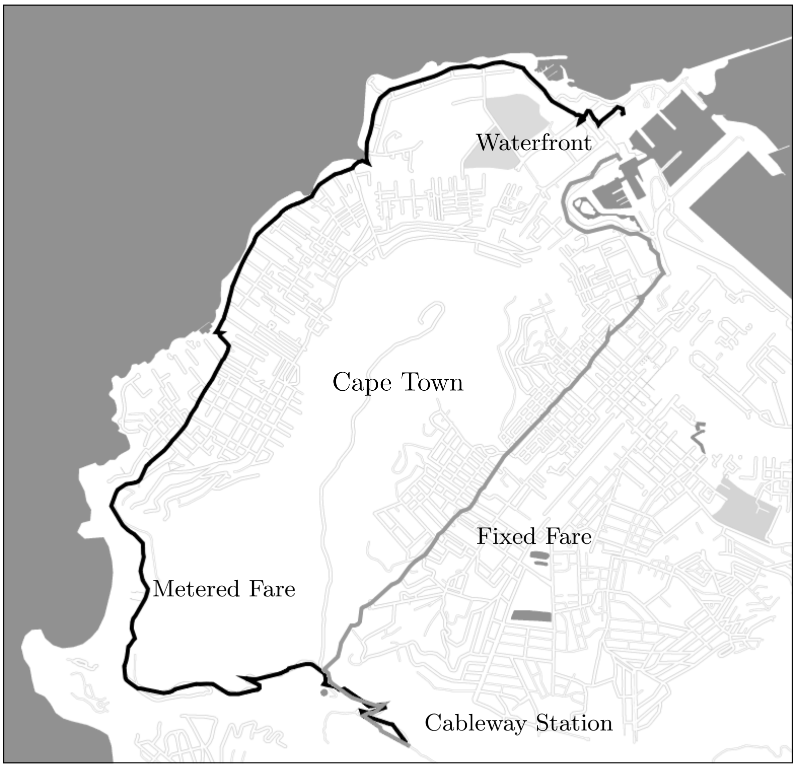

Figure 3: Illustration of the treatment effect. The sea route (solid black line) displays the GPS recording of a trip assigned to the metered fare (M2); the city route (grey line) displays the GPS recording of a trip assigned to the fixed fare (F1). The departure rank and destination were the same. The distance with the metered fare is 11.5 kilometers; the distance with the fixed fare is 7.3 kilometers.

Breaking up the main treatment groups into the subtreatments (Table 5) suggests that there are no significant differential subtreatment effects on distance. Comparing distance between M1 and M2, and between F1 and F2, suggest that it does not matter whether the customer makes counteroffers or actively requests the metered fare; all that matters are the "hard" incentives to increase distance. These results are expected. The theoretical discussion offered no explanation for why excessive driving would differ across subtreatments conditional on the main treatments. Under a fixed price, the incentives to inflate mileage are absent regardless of the price agreed upon before the trip takes place. Assigning a fixed fare with a counteroffer produces the lowest point estimate (F2; $\delta_1 = 0.607$), which differs by almost 0.13 kilometers from F1. However, testing for equality between F1 and F2 reveals that this difference is statistically insignificant (p-value = 0.633). Similarly, the metered fare treatments M1 and M2 differ only in the way they enforced compliance. Because the difference in compliance is small, there is only a small and insignificant differential impact on distance (p-value = 0.843).

Table 5: Experimental evidence of excessive driving under metered fares.

	(1) OLS	(2) OLS	(3) OLS	(4) RE	(5) FE	(6) OLS
Treatment						
Metered fare	-0.563***	-0.566***	-0.570***	-0.453***	-0.404**	
	(0.186)	(0.186)	(0.196)	(0.169)	(0.172)	
Map distance		-0.058	-0.066	-0.071		-0.057
		(0.047)	(0.047)	(0.059)		(0.047)
Subtreatments						
M1: Metered fare, baseline						
M2: Metered fare, probe for meter						0.053
						(0.294)
F1: Fixed fare, accept first offer						-0.480*
						(0.268)
F2: Fixed fare, with counteroffer						-0.607**
						(0.233)
Constant	0.556***	0.922***	1.298*	0.896**	0.477***	0.886**
	(0.147)	(0.334)	(0.729)	(0.414)	(0.086)	(0.396)
Obs.	174	174	174	174	174	174
Additional controls	No	No	Yes	No	No	No

Notes: All columns represents separate regressions. Dependent variable is observed distance minus map distance (in kilometers) in all regression. (1)-(3) OLS, (4) Random effects (GLS), (5) Fixed effects. In model (4) and (5) the group level is equal to the block randomization unit (47 routes). Additional controls are start hour of trip, visible meter in car, radio in car, car type (Sedan/Hatchback/Station Wagon), and independent operator. Robust standard errors in parenthesis. 174 observations (two missing values due to incomplete GPS-recording).

How costly is it to increase mileage? The private costs from driving longer are given by the price of fuel, capital depreciation, service costs and insurance fees. The external costs are mostly given by fuel consumption, which contributes to the emission of carbon-dioxide (CO2) in the atmosphere, and by congestion. Fuel consumption depends, in turn, on vehicle size, speed, and driving style. Importantly, both speed and driving style are affected by the treatment, which implies that a linear extrapolation of the costs might be misleading. For instance, it is possible that drivers choose a route that implies less stop-and-go driving when the meter is running, which could increase fuel efficiency (this is graphically suggested in Figure 3). On the other hand, one can also imagine that the metered fare induces firms to drive faster between stops in order to compensate for the longer route, which could decrease fuel efficiency. There are also important differences across car types that can be exploited in order to improve the precision of cost estimates.

The fuel consumption for each trip is estimated using the US Environmental Protection Agency's (EPA's) fuel efficiency calculations, based on a typical city route (the "Urban Dynamometer Driving Schedule"). EPA's estimates for models built in 2002 are matched with the experimental data based on car type (Midsize Station Wagon, Midsize Car, and Compact Car). The implied fuel consumption is then obtained using the GPS-information of distance traveled, adjusted to reflect that fuel efficiency is a hump-shaped function of average speed (West et al. 1999). From the implied fuel consumption, the costs of the trip are estimated using a petrol price equal to $ 1.5 dollar per liter and an additional service cost of $ 0.2 dollar per kilometer. Finally, one liter of consumed petrol equates to 2390 gram of CO2 emissions.

Table 6: Impact of regulated and unregulated fares on speed, fuel consumption and pollution in the Cape Town taxi market.

	(1)	(2)	(3)	(4)	(5)
	Time	Speed	Fuel con-	Private	CO2
	(seconds)	(km/h)	sumption	costs	emissions
			(liter)	(Rand)	(kilo)
Treatment: Fixed, unregulated fare	-43.548	-1.204	-0.057***	-1.390***	-0.136***
	(31.171)	(0.995)	(0.020)	(0.464)	(0.048)
Map Distance	65.053***	1.825***	0.094***	2.308***	0.225***
	(8.848)	(0.349)	(0.005)	(0.118)	(0.012)
Constant	398.323***	19.534***	0.109***	2.434***	0.260***
	(54.157)	(2.706)	(0.036)	(0.830)	(0.085)
Control: Metered fare	806.75	30.99	0.70	16.92	1.67
Obs.	174	174	174	174	174

Treatment is equal to one if the meter regulations are sidestepped and a fixed fare is renegotiated (subtreatment F1 or F2), zero otherwise. Robust standard errors in parenthesis. Private costs includes fuel and service costs.

The results are displayed in Table 6. The private costs are reduced by 8.2 percent relative the control group mean, and CO2 are reduced by 8.1 percent relative the control group mean. There are no statistically significant effects on the speed of the trip.

4.3 Holdups and ex ante efficiency

Thus far, the evidence implies that a strict enforcement of the regulations increases material and environmental costs. However, the point of most regulatory policies, and the central message in the theoretical section, is that the regulations increase the predictability of the price, which reduces the threat of holdup. In this subsection, we study how the final price was affected by decentralized bargaining.

The results in Table 4 and Figure 4 show that the asking price is quite similar across the main treatment groups. That is, the asking price is slightly higher when the metered fare is used, although the difference is statistically insignificant. Table 7 shows that, although the price under F1 is equal to the metered fares M1 and M2, the renegotiated price with a counteroffer (F2) is somewhat lower than the metered fares. The treatment effect under F2 is not statistically different from the baseline (M1), but it is statistically different from F1 when random or fixed effects estimation is employed.

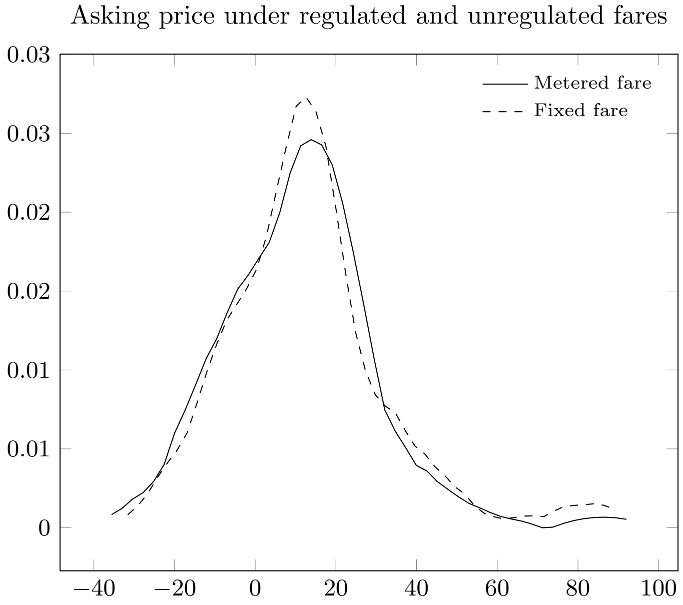

Figure 4: Figure displays the PDFs of the actual (empirical) asking price relative to the regulated fare in South African rands. The regulated price is imputed by using (21). The solid line represents the trips assigned to metered fares (treatment groups M1 or M2, n=87); the dashed line represents the trips assigned to fixed fares (treatment groups F1 or F2, n=87). The PDFs are estimated using kernel (Epanechnikov) density regressions. Wilcoxon rank-sum test of equal distributions: z=-0.431, p=0.6668.

As a supplement, table B.1 displays the estimates based on the price actually paid by the assistant. This “end price” equals the asking price rounded up to the nearest multiple of R10 (regardless of treatment). According to this definition, the renegotiated price with a counteroffer (F2) is significantly lower than the baseline (M1). The effects of the other three treatments are statistically the same, and the differences are small in magnitude. The tipping rule used by the assistant was implemented for practical reasons and is rather arbitrary. It is not certain that the typical passenger abstains from tipping simply because the fare is determined in advance. Therefore, the effects in Table B.1 can be said to represent the upper bound of the profitability of using the metered fare.

The combined results imply that the customer is equally well off by asking for a fixed price and slightly better off by making counteroffers. Customers never become worse off by renegotiating a fixed price. On average, firms thus appear to be equally well off when the customer is making counteroffers, and slightly better off when customers accept the first offer. There is no evidence of the Diamond paradox: customers nor firms are not held up once they sidestep the regulated contracts. The exact impact on profits depends on how much emphasis we put on the direct,

Table 7: Asking price relative advertised price across subtreatments.

	(1) OLS	(2) OLS	(3) RE	(4) FE
M1: Metered fare, no intervention (baseline)				
M2: Metered fare,	1.499	1.268	0.245	-1.197
ask for meter	(4.439)	(4.399)	(2.555)	(2.281)
F1: Fixed fare,	0.386	0.553	0.997	1.289
accept first offer	(3.917)	(3.902)	(3.311)	(3.367)
F2: Fixed fare,	-4.398	-4.654	-4.208	-5.151
with counteroffer	(3.858)	(3.861)	(3.496)	(3.529)
Map distance		0.886		
		(0.792)		
Constant	12.706***	7.242	11.955***	13.287***
	(2.701)	(4.952)	(2.548)	(1.784)

Notes: Prices are in South African Rands. All columns represents separate regressions. Dependent variable is the final price minus an exogenous imputed price based on the advertised kilometer fare and the map distance. All regression controls for departure rank and the map distance. Robust standard errors in parenthesis. 176 observations.

observable costs. Looking only at the private costs, the investigation in Table 6 suggests that the reduction in revenues (Table 4) is offset by the reduction in costs to about 50 percent. However, this is not taking into account time loss nor the environmental externality arising from excessive driving, nor the hazards from driving longer routes, which imply additional penalties from complying with the regulations. The combined results indicate that we cannot falsify the "efficient informal trade"-hypothesis, suggesting the regulations does indeed frame the informal bargaining process, at least in an average sense.

5 Heterogeneity: When are regulations harmful?

The theoretical framework in Section 2 depicted a tradeoff between rigidity and flexibility on regulated markets. This tradeoff is implicit in most theoretical work on incomplete contracts, at least since the work of Williamson (1975). Section 2 concluded with the message that the equilibrium outcome in markets with incomplete regulations could still be constrained Pareto-efficient, in case the regulations serve as threat points in the informal bargaining process, and the previous section provided empirical evidence of efficiency – at least when using the full sample.

The efficiency result is however an equilibrium concept, based on a simple representative-agent model with a specific bargaining strength parameter (β). While the empirical results in Section 5 gives support for the underlying forces behind the proposition in Section 2, they also highlight that the empirical reality does not fully correspond to the model assumptions. For instance, in the experiment we deliberately varied the intensity of the negotiations (β) even though this was presumed to be a fully predictable constant in the model. Additionally, the fact that spontaneous renegotiations appears to occur too seldom suggests that there is some exogenous, non-modelled cost preventing efficient renegotiations from taking place.

With parameter heterogeneity (i.e. different bargaining strengths, β_i), a simple unit price contract is not likely to reproduce the efficient model outcome described in Section 2 for every transaction. In such cases, the outcome of the bargaining process does not guarantee that the parties' market entry requirements are met for all customers and firms with positive match productivity –

which means that the fundamental tradeoff between rigidity and flexibility is still relevant. This heterogeneity can explain why some customers or drivers abstain from renegotiating a fare even in the presence of Pareto gains.

According to equation (38) the importance of preference heterogeneity is given by the strength of moral hazard incentives – the ability to exploit detours that are long enough to improve profits, but short enough to be undetectable by the passengers. This informational asymmetry is the (exogenous) difference between c_H and c_L in the model – the "observable but unverifiable" – , ultimately the threat point in equation (35). It is instructive to consider the special case of $c_H = c_L$. In this case, there is no information asymmetry and no moral hazard incentives for firms (all actions are "verifiable" in a court of law), so the end price under regulations-as-options (38) will coincide with the socially optimal price regardless of whether β_i varies across customers. In such markets, we expect the renegotiated price to be invariant to counter-offers because there are essentially no threat points. By contrast, if c_H is large, even slight individual deviations in bargaining strength $(\beta_i - \bar{\beta})$ can lead to holdups (as seen in equation 38).

One can thus use variation in the "observable but nonverifiable", $c_H - c_L$, as a way to describe the external validity of the estimates. In this empirical context the observable but nonverifiable is defined as "the possibility of taking detours without getting caught". Identifying this variable empirically does not only require information of possible detours but of possible *reasonable* detours. To do this in a transparent and replicable way, I again make use of Google Maps to define alternative routes for each departure-destination pair in the data. Google Maps uses layers of map data (one layer for highway data, another for smaller streets and alleys, on so on) to calculate expected travel times. From this information, Google Maps sometimes offers alternate directions. For some departure and destination choices, there is only one reasonable route suggestion, whereas for others there are several alternative routes which appear adequate. To calculate $c_H - c_L$ for every route, I use the longest route suggestion minus the shortest route suggestion and interpret this difference as the distance that can be exploited by the cabdriver to inflate mileage (with a cap at 2 kilometers to remove a few clearly unreasonable routes). This variable is then used as a binary indicator for whether an alternative route exists, that is, whether $c_H > c_L$.

From the viewpoint of statistical identification, the indicator for alternative routes is orthogonal to the treatments (which were balanced across routes) and, conditional on departure rank, orthogonal to cabdriver-specific features. The latter orthogonality follows from the fact that the route schedule was randomly assigned, so a cabdriver waiting at a rank at, e.g., the Green Garden center was as likely to be assigned to the Sea Point rank as Upper Long Street.[6] Looking for differential responses to fixed and metered fares across the threat point proxy is thus a relatively clean way of addressing to what extent information asymmetries interacts with the regulatory environment.

The results from interacting the treatments across the possibility of taking a detour, displayed in Table 8, tell an intuitive yet important story. When there are few alternative routes there are no distance-gains from sidestepping the meter. The effect on distance travelled is thus driven by the observations for which an alternative route exists. Although this effect is not very surprising, it clarifies the simple incentives involved – that taxis cannot simply drive around in circles to push up the metered fare indefinitely. A more notable result, however, is how the renegotiation affects the price across routes with and without alternative routes. When moral hazard incentives are low – that is, when no alternative routes exists – the price is insensitive to counteroffers. This

[6] However, the threat point variable is not necessarily orthogonal to route-specific characteristics, such as congestion. Note, however, that I control for map distance in all regressions.

Table 8: Moral hazard and the predictability of renegotiations: Subsample analysis across alternative routes

	Dep var: Distance		Dep var: Price	
Alternative route exists?	Yes	No	Yes	No
M1 - Metered fare, no intervention (baseline)				
M2 - Metered fare,	-0.404	0.800*	-4.961	9.129
ask for meter	(0.323)	(0.425)	(4.503)	(6.984)
F1 - Fixed fare,	-0.863**	0.169	-4.324	-2.688
ask for meter	(0.337)	(0.390)	(4.704)	(6.414)
F2 - Fixed fare,	-1.041***	0.133	-16.124***	-3.955
with counteroffer	(0.357)	(0.412)	(4.954)	(6.767)
Obs.	85	89	86	90

Notes: Columns All columns represents separate regressions. Dependent variables are distance travelled in kilometers and the final price, rounded up to the nearest R10. (1)-(2) OLS, (3) Random effects (GLS), (4) Fixed effects. In models (3) and (4) the fixed and random effects are equal to the block randomization unit (47 routes). Robust standard errors in parenthesis. 176 observations.

is completely in line with the theoretical model which says that β is irrelevant since there is no surplus to share. By contrast, when an alternative route exists, the outcome of the renegotiation becomes much less stable – a simple counteroffer can significantly drive down the price. This explains why some cabdrivers appear reluctant to engage in bilateral bargaining ex post, despite the Pareto gains.

The results of Table 8 captures the heart of the fundamental transformation of markets. With complete information, there are no ex post efficiencies from renegotiating the fare, but no ex ante losses either – the end price is predictable. With incomplete information, such as the possibility for firms to take detours, there are ex post gains to extract from renegotiation but these gains come at the expense of less price predictability, depending in this case on the simple renegotiation strategy adopted by the customer. These supplementary results leads to the rather dismal conclusion that regulations are most needed precisely in those markets where they can be most harmful.

6 Concluding remarks

Heavily regulated markets in developing countries are often associated with large informal sectors and low economic output. This paper uses a randomized field experiment on a sample of regulated taxi firms to test whether sidestepping formal entry regulations has a causal effect on efficiency. I find that sidestepping the formal regulations reduces moral hazard and improves ex post efficiency. The empirical results are intuitive. When the regulations are enforced, cabdrivers deliberately increase the costs of production (in this case, by inflating mileage). By contrast, if regulations are sidestepped, drivers take the more direct route, and the costs are minimized. Strict enforcement of the regulations reduces the size of the surplus by 10 percent and the increases the CO2 emissions by 8 percent.

The results appear to provide a case for deregulation. However, I theoretically show that regulations that are never strictly enforced in equilibrium can still frame the informal trade process so that holdups are prevented. In keeping with this prediction, I find empirically that the end price is quite insensitive to the transition bilateral bargaining – on average, no part is worse off from renegotiating the fare. Thus, while the empirical results point to important costs of regulations, the results cannot be interpreted as a general case in favor of complete deregulation.

An important question is to what extent regulated contracts *are* in fact renegotiated when it is mutually beneficial to do so. In this particular setting, some cabdrivers appear to prefer the metered fare. I explain this behavior by introducing parameter heterogeneity. With parameter heterogeneity, a simple unit price contract is not likely to reproduce the efficient model outcome described in Section 2 in every transaction. Even though the size of the cake would increase under informal bargaining, some customers or firms might find it optimal to opt out of such bargaining because (rationally) expect to be shortchanged. In such cases, the regulator does face a tradeoff between rigid regulations (improving ex ante investments) and flexible regulations (improve ex post efficiency). My supplementary empirical results, using quasi-experimental variation in the extent information asymmetries, confirm this tradeoff. In market with strong moral hazard incentives (large threats of fraudulent behavior) informal trade can lead to significant ex ante inefficiencies. However, in such markets, stronger enforcement of incomplete regulations is unlikely to be the preferred policy compared to reforms that more directly address the information asymmetry.

References

Acemoglu, Daron and Robert Shimer (1999) “Holdups and Efficiency with Search Frictions,” *International Economic Review*, Vol. 40, No. 4, pp. 827–49.

Aghion, Philippe, Mathias Dewatripont, and Patrick Rey (1994) “Renegotiation Design with Unverifiable Information,” *Econometrica*, Vol. 62, No. 2, pp. pp. 257–282.

Balafoutas, Loukas, Adrian Beck, Rudolf Kerschbamer, and Matthias Sutter (2011) “What drives taxi drivers? A field experiment on fraud in a market for credence goods,” Working Papers 2011-11, Faculty of Economics and Statistics, University of Innsbruck.

Bardhan, Pranab (1997) “Corruption and Development: A Review of Issues,” *Journal of Economic Literature*, Vol. 35, No. 3, pp. pp. 1320–1346.

Besley, Timothy and Robin Burgess (2004) “Can Labor Regulation Hinder Economic Performance? Evidence from India,” *The Quarterly Journal of Economics*, Vol. 119, No. 1, pp. 91–134.

Castillo, M., R. Petrie, M. Torero, and L. Vesterlund (2012) “Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination,”Technical report, National Bureau of Economic Research.

Chung, Tai-Yeong (1991) “Incomplete Contracts, Specific Investments, and Risk Sharing,” *The Review of Economic Studies*, Vol. 58, No. 5, pp. pp. 1031–1042.

City of Cape Town (2007) “Operating Licences Strategy,” Policy document, Transport Department, City of Cape Town.

Darby, Michael R and Edi Karni (1973) “Free competition and the optimal amount of fraud,” *Journal of law and economics*, Vol. 16, No. 1, pp. 67–88.

de Soto, Hernando (1989) *The Other Path*: New York, NY: Harper and Row, 1989.

Diamond, Peter A. (1971) “A model of price adjustment,” *Journal of Economic Theory*, Vol. 3, No. 2, pp. 156–168.

Djankov, Simeon, Rafael La Porta, Florencio Lopez-de Silanes, and Andrei Shleifer (2002) "The Regulation of Entry," *The Quarterly Journal of Economics*, Vol. 117, No. 1, pp. 1–37.

Dulleck, Uwe and Rudolf Kerschbamer (2006) "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods," *Journal of Economic Literature*, Vol. 44, No. 1, pp. pp. 5–42.

Edlin, Aaron S. and Stefan Reichelstein (1996) "Holdups, Standard Breach Remedies, and Optimal Investment," *The American Economic Review*, Vol. 86, No. 3, pp. pp. 478–501.

Ellingsen, Tore and Magnus Johannesson (2004) "Is There a Hold-up Problem?" *Scandinavian Journal of Economics*, Vol. 106, No. 3, pp. 475–494.

Fisher, Ronald Aylmer (1926) "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture*, Vol. 33, pp. 503–513.

Goldberg, Victor P. (1976) "Regulation and Administered Contracts," *Bell Journal of Economics*, Vol. 7, No. 2, pp. 426–448.

Grether, David M., Alan Schwartz, and Louis L. Wilde (1988) "Uncertainty and Shopping Behaviour: An Experimental Analysis," *The Review of Economic Studies*, Vol. 55, No. 2, pp. pp. 323–342.

Grout, Paul A (1984) "Investment and Wages in the Absence of Binding Contracts: A Nash Bargining Approach," *Econometrica*, Vol. 52, No. 2, pp. 449–60.

Habyarimana, J. and W. Jack (2011) "Heckle and Chide: Results of a randomized road safety intervention in Kenya," *Journal of Public Economics*, Vol. 95, No. 11, pp. 1438–1446.

Hackett, Steven C (1994) "Is Relational Exchange Possible in the Absence of Reputations and Repeated Contact?" *Journal of Law, Economics and Organization*, Vol. 10, No. 2, pp. 360–89.

Harrison, Glenn W. and John A. List (2004) "Field Experiments," *Journal of Economic Literature*, Vol. 42, No. 4.

Hosios, Arthur J. (1990) "On the Efficiency of Matching and Related Models of Search and Unemployment," *The Review of Economic Studies*, Vol. 57, No. 2, pp. pp. 279–298.

Huberman, Gur and Charles Kahn (1988) "Limited Contract Enforcement and Strategic Renegotiation," *The American Economic Review*, Vol. 78, No. 3, pp. pp. 471–484.

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

Iyer, Rajkamal and Antoinette Schoar (2010) "Incomplete Contracts and Renegotiation: Evidence from a Field Audit," MIT LFE Working Paper No. LFE-0713-10.

Keniston, D. (2011) "Bargaining and welfare: A dynamic structural analysis of the autorickshaw market," *Yale University, mimeo. September.*

Klein, Benjamin, Robert G Crawford, and Armen A Alchian (1978) "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law & Economics*, Vol. 21, No. 2, pp. 297–326.

Laffont, Jean-Jacques and Jean Tirole (1986) "Using Cost Observation to Regulate Firms," *Journal of Political Economy*, Vol. 94, No. 3, pp. pp. 614–641.

Lowitt, S. (2006) "The Job-creating Potential of the Metered Taxi Industry in South Africas Urban Areas: Some Preliminary Findings," Employment-oriented Industry Studies, HSRC.

MacLeod, W. Bentley and James M. Malcomson (1993) "Investments, Holdup, and the Form of Market Contracts," *The American Economic Review*, Vol. 83, No. 4, pp. pp. 811–837.

Maskin, Eric and Jean Tirole (1999) "Unforeseen Contingencies and Incomplete Contracts," *The Review of Economic Studies*, Vol. 66, No. 1, pp. pp. 83–114.

Moen, Espen R. (1997) "Competitive Search Equilibrium," *Journal of Political Economy*, Vol. 105, No. 2, pp. pp. 385–411.

Mortensen, Dale T. (1982) "Property Rights and Efficiency in Mating, Racing, and Related Games," *The American Economic Review*, Vol. 72, No. 5, pp. pp. 968–979.

Myrdal, Gunnar (1968) *Asian Drama: An Inquiry into the Poverty of Nations*: New York: Pantheon Books.

Nöldeke, Georg and Klaus M. Schmidt (1995) "Option Contracts and Renegotiation: A Solution to the Hold-up Problem," *The RAND Journal of Economics*, Vol. 26, No. 2, pp. 163–179.

Osborne, Martin J. and Ariel Rubinstein (1990) *Bargaining and Markets*: London: Academic Press.

Pissarides, C. A. (1984) "Efficient Job Rejection," *The Economic Journal*, Vol. 94, pp. pp. 97–108.

Rose-Ackerman, Susan (2003) "Corruption," in Charles K. Rowley and Friedrich Schneider eds. *The Encyclopedia of Public Choice*: Springer US, pp. 67–76.

Shimer, Robert J. (1996) "Contracts in a Frictional Labor Market," Mimeo, Department of Economics, Massachusetts Institute of Technology.

Svensson, Jakob (2005) "Eight Questions about Corruption," *The Journal of Economic Perspectives*, Vol. 19, No. 3, pp. 19–42.

West, B.H., R.N. McGill, J.W. Hodgson, S.S. Sluder, and D.E. Smith (1999) "Development and Verification of Light-Duty Modal Emissions and Fuel Consumption Values for Traffic Models," Department of Transportation, Federal Highway Administration, Washington, DC, March 1999.

Williamson, Oliver (1975) *Markets and Hierarchies: Analysis and Antitrust Implications*: New York: The Free Press.

Williamson, Oliver E. (1979) "Transaction-Cost Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, Vol. 22, No. 2, pp. pp. 233–261.

——— (1983) "Credible Commitments: Using Hostages to Support Exchange," *The American Economic Review*, Vol. 73, No. 4, pp. pp. 519–540.

Williamson, Oliver (1985) *The Economic Intstitutions of Capitalism*: New York: The Free Press.

A Equilibrium framework

A.1 Set up

In this conceptual framework, I discuss the regulator's tradeoff between flexibility and rigidity. Although the model is meant to illustrate the typical taxi market, it is quite general. It draws established ideas from contract theory and search theory (see the end of this section for a literature review). The need for rigid regulations arises from fixed entry costs on both sides of the market, which are sunk at the time customers and sellers meet. The need for flexibility arises because it is not possible to completely regulate the ex post relationship. Herein lies the regulator's tradeoff.

Taxi drivers produce a service worth y to the customer, which is the value of traveling to the destination (instead of, say, walking). To produce the service, taxis use inputs captured by the term $c > c_L$, where c_L is the minimum costs required to produce the service. Trade cannot occur instantly, however. Both customers and firms must make ex ante investments before they can enter the market, which are sunk at the time customers and firms meet. In the taxi market, it is natural to think about these ex ante investments as the cost of search (or waiting time). The driver's exogenous cost of entering the market is k, and the customer's exogenous cost is b.

Firms and customers are matched according to a matching function $m = m(v, n)$, which is homogenous to degree one. v is the share of unmatched firms, and n is the share of unmatched customers. The probability that a firm finds a customer is equal to $q(\theta) = m(v, n)/v$, where $\theta = \frac{v}{n}$ is equal to market "tightness". Conversely, due to the homogeneity of the matching function, the probability that a customer finds a firm is equal to $\theta q(\theta) = m(v, n)/n$. All matched trade relationships are bilateral (one customer, one firm).

The customer will direct his or her search to one out of at least two submarkets and enter the market with the lowest expected costs. The market entry condition is

$$R \geq \theta q(\theta)(y - p) - b \tag{22}$$

where R is the expected cost of entering any other submarket. The number of submarkets are exogenously determined and the total population is fixed.[7] In equilibrium, this means that all submarkets will offer the same price, so (22) will hold with equality.

For the firm, the value of entering the market is equal to

$$V = q(\theta)(p - c) - k. \tag{23}$$

Free entry implies that $V = 0$ in equilibrium such that

$$q(\theta) = \frac{k}{(p - c)}. \tag{24}$$

The endogenous variables are share of unmatched firms v, costs c and the total price p.

A.2 Social efficiency

Social efficiency is characterized by the solution to the social planner's problem of maximizing social welfare in this economy. All customers in this market engage in search and are evenly spread

[7] Only two firms need to compete for an efficient competitive (Bertrand) equilibrium (Acemoglu and Shimer 1999).

across submarkets according to (22). The condition for social efficiency is thus found by asking how many available firms (v) the planner would prefer, given a fixed population (fixed n):

$$\begin{aligned}\max_{v,c} \quad U &= Rn + Vv && (25)\\ &= n\theta q(\theta)(y-p) + vq(\theta)(p-c) - vk - nb && (26)\\ &= vq(\theta)(y-c) - vk - nb. \end{aligned}$$

Notably, the planner puts the same utility weight on firms and customers and does not care about the division of the surplus (that is, the price). The solution to (25) is $c = c_L$ (costs trivially bind at the lowest possible level) and

$$q(\theta) = \frac{k}{(y-c_L)(1-\eta)} \tag{27}$$

where η is a positive constant equal to (the negative of) the elasticity of $q(\theta)$:

$$\eta = -\frac{\partial q(\theta)}{\partial \theta}\frac{\theta}{q(\theta)}. \tag{28}$$

Equation (27) defines the socially efficient level of search frictions in this economy.

A.3 Unregulated bargaining

We now ask under what regulatory regimes the market will internalize the search externalities and minimize costs. Absent regulations, once the two parties have matched and formed a bilateral trade relationship, the price p is determined through bargaining. Firms are residual claimants on productivity so $c = c_L$. The price is given by the solution to the Nash bargaining problem, defined as:

$$\max_p (y-p)^{\beta}(p-c_L)^{1-\beta} \tag{29}$$

Notably, since the ex ante investments k and b are sunk, they will not be internalized in the price. The final price equals

$$p = \beta c_L + (1-\beta)y. \tag{30}$$

Since the final price will be independent of k and b, nothing guarantees that the trading parties will be compensated for their ex ante search investments. If customers have low bargaining power ($\beta = 0$) the consumer surplus (22) will be negative. Conversely, if customers have too high bargaining power ($\beta = 1$) the producer surplus (23) is negative. The problem is due to what Williamson (1985) calls the "fundamental transformation" of markets: the transition from competition to bilateral trade relationships. At the decentralized bargaining stage, the ex ante investments are sunk, which creates a holdup problem. Even though there are mutual benefits from trade, the market does not clear.

If the bargained price happens to lie on a point between the two parties' reservation prices, trade will occur but the search frictions in the market will in general not be efficient. Combining

(30) and the free entry formula (24), the competitive level of tightness is given by

$$q(\theta) = \frac{k}{(y - c_L)(1 - \beta)}. \tag{31}$$

Equation (31) says that an unregulated market will be efficient only under the so-called Hosios-condition (Hosios 1990), that is, when $\beta = \eta$. Both β and η are exogenous constants. Thus, even under competition, there are no market forces assuring this condition holds when the price setting is completely decentralized.

A.4 Rigid and flexible regulations

We will now study how governmental price regulations affects the holdup situation. A common regulation in taxi markets, and indeed in most markets, is that firms are required to post and commit to unit price contracts related to the costs of production ("one dollar per hour", "one dollar per mile", "one dollar per kilo", etc.). In an urban taxi market, this is the familiar *metered fare.* Such a price policy appears promising because it forces firms to commit to the terms of trade ex ante. However, metered fares are similar to procurement contracts (Laffont and Tirole 1986), which are incomplete as they only specify the relationship between the costs and the price, c and p, but omits cost minimization (i.e. that $c = c_L$) from the contract. As such, they might severely circumscribe the gains from trade by creating moral hazard incentives to inflate costs c.

To see this, consider the case where the regulations are strictly enforced, meaning that governments have some technology that allows them to oversee that the unit price is binding ex post (such as installing sealed taximeters in cabs). At the moment of the match, firms and customers take the advertised price p_c as given, so the only variable factor is c. Ex post profits (leaving aside the sunk cost k) is thus $p_c c - c$, and firms have no incentives to decrease costs c as long as $p_c > 1$. While the customer might observe c and c_L, and might infer that costs are not minimized, he cannot prove this in a court of law – being unproductive is not illegal, at least within bounds. Thus, using the jargon of the incomplete contract literature, the problem is that c_L is "observable but nonverifiable".

Suppose firms can increase costs up to a certain level $c = c_H$, after which the costs would be verified as "excessive" in a court of law. The final price will then depend on the structure of the ex ante market. Since firms are residual claimants on productivity, they cannot commit to cost minimization and the final price p will equal $p = pc_H$.

Although costs are unlikely to be minimized, the strictly regulated price will internalize search frictions because the regulations will bridge the competitive stage and the price-setting stage. Following the competitive search literature (Moen 1997; Shimer 1996), the equilibrium unit price p_c is found by maximizing (23) with respect to p_c at $c = c_H$ and $p = p_c c_H$, subject to the customer's entry condition (22). That is, the firms maximize ex ante profits, taking into account that a higher price will increase ex post profits but also attract less customers to their submarket. The solution is

$$p \quad = p_c c_H = \eta c_H + (1 - \eta) y. \tag{32}$$

Using (32) and (24) to solve for $q(\theta)$, frictions are given by

$$q(\theta) = \frac{k}{(y - c_H)(1 - \eta)}. \tag{33}$$

Comparing (33) with (27), strictly enforced price contracts will not maximize social welfare. The reason is that the costs will be too high $c_H > c_L$.

A special case occurs if customers cannot verify productivity at all. As $c_H \to y$, firms monopolize all consumer surplus from trade. However, customers anticipate this, which implies that firms must lower the per-unit price p_c in order to attract any customers to their submarket. The competitive unit price p_c will therefore tend to unity. As this happens, the surplus from trade approaches zero in equilibrium. The outcome is similar to the paradoxical result found in Diamond (1971) except that market failure is double-sided in this case. Even if the entry costs k and s are trivial, neither the firms nor the customers find it worthwhile to enter the market.

The above scenario can be seen as a case of regulatory failure. In an attempt to solve the problems associated with search frictions and holdups, the regulator reduces the size of the cake without assuring that the allocation of the cake will guarantee that trade occurs. The problem lies in the incomplete nature of the regulations. The (perhaps obvious) solution would be to force firms to post a complete price contract, relating p_c to y instead of c, but let us suppose this is not possible.[8] However, there is a regulatory environment that achieves both ex ante and ex post efficiency, even when the regulator cannot dictate the ex post allocation. The key is to construct the regulations in such a way that the incomplete price regulation becomes an *option contract*, that stipulates the fallback positions in case the renegotiation breaks down.

Under the regulations-as-options perspective, the trading parties will sidestep the regulations and engage in informal bilateral bargaining when they meet (notice that the model is agnostic about whether doing so is illegal or not). The solution to the regulations-as-options case is found by backwards induction. First, we ask what the end price p is, given the posted price p_c. The solution to the bargaining problem is given by

$$\max_p (y - p - [y - p_c c_H])^\beta (p - c_L - [p_c c_H - c_H])^{1-\beta}, \tag{34}$$

where the threat point terms (in brackets) have been added to the bargaining problem in (29). If the bargaining is successful and a fixed price is agreed upon before the costs are realized, the regulated price p_c is no longer dictating the outcome. In this case, the firm have no incentive to maximize costs, so $c = c_L$. By contrast, in case bargaining fails, the cabdriver will turn on the meter and maximize costs ($c = c_H$).

The solution to (34) is

$$p = p_c c_H - \beta(c_H - c_L). \tag{35}$$

Both customers and firms correctly anticipate that (35) will determine the final price. Therefore, firms maximize profits (23) with respect to p_c, subject to both the demand constraint (22) *and*

[8] In the taxi market, this means that the taxis post a complete list of fixed departure and destination sites (like a bus shuttle company). More generally, exactly why incomplete price contracts exists has been an active area of contract theoretical research since Maskin and Tirole (1999) argued that uncertainty alone is not a satisfactory explanation as long as agents are forward-looking and rational. One possible explanation for incomplete regulations is that it is simply too costly to oversee and monitor more sophisticated regulations (i.e. a menu cost-problem). A more convincing argument, however, is that that the regulator might not *need* to provide more complete regulations. The theoretical point made in this section is that the resulting allocation of the trade surplus can be efficient anyway.

the bargaining outcome (35). The per unit price is thus given by

$$p_c = \frac{y(1-\eta) + \eta c_L + \beta(c_H - c_L)}{c_H} \tag{36}$$

Using expression (36) in (35), the end price equals

$$p \quad = p_c c_H = \eta c_L + (1-\eta)y. \tag{37}$$

Using (37) and (24) we see that the market equilibrium level of tightness $q(\theta)$ coincides with the condition for the socially optimal level of tightness (27). The regulations-as-options solution thus guarantees that trade occurs *and* that the costs are minimized ex post.

The option contract perspective offers a new perspective on informal trade and the act of sidestepping regulations. More specifically, it implies that "bad" regulations might be necessary for efficiency, even though they are not strictly followed in equilibrium. This finding has stark policy implications. Following the work of de Soto (1989), a number of papers have found that heavier regulations are associated with more informal trade (Djankov et al. 2002; Besley and Burgess 2004). At the face of it, these empirical findings appear to provide a strong argument for regulatory reform, perhaps in particular for deregulation: if the rules are sidestepped and renegotiated, what harm can be caused by removing them? In the model presented in this section, this conclusion is not justified. Deregulation would just take us back to the holdup problem described in equation (29).

A.5 Match-specific bargaining strengths

The socially optimal result described in the previous sections critically hinges on the ability of a single posted price to internalize the preference and technology parameters. With parameter heterogeneity, every single transaction is not likely to be efficient. Suppose, for instance, that customers differ in bargaining strength β_i (using i as subscript to denote that the β is match-specific). Unable to post a menu of prices over different bargaining strengths, the firms must post a single unit price based on some representative measure of bargaining strength, say $\bar{\beta}$. Under regulations-as-options, the match-specific end price will then equal

$$p \quad = p_c c_H = \eta c_L + (1-\eta)y + (\bar{\beta} - \beta_i)(c_H - c_L). \tag{38}$$

Thus, in the case with heterogenous bargaining strength and informal bargaining, the end price can drive some firms and customers out of the market, and the regulator's dilemma is again whether the problems with holdups are worse than the problems of cost maximization. Since we shall deliberately vary the intensity of counteroffers, this variation of the model is more relevant when interpreting some results from the experiment. More specifically, equation (38) suggest a route for interaction variables, as the impact of making counteroffers – i.e. the instability of the predicted price – will be stronger if the firm's moral hazard opportunities $(c_H - c_L)$ are stronger. Notably, an efficient outcome, independent of an individual's bargaining strength, is approached as $c_H \to c_L$ (that is, when the possibility to inflate costs decreases).

B Tables

Table B.1: The tipping factor. End price relative the advertised price across subtreatments – all fares rounded up to the nearest multiple of R10.

	(1) OLS	(2) OLS	(3) RE	(4) FE
M1 - Metered fare,				
no intervention (baseline)				
M2: Metered fare,	2.122	1.889	0.855	-0.655
ask for meter	(4.440)	(4.402)	(2.634)	(2.361)
F1: Fixed fare,	-3.647	-3.479	-2.987	-2.671
accept first offer	(3.932)	(3.916)	(3.224)	(3.272)
F2: Fixed fare,	-8.537**	-8.796**	-8.271**	-9.229**
with counteroffer	(3.858)	(3.862)	(3.486)	(3.529)
Map distance		0.893		
		(0.796)		
Constant	16.961***	11.451**	16.207***	17.528***
	(2.720)	(4.984)	(2.531)	(1.749)

Notes: All columns represents separate regressions. Dependent variable is the final price, rounded up to the nearest R10, minus an exogenous imputed price based on the advertised kilometer fare and the map distance. (1)-(2) OLS, (3) Random effects (GLS), (4) Fixed effects. In models (3) and (4) the fixed and random effects are equal to the block randomization unit (47 routes). Robust standard errors in parenthesis. 176 observations.

Table B.2: Intention-to-treat, per protocol, on-treatment and IV estimates.

	(1) First-stage	(2) ITT (reduced-form)	(3) Per-protocol (dropping noncompliers)	(4) On-Treatment	(5) IV
Dep. variable:	Used meter	Excessive driving			
Metered fare treatment	0.875***	0.563***	0.619***		
	(0.036)	(0.186)	(0.196)		
Did driver use the taximeter?				0.603***	0.636***
				(0.190)	(0.210)
Constant	0.034*	-0.007	-0.003	-0.006	-0.021
	(0.019)	(0.115)	(0.116)	(0.108)	(0.118)
Obs.	174	164	174	176	174

Notes: All columns represents separate regressions. Excessive driving is observed distance minus map distance. The IV model is estimated with treatment as an instrument for meter usage. Robust standard errors in parenthesis (F-statistic from first stage regression 635.78). Two observation with incomplete GPS-recording omitted.