Optimal Monitoring with External Incentives: The Case of Tipping

Ofer H. Azar*
Northwestern University

September 17, 2003

*Ofer H. Azar, Department of Economics, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA. Tel.: +1-847-491-9593, fax: +1-847-491-7001, e-mail address: o-azar@northwestern.edu.

I thank Eddie Dekel, Jaehong Kim, James MacDonald, Robert Porter, William Rogerson, Michael Whinston, Asher Wolinsky, participants in the Industrial Organization Society Conference in Boston (2003) and especially James Dana for helpful discussions and comments. I am also grateful to two anonymous referees for their valuable comments that helped improve this article. Financial support from the Center for the Study of Industrial Organization at Northwestern University is gratefully acknowledged.

Optimal Monitoring with External Incentives: The Case of Tipping

Abstract

The article examines the optimal choice of monitoring intensity when workers face external incentives (incentives that are not provided by the firm), such as tips, satisfaction from working well, or the desire to build reputation in order to be more attractive to other employers. Increase in such external incentives reduces optimal monitoring intensity but nevertheless increases effort and profits unambiguously. The model explains why US firms supported the establishment of tipping in the late 19th century but raises the possibility that European firms make costly mistakes by replacing tips with service charges.

1. Introduction

Tipping is a significant economic activity, and yet its economic implications have hardly been explored. Tips in US restaurants alone are around $27 billion a year.[1] Obviously, adding tips in other establishments such as hotels and taxis, and in additional countries, results in a much higher figure. Millions of workers depend heavily on tip income. Wessels (1997), for example, reports that in the United States alone there are over two million people who are servers as their primary occupation, and the number may be 60 percent higher if we add those who are servers as a secondary occupation. He adds that tips represent 58 percent of servers' income in full-course restaurants and 61 percent in counters, and that these figures are likely to be understated because servers often underreport their tip income. Finally, tipping has become a source of income in many different occupations: Lynn, Zinkhan and Harris (1993), for example, consider 33 service professions that are tipped.

How has tipping become such a prevalent social norm? Who has an incentive to support it? Do firms benefit from tipping, and in what ways? I analyze the interaction between tipping, which can be thought of as buyer monitoring, and monitoring by the firm. The analysis suggests that by motivating workers to provide better service, tipping enables the firm to reduce its costly

---

[1] The extent of tipping has to be estimated because tips are often unreported for tax purposes (according to Hemenway (1993) the only income with a lower compliance rate is illegal income). Sales in the United States in 2002 of food and alcoholic beverages to consumers in full-service restaurants, bars and taverns, and lodging places, were $146.7, $13.3 and $18.6 billion, respectively (U.S. Census Bureau, 2002; the numbers for 2002 are a projection). Summing the three numbers and multiplying by an average tip of 15 percent yields annual tips of $26.8 billion.

monitoring of workers and to increase the price it charges (because of the increased service quality). Therefore, tipping increases the profits of the firm, so firms have an incentive to support the tipping custom.

While the article focuses on the case of tipping[2], the theoretical model is applicable to additional examples in which workers face external incentives (incentives that are not provided by the firm). One such example is the satisfaction that workers derive from doing their job well, especially in jobs that require initiative and creativity. This satisfaction (often referred to as "intrinsic motivation") motivates workers to excel even when they face no monetary incentives to do so.

Another example is that of military pilots: their future prospects and expected salaries as civil pilots later in life depend on their performance in the military, thus providing them additional incentives to do their job well beyond the incentives provided by the military.[3] Similarly, anyone who thinks he may change employers in the future (whether voluntarily or not) has an incentive to work well in order to be more attractive to the next employer. Potential employers receive information about previous performance of the candidate from various sources, such as letters of reference and items on the curriculum vitae. Consequently, current performance affects the candidate's reputation and his prospects with other employers, giving him incentives to work well that are not provided by the firm.

---

[2] In particular, I find it more concrete to talk about a specific tipping occasion, although the analysis and the ideas are applicable to tipping in general. Tipping in restaurants is the natural candidate, as it is the most common form of tipping. I therefore use firms and restaurants interchangeably; the same applies to workers and waiters.

[3] I thank James MacDonald for this example.

The common theme in all the above examples is that the worker faces external incentives to do what the firm also wants to achieve. In the case of tipping, tips promote higher service quality, and the firm wants to encourage high service quality as well; similarly, self-fulfillment and satisfaction from being successful, or reputation building in order to improve one's value in the job market, motivate the worker to work harder, which is also what the firm wants.

The existing literature about tipping is mostly empirical, and includes two main types of studies. One type interviews customers when they leave a restaurant and tries to evaluate which variables affect the tip size (for example, does food quality affect tips?). Major contributions of this type include Bodvarsson and Gibson (1994, 1997). A second type of study asks waiters to behave in a certain way (for example to touch the customer lightly or to write "Thank you!" on the bill) and records the effect of this behavior on tips, using a control group as a benchmark (see for example Crusco and Wetzel, 1984).[4] A unique and interesting study about tipping is the experimental article of Ruffle (1998), in which participants in dictator and ultimatum games acted in a way that resembles tipping.

The theoretical work on tipping started with the pioneering work of Ben-Zion and Karni (1977), who show that tipping is consistent with a selfish customer only for the case of a repeated customer. They suggest that in order to explain why one-time customers tip one should consider altruistic behavior and social norms, which are not included in their model. Jacob and Page (1980) suggest that optimal monitoring may involve monitoring by both the owner and the buyer who interacts with the monitored employee. Sisk and Gallick (1985) argue that tips ultimately protect the buyer from an unscrupulous seller (or his agent) when the brand-name mechanism for assuring contractual performance is insufficient. Schwartz (1997) suggests that

[4] For an excellent review of the empirical literature on tipping see Lynn and McCall (2000a).

tipping can increase the firm's profits when it enables price discrimination between two consumer segments that differ in their demand functions and their propensity to tip. Ruffle (1999) presents a theoretical model about gift giving and discusses briefly how the model can be applied to tipping as well. Azar (2003a) presents a model of the evolution of social norms. When a norm is costly to follow and people do not derive benefits from following it except for avoiding social disapproval, the norm erodes over time. Tip percentages, however, increased over the years, suggesting that people derive benefits from tipping, such as impressing others and improving their self-image as being generous and kind.

In this article, I analyze the optimal choice of monitoring and incentives by the firm when the worker faces external incentives that encourage him to do what the firm also wants to achieve. The theoretical analysis suggests that firms benefit from higher sensitivity of tips to service quality, because it enables them to reduce the cost of monitoring. This implies that firms should encourage customers to tip badly (or not at all) for bad service, rather than to always tip. In addition, as long as tips are positively correlated with service quality, firms benefit from the existence of tipping. This result is consistent with historical evidence that suggests that US firms promoted the custom of tipping in the late 19th century, despite attempts of several consumer groups, and even workers, to abolish the custom (Segrave, 1998; Azar, 2003b). This result, however, also suggests that numerous European firms that replaced tips with service charges possibly made a costly mistake. I discuss, however, why this might not be a mistake after all. The model also implies that in countries in which tipping is not prevalent, for example in Japan, Australia and the Scandinavian countries, firms may do better by trying to promote the custom of tipping.

The discussion above suggests that the main contributions of this article can be categorized

as follows: first, it addresses the issue of optimal monitoring in the presence of external incentives. Tips, intrinsic motivation, and reputation building are a few examples of such incentives. Second, the article contributes to the literature about tipping, analyzing the relationship between tipping and monitoring by the firm. Finally, the article compares the theoretical predictions to the behavior of firms in the United States and Europe and offers a potential explanation to the puzzle regarding the choices of European firms.

2. The model

The game involves two players (a firm and a worker) and two stages. In the first stage the firm chooses how intensely to monitor the waiter, which in turn determines also the incentives to provide good service that the waiter faces. In the second stage the waiter chooses the service quality to provide, and then receives both his tip and the incentives from the firm according to the service quality chosen. The tip is potentially increasing in the service quality provided. This may follow from the social norm being that better service should be rewarded by a higher tip. Alternatively, it may follow from the customer trying to discipline the waiter in a repeated-interaction scenario: the customer gives better tips for better service in order to motivate the waiter to give good service in future encounters. The task that the waiter has to perform is to serve a single customer. Serving a table of four can be considered as having four identical tasks; the effort and incentives are simply four times those for serving a single customer. I assume for simplicity that the bill size per customer is constant.

2.1. Service quality

Let us denote service quality by s and define s = 0 to be the service quality that minimizes

the waiter's effort. The assumption that such quality exists follows from the observation that below some quality level, reducing quality is in fact costly for the waiter. For example, being too slow and bringing the food cold may result in a requirement to heat the food, which causes the waiter more effort than bringing the food hot in the first place. Similarly, being rude may be more costly than just being unfriendly.

Since service quality has no natural scale, we can scale it as we wish. I choose to scale it in a way that makes the tip linear in service quality.[5] That is, choose $s = 1$ to represent an arbitrary quality level that is better than $s = 0$. Denote the tip left for $s = 0$ as T_0 and the tip left for $s = 1$ as $T_0 + T$. Now define $s = 2$ to be the quality level that results in a tip of $T_0 + 2T$ and so on. As a result, the tip is linear in service quality. Let $T_0 + sT$ be the tip in dollars given for service quality s, where $T_0 \geq 0$ and $T \geq 0$.[6]

2.2. The firm

Monitoring by the firm provides incentives for the waiter to give good service, in addition to the incentives provided by tips. The firm can punish bad service by dismissing the waiter or giving him bad shifts, bad tables, or fewer tables to serve. On the other hand, it can reward good service by giving the waiter more tables to serve and better shifts and tables. Whatever the incentives are, the waiter cares about their monetary implication. For simplicity, I assume that

---

[5] This is done in order to make the model traceable and to provide a precise solution to the model. The same qualitative results, however, hold more generally as long as the tip is weakly increasing in service quality.

[6] When $T = 0$ we cannot scale the service quality according to the tips given, because the tip is always T_0. In this case I scale the service quality according to the incentives provided by the firm for different quality levels in such a way that these incentives are linear in quality.

the monetary value of the incentives for the waiter is linear in service quality. In addition, the firm may have to pay the waiter a wage regardless of service quality and his tip income, for example because of minimum wage laws.[7]

From the waiter's perspective, this means that the wage and the incentives provided by the firm have a total value of $w + \mu s$; w is given exogenously (for example it may be the minimum wage), s is chosen by the waiter, and μ is chosen by the firm. The value of μ represents the intensity of monitoring by the firm. When the firm monitors the waiters more closely (higher μ), its ability to punish bad waiters and reward good waiters increases, and therefore the monetary value of the incentives (from the waiters' perspective) becomes steeper in service quality; that is, the waiters' incentives to provide good service are increasing in μ.

For example, a small investment in monitoring may be to dismiss waiters whom a customer complains about. It is very cheap, but does not provide many incentives for excellent service. Waiters would probably be careful not to be too rude or careless (assuming that their utility is strictly above their reservation utility, so they strictly prefer to keep their job), but they would not try very hard to provide the best service possible. A higher investment in monitoring can be to test the waiters' knowledge of the menu occasionally and to employ a worker whose job is to watch the waiters and rate their service quality. This enables the firm to rank the performance of the waiters, and as a result to reward the best waiters by giving them better tables or shifts or by other means.

---

[7] In the United States the current federal law says that tipped employees should receive at least $2.13 an hour, and their wage and tips together should be at least equal to the minimum wage (which is currently $5.15 an hour). Several states adopted laws that require paying tipped workers the regular minimum wages regardless of the tips they earn.

The important thing to notice is that the cost of monitoring for the firm is not equal to the monetary value of the incentives from the worker's perspective. Although the waiter faces the compensation scheme $w + \mu s$, this is not the labor cost for the firm, because of two reasons. First, some of the expenses associated with monitoring, such as employing workers to monitor the waiters, are costly for the firm but are not an income for the waiters. Second, some of the incentives faced by the waiters are not an expense for the firm, for example giving better shifts or tables to the best waiters.

The firm has also variable costs, for example the cost of food and the wages of cooks and managers. I assume that the total cost of producing a quantity q (the number of customers served) when monitoring intensity is μ is equal to:

$$(1) \qquad cq + \delta\mu^x q,$$

where $c > 0$, $\delta > 0$, and $x > 1$. The costs that are not related to monitoring or incentives, such as minimum wages for the waiters, the cost of food, and wages of cooks and managers, are included in cq. The cost of monitoring and providing incentives is $\delta\mu^x q$; this cost includes for example the wages of workers who monitor the waiters. Serving more customers (higher q) requires additional waiters and therefore increased monitoring costs (if monitoring intensity is to remain constant). The cost function is based on the assumptions that total monitoring cost is proportional to q and that the cost of monitoring is strictly convex in monitoring intensity (therefore $x > 1$).

I assume that the demand faced by the firm is continuous and downward sloping in price.[8] In addition, the customers' willingness to pay is strictly increasing in service quality, and I allow

[8] This precludes the case of perfect competition, but is consistent with many industry structures, for example if the

the willingness to pay to be either linear or concave in service quality. The inverse demand is therefore a function of both the quantity sold and service quality. The inverse demand faced by the firm is assumed to take the following form:

$$p(q, s) = \alpha - \beta q + \phi s^{y}, \quad (2)$$

where $\beta > 0$, $\phi > 0$, $\alpha > c > 0$, and $0 < y \leq 1$.

As a result, the firm's profit function is:

$$\pi(q, \mu, s) = (\alpha - \beta q + \phi s^{y} - c - \delta \mu^{x})q. \quad (3)$$

2.3. The waiter

The waiter derives income both from tips and from the firm: his total income is $T_0 + sT + w + \mu s$. His effort is a function of the service quality he provides. I assume that the effort function is strictly convex and takes a quadratic form, $e(s) = E_0 + E_1 s + E_2 s^2$. Since we defined $s = 0$ to be the service quality that minimizes effort, it follows that $E_1 = 0$. Strict convexity of e implies $E_2 > 0$. Assuming that the waiter's utility function is quasi-linear in money, his utility is equal to:[9]

$$v(s) = T_0 + sT + w + \mu s - E_0 - E_2 s^2. \quad (4)$$

---

firm is a monopoly, or if the industry is an oligopoly with differentiated products. Restaurants differ in their location, the food they serve, their quality level, and sometimes in their opening hours, so the assumption of product differentiation is clearly reasonable in the restaurant industry.

[9] In the context of intrinsic motivation, s represents how well the worker performs his task and T stands for the degree with which good performance increases the worker's satisfaction. T_0 is the degree of this satisfaction when $s = 0$.

The waiter chooses s to maximize his utility and takes T_0, T, w and μ as given. To ensure that the individual rationality constraint (IRC) is satisfied in equilibrium (so the waiter prefers to work as a waiter rather than to quit and find another job), a sufficient condition is that working and providing zero service quality is better for the waiter than his outside option. If in equilibrium he chooses to provide a strictly positive service quality, it means that his utility from doing so is at least the utility from choosing zero service quality, and therefore the IRC is satisfied. If we denote the waiter's reservation utility by v_0, then the following assumption gives a sufficient condition for the IRC to hold:

ASSUMPTION 1. $T_0 + w - E_0 \geq v_0$.

3. The equilibrium

The equilibrium can be solved for using backward induction. In the second stage, the waiter chooses which service quality to provide, given the tip he expects and the incentives provided by the firm. The following proposition describes his optimal choice:

PROPOSITION 1.[10] Service quality in equilibrium is $s = (T + \mu)/2E_2$.

Thus, service quality is strictly increasing in both tips (T) and monitoring intensity (μ); it is strictly decreasing in E_2 because a higher E_2 corresponds to a higher marginal cost of increasing service quality. Given the choice of service quality by the waiter, the firm chooses the quantity it

[10] All the proofs are in the appendix.

wants to sell and the intensity of monitoring. The following proposition characterizes its optimal choices:

PROPOSITION 2. i) The firm's optimal choice of monitoring intensity is given by the value of μ that solves $\phi y(T+\mu)^{y-1}/(2E_2)^y - \delta x\mu^{x-1} = 0$. Denote this value by μ^*. There exists a unique value of μ^*, which is strictly positive.

ii) Optimal q is given by $q^* = [\alpha - c - \delta(\mu^*)^x + \phi(T + \mu^*)^y/(2E_2)^y]/2\beta$, and $q^* > 0$.

The condition $\phi y(T+\mu)^{y-1}/(2E_2)^y - \delta x(\mu)^{x-1} = 0$ that defines the optimal value of μ may seem complicated, but is in fact intuitive. The marginal benefit from increasing μ is the increase in price that results from the improved service quality, times the quantity sold. This equals to $q^*\phi y(T+\mu)^{y-1}/(2E_2)^y$. The marginal cost of increasing μ is $q^*\delta x\mu^{x-1}$. The condition above equates the marginal benefit and the marginal cost of increasing μ. How do the monitoring intensity chosen by the firm and equilibrium service quality depend on the tipping function of the customer? Corollary 1 provides the answer:

COROLLARY 1. i) $\partial\mu^*/\partial T = -\mu^*(1-y)/[(x-1)(T+\mu^*)+\mu^*(1-y)]$. It follows that if $y = 1$ then $\partial\mu^*/\partial T = 0$, otherwise $-1 < \partial\mu^*/\partial T < 0$.

ii) Let s^* be the service quality chosen by the waiter; then $\partial s^*/\partial T = (1 + \partial\mu^*/\partial T)/2E_2 > 0$.

Part (i) of Corollary 1 suggests that when tips are more sensitive to service quality (higher T), the firm chooses to reduce monitoring intensity (strictly if $y < 1$). The increase in the incentives provided by the customer when T goes up exceeds the effect of the reduced

monitoring, so that in total the waiter faces more incentives to provide good service. This is the reason for part (ii) of the corollary, which suggests that service quality increases when T goes up despite the reduction in the monitoring intensity.

Figure 1 illustrates the optimal choice of μ (the figure corresponds to $y < 1$). The increasing curve (MC) is the marginal cost of increasing μ, per unit of output. It starts at the origin and increases without bound. It is strictly concave if $1 < x < 2$ and strictly convex if $x > 2$; the figure corresponds to $x < 2$, but the only important thing for the analysis below is that MC is increasing in μ. MB^0 is the marginal benefit per unit of output from increasing μ, when $T = T^0$. It is equal to the increase in the price (holding the quantity sold unchanged) that results from increasing μ by one unit (the increase in price is a result of the higher service quality chosen by the waiter when μ is higher). MB^1 is the corresponding graph for $T = T^1$, where $T^1 > T^0$. Since $y < 1$, MB^1 is below MB^0. Notice that MB is strictly positive when $\mu = 0$. In addition, when $y < 1$, MB is strictly decreasing in μ and it approaches zero as μ approaches infinity (when $y = 1$, $MB = \phi/2E_2$ for all values of μ). Since MC is strictly increasing, MB is non-increasing, and MB is higher than MC for $\mu = 0$, there is a unique intersection between MB and MC at a strictly positive value of μ (defined as μ^{*0} for T^0 and μ^{*1} for T^1), which is the optimal value of μ. Figure 1 illustrates that when T increases, μ^* decreases (for $y < 1$), as suggested by Corollary 1.

[Figure 1 here]

How does a change in T affect the equilibrium quantity and price? The following corollary provides the answers:

COROLLARY 2. i) Equilibrium quantity is strictly increasing in T: $\partial q^*/\partial T = \delta x(\mu^*)^{x-1}/2\beta$

> 0.

ii) $\partial p^*/\partial T = [(1 + 2\partial \mu^*/\partial T)\delta x(\mu^*)^{x-1}]/2$, which can be either positive or negative.

How does T affect profits? The following proposition suggests that profits are unambiguously increasing in T:

PROPOSITION 3. Equilibrium profits are strictly increasing in T.

4. Discussion

It follows from Proposition 3 that the firm wants T to be as high as possible. This means that the firm should encourage customers to tip badly for poor service rather than to always tip generously. Moreover, any strictly positive T yields higher profits than T = 0. Assuming that people tip more for good service (which is supported by empirical evidence, see Lynn and McCall, 2000b), this implies that if the firm has the option whether to implement tipping or not, it should choose to use tips.[11] The reason is twofold: first, tipping provides incentives to the waiters and enables the firm to reduce its costly monitoring of them. Second, even after the firm reduces its monitoring intensity, the incentives faced by the waiters are higher than without tips, as suggested by Corollary 1. As a result, equilibrium service quality is higher when tipping exists, increasing the consumers' willingness to pay and the firm's profits.

This observation is consistent with evidence from the history of tipping. In the late 19th century, when tipping began to be established in the United States, the owners of restaurants and

[11] An exception to this rule may occur when workers enjoy economic rents if tips are used, and the firm can capture the rent by imposing a service charge that replaces tips. More on this below.

hotels were often blamed (by those who disliked tipping) to be the ones who promoted the custom (Segrave, 1998). An editorial in the *New York Times* in 1899, for example, claimed that the tipping practice is a wretched system that was originated and perpetuated

> not by its victims, the men who give and take tips, but by those who profit by it every year to the extent of millions more than a few. The real takers of tips are the hotel and restaurant proprietors, the owners of steamships, the offices and stock-holders of railways, and a dozen other classes of employers… every tip saves the payment of wages to an equal amount… This throws a flood of light on the frequent assertions that the abolition of the tipping system is impossible.[12]

Indeed, the evidence in several industries implies that where tipping became common, wages were reduced to reflect the presence of tipping, although it is not clear whether the reduction in wages was at the same amount as the tips (Segrave, 1998).[13] But even if the claim that wages were driven down by the amount of tips is true, this still does not explain why restaurants, hotels and others had an incentive to implement tipping. For the customer, having to add a tip is the same as an increased price (when the increase is by the same amount as the tip).[14] Consequently, the owner could increase prices instead of encouraging people to tip, and get the increased revenues directly rather than by reducing the workers' wages. The analysis in this

---

[12] "Topics of the Times," *The New York Times*, November 21, 1899, p. 6. The quote is adopted from Segrave (1998).

[13] One might expect prices to fall when restaurants' costs go down due to lower wages. Unfortunately, I am not aware of any available data on restaurant prices before and after tipping was implemented in that restaurant. Customers' reactions to tipping from that period, however, suggest that prices did not fall (see Scott, 1916; Segrave, 1998; Azar, 2003b).

[14] If people exhibit bounded rationality, framing effects and mental accounting may make tips seem less expensive than increased prices (when the increase is in the same amount as the tips). I assume here that people are rational and treat tips and increased prices in the same way.

article suggests that the reason why firms chose to support tipping rather than to increase prices may be that they realized that tipping provides incentives for good service, resulting in better service quality and reduced monitoring costs.[15]

The model may also provide the explanation for another puzzle. Waiters often earn income (tips and wages) that exceeds their reservation wages. The reason is that usually the firm cannot take the tips away from the waiters, and wages are often required by law to be above a certain minimum. The firm can increase its profits if it can take this economic rent from the waiters. Different firms tried different methods to extract this economic rent: in some cases, for example, waiters paid for the privilege to work and receive tips (Segrave, 1998; Seligman, 1998).

Today, many restaurants require that the waiters split the tips with other workers, for example with the busboys. This enables the restaurant to pay lower wages to the busboys and is a way of extracting some of the economic rent from the waiters. In the United States, however, these arrangements, called "tip outs," are limited by the Fair Labor Standards Act. Tipped employees cannot be forced by employers to share tips with employees who do not ordinarily participate in tip-pooling arrangements (such as janitors and dishwashers). In addition, if tip-pooling exceeds 15 percent of the tips, the Department of Labor will investigate to assure that the pooling agreement is "customary and reasonable" (Wessels, 1997).

---

[15] Today, another benefit for the firms of implementing tipping is that the waiters usually do not report their entire tip income to the tax authorities (see Hemenway, 1993). Consequently, the net income of the waiters is higher than what it would be if the firm paid them the same amount (of the tips) as wages. The firm, in turn, can capture at least some of this additional surplus by various means (see the discussion below). While this reason might support tipping today, it cannot explain why US firms supported tipping in the 19th century, since income taxes were introduced in the US only in the 20th century. I thank an anonymous referee for raising this important point.

Similarly, the firm can reduce the waiters' economic rent by giving each waiter less tables to serve or requiring him to do non-tipped activities as well. Giving waiters only a few tables, however, does not contribute to the restaurant's profits beyond a certain point, so it reduces the waiters' economic rents but does not transfer them to the restaurant (instead, it simply reduces efficiency). Similarly, asking a skilled waiter to either perform tasks that can be done by cheaper labor (e.g. clean tables or dishes) or that require different skills (e.g. being responsible for purchasing and inventory) is inefficient and therefore can transfer the economic rents of the waiters to the restaurant only partially.

There is a simpler way, however, to extract all the rent: add a 15-percent service charge to the bill (that replaces the tip), pay the waiters their reservation wages, and keep the rest (or alternatively, increase prices by 15 percent and declare that the restaurant has a policy of no tipping). Why do US restaurants retain tipping and do not use this method? The answer may be that they are aware of the incentives that tips provide for good service and they realize that tipping saves costly monitoring. Customers, at least, seem to be aware of the positive effect of tips on service quality; most of them want to retain their freedom in choosing how much to reward the waiter, and oppose replacing tipping with service charges.[16]

In Europe, however, the situation is different. Many restaurants impose a service charge instead of allowing the customer to choose the tip. The model suggests that this policy increases

---

[16] In an on-line poll in www.tipping.org, a website dedicated to tipping, one of the questions posted was "Would you tolerate higher prices at a restaurant in order to do away with tips?" On April 2, 2003, out of 1633 voters, 26% answered positively, 56% negatively, and the rest answered "maybe" or "unsure." Service charges and higher prices are equivalent from the customer's perspective since both are an expense for the customer and do not provide incentives to the waiters.

monitoring costs and reduces profits. Why do European restaurants adopt this policy? There are at least five possible explanations.[17] First, European firms might make a mistake. Second, they may be afraid that too many customers would act opportunistically and will not tip.[18] Third, it seems that in Europe pay is less linked to performance than in the United States more generally, possibly reflecting differences in values and attitudes. This might be another reason why tipping is more popular in the United States than in Europe.[19] Fourth, tipping is often thought to create different classes (see for example Segrave, 1998); European countries may have more resistance to this because of different values.[20] Lynn, Zinkhan and Harris (1993), for example, found that tipping was less prevalent in countries with a low tolerance for status and power differences between people.

Finally, the reason for implementing service charges in European restaurants may be to extract an economic rent enjoyed by the waiters when they receive tips. This reason is more

---

[17] These explanations may also be the reason why other countries in which tipping is not common (such as Japan, Australia and the Scandinavian countries) do not adopt tipping.

[18] This explanation is not very plausible, as it raises the question why would European customers behave so differently from American customers, who rarely stiff according to empirical evidence (Bodvarsson and Gibson, 1997).

[19] I owe this idea to an anonymous referee.

[20] I thank an anonymous referee for making this point. As this referee suggests, status may also be the reason why some occupations are tipped while others are not: those who receive tips tend to work in lower-status jobs, and this might explain why we do not tip doctors, lawyers, and university lecturers, among others, despite the advantages of tipping in improving service quality and increasing the firm's profits. A full discussion of the reasons for different tipping practices in different occupations is very interesting but is beyond the scope of this article; the interested reader is referred to Azar (2003c), who discusses this topic in detail.

likely if the firm is obligated to pay minimum wages regardless of tips, because then the waiter's income (and therefore the potential economic rent) is much higher. Indeed, in Israel when a court decided that waiters should receive minimum wages in addition to tips, many restaurants replaced tipping with service charges (Sinay, 2001). Whether minimum wage laws are indeed the reason for the differences between the tipping practices in the United States and in Europe, however, remains a topic for future research.[21]

5. Conclusion

The article explores how the optimal choice of monitoring by the firm is affected by external incentives that the worker faces. The theoretical model uses the example of tipping, but applies also to the cases of intrinsic motivation that causes the worker to derive satisfaction from a job well done and of reputation building in order to increase one's value in the job market.

The analysis suggests that tips have the potential to motivate workers to provide good service quality, and by doing so also to reduce the need for costly monitoring of workers by the firm. The extent to which tips realize this potential depends on the sensitivity of tips to service quality.

---

[21] Ideally, we would like to compare laws, court decisions (if applicable) and actual practices regarding whether tipped employs should receive minimum wages in addition to tips between the United States and European countries (as well as other countries). If minimum wages are imposed in tipped occupations, we should also consider the relative level of minimum wages. We would then want to estimate the total income of waiters and compare it to income in similar occupations to see whether service charges are used in those countries in which waiters enjoy particularly high economic rents (if tips are to be used). In addition, it would be interesting to examine whether changes in tipping practices followed changes in legislation about minimum wages of tipped employees. This project is beyond the scope of the current article, and will most likely require a team of authors who have control of all the relevant languages.

The higher this sensitivity is, the more motivation tips provide for the workers and the more the firm can reduce its costly monitoring. As a result, the firm's profits are increasing in the sensitivity of tips to service quality, meaning that the firm should encourage customers not to tip for bad service. In addition, the firm's profits are higher when tips are used than when a fixed service charge is imposed. An exception to this rule occurs when tipped workers receive income that exceeds their reservation wages. By replacing tips with a service charge the firm may then capture the workers' economic rent, and in this case a service charge may increase the firm's profits compared to tips.

In the context of intrinsic motivation, the conclusion is that firms should try to increase the sensitivity of the worker's satisfaction to his performance level, for example by providing him more feedback about his performance. This will improve the worker's effort, reduce the costs of monitoring, and increase profits.

The theoretical contribution in this paper is a first step in a direction that warrants future research. One interesting idea is to examine the case of a worker who has two tasks, one of which carries external incentives and the other does not. For example, professors are required to teach and to do research. While research output affects significantly the professor's reputation and therefore his salary (whether he stays in the same institution or not), teaching quality does not affect his salary significantly in most research-oriented institutions. It would be interesting to examine how the equilibrium and optimal monitoring of the two tasks look like in this case. Additional interesting questions are whether professors spend too much time on research relative to teaching because of the external incentives mentioned, and whether business schools, in which

teaching quality is considered very important, monitor teaching more carefully than other departments.[22]

Another interesting idea for future research is to test empirically the predictions of the model. This seems to be easier with respect to the tipping example than with intrinsic motivation. How do monitoring costs and service quality compare between restaurants that use tipping and those that use service charges?[23] In a restaurant that imposes a fixed-percentage gratuity for large parties, as is common in US restaurants, do waiters give small parties better service than they provide to large parties?

It is also interesting to examine more closely the policy of firms in several countries to replace tips with service charges. When did European restaurants start to replace tips with service charges? Why do they adopt this policy? Does it enable the restaurant to capture an economic rent enjoyed by the waiters when they receive tips? If so, what are the main differences that cause tipping to be prevalent in the United States but not in Europe? Different minimum wage laws? Different attitudes of customers toward the tipping custom? As a challenging economic phenomenon that has hardly been explored by economists, tipping offers many opportunities for future research; the above questions are only a partial list (for a more complete list, see Azar, 2003d).

---

[22] I owe these interesting ideas to an anonymous referee.

[23] One should take into account, however, that the decision whether to impose service charges or to use tips is endogenous and therefore may reflect the characteristics of the restaurant in terms of its service quality, monitoring costs, and so on.

Appendix: Proofs

PROOF OF PROPOSITION 1. The waiter chooses s to maximize $v(s) = T_0 + sT + w + \mu s - E_0 - E_2 s^2$. The first-order condition is given by $T + \mu - 2E_2 s = 0$, or $s = (T + \mu)/2E_2$. The second-order condition is satisfied because $E_2 > 0$. □

PROOF OF PROPOSITION 2. Substituting $s = (T + \mu)/2E_2$ in (3) we get:

$$(5) \qquad \pi(q, \mu) = [\alpha - \beta q + \phi((T + \mu)/2E_2)^y - c - \delta\mu^x]q.$$

The optimal values μ^* and q^* have to satisfy the following first-order conditions (subscripts denote partial derivatives):

$$(6) \qquad \pi_\mu = q^*\phi y(T + \mu^*)^{y-1}/(2E_2)^y - q^*\delta x(\mu^*)^{x-1} = 0, \text{ and}$$

$$(7) \qquad \pi_q = \alpha - 2\beta q^* + \phi(T + \mu^*)^y/(2E_2)^y - c - \delta(\mu^*)^x = 0.$$

The second-order sufficient conditions are:

$$(8) \qquad \pi_{qq}(q^*, \mu^*) = -2\beta < 0,$$

$$(9) \qquad \pi_{\mu\mu}(q^*, \mu^*) = q^*\phi y(y - 1)(T + \mu^*)^{y-2}/(2E_2)^y - q^*\delta x(x - 1)(\mu^*)^{x-2} < 0, \text{ and}$$

$$(10) \qquad \pi_{\mu\mu}(q^*, \mu^*)\pi_{qq}(q^*, \mu^*) - [\pi_{\mu q}(q^*, \mu^*)]^2 > 0.$$

Since $\alpha > c$, the firm can always make strictly positive profits by choosing $\mu = 0$ and a small positive q; therefore, $q^* \neq 0$, because choosing $q = 0$ yields zero profits. Since $q^* \neq 0$, divide (6) by q^* to get that $Z \equiv \phi y(T+\mu^*)^{y-1}/(2E_2)^y - \delta x(\mu^*)^{x-1} = 0$. To see that the value of μ^* that solves this equation exists and is unique and strictly positive, notice that $Z(\mu^* = 0) > 0$, and as μ^*

approaches ∞, Z goes to $-\infty$ (recall that $x > 1$ and $0 < y \leq 1$). It is easy to verify that Z is continuous and strictly decreasing in μ^*; it follows that there is a unique and strictly positive value of μ^* for which $Z = 0$.

From (7) it follows that $q^* = [\alpha - c - \delta(\mu^*)^x + \phi(T + \mu^*)^y/(2E_2)^y]/2\beta$. Since we did not incorporate the restriction $q \geq 0$ in the maximization problem, we have to make sure that this value of q^* is not negative. Since $\alpha > c$, a sufficient condition for $q^* > 0$ is $\phi(T + \mu^*)^y/(2E_2)^y \geq \delta(\mu^*)^x$. Using (6), $\phi(T + \mu^*)^y/(2E_2)^y = \delta x(\mu^*)^{x-1}(T + \mu^*)/y$. That is, the sufficient condition becomes $\delta x(\mu^*)^{x-1}(T + \mu^*)/y \geq \delta(\mu^*)^x$. Divide both sides by $\delta(\mu^*)^{x-1}$ to get the condition $x(T + \mu^*)/y \geq \mu^*$. Since $x > 1 \geq y$ and $T \geq 0$, this is satisfied.

Next, consider the second-order conditions. Notice that (8) is satisfied because $\beta > 0$. In addition, $\pi_{\mu q}(q^*, \mu^*) = \phi y(T + \mu^*)^{y-1}/(2E_2)^y - \delta x(\mu^*)^{x-1}$; since $q^* \neq 0$, it follows from (6) that $\pi_{\mu q}(q^*, \mu^*) = 0$. Therefore, if (9) is satisfied, the inequality in (10) follows immediately. Using (6), we get that $\pi_{\mu\mu}(q^*, \mu^*) = q^*\delta x(\mu^*)^{x-1}[(y - 1)/(T + \mu^*) - (x - 1)/\mu^*]$. Therefore, the sign of $\pi_{\mu\mu}(q^*, \mu^*)$ is equal to the sign of $[(y - 1)/(T + \mu^*) - (x - 1)/\mu^*] = [(y - 1)\mu^* - (x - 1)(T + \mu^*)]/\mu^*(T + \mu^*) = [\mu^*(y - x) - T(x - 1)]/\mu^*(T + \mu^*) < 0$, where the last inequality follows from $T \geq 0$, $\mu^* > 0$ and $x > 1 \geq y$. Therefore, all the second-order conditions are satisfied. This completes the proof. □

PROOF OF COROLLARY 1. From Proposition 2 it follows that $H(\mu^*, T) \equiv \phi y(T + \mu^*)^{y-1}/(2E_2)^y - \delta x(\mu^*)^{x-1} = 0$. Using the Implicit Function Theorem, $\partial\mu^*/\partial T = -H_T/H_{\mu^*}$. Notice that $H_T = \phi y(y - 1)(T + \mu^*)^{y-2}/(2E_2)^y = (y-1)\delta x(\mu^*)^{x-1}/(T+\mu^*)$, using Proposition 2(i). Similarly,

$H_{\mu^*} = (y-1)\delta x(\mu^*)^{x-1}/(T+\mu^*) - \delta x(x-1)(\mu^*)^{x-2}$. Substituting and simplifying we then get that $\partial\mu^*/\partial T = -H_T/H_{\mu^*} = -\mu^*(1-y)/[(x-1)(T+\mu^*)+\mu^*(1-y)]$, which is equal to zero when $y = 1$ and is between -1 and 0 (not including the endpoints) when $y < 1$ (recall that $x > 1$). This completes part (i). Part (ii) follows immediately: $s^* = (T + \mu^*)/2E_2$ implies that $\partial s^*/\partial T = (1 + \partial\mu^*/\partial T)/2E_2 > 0$. □

PROOF OF COROLLARY 2. i) Recall that $q^* = [\alpha - c - \delta(\mu^*)^x + \phi(T + \mu^*)^y/(2E_2)^y]/2\beta$, and that μ^* is derived from an equation that does not involve q^* (see Proposition 2(i)). This implies that $\partial q^*/\partial T = [\phi y(T + \mu^*)^{y-1}(1 + \partial\mu^*/\partial T)/(2E_2)^y - \delta x(\mu^*)^{x-1}(\partial\mu^*/\partial T)]/2\beta$. Substituting $\phi y(T+\mu^*)^{y-1}/(2E_2)^y = \delta x(\mu^*)^{x-1}$ (this equation is a simple rearrangement of Proposition 2(i)), it follows that $\partial q^*/\partial T = \delta x(\mu^*)^{x-1}/2\beta > 0$.

ii) Notice that equilibrium price satisfies $p^*(q^*, s^*) = \alpha - \beta q^* + \phi(s^*)^y = [\alpha + \phi(s^*)^y + c + \delta(\mu^*)^x]/2$. Consequently, $\partial p^*/\partial T = [\phi y(s^*)^{y-1}\partial s^*/\partial T + \delta x(\mu^*)^{x-1}\partial\mu^*/\partial T]/2$. Substituting $s^* = (T + \mu^*)/2E_2$ and $\partial s^*/\partial T = (1 + \partial\mu^*/\partial T)/2E_2$ we get $\partial p^*/\partial T = [(1 + \partial\mu^*/\partial T)\phi y(T+\mu^*)^{y-1}/(2E_2)^y + \delta x(\mu^*)^{x-1}\partial\mu^*/\partial T]/2$. Using again $\phi y(T+\mu^*)^{y-1}/(2E_2)^y = \delta x(\mu^*)^{x-1}$ (from Proposition 2(i)), we get $\partial p^*/\partial T = [(1 + 2\partial\mu^*/\partial T)\delta x(\mu^*)^{x-1}]/2$. The sign of this expression is equal to the sign of $(1 + 2\partial\mu^*/\partial T)$. Recall from Corollary 1(i) that $\partial\mu^*/\partial T = -\mu^*(1-y)/[(x-1)(T+\mu^*)+\mu^*(1-y)]$. When y is close to 1 and x is not, $\partial\mu^*/\partial T$ is close to zero, and $\partial p^*/\partial T > 0$. When x is close to 1 and y is not, $\partial\mu^*/\partial T$ is close to -1 and $\partial p^*/\partial T < 0$. This completes part (ii). □

PROOF OF PROPOSITION 3. Recall that $\pi(q, \mu) = [\alpha - \beta q + \phi((T + \mu)/2E_2)^y - c - \delta\mu^x]q$.

Using the Envelope Theorem, $d\pi^*/dT = \partial\pi^*/\partial T = q^*\phi y(T + \mu^*)^{y-1}/(2E_2)^y > 0$. □

References

Azar, O.H., 2003a, What sustains social norms and how they evolve? The case of tipping, Journal of Economic Behavior and Organization (forthcoming).

Azar, O.H., 2003b, The history of tipping – from sixteenth-century England to United States in the 1910s, Journal of Socio-Economics (forthcoming).

Azar, O.H., 2003c, Why do we tip waiters but not flight attendants? An empirical investigation, working paper, Department of Economics, Northwestern University.

Azar, O.H., 2003d, The implications of tipping for economics and management, International Journal of Social Economics (forthcoming).

Ben-Zion, U., and E. Karni, 1977, Tip payments and the quality of service, in: O.C. Ashenfelter and W.E. Oates, eds., Essays in labor market analysis (John Wiley & Sons, New York) 37-44.

Bodvarsson, O.B., and W.A. Gibson, 1994, Gratuities and customer appraisal of service: evidence from Minnesota restaurants, Journal of Socio-Economics 23, 287-302.

Bodvarsson, O.B., and W.A. Gibson, 1997, Economics and restaurant gratuities: determining tip rates, American Journal of Economics and Sociology 56, 187-204.

Crusco, A.H., and C.G. Wetzel, 1984, The Midas touch: the effects of interpersonal touch on restaurant tipping, Personality & Social Psychology Bulletin 10, 512-517.

Hemenway, D., 1993, Prices & choices: Microeconomic vignettes, 3rd edition (University Press of America, Lanham).

http://www.tipping.org

Jacob, N.L., and A.L. Page, 1980, Production, information costs and economic organization: the buyer monitoring case, American Economic Review 70, 476-478.

Lynn, M., G.M. Zinkhan, and J. Harris, 1993, Consumer tipping: a cross-country study, Journal of Consumer Research 20, 478-488.

Lynn, M., and M. McCall, 2000a, Beyond gratitude and gratuity: a meta-analytic review of the predictors of restaurant tipping, Working paper, School of Hotel Administration, Cornell University.

Lynn, M., and M. McCall, 2000b, Gratitude and gratuity: a meta-analysis of research on the service-tipping relationship, Journal of Socio-Economics 29, 203-214.

Ruffle, B.J., 1998, More is better, but fair is fair: tipping in dictator and ultimatum games, Games and Economic Behavior 23, 247-265.

Ruffle, B.J., 1999, Gift giving with emotions, Journal of Economic Behavior and Organization 39, 399-420.

Schwartz, Z., 1997, The economics of tipping: tips, profits and the market's demand-supply equilibrium, Tourism Economics 3, 265-279.

Scott, W.R., 1916, The itching palm: A study of the habit of tipping in America (The Penn Publishing Company, Philadelphia).

Segrave, K., 1998, Tipping: an american social history of gratuities (McFarland, Jefferson).

Seligman, D., 1998, Why do you leave tips? Forbes, December 14.

Sinay, R., 2001, Trying to protect waiters hurts their income, *Ha'aretz*, June 4 (in Hebrew).

Sisk, D.E., and E.C. Gallick, 1985, Tips and commissions: a study in economic contracting, Federal Trade Commission Bureau of Economics Working Paper 125.

The New York Times, 1899, Topics of the times, November 21, 6.

U.S. Census Bureau, 2002, Statistical Abstract of the United States: 2002, on-line at http://www.census.gov/prod/2003pubs/02statab/services.pdf.

Wessels, W.J., 1997, Minimum wages and tipped servers, Economic Inquiry 35, 334-349.

Figure 1: Optimal Choice of μ ($T^0 < T^1$; $y < 1$)