REFERENCE-DEPENDENT PREFERENCES AND THE ALLOCATION OF EFFORT OVER TIME:

Evidence from natural experiments with bike messengers

By

Lorenz Goette* and David Huffman**

Abstract

We use natural experiments at two firms – plausibly exogenous increases in the piece rate – to study the impact of incentives on within-day effort profiles. Our first finding is that raising the commission rate has zero effect on total effort over the day. This is similar to other studies, which show no effect, or even a negative effect of incentives on effort. However, this evidence is consistent with two competing explanations: lack or response to incentives could be due to fatigue, or it could be due to reference dependent preferences (income targeting). Distinguishing between these explanations is important for understanding what can be done to improve the effectiveness of incentives, and is important because reference dependent preferences may affect the allocation of effort in a broader array of work environments. In this paper we are able to distinguish between these explanations, because our data are richer than previous studies' and allow us to look closely at within-day effort profiles. Although we find that a higher piece rate has zero net effect on daily effort, we find a strong effect on the way that effort is allocated within the day. On the higher piece rate, workers exert significantly more effort early in the day, but work significantly less hard later in the day. We show that this pattern is difficult to explain with fatigue: a broad class of fatigue functions predicts a non-negative response of effort to incentives at all points during the day. Furthermore, some plausible forms of fatigue predict that effort should increase most strongly at the end of the day, which is the opposite of what we observe. The effort profiles we observe are, however, consistent with a model in which workers have reference-dependent preferences and a salient daily income goal. We conclude that reference-dependent preferences are important for understanding daily labor supply.

---

* University of Zurich and IZA, Bonn. Email: lorenz@iew.unizh.ch

** IZA Bonn. Email: huffman@iza.org

I. Introduction

The standard economic model captures a strong intuition: workers should work harder when wages are high. This intuition is appealing, because it seems to offer a solution to a fundamental problem facing employers, namely the problem of encouraging workers to choose the optimal timing of effort. However, this intuition is not as general as it is often supposed. An obvious objection is that costs of effort, or fatigue, may dampen the response of effort to incentives. There is also a competing intuition, based on the observation that the motivation to exert effort often comes from a desire not to fall short of a goal, or reference point. In the case that workers have a daily income goal in mind, it is intuitive that a higher wage can actually lead to lower total effort, because it causes the worker to reach their goal more quickly and removes this important source of motivation for the rest of the day.

A recent literature has tested the response of effort to incentives in "neoclassical environments," in which workers have relative freedom to choose effort. Despite these favorable environments, there is little evidence to support the prediction of the standard model. The basic stylized fact from this literature is that wage elasticities of effort are typically zero or negative (Camerer et al, 1997; Chou, 2002; Shearer, 2002; Treble, 2003; Farber, 2003; Fehr and Goette, 2003). Some of these studies are not entirely conclusive, because the variation in wages may not be exogenous to the choice of effort (e.g. Camerer et al, 1997; Chou, 2003; Farber, 2003). In other studies, it is clear that effort responds negatively to higher wages (Fehr and Goette, 2002). However, to identify the source of small, and sometimes perverse, effect of incentives on effort one needs further information. The results from these studies are in most cases consistent with three possible explanations: inelastic marginal costs, fatigue, or reference-dependent preferences. Distinguishing between these explanations is important, for understanding what can be done to improve the effectiveness of incentives and because reference-

dependent, or goal-based preferences may affect the allocation of effort over time in a broader array of work environments than inelastic marginal costs or fatigue.[1]

This paper uses a natural experiment occurring at two bicycle messenger firms to compare the within-day effort profiles of workers before and after a wage increase. The data we use are the most detailed data we are aware of for studying the response of effort to incentives within a firm; we observe hourly productivity data for all workers on a given day, over the course of several years, at multiple firms. We also observe natural experiments at these firms, which are changes in the piece rate that are plausibly exogenous. The combination of exogenous variation in wages and the ability to observe within-day effort profiles contributes extra information that makes it possible distinguish between fatigue and RDP. By contrast, previous studies have typically relied on measures of labor supply that are aggregated to the day.

We derive the predictions of three different models regarding the impact of a wage increase on a worker's within-day effort profile. These are novel predictions; there has been little consideration of the optimal allocation of effort over a day, partly because of the data limitations of previous studies. A prototypical model with time-separable utility predicts a positive response of effort to incentives throughout the day, and an increase in total effort, except in the case that the marginal cost of effort is sufficiently steep as to make zero response optimal in all work episodes. We show that a model of fatigue, in which utility is no longer separable in the sense that effort in one work episode increases costs in the next, generally predicts a non-decreasing effort profile over the day and a non-negative response to incentives at all points during the day.. We also show that a model with RDP can predict a negative impact of incentives on total effort, because the wage increase causes the worker to reach their target earlier in the day, at which point they reduce effort for the rest of the day.

---

[1] Reference-dependent preferences have implications beyond piece-rate environments, in settings where fatigue is less likely to be important, e.g. RDP may affect the allocation of effort over time in white-collar jobs.

The evidence we find is consistent with the RDP model, but not with inelastic marginal costs or fatigue. At both firms, the wage increase has zero effect on total effort. This finding alone is consistent with all three of the alternative explanations. However, the wage increase has a strong effect on the way that effort is allocated within the day; on the higher commission rate, workers exert significantly more effort early in the day, but work significantly less hard at the end of the day, resulting in a net increase of zero. Steep marginal costs cannot explain this result: this explanation only predicts zero response of total effort if there is zero response at all points during the day, but this is not at all what we observe. A model of fatigue cannot explain the reduction in effort at the end of the day, because it also predicts a non-negative response of effort in all work episodes following a wage increase. In fact, some versions of the fatigue model predict the exact opposite of the pattern we observe: an especially strong increase in effort at the end of the day. The profile we observe is, however, consistent with a model of reference-dependent preferences, in which workers on a higher commission rate work less hard later in the day because they surpass their income goal more quickly. Based on our evidence on within-day effort profiles, we conclude that reference-dependent preferences are the most likely mechanism behind the perverse effect of incentives observed in this and other studies.

Having found evidence that RDP play a role in daily labor supply, it is interesting to ask whether this behavior is robust to experience. This question echoes a recurrent theme in the experimental economics literature, which is that many departures from the standard economic model disappear if individuals are given enough time to learn. The question is also particularly relevant for RDP: there is strong disagreement about whether RDP reflects a true preference, or is a bias in judgment that can be eliminated over time. One difficulty with testing the learning hypothesis outside of the laboratory is determining the relevant amount of time for separating experienced from inexperienced individuals. However, our data are very clear in this regard: the earnings-experience profile for bicycle messengers shows a steep learning curve for the first four months of employment, but essentially no improvement in earnings thereafter. Using four months as the relevant cutoff, we compare the responses of experienced and inexperienced messengers to a wage increase. We find that the perverse response to the wage increase is actually more pronounced for the experienced messengers, in terms of the initial increase in effort and

also the decrease in effort at the end of the day. This goes in the opposite direction of the learning hypothesis, providing no evidence that experience eliminates RDP.

The rest of the paper proceeds as follows. Section II derives theoretical predictions for the three different models, regarding the impact of incentives on effort profiles. Section III gives the empirical setup, describing the firms in our study, our data, the empirical strategy we use, and the specifics of our estimation methods. Section IV presents our results. Section V concludes.

II. Theory

The standard economic model captures a strong intuition regarding the impact of incentives: Workers work harder when wages are high. The aim of this section is to point out the circumstances in which this intuition is correct, and the circumstances in which it is misleading. We focus on three different models: (1) time-separable utility and inelastic marginal costs; (2) inseparable utility due to fatigue effects, i.e. exerting effort early in the day raises the marginal cost of effort for the rest of the day; (3) preferences that are reference-dependent. We explore the implications of these models for within-day effort profiles. We show that the first model always predicts that individuals exert more effort, except if the marginal costs of effort are very inelastic. The second model is useful mainly because it incorporates realistic fatigue effects, and shows how these affect within-day effort profiles. However, we show that fatigue effects alone are insufficient to overturn the prediction of the standard model.[2] Finally, the third model shows how reference-dependent preferences (RDP) can reverse the intuition that higher wages should lead to more effort.

---

[2] We thus correct a misleading intuition, which is that fatigue effects can cause a worker to work harder initially during the day, and then reduce effort, in response to a wage increase.

A. Time-Separable Preferences and Inelastic Marginal Cost

The prototypical economic model assumes that individuals maximize a time-separable utility function over their lifetime, or more realistically, over an extended period of time such as several years. In this paper, we are interested foremost in the allocation of effort within a given day, during such an extended period of time.

To fix notation, divide a workday k into m work episodes of length Δ. We adopt the convention that episode t lasts from t to $t + \Delta$ on day k. To characterize the allocation of effort to different work episodes within a day, we do not need to solve the full lifetime optimization problem; intertemporal maximization implies that the optimal choice of effort within each work episode t has an equivalent representation as

$$\underset{e_t}{Max}\ V_t = \lambda w_k e_t - c(e_t) \qquad (1)$$

The parameter λ is the marginal utility of lifetime income at the optimum, w is the (discounted) wage on workday k, and e_t – in a slight abuse of notation - denotes labor supply in episode t (on workday k). $c()$ is the convex cost of effort in terms of utility in that work episode.[3] A noteworthy feature of the individual's choice problem is that small changes in w lead to only small changes in lifetime wealth, and thus leave λ essentially unchanged. Hence, the valuation of income in each episode should be linear. We discuss below how reference-dependent preferences may overturn this prediction and lead to non-linear valuation of income over the course of a day.

The optimality condition for effort in episode t is

$$c'(e_t) = \lambda w_t \qquad (2)$$

It its simplest form, this is where the intuition that higher wages should lead to higher effort comes from: Because a small change in the wage rate leaves lifetime wealth essentially unchanged, λ remains constant. It is then easy to see why higher wages call for higher effort. There is an even clearer prediction in the case where the individual

---

[3] The convexity of $c()$ follows from the concavity of the underlying intertemporal preferences (see the appendix in Fehr and Goette, 2002, for a proof).

knows that at some point in time, wages will increase. In this case, any income effects are entirely absorbed in λ as soon as the individual learns about the wage profile. This prediction holds for any size wage increase, as long as the increase is anticipated (see also Browning, Deaton, and Irish, 1985).

However, the first qualification to the standard intuition is also apparent in this simple optimality condition. The curvature of $c'()$ limits the response of effort to the wage in the time separable model. In particular, if marginal cost increases very rapidly in effort, this dampens the response of effort to a wage increase. In the limit, not responding to the higher wage may be optimal.

B. Fatigue

While the time-separable model is useful for illustrating the mechanism behind the intuition that is the focus of this section, it is clearly incomplete. For example, it seems important to consider the possibility that individuals get tired within a workday. Little attention has been paid to how this affects the optimal allocation of effort over the workday. In this subsection, we consider two broad classes of fatigue, and show how the presence of fatigue affects the effort profile, and, more importantly, the response to incentives. By fatigue, we mean that effort earlier in the day raises the marginal cost of effort later in the day.

Fatigue with Recovery: We say that an individual "recovers" from fatigue if, having stopped working for one episode t, the individual enters the next episode with lower marginal costs. Formally, effort costs in period t depend on effort e_t and the "stock" of fatigue $k_t = e_{t-1} + \delta k_t$ with $0 < \delta < 1$. We assume that

$$c(e_t,k_t) = c(e_t + \delta k_t)$$

Thus, the stock of fatigue decays exponentially, capturing the idea of recovery; if an individual rests, this will lower the marginal costs in the next period. In an appendix (available on request) we show that in this case the optimal effort profile is U-shaped, without any further assumptions. There is a strong intuition behind this u-shaped profile: Optimality dictates that the marginal costs of effort are constant across periods. Now

consider the second work episode: Because a fraction δ of the initial effort level e_1 has decayed, keeping the marginal cost of effort constant implies setting $e_2 = (1 - \delta)e_1$ in the second episode. Therefore, effort is higher in the first episode than in the second. The condition that describes second-period effort also applies to all subsequent periods, until the last, dictating a constant effort-level during these periods. In the last period, effort increases, because it has no further effect on future work episodes. In an appendix, available on request, we derive this formally.

In the case of fatigue with recovery, an increase in the wage unambiguously causes effort to rise in all work episodes. Fatigue dampens the effect of a wage increase, and the dampening increases with δ. But it is clearly optimal to increase effort in all episodes, in order to equate the marginal costs of effort across the day.

Fatigue without Recovery: We model fatigue without recovery along the same lines. The key difference is the stock of fatigue, which in this case is given by $k_t = e_{t-1} + k_t$ because the stock never decays. We assume that

$$c(e_t,k_t) = c(e_t + \beta k_t)$$

with $0 < \beta < 1$. In this case, the optimal effort profile is strictly increasing, provided the marginal cost of effort is "not too convex".[4] The intuition behind the result is straightforward: Because the fatigue stock never declines, there is an incentive to postpone effort towards the end of the day, when there are no future consequences of effort in terms of fatigue. But then, equating the marginal costs of effort calls for a smooth increase in effort over the day, provided that the marginal cost of effort does not increase too rapidly.

Effort increases at all points in time in response to a wage increase. We derive this result in the appendix. There is no robust prediction other that effort increases in each episode. However, in many cases, e.g. with linear marginal costs, individuals respond more to

---

[4] In the appendix we prove that the effort profile is strictly increasing if the *marginal* costs of effort are concave or if they are linear. Because the results hold for the linear case, they must hold for some degree of convex curvature as well.

wages later in the day than early in the day, for the same reason that the effort profile is increasing (but if there is significant convex curvature in marginal costs, this works in the opposite direction). Again, fatigue effects dampen the response to a wage increase.

To summarize, both forms of fatigue predict a dampened response of effort over the day to wages, but effort can never decline with wages. Fatigue without recovery can generate a back-loaded response in some cases, while fatigue with recovery predicts an even, but probably small increase throughout the day.

C. Reference-Dependent Preferences

The psychology literature suggests that the previous models leave out a fundamental aspect of preferences: Individuals tend to evaluate outcomes as gains and losses relative to a reference point, and thus may especially dislike a low daily income, because it feels like a loss.

More formally, reference-dependent preferences have two important features that influence the valuation of outcomes: *Loss Aversion*, which is the tendency for individuals to feel more strongly about avoiding a loss of one unit than making a gain of one unit, and *diminishing sensitivity*, which is the decrease in the marginal valuation of another unit of the outcome as the distance from the reference point increases.

These two features are captured by the Kahneman-Tversky (KT) value function. The value function exhibits a kink at the reference outcome, reflecting loss aversion, and is concave in gains, and convex in losses, reflecting diminishing sensitivity.

The KT value function is relevant for labor supply if workers have a reference income, or income goal in mind. Experimental evidence shows that goals are a pervasive aspect of human decision-making, and furthermore that goals, even if arbitrarily assigned, inherit the properties of the value function (Heath, Larrick, and Wu, 1999). The particular work environment we study in this paper is one where there is a salient link between effort and earnings at any point in time, and individuals have discretion over how much effort they want to exert. While this is a situation that corresponds most closely to the standard

economic model of labor supply, the salience of income might also trigger behaviorally relevant sensations of gains and losses relative to a daily reference income.

The results in Heath, Larrick and Wu (1999) strongly suggest that if an individual is working towards a goal, it is not so much the overall likelihood of reaching the goal that determines sensations of gains and losses, but rather the current distance from the goal that triggers these sensations.[5] Loewenstein, O'Donoghue and Rabin (2003) argue that this pattern is more general. They develop a formal model of *projection bias* to capture the intuition that when evaluating preferences about future states, individuals' evaluation is biased towards the current state: E.g., when shopping while hungry, an individual will buy too much food, because she can't imagine she will ever be much less hungry that she is now. In the application that we consider, this bias implies that the worker expects the future value of income to be similar to the current valuation, e.g. early in the day she does not fully anticipate the added urgency she will feel when she approaches her reference income. This implies that the value of income in the current period is largely determined by the income earned so far, and not so much by expectations about earnings over the rest of the day.

A simple way to incorporate this into our model is to assume that income earned up to episode k in a day *entirely* determines sensations of gains and losses.[6]

---

[5] See problems 8,9 and 10 in their paper. The problems compare scenarios where two individuals have a goal for an activity over the next month. The difference is that, in one case it is framed as a monthly goal, whereas in the other case it is framed as a weekly goal equal to one fourth of the monthly goal. The results indicate that in the first week, the individual with the monthly goal has a harder time motivating herself to work towards that goal. If all that mattered for the monthly goal was whether a goal was eventually achieved, this should not be the case.

[6] This corresponds to the case of full projection bias in O'Donoghue, Loewenstein, and Rabin (2003). All our results also hold for partial projection bias, in which the predicted future state is “between” the current state and the true future state.

$$V_t = v(we_t + y_t - r) - c(e_t)$$

where y_t is income earned up to episode t. For simplicity, we ignore fatigue effects, as they are not essential to our argument. The last equation has a straightforward implication for the allocation of effort over time: The first-order condition for effort is

$$wv'(we_t + y_t - r) = c'(e_t)$$

This produces a distinct pattern of effort over the day: Early in the day, the messenger is relatively far from the income target. Diminishing sensitivity implies that the valuation of income is relatively low, and consequently effort is relatively low. As the messenger approaches the income target, the marginal valuation of money increases, leading to higher effort. Once the income target is surpassed, the marginal valuation of income drops, because losses are no longer a concern.

Reference-dependent preferences also lead to a distinct profile following an increase in incentives. Unlike the standard model, which unambiguously predicts an increase in total effort over the day following a wage increase, the RDP model predicts that the effect is ambiguous: total daily effort may increase, or it may decrease, in response to higher wages. With RDP, a wage increase has two effects, which work in opposite directions. Higher wages lead to an increase in effort at all distances from the target.[7] At the same time, higher wages cause the individual to reach the target earlier in the day, and thus cause the drop in the marginal valuation of effort associated with surpassing the target to occur earlier in the day.[8] If the drop in valuation at the target is sufficiently large, the

---

[7] In a model with partial projection bias, this is not necessarily true. While this makes it more likely that the overall effect of wages on effort is negative, it also attenuates the time profile.

[8] An alternative way to model this is to assume that an individual experiences anxiety (has psychological costs) each moment that they are below the reference income, along the lines of Caplin and Leahy (2001). This creates a motive to exert more effort early in the day, in order to minimize the time spent below the reference level. Higher wages increase effort at any given distance from the reference income. But higher wages also

effect of being beyond the target for a larger portion of the day outweighs the fact that the individual works harder at all distances from the target, leading to a decrease in total effort.

III. Empirical Setup

A. Bicycle messenger firms

Bicycle messenger firms offer same-day, or same-hour delivery of packages in most cities around the world. In urban areas that are congested with traffic, bicycle messengers offer an advantage, in terms of speed, over vehicle couriers. Demand for bicycle messenger services comes from various sources, including law firms, advertising agencies, real-estate companies, and scientific laboratories, all of whom require frequent deliveries, e.g. legal documents filed in court or blood samples taken from a doctor to a laboratory.

Compensation for bicycle messenger is typically based on a piece rate, which is a fixed fraction of the price of each delivery. Messenger firms price deliveries based on the distance between the pick-up and drop-off locations for the delivery, and on how quickly the customer needs the delivery to be moved from one location to the other.

Bicycle messengers are attractive subjects for the study of intertemporal substitution, because they have substantial discretion over their choice of effort during the day. Deliveries are announced over the airwaves by a dispatcher, and the announcements are heard by all of the company's messengers who are working that day. Messengers have several ways to vary effort in this setting: they can work hard to finish deliveries quickly, and lobby the dispatcher for more deliveries, or they can make deliveries slowly, and respond slowly to the dispatchers calls on the radio.

---

make it more attractive to surpass the income target earlier in the day, and avoid some of the psychological costs for the rest of the day. This leads to the same ambiguous effect on total effort that is captured by our model.

Importantly, dispatchers at messenger firms typically have another duty besides calling out deliveries. As soon as a delivery is requested by a customer, the dispatcher creates a record for that delivery in a computer database. They update this record as new information becomes available, e.g. recording the timing of the delivery, and which messenger actually makes the delivery.

B. Data

The data that we have obtained from several bicycle messenger firms is much more detailed than the data used in previous studies of intertemporal substitution, allowing us to compare the within-day effort profiles of many individuals working on a given day.[9] We use the electronic delivery records created by dispatchers as they receive calls from customers, dispatch deliveries, and receive confirmation of deliveries. These records include the price and timing of each delivery, made by each messenger on a given day. This paper focuses on two bicycle messenger firms, one in Basel, Switzerland (Firm A) and one in San Francisco, California (Firm B). The records at Firm A start in June 1998 and end in July 2003. The records at Firm B start in June, 2001 and end in May, 2003.

D. Natural experiments

Firms A and B promise to give messengers a higher piece rate after they have worked for several months. The firms use this deferred compensation scheme to retain messengers and reduce turnover. Importantly, this strategy requires that firms make messengers fully aware of the future increase in the piece rate. The data show that this is implemented in a similar way at the two firms: at Firm A the average time a messenger works before the

---

[9] With one exception, previous studies have used effort data that is aggregated to the day. Farber (2003) looks at within-day effort profiles, but has no exogenous variation in wages to distinguish between different models of labor supply. See Goette, Huffman, and Fehr (2004) for a discussion of the identification problem that arises in this case.

piece rate is increased is 12 weeks, and at Firm B the average is 14 weeks. One noteworthy difference, in terms of the data that we have for the two firms, is the number of messengers who switch piece rates during the sample period. In the data for Firm A, we observe 110 out of 260 messengers switching from the low piece rate to the high piece rate, whereas in the data for Firm B we only observe 8 out of 25 switching. This means that in the empirical analysis below, the effects of increasing the piece rate at Firm B will be identified using substantially fewer observations than at Firm A

E. Descriptive Statistics

This section uses descriptive statistics to give a sense of how the workday is organized at Firms A and B, providing a backdrop for our empirical analysis in the next section. Overall, the two firms are quite similar, considering that they are operating in different countries. Both firms share the same essential institutional ingredients that are common to bicycle messenger firms, such as piece-rate pay for messengers, and relative freedom to choose effort during the day. There are some notable differences, however, which we discuss as we go along.

The main difference between the two firms is the length of the workday for a messenger. At Firm A, the workday is divided into two shifts. At Firm B, messengers work for a whole day. Panel A of Figure 1a shows the distribution of starting and quitting hours for the early shift at Firm A, and Panel B shows the same distributions for the later shift. The early shift goes from 7:00 or 8:00 in the morning until 12:00 or 13:00 in the afternoon. The later shift starts at 12:00 or 13:00 and goes until 17:00 or 18:00. Figure 1b shows the distributions of starting and quitting hours for a workday at Firm B: The typical workday at Firm B begins at 7:00 or 8:00 in the morning and lasts until 17:00 or 18:00.

We measure the time spent working in two ways. Hours-on-duty are defined as hours between the first delivery and the last, including hours during which the messenger had no deliveries (breaks). An alternative measure is active hours, i.e. only those hours in which the messenger had positive earnings, which comes closer to capturing the time spent actively working. At Firm A, the average shift length, measured as hours-on-duty, is 5.05 hours. The average number of active hours is 4.85. At Firm B, the average

workday lasts 9.84 hours, and the average number of active hours is 8.28. By either measure, messengers at Firm A work about half as many hours, on average, as messengers at Firm B.

Tables 1 and 2 give distributions from Firms A and B, for both of our measures of hours-worked. At Firm A, the majority of shifts last five hours, but roughly 20% of shifts are four hours long, and 25% of shifts entail six or more hours on-duty. At firm B, the majority of workdays are ten hours long, but roughly 20% are nine hours long, and 20% are eleven hours long. Looking at the distributions for active hours, it is clear that the longer workday is associated with more time spent inactive during the day. At Firm A, the distribution for active hours is quite similar to hours-on-duty, showing only slight evidence of inactivity. At Firm B, however, 70% of workdays involve ten or more hours-on-duty, but only 10% of workdays have this many active hours. Most workdays at Firm B -- about 66% -- have between eight and nine active hours, and the remaining 20% of the workdays range from five to seven active hours. Part of the explanation for this within-day inactivity, roughly one hour, is the fact that the workday at Firm B includes a lunch break, whereas the messengers at Firm A typically have lunch outside of working hours, i.e. after the morning shift has ended, or before the start of the afternoon shift.

Average daily revenues are approximately 245 CHF at Firm A (≈ $200) and $300 at Firm B. Using the average commission rate at each of the firms, this translates into daily earnings of roughly $80 and $155, respectively. Given that messengers at Firm A work half as many hours as messengers at Firm B, hourly earnings are about $1 higher at Firm A.

Figures 2a and 2b show the distributions of daily revenues at Firms A and B. Clearly there is substantial variation in revenues at both firms. Table 3 presents an analysis of variance for daily revenues. At Firm A, stable messenger characteristics explain more of the variance than day effects. At Firm B, day effects explain slightly more than messenger effects. Together, day and messenger effects explain approximately 42% of the variation in daily revenues at Firm A, and 53% of the variation at Firm B. This leaves substantial variation in revenues due to messenger-specific, day-specific shocks.

F. Empirical Strategy

Our empirical strategy is to compare the labor supply of individuals working on a given day. Our measure of labor supply is defined as follows: We follow each messenger working at Firm A on a particular day for seven hours (where the convention is a five-hour workday) and follow each messenger for 12 hours at Firm B (where the convention is closer to ten hours); we define hourly labor supply as hourly revenues. This creates seven measurements of hourly labor supply for each workday at Firm A, and twelve for each day at Firm B. If a messenger had zero revenues during a particular hour, we set labor supply to zero in that episode.

This measure of labor supply is the broadest possible, and is precisely as standard theory suggests it should be. It captures (i) how hard a messenger is working, (ii) whether he is taking breaks during the day, and (iii) when the messenger quits for the day (after the messenger quits, we set labor supply to zero for the remaining hours in the workday). It will sometimes be useful to contrast this measure with a measure that only reflects the choice of work hours, exluding the choice of effort. In this case, the dependent variable is a binary variable, equal to one if hourly revenues are positive, and zero otherwise.

We estimate equations of the form

$$e_{iht} = \gamma^1 high^1_{iht} + \gamma^2 high^2_{iht} \;..\; + \gamma^m high^m_{iht} + \; x_{iht}b + a_i + d_{ht} + u_{iht}$$

e is labor supply (as defined above) of messenger i at hour h on date t. Our coefficients of interest are the γ^k's. The variable $high^k$ is a dummy variable equal to one if it is the messenger's kth hour of work and he is on the high commission rate, and zero otherwise. We want the γ^k coefficients to reflect the messengers' change in labor supply in work hour k due to the increase in the commission rate. For the coefficients to have this interpretation, we need to control for other factors, correlated with work hour and the commission rate, that do not affect labor supply through this interaction.

The vector x contains time-varying individual control variables. Its most important subset consists of dummy variables controlling for the number of hours worked so far. Thus, the baseline effort profile does not affect the estimates of γ. x also contains a set of dummy variables controlling for experience. Especially at firm A, workers entering the firm are "rookies" with no prior experience as messengers. Hence, there could be significant

learning as they acquire more courier skills during their employment. Because there is variation in the timing of the promotion, we are able to separate learning from the change in incentives. We also include a messenger fixed effect a_i to avoid bias arising from the possibility that below-average messengers drop out before they get promoted. Finally, we include an hourly fixed effect d_{ht}, which we estimate separately for each hour on each day to control for the time profile of the availability of deliveries.

With these controls in place, γk indicates by how much the messenger changed labor supply in work hour k in response to the increase in the commission rate. As pointed out earlier, the standard model says that these coefficients should only reflect the intertemporal substitution effects, since income effects, if they have occurred, materialized at the time the messengers started working for the firms, and are absorbed in the fixed effects. According to the standard model, effort must increase after the change to the higher commission rate, and thus the γ^k 's should be strictly positive.

We estimated the equation above using OLS. An important issue is how one should calculate the standard errors of the estimated coefficients. Given the hourly frequency of our measures, there are various ways in which u_{it}, the error term, departs from the i.i.d. assumption of OLS. First, the way we construct our measure of labor supply makes the error term inherently heteroskedastic.[10] We correct for this by estimating robust standard errors. Second, there are two potential sources of correlation between the error terms. Within a given day, if one messenger was assigned a delivery, another messenger will end up with one less delivery. This leads to negative correlation of the residuals within a day, rendering OLS standard errors too large. On the other hand, there could be positive correlation in u_{it} for observations coming from a given messenger, rendering OLS standard errors too small (see Bertrand, Duflo and Mullainathan, 2002, for an extensive discussion). We are left with no alternative but to report two sets of standard errors. One set is adjusted for clustering on days. Because this ignores the (potentially) positive correlation within individuals, we consider these standard errors the lower bounds. The

[10] Because our dependent variable is bounded below by zero, this necessarily implies that the variance of the error term differs between observations.

other is adjusted for clustering on messengers. We consider this the upper bound on the standard errors, because it ignores the (potentially) negative correlation within days. However, most of our conclusions do not depend on which adjustment of standard errors we use, and if they do, we point it out in the discussion.

IV. Results

We begin by comparing our results to earlier studies. These studies rely on labor supply aggregated to the day, so for ease of comparison, the first row of Table 4 presents the aggregate response of effort to the higher commission rate, i.e., the sum over all estimated γ^k's. In the first columns for Firms A and B, we show that higher commission rates are associated with higher total revenues per day, although the effect is not robust to clustering on messengers. But since we have not yet controlled for messenger-specific differences, this should not be interpreted as the change in behavior due to the higher commission rate. In fact, the second columns for Firms A and B show that when we include messenger fixed effects, the relationship between higher commission rates and total revenues vanishes. At Firm A, the point estimate is CHF 1.58, which is tiny relative to average daily revenues of CHF 250. The point estimate at Firm B is actually negative, although it is not significant at the five percent level. Again, the point estimate is quantitatively small (USD 8.25 relative to average revenues of USD 290 per day). The last two columns in Table 4 display the effect of the higher commission rate on the total number of work episodes with positive revenues. They also show little or no adjustment in working time: At Firm A, active work time increases by 0.014 (approximately 50 seconds), and at Firm B it decreases by 0.25 (15 minutes).

The results in Table 4 are broadly consistent with the findings from other studies that examine how wage changes affect effort per day. Shearer (2002) finds no effect of higher piece rates on the daily effort of tree planters. Treble (2003) uses variation in wages for coal miners in the 19th century and finds little response of effort to the wage increase. However, as we discussed in Section II, the results in our study and in these two others

might be due to fatigue effects that dampen effort responses, or simply because labor supply is very inelastic (because the marginal costs of effort rise sharply).

A. The Time Profile of the Change in Incentives

We are able to go beyond previous studies, by comparing the within-day effort profile before and after an increase in the commission rate. The evidence we find lends no support to the hypothesis that it is too costly to respond to the wage increase. Referring again to Table 4, the second row shows that the hypothesis that all γ^k are equal to zero is overwhelmingly rejected in all cases. Figures 3a and 3b show the γ^k's over time, with error bars for plus and minus two standard errors of the estimate (thus, the 95% confidence interval of the estimate coefficient). These graphs show a clear response to changes in the commission rate. At Firm A, we see a distinct pattern of initially increasing and then decreasing effort over time, in response to the increase in the commission rate. Because the large increase in effort is offset by an almost equally large decrease of effort towards the end of the day, the result is a tiny net change in total labor supply for the day. The distortions in labor supply throughout the day are quantitatively significant. Average revenues are roughly CHF 50 in each hour if a messenger is active: Thus, early in the day the elasticity of revenues with respect to the commission rate is more than one, but later in the day, it becomes almost minus one. At Firm B, there is a similar time profile over the day. There is a clear decline in the effort profile after the increase in the commission rate, resulting in a significantly lower effort towards the end of the day with negative elasticities in the same order of magnitude observed at firm A. There is somewhat less evidence that the messengers exert more effort early in the day. A test of the hypothesis that the first four coefficients add up to zero is inconclusive ($p < 0.05$ when clustering on days, but the effect is not significant when we cluster on messengers). In all cases, we reject that the last four hours add up to zero, at the 1% significance level.

It is interesting to examine whether these changes in labor supply are due to changes in the pace of work, or to changes in the number of active work episodes (because, e.g., the messengers might take more breaks). Figure 4a and 4b display the γ^k's from the

regressions in the last two columns in Table 4, where our measure of labor supply is whether or not a messenger is active in a particular hour. Both figures are scaled so that a bar for a given hour reflects the proportion of the total change in labor supply during that hour that is due to inactivity. Thus, if the bar in Figure 4 is the same size as the bar for the corresponding hour in Figure 3 (which shows the total change in hourly labor supply resulting from the higher commission rate), this means that the entire change in labor supply in that hour is due to a change between the active and inactive states, rather than from a reduction in the intensity of work. For Firm A, the bars in Figure 4a are generally one-third to one-half of the overall effect, suggesting that about two thirds of the change in labor supply after the commission rate increase are due to changes in work intensity. At Firm B, the picture is less clear-cut. As pointed out earlier, there is little evidence that the messengers respond to the higher commission rate earlier in the day with the broad measure of labor supply, and this picture is reinforced in Figure 4b. Because of the lack of precision in the estimates at Firm B, it is also hard to interpret the coefficients for later hours, but the emerging picture is that more of the reduction in effort is due to quitting work early at Firm B, relative to Firm A.

As explained in section II.B, this pattern is inconsistent with the two very general forms of fatigue effects that we examined. These types of fatigue predict a constant, or increasing, response to the commission rate over the day. Further, they predict a strictly non-zero response at each point in time. We reject both of these implications for both firms.

The time profile observed at both firms is, however, consistent with the model of reference-dependent preferences that we discussed in section II.C. Most importantly, the RDP model predicts the decline in effort towards the end of the day resulting from an increase in the commission rate. Because the messengers on high commission rates reach their income target more quickly, the concern over falling short of the income target impacts behavior for a shorter period of the day. The drop in marginal valuation relatively early in the day can explain why messengers on the high commission rate are working less hard towards the end of the day than messengers on the low commission rate. The model also explains why the messengers exert more effort early in the day;

there is nothing in the model that makes incentives ineffective, provided that the messenger is below his income target.

Overall, the evidence from the time profiles clearly shows that individuals respond to changes in the commission rate, but this response is only consistent with the reference-dependent model. This evidence is more conclusive than that of previous studies, because within-day effort profiles bring additional information that makes it possible to distinguish between a reference-dependent model and fatigue or inelastic effort costs. By contrast, the results in Treble (2003) and Shearer (2002) are consistent with either the RDP model or these other models, and other studies that find a negative association between wages and effort have difficulties establishing that this relationship is causal[11] (Camerer et al. 1997; Chou, 2002; Farber, 2003) or need further information to distinguish RDP from other models of behavior (Fehr and Goette, 2002).

B. Does Experience Eliminate Reference-Dependence?

It is a recurrent theme in the experimental economics literature that many departures from the standard economic model disappear if individuals are given enough time to learn. The implicit assumption behind this conjecture is that the departure under consideration is caused by a lack of information, or a cognitive limitation which can be overcome through experience.

---

[11] Three studies examine the daily labor supply of cab drivers, either in New York (Camerer et al. 1997, Farber, 2003) or Singapore (Chou, 2002). They regress hours worked on a particular day on that day's wage. A robust finding is that the estimated elasticities are negative, and not caused by measurement error (see Camerer et al., 1997, and Chou, 2002). However, it is unknown where the variation in daily wages comes from. While some assert that it is daily changes in demand that drives variation in wages, there is a possibility that supply-side shocks also cause the average daily wage to change. Since none of the studies has been able to find an instrument to break this simultaneity, there still is an ambiguity in how to interpret the negative elasticities. See also Farber (2003) for a discussion of this point.

Reference-dependent preferences potentially fall into this category. There is strong disagreement over whether RDP constitute "real" preferences, in the sense that individuals truly care about sensations of gains and losses, or whether they are a temporary bias in perception that will go away with experience in the task (see, e.g., Rabin, 1997). There is evidence that is consistent with the view that RDP are a bias rather than a preference: List (2003) finds that individuals with trading experience in a real-world market are less loss averse in a laboratory experiment, and it is tempting to interpret this as a causal effect of market experience on loss aversion.[12] More relevant to our application are the findings by Camerer et al. (1997): they conclude that the behavior of more-experienced cab drivers is less driven by income-targeting than that of inexperienced drivers. However, the evidence they find is not overwhelming: In one out of three cases, the point estimates go in the wrong direction. Furthermore, Chou (2002) could not replicate this finding in his data from cabdrivers in Singapore.

In our application, there is clear scope for learning from experience. In Figure 5, we show the relationship between revenues per hour and experience. There is a dramatic and quantitatively large increase in revenues as a messenger's tenure increases. Hourly revenues increase by CHF 2.40 after one month of experience. In the fourth month, revenues are on average CHF 6/hour higher than for rookies. It is also clear that the learning curve flattens out: Increasing experience from four months to three years increases revenues by less than CHF 5/hour. The figure strongly suggests that when it comes to acquiring revenues, messengers have a lot to learn when they first begin employment, and that more than half of the overall learning takes place in the first three month. Hence, it is interesting to ask whether more experience helps to eliminate, or

---

[12] The issue here is that market experience is itself a choice variable of these individuals, so that the level (and change) in market experience cannot be viewed as truly exogenous to loss aversion. What would be needed to make this claim is exogenous variation in trading experience, such as, e.g., a trading subsidy for one group of subjects but not the other. One could then use this as an instrument for trading to see whether there is a decrease in loss aversion in the laboratory experiment.

attenuate, the RDP pattern we find in the data. Because of the long panel dimension at Firm A, we can examine whether more experienced messengers respond to the increase in the commission rate differently than less experienced messengers. We define a messenger as inexperienced if he has worked as a messenger for less than four months (because Figure 5 suggests that learning is most intense during this time), and estimate separate γ^ks for the two groups. Table 5 presents an overview of the results. In the first row, we display the impact on daily labor supply. We find that the point estimate for the more experienced messengers is smaller than the one for less experienced messengers, and is in fact negative. The difference is only marginally significant. However, when we perform a Wald test of whether the two effort profiles are identical, this hypothesis is strongly rejected, no matter how we calculate the standard errors. Figure 6 displays the two response profiles. The response of the more experienced messengers to the change in the commission rate is much stronger than the response of less experienced messengers. The more experienced messengers increase effort by more early in the day, but subsequently reduce effort by even more. Thus, if anything, the pattern predicted by RDP grows stronger as the messengers become more experienced. We conclude that our data provide no evidence that experience eliminates reference-dependent preferences.

V. Conclusion

At two different firms, we find that a wage increase has zero effect on total daily effort, in contrast to the prediction of the standard model. However, we also find that the wage increase has a strong effect on the way that effort is allocated within the day; on the higher commission rate, workers exert significantly more effort early in the day, but work significantly less hard at the end of the day, resulting in a net increase of zero. We show that inelastic marginal costs, and a broad class of cost functions with fatigue, cannot explain this pattern. However, a simple model with reference-dependent preferences can account for all the facts. We conclude that reference-dependent preferences are important for determining the allocation of effort over a day, and for explaining the lack of response of total daily effort to incentives. Finally, we find no evidence that experience eliminates the pattern of behavior generated by reference-dependent preferences.

Our findings are important for at least three reasons. First, our evidence shows convincingly that a central prediction of the standard model does not always hold, and they highlight the predictive power of a simple, alternative model that is psychologically more-realistic. Second, our findings are important for understanding what, if anything can be done to improve the effectiveness of incentives. If fatigue were the main factor explaining the lack of response to incentives, then options for the firm would consist of finding ways to reduce fatigue. However, we show that the data are consistent with goal-motivated behavior, which suggests that firms may want to learn about, and if possible try to manage workers' goals. Third, our evidence is important because reference-dependent preferences are likely to have implications for a broader array of work environments than piece-rate work. In any work task for which there is some scope for individuals to choose effort and affect the level of output, it is possible that workers have a daily goal and vary effort around this goal. Goals may thus have implications for the timing of effort, and total daily effort, in many settings besides piece rates, for instance white-collar jobs with yearly salaries. Our results raise the possibility that workers in these settings may not be allocating effort over time in the way preferred by the firm, and point to reference-dependent preferences as the mechanism behind this problem. Although incentives such as piece rates are typically not practical in these types of jobs, understanding the role of reference-dependent preferences may give firms other options for influencing effort.

References

Browning, Martin, Angus Deaton and Margaret Irish (1985), "A Profitable Approach to Labor Supply and Commodity Demands Over the Life Cycle", *Econometrica*, 53, pp. 503-43.

Camerer, Colin, Linda Babcock, George Loewenstein and Richard Thaler (1997). "Labor Supply of New York City Cabdrivers: One Day at a Time." *Quarterly Journal of Economics* 112(2), 407-41.

Caplin, Andrew and John Leahy (2001), "Psychological Expected Utility and Anticipatory Feelings", *Quarterly Journal of Economics* 116(1), pp. 55 – 80.

Chou, Yuan K. (2002). "Testing Alternative Models of Labor Supply: Evidence from Cab Drivers in Singapore." *The Singapore Economic Review 47(1),* pp. 17 – 47.

Farber, Henry S. (2003). "Is Tomorrow Another Day? The Labor Supply of New York Cab Drivers." NBER Working Paper 9706.

Fehr, Ernst and Lorenz Goette (2002). "Do Workers work more when Wages are High? Evidence from a Randomized Field Experiment." IEW Working Paper 144.

Heath, Chip, Richard Larrick, and George Wu (1999). "Goals as Reference Points." *Cognitive Psychology* 38, pp. 79-109.

List, John (2003), Does Market Experience Eliminate Market Anomalies?," *Quarterly Journal of Economics* 118(1), pp. 41-71.

Loewenstein, George, Ted O'Donoghue and Matthew Rabin (2003), "Projection Bias in Predicting Future Utility, *Quarterly Journal of Economics* 118(4), pp. 1209 – 1248.

Rabin, Matthew (1998), Psychology and Economics, *Journal of Economic Literature* pp. 1 – 46.

Shearer, Bruce (2002), "Piece Rates, Fixed Wages and Incentives: Evidence from a Payroll Experiment.", forthcoming *Review of Economic Studies*.

Treble, John (2002), "Intertemporal Substitution of Effort: Some Empirical Evidence", *Economica* 70, pp. 579 – 595.

Tversky, Amos and Daniel Kahneman (2000). *Choices, Values, and Frames*. Cambridge, MA: Cambridge University Press.

Table 1: Hours On-duty and Hours Active, Firm A

Hours On-duty	Frequency	Percent	Hours Active	Frequency	Percent
1	64	0.34%	1	72	0.38%
2	83	0.44%	2	160	0.84%
3	460	2.43%	3	877	4.63%
4	3,949	20.84%	4	4,699	24.80%
5	9,571	50.50%	5	9,233	48.73%
6	3,262	17.21%	6	3,424	18.07%
7+	1,562	8.24%	7+	483	2.55%

Note: final category includes seven or more hours or more

Table 2: Hours On-duty and Hours Active, Firm B

Hours On-duty	Frequency	Percent	Hours Active	Frequency	Percent
5	0	0%	5	91	2.68%
6	22	0.65%	6	156	4.59%
7	49	1.45%	7	472	13.88%
8	154	4.55%	8	1,058	31.11%
9	689	20.34%	9	1,244	36.58%
10	1,817	53.63%	10	365	10.73%
11	639	18.86%	11	15	0.44%
12	18	0.53%	12	0	0%

Table 3: Analysis of Variance of Daily Earnings

	Firm A	**Firm B**
	R^2 from regression	
Date fixed effects	0.17	0.29
Messenger Fixed Effects	0.29	0.25
Date and Messenger Fixed Effects	0.42	0.53

Table 4: Baseline Results. OLS Estimates

	Dependent Variable: Revenues in work hour *k*				Dependent Variable: Worked in hour *k* (dummy variable)	
	Firm A		Firm B		Firm A	Firm B
Change in daily labor supply due to increase in commission rate sum over all γ^k	4.59** (1.019) [7.136]	1.829 (2.244) [5.25]	38.724** (4.751) [20.777]	- 8.257* (5.097) [10.662]	0.014 (0.041) [0.061]	-0.250$^{**,\#}$ (0.079) [0.144]
F-test that all γ^k are zero (p-value) first : clustering on date second: clustering on messenger	p <0.001 p <0.001	p <0.001 p <0.001	p <0.001 p <0.001	p <0.001 p <0.001	p <0.001 p <0.001	p <0.001 p <0.001
Control Variables						
Work hour profile	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$
Experience profile	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$	Yes$^{**,\#\#}$
Messenger fixed effect (# of messengers)	No (260)	Yes$^{**,\#\#}$ (260)	No (25)	Yes$^{**,\#\#}$ (25)	Yes$^{**,\#\#}$ (260)	Yes$^{**,\#\#}$ (25)
Hourly fixed effects (# of hours)	Yes$^{**,\#\#}$ (17,964)	Yes$^{**,\#\#}$ (17,964)	Yes$^{**,\#\#}$ (5,499)	Yes$^{**,\#\#}$ (5,499)	Yes$^{**,\#\#}$ (17,964)	Yes$^{**,\#\#}$ (5,499)
R^2	0.592	0.612	0.258	0.388	0.747	0.602
Number of work episodes	131,573	131,573	36,277	36,277	131,573	36,277

Notes:
a) Full regression results available from the authors upon request.
b) Robust standard errors, adjusted for clustering on date, in parentheses. *,** indicates significance at the 10 and 5 percent level, respectively.
c) Robust standard errors, adjusted for clustering on messengers, in brackets. $^{\#,\#\#}$ indicates significance at the 10 and 5 percent level, respectively.

Table 5: Experience and the RDP Pattern in Effort

	Dependent Variable: Revenues in work hour k	
	Firm A	
	Inexperienced messengers	Experienced messengers
Change in daily labor supply due to increase in commission rate sum over all γ^k	5.321* (2.817) [3.914]	-1.99 (3.143) [7.713]
Difference between the two groups in overall response	7.311* (3.914) [5.941]	
F-test that all γ^k are zero (p-value) first: clustering on date second: clustering on messenger	$p < 0.001$ $p < 0.001$	$p < 0.001$ $p < 0.001$
Wald test for equality of time profiles (p-value) first: clustering on date second: clustering on messenger	$p < 0.001$ $p < 0.001$	

Notes: Same control variables and convention to calculate standard errors as in Table 4

Figure 1a

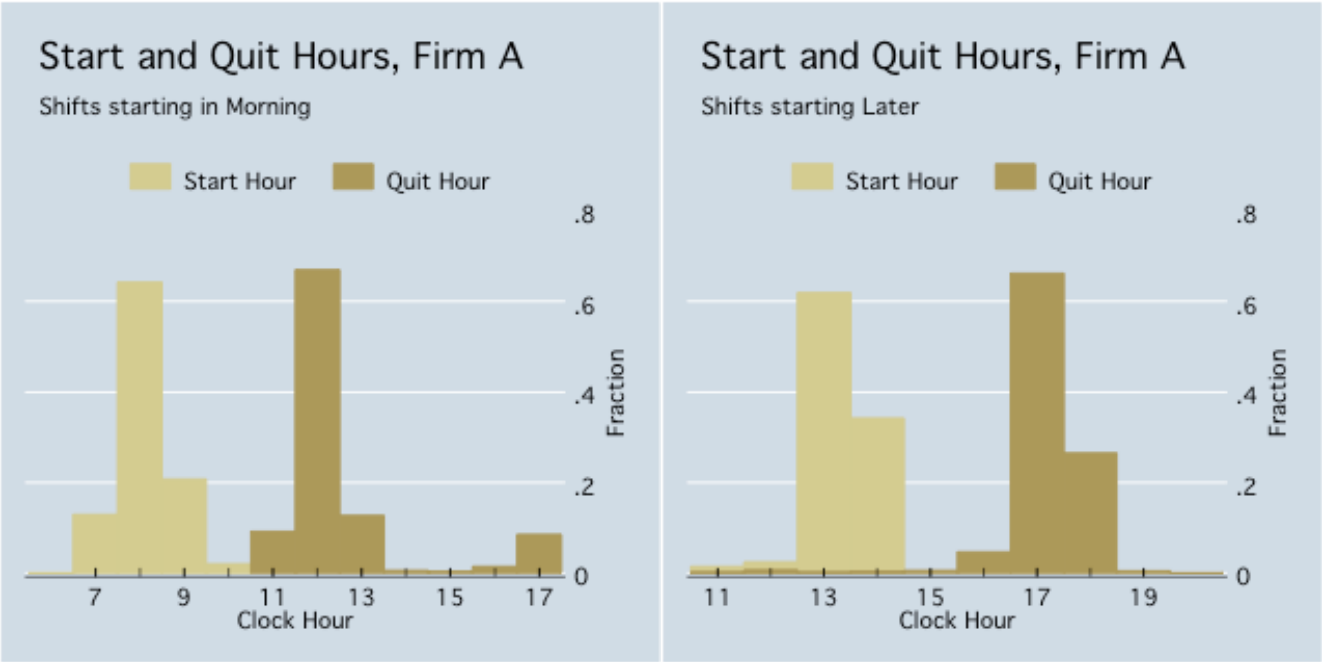

Figure 1b

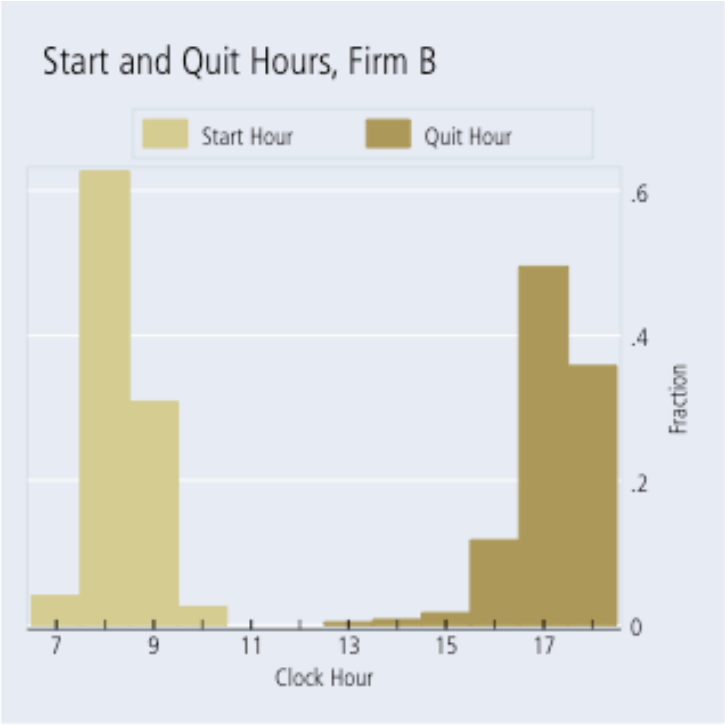

Figure 2a

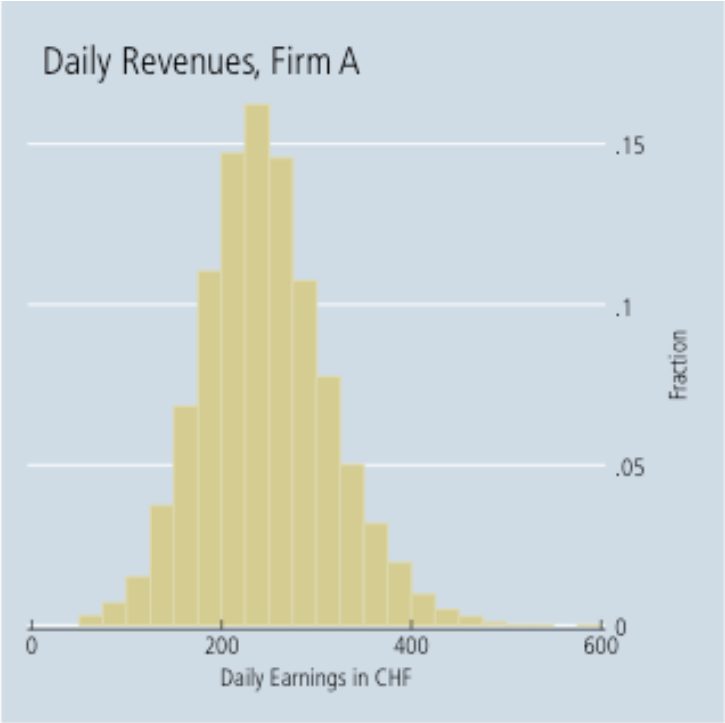

Figure 2b

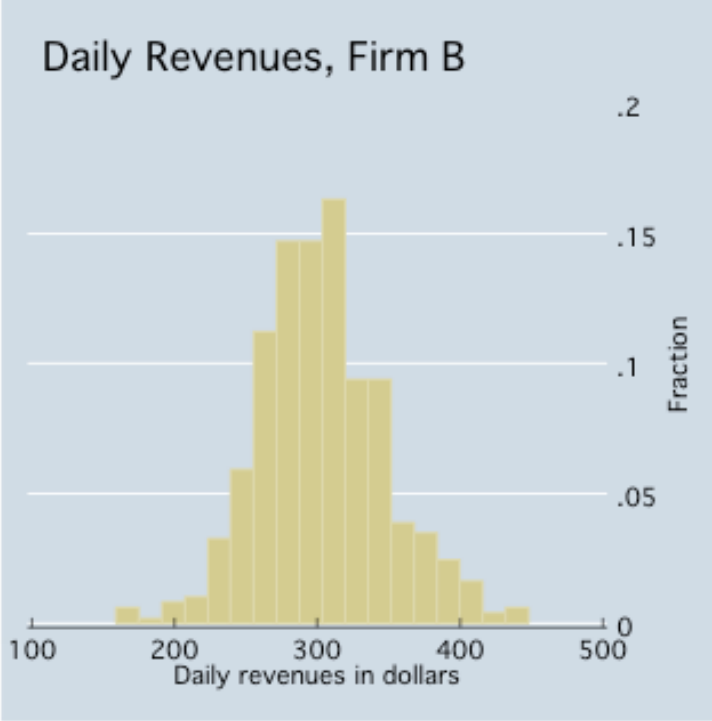

Figure 3a: The Impact of a 5 Percentage Point Increase in the Commission Rate on Messengers' Hourly Revenues, Firm A
(+/- 2s.e. of estimate, adjusted for clustering on days)

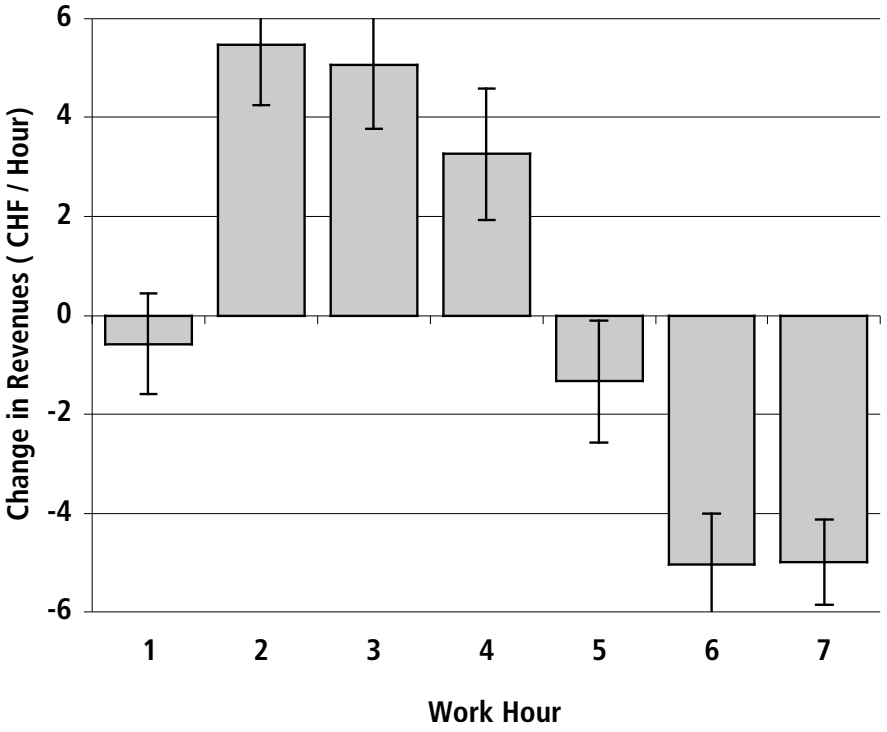

Figure 3b: The Impact of a 5 Percentage Point Increase in the Commission Rate on Messengers' Hourly Revenues, Firm B
(+/- 2s.e. of estimate, adjusted for clustering on days)

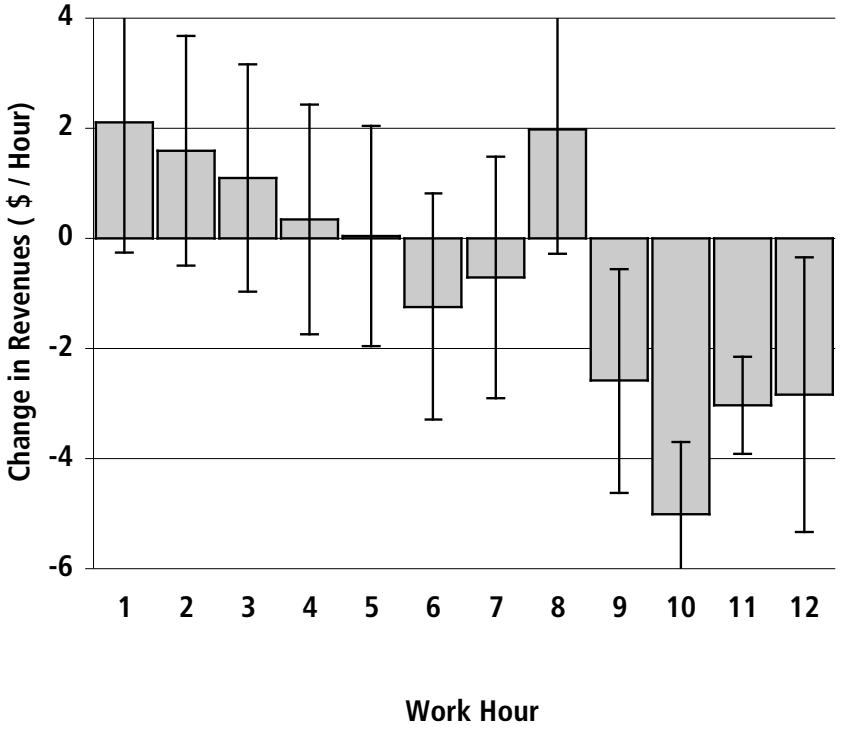

Figure 4a: The Impact of a 5 Percentage Point Increase in the Commission Rate on Hourly Activity Measure, Firm A
(+/- 2s.e. of estimate, adjusted for clustering on days)

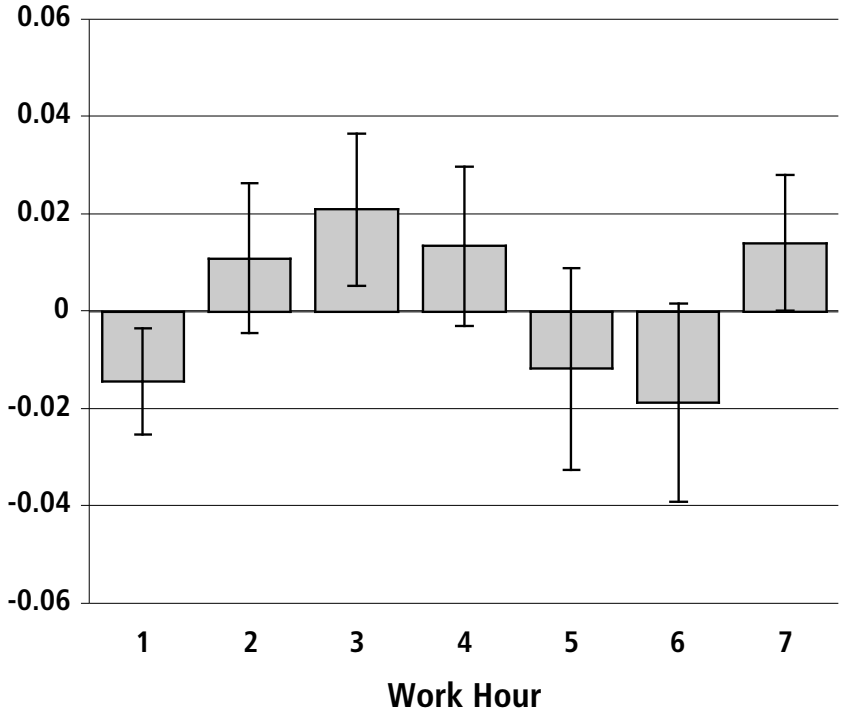

Figure 4b: The Impact of a 5 Percentage Point Increase in the Commission Rate on on Hourly Activity Measure, Firm B
(+/- 2s.e. of estimate, adjusted for clustering on days)

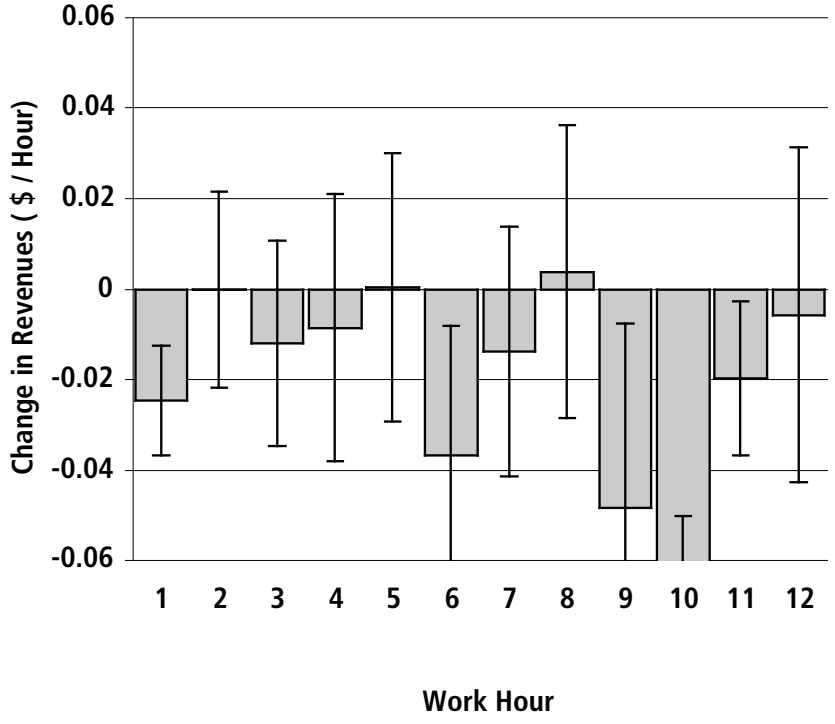

Figure 5: Experience and Revenues per Hour:
Regression Estimates of impact on hourly revenues (Regression from Table 4, column 3, +/- 2s.e. of estimate, adjusted for clustering on day)

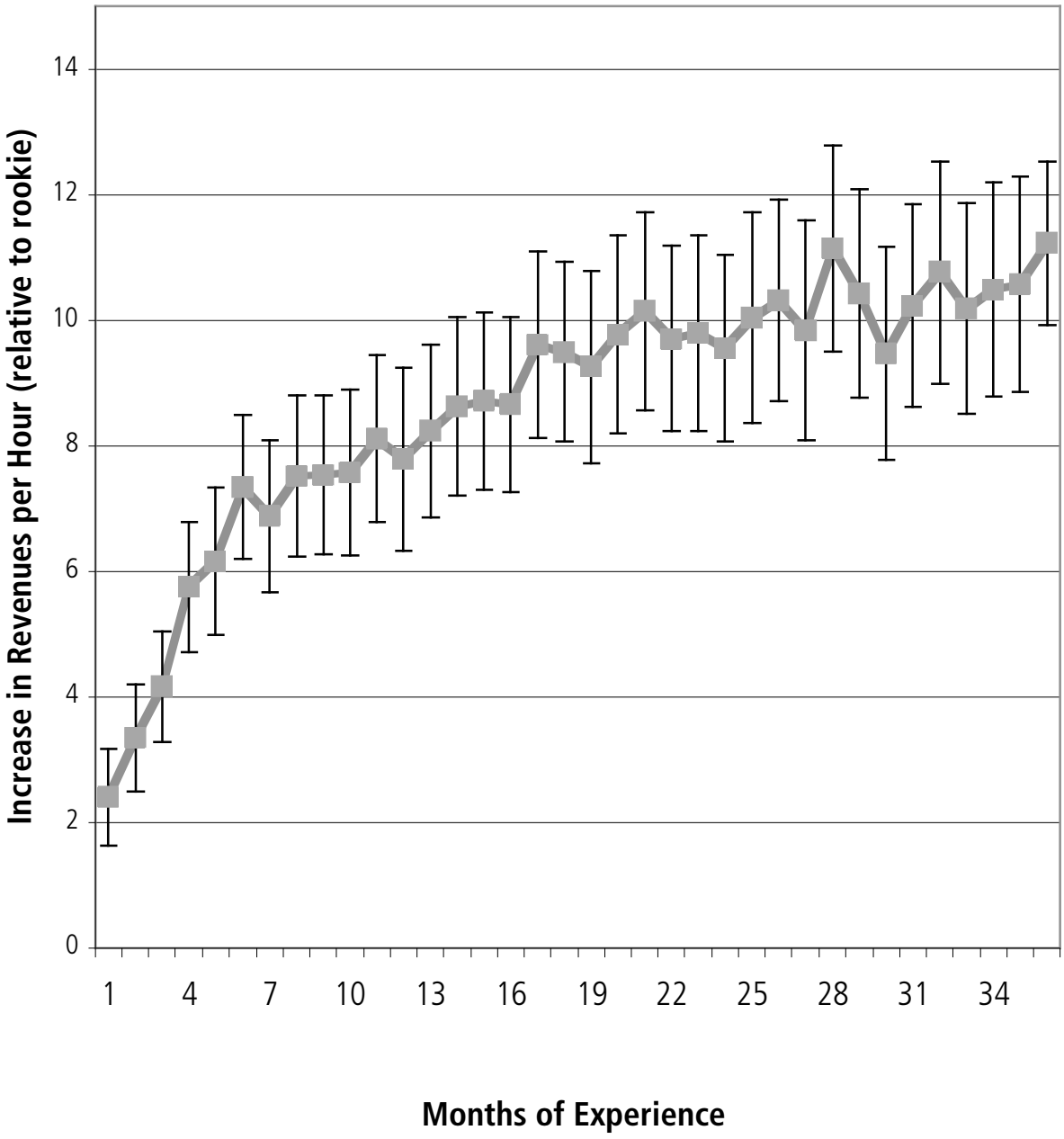

Figure 6: The Impact of a 5 Percentage Point Increase in the Commission Rate on Messengers' Hourly Revenues, Inexperienced vs. Experienced Messengers, Firm A (+/- 2s.e. of estimate, adjusted for clustering on days)