

# Do Stronger Incentives Increase Effort?

## Evidence from a Field Experiment

Ernst Fehr and Lorenz Götte

University of Zürich

This is a very preliminary version the main purpose of which is to provide material for a presentation at the European Summer Symposium on Labour Economics

April 2001

Comments are very welcome

**Abstract:** The standard economic theory of inter-temporal choice predicts that a temporary increase in the returns from working raises the supply of working hours and the supply of effort per hour. We examine these predictions with a unique high frequency data set from a field experiment. We implemented a large exogenous and fully anticipated temporary increase in the returns from working in a firm where workers could freely choose their working time and their effort. **Our results indicate that employees indeed work more hours. However, they also provide *less effort per hour*.** We show that this result cannot be attributed to the exhaustion of workers. The reduction in effort is therefore inconsistent with standard economic theory. It is, however, consistent with a view portraying workers as applying different choice brackets to different decisions. Workers have to commit themselves in advance when choosing their working hours. For this decision they apply a wide choice bracket taking into account the full benefits and costs of their decision. The effort decision, however, is made during working time which favors the application of a narrow choice bracket, i.e., the workers take into account only the narrow costs and benefits accruing instantaneously.

# 1. Introduction (to be done)

## 2. The Experimental Set-up

In this section, we describe our experimental set-up. Our study is based on the complete records of two large messenger services in Zurich, *Veloblitz*, and *Flash Delivery Services*. Around the time period of the experiment 58 messengers worked at Veloblitz and 55 at Flash. We first describe the organisation of work at a bicycle messenger service. There are three important features. First, as we will explain in more detail below, messengers can freely choose how many hours to work and how much effort to exert. Second, in both firms messengers receive no fixed wage. Instead, each individual receives a fixed share of the revenue that he or she generates. Third, demand for messenger services is highly volatile across days. This is important, because it implies that messengers are familiar with substantial variations daily earnings. Hence, if learning is important to understand the logic of intertemporal substitution, then our subjects are a well-trained set of subjects and should respond accordingly during the experiment. We then briefly discuss some of the messenger characteristics. We show that there is no tendency for particularly ill-educated individuals to end up at a messenger service.

Equipped with the institutional features, we can then discuss the experimental design. At Veloblitz, we implemented a fully anticipated large-scale exogenous variation in revenue shares. Messengers were randomly assigned to one of two treatment groups, A or B. For group A, we implemented a 25 percent wage increase during four weeks in September 2000, for group B in November 2000. During both treatment periods, this leaves the other messengers at Veloblitz and all messengers at Flash as a control group.

### 2.1 Work at a Messenger Service

Unless pointed out below explicitly, the arrangements are the same for the two messenger services, Veloblitz and Flash. In order to work at the messenger service, applicants are evaluated by an experienced messenger in order to assess the applicants quality (e.g., his fitness, knowledge of location and names of streets, courtesy, and skills handling the CB radio).

## *Hours and Effort*

Once accepted as an employee, messengers can choose freely how many hours to work. Hours are chosen in the form of shifts: One shift lasts five hours. On each weekday, there are about 30 shifts available at Veloblitz, and about 22 at Flash. At the messenger service's office, the shifts are displayed on a shift plan for every calendar week. There are two types of shifts, called "fixed" and "variable". A "variable" shift simply means that a shift is vacant on a particular day. Any messenger can sign up to work that shift, e.g., on Wednesday from 8 am to 1 pm. If a messenger commits to a "fixed" shift, this means that he will work that shift every week. For example, if a messenger chooses Wednesday, 8 am – 1 pm to be a fixed shift, he will have to fill that shift on every Wednesday. Fixed shifts can only be cancelled upon a four weeks notice period. Roughly two thirds of all shifts are fixed. All other shifts are variable and available for any messenger to sign up. Two additional points are worth mentioning. First, at both messenger services, there is no minimum number of shifts that the messengers have to work. Second, both messenger services have found it difficult to fill the available shifts. On four out of five weekdays, there is at least one unfilled shift. This implies that messengers are unlikely to be rationed in the choice of shifts. Any almost any date, there is at least one shift vacant.

Messengers' earnings are given solely by a percentage  $w$  of their daily revenues. Hence, if a messenger completes deliveries that generate revenues  $r_i$ , his earnings on that day will be  $wr_i$ . For brevity, we will refer to  $w$  as the wage henceforth. There is no lower bound on the number of deliveries that they have to complete during a five hours shift. Importantly, messengers have substantial discretion over how much effort to put into work. The messengers deployed stay in contact with the dispatcher at the messenger service's office only through CB radio. In order to allocate a delivery, say, from location A to location B, the dispatcher will contact the messenger that he thinks is closest to A to pick up the delivery. All messengers can follow the radio. If they believe that they are closer to A than the messenger that was originally contacted, they can get back to the dispatcher and say so and will then be allocated to that delivery. Conversely, if the messenger does not want to carry out the delivery from A to B, the dispatcher will not force him to do so. Thus, messengers have several means to increase the number of deliveries they complete. They can drive at higher speed, follow the radio more actively, lobby for additional deliveries aggressively, or find the shortest possible ways to carry out a delivery.

Overall, work at a bicycle messenger service comes very close to a model where individuals are completely unconstrained in choosing how many shifts (hours) they work, and how hard they work (how many deliveries they complete during a shift).

### ***The Demand for Messenger Services***

As part of the experimental setup, we obtained the complete records of all deliveries at Veloblitz and Flash between January 1999 and November 2000. These records contain every single delivery that a messenger carried out on a particular date. They allow us to track precisely when a messenger worked a shift and contain all deliveries and their prices.

Figure 1 displays the evolution of the total number of normalised deliveries per day, carried out by Veloblitz and Flash. The time period spans weekdays from January 1999 to November 2000, with the exception of a few days in October 2000, where the Flash records are missing. Since Flash also employs car messengers, we distinguish between total deliveries and deliveries carried out by bicycle messengers at Flash. All three series are normalised by the value of their first observation, because the messenger services requested that the number of deliveries be not available to their competitors. Figure 1 shows that both firms grew by approximately 60 percent over the two years considered. It is striking how strongly total deliveries are correlated at the two messenger services. The correlation amounts to  $r = 0.72$ , implying a common component of variance of 52 percent<sup>1</sup>. The figure also shows that the share of deliveries carried out by bicycle messengers at Flash steadily decreases. Nevertheless, it displays strong swings at exactly the same dates when Veloblitz experiences swings and the correlation amounts to  $r = 0.45$ , which is still very high. Figure 1 makes it plain that Flash and Veloblitz operate in the same market. Hence, in the econometric estimates below, we can use the Flash bicycle messengers as a useful control group to reliably control for daily changes in demand.

### ***Variability in Earnings***

Figure 1 also suggests that the messengers' daily earnings are highly variable. The implied average hourly earnings for messengers are roughly CHF 25 at Veloblitz and slightly less at

---

<sup>1</sup> Of course, the correlation coefficient is unaffected by the normalization of the series, because it enters the covariance and the variance of each series alike.

Flash. But at both messenger services, earnings are highly variable, as Figure 2 shows. We calculated the monthly standard deviation of daily earnings for every messenger and every month, and then averaged over messengers in every month. We plot this average standard deviation of daily earnings that a messenger faces as a percentage of mean earnings. The figure shows that messengers' earnings are highly volatile. It implies that hourly earnings regularly vary between CHF 18.5 and CHF 32.5.

The important point here is that, obviously, messengers are very familiar with variations in earnings opportunities over time. Hence, the wage change implemented by our experiment that we will describe below, varies wages in a range that is very familiar to the messengers.

## **2.2 The Experimental Design**

In order to evaluate the impact of an anticipated wage increase on behavior, we conducted the following field experiment at Veloblitz: All messengers were randomly assigned to one of two groups, A or B. The randomization was based on the administrative codes that the messenger service uses to identify a messenger in its accounting system. The first messenger that worked at Veloblitz was assigned the number 1, the second 2, and so forth. Messengers with odd numbers were assigned to group A, messengers with even numbers to group B.

Figure 3 summarizes the design of the experiment. Recall that the messengers' compensation is a percentage  $w$  of his daily revenues. Currently,  $w = 0.39$  for males, and  $w = 0.44$  for females at Veloblitz. Male and female messengers in group A had the opportunity of receiving a (roughly) 25 percent higher wage of  $w = 0.49$ ,  $0.54$ , respectively, during the four weeks between September 11<sup>th</sup> and October 6<sup>th</sup> 2000. For members of group B, we increased the wage by the same amount during the four weeks between October 30<sup>th</sup> and November 24<sup>th</sup> 2000. The additional earnings were all paid out on December 8<sup>th</sup> for both groups.

We refer to the two sequences as the treatment periods. For brevity, we refer to the members of group A participating in the experiment during treatment A, as the "treatment group A". Similarly, members of group B who participate in the experiment are referred to as the "experimental control group A" during treatment A. Non-participating messengers at Veloblitz and messengers at Flash are referred to as the "field control group", as can be seen in Figure 3.

The only participation constraint was that, in order to receive the wage increase, messengers had to fill in a questionnaire at the beginning and at the end of each of the

treatment periods, as shown in Figure 3. Hence, in order to participate, messengers had to fill in four questionnaires in total. There was a deadline indicated for each questionnaire to be returned completed. After distribution, messengers had roughly ten days to complete the questionnaire. If a messengers worked within that period of time, but failed to return the questionnaire, he or she was excluded from the experiment and received no payoff.

The wage increase and the participation rules were communicated to the messengers on August 29<sup>th</sup> in a presentation at the Veloblitz office. Moreover, posters at the Veloblitz office and handouts that were placed throughout the office ensured that all messengers were informed about the experiment even if they did not attend the presentation. One of the authors was available for questions regarding the questionnaires every Monday and Friday throughout the experimental period.

The messengers did not know that a purpose of the experiment was the study of labor supply behavior. They also did not know that we received the full (anonymous) records of each messenger about the number of shifts and the number of deliveries completed. We told the participants that we wanted to study the relation between wages and job satisfaction. The announced purpose of our study was credible because the questionnaires contained several questions related to job satisfaction.

The experiment as such represents an important innovation to the existing literature for several reasons. First, it implements an anticipated and exogenous variation in the (output based) wage rates of the messengers, which is key to studying the intertemporal substitution of labor. The experimental wage increase was massive. It amounts to a roughly 25 percent higher wage during four weeks, and provides a clear incentive to work more and work harder. Moreover, the participating messengers are experienced, and daily fluctuations in their earnings are common. Hence, we experimentally implement a wage change into an otherwise familiar environment. Second, the data we obtained from Veloblitz allows us to study two dimensions of labor supply: Hours (shifts) and effort (deliveries). No other study that I am aware of can look at the two dimensions simultaneously. Third, we supplement the data obtained from Veloblitz with four surveys that we conducted among the participating messengers. The surveys allow us to investigate the prediction of the economic model in more detail. Fourth, we can combine the data set with the full records from a second messenger service operating in the same market. This will prove useful to investigate any effect that the experiment might have had on the control group at Veloblitz, and helps to reliably eliminate daily fluctuations in demand.

## *Treatment Effects*

Three effects will play a key role in our analysis below. We call them the direct treatment effect, the indirect treatment effect, and the announcement effect. Figure 3 helps to see how each of these effects is identified.

(i) The **direct treatment effect** describes the impact that the experiment has on behavior (shifts worked, deliveries per shift) of the **treatment group relative to the experimental and the field control group** during the treatment periods. In terms of Figure 3, the direct treatment effect measures how the behavior of the treatment group differs from the experimental and the field control group during the treatment. Specifically, we use a dichotomous variable that equals 1 for all messengers of the treatment group during the treatment period and equals zero otherwise.

(ii) The **indirect treatment effect** describes the impact that the experiment has on behavior of **all messengers at Veloblitz relative to the field control group** (messengers at Flash) during the treatment periods. In terms of Figure 3, the indirect treatment effect measures how the behavior of messengers at Veloblitz differs from the field control group during the treatment. Specifically, we use a dichotomous variable that equals 1 for all messengers at Veloblitz during the treatment period and equals zero otherwise. The implicit assumption here is that the experiment had no effect on the behavior of the messengers at Flash.

(iii) The **announcement effect** describes the "impact" of the announcement of the experiment on the **participating messengers relative to all other messengers** (non-participating messengers at Veloblitz and all messengers at Flash). In terms of Figure 3, the announcement effect measures how the behavior of participating messengers differs from all other messengers as of the announcement of the experiment on August 29<sup>th</sup>. We use a dichotomous variable that equals 1 for all participating messengers at Veloblitz as of August 29<sup>th</sup>.

Under different behavioral hypothesis, the direct and indirect treatment effect have different interpretations and these will be explained in the next section.

The announcement effect can capture two possible effects: First, it can capture a possible income effect of the experiment. Any wage increase has two, potentially conflicting, effects. An individual's wage is its price of leisure. Hence, if the wage is increased, the price of leisure rises and this acts to reduce an individual's leisure consumption. This is the substitution effect

of wages. At the same time, a higher wage makes the individual richer for given levels of leisure. The higher wage generates more consumption of goods in every period. This decreases the marginal utility of consumption goods and shifts an individual's preference towards more leisure after the announcement of a wage increase. Second, good and leisure consumption can be complements. In that case, there is a component in the income effect shifting the preference towards more work again. In general, we are left with no prediction as to whether the income effect will increase or decrease labor supply. Key to our analysis, as we will explain in more detail below, is that the income effect will become operative immediately after we announced the experiment, i.e., immediately after the *new information about the future income stream is released*. The income effect constitutes part of the announcement effect. Yet, the announcement effect may also capture selection effects. While our experiment offers the advantage of implementing a large anticipated wage change in a real-life setting, we cannot force individuals to participate. Out of the 58 messengers at Veloblitz, 45 participated in the experiment. One of the 45 subjects ceased to participate during the experiment. Of the 13 non-participating messengers only one individual explicitly refused to participate.

We conjecture that the non-participants did not find it worthwhile to participate because they were already relatively detached from the company compared to the participating messengers. This is indicated by the low number of shifts they worked since July 2000. Between July and November 2000 the non-participants worked on average only one shift per week. Note that this potential selection effect poses no problem for the main purpose of our study, i.e., the comparison of the number of shifts and deliveries per shift across treatment conditions for the participating messengers.

### **3. Behavioral Hypotheses**

In this section, we contrast two behavioral hypothesis that lead to distinct predictions of the sign of the direct (and indirect) treatment effect for the experiment. The first behavioral hypothesis is the Anticipated Wage Hypothesis, as it is implied by a standard economic model of intertemporal substitution in the spirit of Lucas and Rapping (1970). The second is the Target Income Hypothesis, as put forward, e.g., by Camerer et al. (1997).

### 3.1 The Anticipated Wage Hypothesis

The two key implications from forward-looking optimizing model in the spirit of Lucas and Rapping (1970) can be summarized by two equations.

The first equation describes how labor supply  $e$  in period  $t$  should be chosen:

$$c'(e_t) = \lambda_t w_t r'(e_t) \quad (1)$$

where  $c'(e)$  is the marginal disutility of work,  $\lambda_t$  is the marginal utility of income,  $w_t$  is the wage in period  $t$ . To fit the above description to the experimental setup,  $r(e)$  relates the messengers effort to his revenues, and  $r'$  is the marginal revenue product. Equation (1) has the straightforward interpretation that the marginal disutility of effort should be equated to its marginal utility  $\lambda_t w_t r'(e_t)$ . The second key equation describes the movement of  $\lambda_t$  over time:

$$\lambda_t = (1+r_t)\beta E_t \lambda_{t+1} \quad (2)$$

Equation (2) dictates that the marginal utility of income in period  $t$  must be equated to the expected discounted marginal utility of income in period  $t+1$ , where  $E_t$  denotes the conditional expectation,  $r_t$  is the interest rate, and  $\beta$  is the subjective discount factor.

To fix ideas compare two identical messengers, a and b. Both learn that at some future date, their wage will be increased from  $w$  to  $w_H$ , and, as in our experiment, suppose that a's wage is  $w_H$  in period A and  $w$  in period B while for b it is the other way round. The key prediction that we will exploit is the following. During the period of time where the wage is increased to  $w_H$ , the messenger *must* exert more effort and work more hours<sup>2</sup> (i) relative to earlier periods and (ii) relative to the other messenger whose wage does not change. This is commonly referred to as the *Anticipated Wage Effect*. It holds irrespective of any income effect that the wage increase might have. The reason is that because both individuals also choose consumption optimally to satisfy (2), any level effect on consumption that acts through  $\lambda_t$  must occur immediately when the information about the wage increase is released. Hence, by (2)  $\lambda_t$  changes instantly and identically for both individuals. The higher wage is perfectly anticipated and, upon impact of the wage increase,  $e$  must go up. For  $\lambda_t$  constant, (1)

---

<sup>2</sup> For simplicity of the exposition, we have so far ignored the choice of hours as a separate variable. However, this is easily incorporated. Suppose that the messenger must either choose  $e = 0$ , reflecting leisure, or  $e > \hat{e}$ , where  $\hat{e}$  reflects a minimal required effort level. Then, equation (1) describes the optimality condition if the messenger works. The messenger will work (choose  $e > 0$ ) if the marginal benefit of working one shift (at the optimal intensity  $e$  that solves (1)) covers the fixed costs of foregoing leisure. The details are worked out in Götte (2001).

implicitly defines the labor supply curve of the messenger, often called (for obvious reasons) the  $\lambda$ -constant labor supply curve.

In the discussion above, the distinction between the comparative static effect of the wage increase (how  $e$  changes relative to a reference period) and the comparison between messengers (how  $e$  differs from the other messenger's effort, whose wage was not increased) is inessential.

However, this is not true for the experiment. In our setting, revenues of messenger  $i$  also depend on the choices of shifts and effort level by the other messengers. The more messengers are working, and the more effort they exert, the lower are the returns to increasing  $e$  for the messenger who is on the treatment and receives a higher wage. A rational messenger considers this fact when choosing  $e$ . Formally, this turns the situation into a game with strategic substitutes between messengers. It is well known that the comparative statics for equilibrium strategies are not easily characterized for this class of games<sup>3</sup>. Hence, we lose our prediction about the comparative static effect of the higher wage on  $e$ . However, in this class of games and under relatively weak conditions, one can still characterize the equilibrium choices of the messengers (item (ii) above). This is a direct consequence of a result in Athey and Schmutzler (2001)<sup>4</sup>. Using this result, we obtain the following prediction for our experiment:

***The Anticipated Wage Hypothesis:** If, in the presence of strategic substitutability between shifts and between individual deliveries per shift,  $c'(e)$  is increasing and  $r'(e)$  is decreasing in effort, then messengers in the treatment group must work more shifts and complete more deliveries during a shift, than otherwise identical messengers in the control group.*

A positive direct treatment effect on, both, the number of shifts and the number of completed deliveries, is evidence for the Anticipated Wage Hypothesis. The indirect treatment effect will capture any effect that the change in effort of the treatment group has indirectly on all messengers at Veloblit, because the messengers choices of effort and shifts are strategic substitutes. But as is highlighted in Athey and Schmutzler (2001), we are unable to predict the

---

<sup>3</sup> XXX Add footnote: this is because of the indirect effects.

<sup>4</sup> XXX Add technical footnote.

sign of this effect, unless one puts severe restrictions on the functional form of how effort translates into additional deliveries.

As already mentioned in the previous section, the announcement effect also contains any possible income effect that the experiment might have had.

### **3.2 The Target Income Hypothesis**

Camerer et al. (1997) argue that individuals set an income target of how much the individual wants to earn in, say, a month. Moreover, individuals are loss averse around that income target, in line with a Kahneman-Tversky value function. Hence, the value function exhibits a kink around the target, and is convex and steep in the domain of losses (relative to the target), but concave and flat in the domain of gains. Camerer et al. argue that the income target is not very responsive to labor market conditions.

Hence, when wages are high, individuals will put less effort into work and quit earlier, because the higher wage allows them to achieve the income target more quickly. This strong version of the downward sloping supply curve implies that individuals will always achieve exactly the income target, which implies that the  $\lambda$ -constant labor supply curve is downward sloping and the elasticity is -1: If wages are one percent higher, the individuals will reduce effort by one percent as to achieve the income target.

Similarly, this produces the following prediction

***The Target Income Hypothesis:** If the messengers set an income target, then those in the treatment group must work fewer shifts and complete fewer deliveries during a shift, than otherwise identical messengers in the control group.*

A *negative direct treatment effect* on, both, the number of shifts and the number of completed deliveries, is evidence of the Anticipated Wage Hypothesis. Again, we have no prediction for the sign of the indirect treatment effect.

## **4. Results**

This section reports the results from our field experiment. Before proceeding, we briefly explain the data that we used. We include all observations that occur on weekdays (Monday

through Friday) between January 1999 to November 2000 for most of our estimates. We include all observations where messengers complete more than one delivery per shift, but less than 26. "Shifts" with only one delivery involve corrections of booking errors. Shifts with more than 26 deliveries involved erroneous booking in all cases that could be verified. Each restriction excludes roughly two percent of the observation. Moreover, I exclude all observations of messengers who were not working for at least six weeks. Jointly, these restrictions exclude 5.29 percent of the data, but our results are not sensitive to these exclusions.

#### **4.1 The Impact on Shifts**

This subsection presents the results for the impact of the experiment on the number of shifts worked. In addition we also examine other determinants of the choice of shifts. We will proceed in the following way. First, we provide a simple comparison of the number of shifts in the treatment group to the number of shifts in the experimental control group. The advantage of this test is that it only compares the choices of participating messengers working under different wage levels. This simple comparison gives a first indication of the direct treatment effect. After this we present a more elaborate statistical model that controls for the other determinants of the choice of shifts.

Accidentally, the number of participating messengers is 22 both in group A and group B. During treatment A, the treatment group worked 287 shifts, while the control group worked 192. During treatment B, the treatment group worked 251 shifts, and the control group worked 192. A simple t-test with the number of shifts of the individual messengers as observations confirms that these differences are significant ( $t = 2.38$ ,  $p < 0.05$ ). This seems to suggest that the direct treatment effect is positive.

However, testing for a significant difference between the two groups is not that straightforward. The reason can be gauged from Figure 4, which displays the working hazard, i.e. the probability of working a shift, conditional on the number of days that have elapsed since the last shift. The figure indicates a pronounced duration dependence in our data, i.e. if a messenger worked yesterday he or she is much more likely to work today, too, compared to a messenger who did not work yesterday. This means that a simple t-test does not give us the direct treatment effect because duration dependence artificially amplifies the effect of a wage increase: Individuals who – due to the wage increase - worked yesterday are more likely to

work today even if they would not experience a higher wage today. To rule out this confound we base our test on the survivor function, i.e. the share of messengers who have not worked for at least  $T$  days. If the direct treatment effect is positive, the survivor function of the treatment group should lie below the survivor function of the control group: For any time interval that elapsed since the last shift, more messengers in the control group should chose to *not* work a shift (hence, more messengers of the control group "survive" as "non-workers"). Figure 5 shows that this is indeed the case. The figure plots  $-\log(-\log(\cdot))$  of the survivor function against  $\log(\text{days since last shifts})$ , as it is conventional to do. The difference in the survivor functions is significant (log-rank test for equality of the survivor function  $X^2(1) = 4.84$ ,  $p < 0.05$ ).

The above test does not exploit the whole variation in the data. To examine the determinants of shifts in more detail we use a proportional hazard model, which is also known as a Cox regression (Cox, 1972). It models the probability of working a shift at date  $s$  conditional on characteristics of messenger  $i$  and the duration dependence that specifies how the conditional probability of working varies with the number of days since the last shift. Formally

$$\Pr(i \text{ works on date } s \mid \text{hasn't worked } T \text{ days}) = \exp(x_{is}\beta + \gamma \text{Treat}_{is}) \Psi(T) \quad (3)$$

$\Psi(T)$  is the unknown time dependence, i.e. a function that indicates the baseline probability of working a shift, if the messenger has not worked for  $T$  days. As can be seen in Figure 4, this time dependence is highly complex. It is an advantage of the Cox Regression that it need not be specified (see Cox, 1972). *Treat* summarizes the treatment variables that we discussed in section 2. Finally,  $x$  contains a number of control variables that we discuss below.

We estimate two versions of (3), one in which we stratify by firm and one in which we stratify by every messenger (see Ridder and Tunali, 1999, for details). For simplicity, we directly display the proportionate change in the hazard, i.e. by how much an increase in  $x$  shifts the baseline hazard up. An entry greater than one increases the working hazard, while an entry between zero and one lowers it.

Table 1 displays the results. The direct treatment effect is positive and significant in both specifications, and of the same order of magnitude. The coefficient of 1.17, e.g., in column (2) implies that the conditional probability of those in the treatment group during the experiment is 17 percent higher than for comparable members of the control group. Notice that one should be cautious not set the intertemporal elasticity of substitution equal to this estimate,

because it is the outcome of a strategic interaction. However, this treatment elasticity of  $0.17/0.25=0.68$  is substantial and similar to the results obtained by Oettinger (1999).

The indirect treatment effect seems to be unimportant, suggesting that no rationing of shifts occurred because of the treatment. While the point estimate is slightly below one, i.e., it reduces the conditional probability of work a little bit, this estimate never becomes significant. The announcement effect is positive and highly significant.

The other control variables are also interesting. In Table 1, we distinguish between tenure, i.e. the time elapsed since the messenger joined the company, and experience, i.e., the number of shifts that the messenger has worked during his employment at the messenger service. Longer tenure decreases the probability to work a shift significantly. Conversely, more experience with work increases the working hazard, holding tenure constant. Both variables are highly significant. The results also suggest that female messengers work less frequently. Finally, the dummy variable "Last month" is 1 if the messenger is in his last month of employment. It shows that messengers work significantly fewer shifts as the end of employment at the messenger service approaches. By comparing the log-likelihood between columns (2) and (1), we also see that stratifying according to messengers greatly increases the log-likelihood and shows that there are significant individual differences in working habits.

### ***Interactions***

The above results show that the direct treatment effect on the number of shifts is positive and precisely in line with economic theory. They clearly reject the Target Income Hypothesis. Our data allow us to explore this effect in more detail. We present a number of interactions with variables that might potentially be important for the magnitude of the direct treatment effect, but that are typically not easily observable. For instance, economic theory implies that individuals with higher marginal utility should increase shifts more during the treatment than other individuals. Conversely, messengers who find work more exhausting should have a lower direct treatment effect.

We use two questions from our survey to obtain a measure of marginal utility. In the first question, we asked the experimental subjects the following:

Suppose you received CHF  $x$  of additional income this month. How urgently would you need that money?

The variable  $x$  in this question was tailored to be roughly one fourth of the messenger's monthly income. Individuals could answer on a 1 to 5 scale, where 1 was labelled "not urgently at all" and 5 was labelled "very urgently". If individuals behave according to economic theory, the response to this question provides a measure of the marginal utility of income. The more urgently they need the money, the higher the marginal utility of consumption must be. We use the share of income spent on non-durable consumption goods as a second proxy for the marginal utility of income. A higher share of non-durable consumption goods indicates a higher marginal utility (because otherwise, the individual would save more money).

Table 2 displays the results. In column (1), we report the result for the "need money urgently" question. The first row displays the estimate for the treatment effect when no interactions are included<sup>5</sup>. Column (1) shows that when both the direct treatment effect and the interaction with our proxy for marginal utility are included, both point estimates are positive. But while each fails to be significant individually, they are jointly. The third row shows that when the interaction is included alone, it does almost as well as the direct treatment effect. The point estimate in the third row suggests that the direct treatment effect varied between 1.04 for those indicating 1 and 1.2 for those indicating 5 (i.e., those needing the money very urgently). Hence, there is some evidence that higher marginal utility of income induces messengers to work more shifts during the treatment.

We obtain a similar result when we use the share of income spent on non-durables as a proxy for the marginal utility of income. The median share in our sample was 60 percent. Again, the estimates are very similar and those with a higher share of expenditures for non-durables tend to work more shifts.

Finally, we interact the treatment effect with the time the messenger indicated he was above the anaerobic level. This variable was constructed using the Borg scale. The Borg scale is a familiar self-rating device to measure physical exertion. It is routinely used in sports physiology and known to produce reliable results. The Borg scale offers different categories of intensity levels, where each category is described by standardized instructions. Messengers had to indicate how long they felt they were working at each intensity level over a typical

---

<sup>5</sup> Because not all participating messengers have filled in all questionnaires, the sample is not entirely the same as in Table 1: We excluded all observations of participating messengers for whom the information on the interacting variable was not available. To convince the reader that this matters little, and as a further indication that the results are robust, we redisplay the direct treatment effect for each sample again in the first row of Table 2.

shift. We calculated the time they spend above the level that is associated with the anaerobic level. The median response was 12 percent out of three hours effective riding time, i.e., roughly 20 minutes over five hours, but it varied between zero and 60 percent. The estimates in column 3 of Table 2 reveal a weakly significant relationship between perceived exertion and the magnitude of the direct treatment effect.

#### **4.2 The Impact on Deliveries**

We proceed the same way in discussing the results for the number of deliveries per shift. First, we offer a simple statistical test of the direct treatment effect, and we then explore various potential problems in a multiple regression to assess potential alternative explanations and the robustness of our result.

Again, we simply compare the average number of deliveries of the treatment group with the average number of deliveries of the control group during the two treatment periods. In treatment A, the treatment group carried out an average of 14.4 deliveries, while the control group carried out 15 deliveries on average. In treatment B, the treatment group carried out 14.6 deliveries, while the control completed 15.4 deliveries on average. In both treatments, the treatment group, even though on a higher wage, carries out fewer deliveries. Taken together, this difference is significant ( $t=2.554$ ,  $p < 0.05$ ). The results, as all the following, also hold if we use revenues instead of deliveries: members of the treatment group have on average lower revenues ( $t=2.328$ ,  $p < 0.05$ ). Hence, these results lend no support to the prediction of the economic model. The treatment group carried out fewer, instead of more deliveries.

Three important issues cannot be addressed using this simple test. First, we already saw that messengers in the treatment group work more shifts than messengers in the control group. But if working yesterday hurts today's performance, this could be one explanation for the negative treatment effect. Second, it could be that the two groups worked on different days. Maybe the member of the treatment group just filled any vacant shift, even on days where many other messengers were working or where earnings were expected to be low. Third, the derivation of the theoretical result that messengers in the treatment group should complete more deliveries relies crucially on the assumption that the messengers' strategies are strategic substitutes, and this needs to be validated.

We therefore estimate the following regression:

$$\ln(\text{deliveries}_{it}) = x_{it}\beta + \gamma\text{Treat}_{it} + d_t + e_{it} \quad (4)$$

Again, the variables of key interest are the direct and indirect treatment effect, as well as the announcement effect, summarised in *Treat*.  $d_t$  is a daily fixed effect to control for variation in messenger demand (recall from Figure 1 that it fluctuates strongly). In addition to the daily fixed effects, we include as control variables  $x$  all control variables from the Cox regression discussed earlier. In addition, we include the number of bicycle messengers deployed by each messenger service in each shift. Moreover, we control for the composition of deliveries that a bicycle messenger was carrying out. Flash and Veloblitz use similar schemes to price a delivery, which mainly reflects the distance between the pick-up and final destination of a delivery. Roughly 80 percent of all deliveries can be coded into six categories. We use the fraction of each category as an explanatory variable, with the uncodeable deliveries being the reference category. We also add a dummy variable indicating whether the messenger is a member of Veloblitz or Flash.

The results are displayed in Table 3. Consider first the treatment effects. Column (1) shows that the direct treatment effect is indeed negative and significant. It implies that while on the treatment, messengers did 5.5 percent fewer deliveries, which is roughly of the same order of magnitude as in the simple statistical test. In column (2), where individual fixed effects are included, the point estimate remains negative, though slightly smaller in absolute value and significant at the 10 percent level. Hence, this reinforces the picture obtained by the simple comparison of treatment and control group given previously.

Turning to the indirect treatment effect, we see that the point estimate is positive and roughly of the same magnitude as the direct treatment effect, and significant at the 5 percent level. Recall that the indirect treatment effect measures how the number of deliveries of messengers at Veloblitz differs from the number of deliveries of messengers at Flash during the experimental treatment. A plausible explanation is that *because* the treatment group exerted slightly less effort, other messengers at Veloblitz completed somewhat more deliveries during the experiment. The implied restriction that the indirect treatment effect and the direct treatment effect are equal in absolute value is not rejected by the data. Columns (3) and (4) present the restricted estimate, which, again, is negative and significant.

As in the estimates for the working hazard, the announcement effect is positive and significant. Notice that in passing from columns (1) to (2) or (3) to (4), the estimate of the announcement effect is somewhat reduced, which indicates selectivity on unobservable

messenger characteristics. Again, neglecting the announcement effect would have caused an upward bias into the direct and indirect treatment effects.

Turning to the control variables, we see that having worked yesterday does not hurt today's effort. It even increases today's effort because the dummy variable is positive and significant. This is in line with what the messengers themselves say: They perceive work as less exhausting have they worked the previous day, but generally say that it does not matter for the number of deliveries at all. When moving from (1) to (2), the point estimate drops by two thirds. It indicates that more productive messengers choose to work on two subsequent days more often. Indeed, once we control for messenger fixed effects, the impact of having worked the previous day becomes very small and probably hardly noticeable.

Can the negative direct treatment effect be explained because members of the treatment group work on average on worse days? Recall that all specifications include daily fixed effects. Hence, the negative coefficient on the direct treatment effect does not arise because members of the treatment group also work on predictably bad days. Thus, our results reject the proposition that messengers in the treatment group exert more effort than messengers in the control group. The two other explanations for a negative treatment effect that we mentioned above are also not supported by the data.

We now turn to the issue of strategic substitutes. We included two measures of strategic substitutability. The first variable, called "Direct Competitors", is a proxy for the number of bicycle messengers who work during the same shift within the same firm. The second variable, called "Other Competitors" captures the pronounced increase in the number of competing car messengers at Flash. Recall that one shift lasts five hours. Several shifts run from 8 am to 1 PM, and from 1 pM to 6 pM, but not all. Roughly one third of the shifts start around 10 am and last until 3 pm in order to smooth out the change in the shifts at 1 pm. However, we do not know which messenger worked at which shift and cannot determine the exact number of competing messengers. As an approximation, we divide the number of messengers by two, hence the number of competitors will be right on average. Strategic substitutability implies that the number of deliveries must be decreasing in the number of competitors. The point estimates of the coefficients on both variables support strategic substitutability. Both coefficients are negative, highly significant and quantitatively large. The point estimate of -0.035 implies that adding one additional messenger to the shift reduces deliveries of each messenger deployed by 3.5 percent. Given the variation in the data, differences of plus or minus two messengers per shift are not rare. The estimates imply that

this decreases the number of deliveries of each messenger by 14 percent. The estimates imply that adding a car messenger depresses the bicycle messengers' deliveries by roughly 3 percent. Overall, this is clear evidence of strategic substitutability between messengers, even between bicycle messengers and car messengers. We also find a positive and large effect of experience with work on deliveries. The estimated coefficient of 0.048 in column (1) is expressed as an elasticity, but must be interpreted with caution, because the interpretation of an elasticity is only valid for small variation in tenure. The impact must be evaluated at sample means. It implies that increasing tenure from 1 to 100 days increases the number of deliveries per shift by approximately 3.

There are two remaining control variables. First, gender has a negative sign, but is not very large quantitatively. It implies that female messengers complete one less delivery, which is a moderate gender difference. Second, the dummy indicating that the messenger is employed at Veloblitz is positive and large. However, to interpret the dummy as a difference in productivity, one must be sure that the reference categories of the composition of deliveries are comparable. Differences therein will alter the Veloblitz dummy, making a structural interpretation difficult.

### ***Interactions***

The above results show that messengers significantly decreased the number of deliveries they completed during a shift. This represents a sharp contradiction to the prediction of economic theory and is worthwhile to be explored in more detail. In particular, the results are compatible with choice bracketing, as explained in the introduction. When working on a shift, individuals take a narrow decision frame and evaluate a day at a time. The experimental treatment increases their daily income at any level of effort. Individuals display daily income effects (recall that, by (2), they shouldn't) of such a magnitude that it reduces their overall effort level.

To further explore the determinants of the negative direct treatment effect, we add the same interactions to the estimates as in the previous section. However, choice bracketing suggests a different interpretation of the variables. It is plausible that if an individual is prone to choice bracketing he or she will exhibit this behaviour in several domains and not just when choosing effort. An important example where choice bracketing also may occur is the choice of consumption plans. If individuals narrowly bracket the choice how much of their current income to consume, they will spend too much. This means that the responses to our question

regarding the urgency of additional money can be interpreted as a proxy for the degree of choice bracketing. Similarly individuals who spend a large fraction of their current income on non-durable consumption, are more likely to bracket their decisions narrowly. Consequently, these individuals should have a more negative direct treatment effect. The results are reported in Table 4. The presentation of the results on the interaction effects is organised in exactly the same way as in the previous section.

Consider the interaction with the urgency question first. In the first column of Table 4, we see that the point estimate on the interaction term is negative in all specifications. When we only include the interaction term, the point estimate is negative and always significant at the 5 percent level. The same is true when we use the expenditure share for non-durable consumption goods in the interaction. Individuals with a high share of non-durable expenditures have a more negative direct treatment effect.

Our interpretation is that individuals who indicate that they need additional money more urgently apply a more narrow choice bracket. Note that this interpretation is also consistent with the negative and significant impact of the expenditure share for non-durables on the direct treatment effect, which is also present in the data.

Finally, we present the results for the interactions with our measure of perceived exertion, as measured by the fraction of time spent above the anaerobic level. The interaction term again is negative and significant. Note that a negative coefficient on the interaction term is consistent with economic theory only if the point estimate of the direct treatment effect is resolutely positive, which is never the case. But again, choice bracketing offers a plausible alternative explanation: Those individuals who find the work particularly exhausting are the most tempted to reduce effort in response to the treatment.

## **5. Concluding Remarks (to be done properly)**

The results of our field experiment provide a clean test of the standard economic theory of inter-temporal labour supply. Our results on the choice of working hours (shifts) are consistent with the prediction of this model. We refute, however, the prediction that stronger incentives will cause an increase in effort per shift. Instead, the average worker in our sample decreases effort in the face of stronger incentives. This “anomalous” average effect hides considerable heterogeneity among the workers. Since we control for this heterogeneity by individual fixed effects they pose no problem for our conclusions. It is nevertheless

interesting to know that there is a large group of workers who decreases effort in response to stronger incentives while a non-negligible minority increases effort. Our results are consistent with the idea that workers apply different choice brackets when deciding on hours and effort. However, whatever the interpretation of our results will be, they indicate that even in an environment like ours, that is conducive for incentive effects it should not be taken for granted that stronger incentives increase effort.

## References

- Blundell, Richard; MaCurdy, Thomas E. (1999), Labor Supply: A Review of Alternative Approaches, in: Ashenfelter, Orley and David Card (Eds.): Handbook of Labor Economics.
- Borg, G. (1985). An introduction to Borg's RPE-scale. Ithaca, NY: Mouvement Publications.
- Camerer, Colin et al. (1997), Labor Supply of New York City Cabdrivers: One Day at a Time, Quarterly Journal of Economics; 112(2), May 1997, pages 407-41.
- Cox, D. R. (1972). Regression Models and Life-Tables (with discussion). Journal of the Royal Statistical Society Series B 34: 187-220.
- Lucas, Robert E., Jr.; Rapping, Leonard A. "Real Wages, Employment, and Inflation" Journal of Political Economy; 77 (5) Sept./Oct. 1969, pp. 721-54.
- Oettinger, Gerald S. (1999), An Empirical Analysis of the Daily Labor Supply of Stadium Vendors, Journal of Political Economy; 107(2), April 1999, pages 360-92.
- Read, Daniel; Loewenstein, George; Rabin, Matthew (1999), Choice Bracketing, Journal of Risk and Uncertainty; 19(1 3), December 1999, pages 171-97.
- Ridder, Geerd and Tunali, Insan (1999). Stratified Partial Likelihood Estimation. Journal of Econometrics 92(2): 193-232.
- Tversky, Amos; Kahneman, Daniel (1991), Loss Aversion in Riskless Choice: A Reference Dependent Model, Quarterly Journal of Economics; 106(4), November 1991, pages 1039-61.

TABLE 1: BASELINE RESULTS FOR THE CHOICE OF SHIFTS

COX REGRESSIONS: PROBABILITY OF WORKING, CONDITIONAL ON DAYS SINCE  
LAST SHIFT (CHANGES IN HAZARD RATES DISPLAYED)

|                           | (1)                  | (2)                  |
|---------------------------|----------------------|----------------------|
| <b>Treatment Effects</b>  |                      |                      |
| Direct Treatment Effect   | 1.162**<br>(2.382)   | 1.17**<br>(2.439)    |
| Indirect Treatment Effect | 0.945<br>(-1.015)    | 0.919<br>(-1.443)    |
| Announcement Effect       | 1.447***<br>(7.911)  | 1.376***<br>(5.867)  |
| <b>Control Variables</b>  |                      |                      |
| Log(Experience)           | 1.134***<br>(15.821) | 1.254***<br>(9.954)  |
| Log(Tenure)               | 0.855***<br>(15.427) | 0.828***<br>(-9.547) |
| First Month (DV)          | 0.958<br>(-1.018)    | 1.0154<br>(0.226)    |
| Last Month (DV)           | 0.879***<br>(-3.93)  | 0.889***<br>(-2.923) |
| Female (DV)               | 0.85***<br>(-4.698)  |                      |
| Controls for Months (DVs) | Yes***               | Yes***               |
| Stratified according to   | Firm                 | Messenger            |
| Log Likelihood            | -188,531             | -93,845              |
| Number of Observations    | 22,064               | 22,064               |

- Notes:
- \*, \*\*, \*\*\* denotes significance at the 10, 5, and 1 percent level, respectively
  - z-statistics in parenthesis. Z- statistics test whether the change in the hazard rate is significantly different from 1.
  - DV indicates dummy variable

TABLE 2: EXPLAINING THE DIRECT TREATMENT EFFECT ON SHIFTS

Dependent Variable in All Models: Probability of Working, conditional on Days since last Shift  
(Changes in Hazard Rates displayed)

COX REGRESSIONS

|   | Model (1)<br><br><b>Interaction with:<br/>"need the<br/>additional money<br/>urgently"</b><br><br><i>N</i> =21,603 | Model (2)<br><br><b>Interaction with:<br/>Non-Durable<br/>Consumption<br/>Share</b><br><br><i>N</i> =21,519 | Model (3)<br><br><b>Interaction with:<br/>Time Above<br/>Anaerobic Level</b><br><br><i>N</i> =21,236 |
|---|--|---|--|
| <b>Direct Treatment Effect alone</b><br>Direct Treatment Effect | 1.18**<br>(2.469)  | 1.178**<br>(2.439)  | 1.156**<br>(2.157)   |
| <b>Interactions</b><br>Direct Treatment Effect                  | 1.171<br>(1.311)   | 1.184<br>(1.336)  | 1.125<br>(1.439)   |
| Interaction Term  | 1.003<br>(0.074)   | 1.001<br>(0.04)   | 1.136<br>(0.571)   |
| Test for joint significance                                     | significant**  | significant**   | Significant*   |
| <b>Interaction alone</b><br>Interaction Term                    | 1.04**<br>(2.114)  | 1.201**<br>(2.061)  | 1.364*<br>(1.744)  |

- Notes:
- a. \*, \*\*, \*\*\* denotes significance at the 10, 5, and 1 percent level, respectively.
  - b. z-statistics in parenthesis. Z- statistics test whether the the change in the hazard rate is significantly different from 1.
  - c. same controls as in Table 2, column (1).
  - d. Tests for joint significance are F-tests testing whether both coefficients equal to 0.

TABLE 3: BASELINE RESULTS FOR THE CHOICE OF EFFORT

DEPENDENT VARIABLE: LOG(#DELIVERIES)

OLS REGRESSIONS

|                                   | (1)                    | (2)                    | (3)                    | (4)                    |
|-----------------------------------|------------------------|------------------------|------------------------|------------------------|
| <b>Treatment Effects</b>          |                        |                        |                        |                        |
| Direct Treatment Effect           | -0.055**<br>(-2.279)   | -0.038*<br>(-1.802)    | -0.058***<br>(-3.075)  | -0.05***<br>(-2.887)   |
| Indirect Treatment Effect         | 0.0617**<br>(2.517)    | 0.064***<br>(2.778)    | (restricted)           | (restricted)           |
| Announcement Effect               | 0.058***<br>(2.647)    | 0.053**<br>(2.421)     | 0.061***<br>(3.712)    | 0.06***<br>(2.721)     |
| <b>Control Variables</b>          |                        |                        |                        |                        |
| Worked Yesterday (DV)             | 0.069***<br>(10.933)   | 0.02***<br>(3.292)     | 0.069***<br>(10.933)   | 0.02***<br>(3.296)     |
| Log(Experience)                   | 0.048***<br>(14.342)   | 0.047***<br>(5.532)    | 0.05***<br>(15.01)     | 0.045***<br>(5.537)    |
| Log(Tenure)                       | 0.019***<br>(3.708)    | 0.044***<br>(4.833)    | 0.019***<br>(3.87)     | 0.044***<br>(4.833)    |
| # Competing<br>Bicycle Messengers | -0.035***<br>(-10.686) | -0.033***<br>(-11.085) | -0.035***<br>(-12.359) | -0.033***<br>(-11.085) |
| # Competing<br>Car Messengers     | -0.031***<br>(-6.687)  | -0.032***<br>(-7.625)  | -0.01***<br>(-6.687)   | -0.032***<br>(-7.625)  |
| First Month (DV)                  | -0.05***<br>(-2.86)    | 0.014<br>(0.709)       | -0.046***<br>(-2.692)  | 0.013<br>(0.684)       |
| Last Month (DV)                   | -0.037***<br>(-2.836)  | -0.016<br>(1.163)      | -0.037***<br>(-2.832)  | -0.016<br>(1.13)       |
| Female (DV)                       | -0.066**<br>(-4.826)   | --                     | -0.066**<br>(-4.828)   | --                     |
| Member of Veloblitz (DV)          | 0.175***<br>(4.971)    | --                     | 0.175***<br>(4.988)    | --                     |

(continued on next page)

|   |        |        |        |        |
|---|--------|--------|--------|--------|
| Daily Fixed Effects                             | Yes*** | Yes*** | Yes*** | Yes*** |
| Controls for Composition                        | Yes*** | Yes*** | Yes*** | Yes*** |
| Messenger Fixed Effects                         | No     | Yes**  | No     | Yes*** |
| Within Days $R^2$                               | 0.191  | 0.394  | 0.191  | 0.394  |
| Fraction of Variance due to Daily Fixed Effects | 0.273  | 0.34   | 0.288  | 0.34   |
| Number of Observations                          | 22,064 | 22,064 | 22,064 | 22,064 |

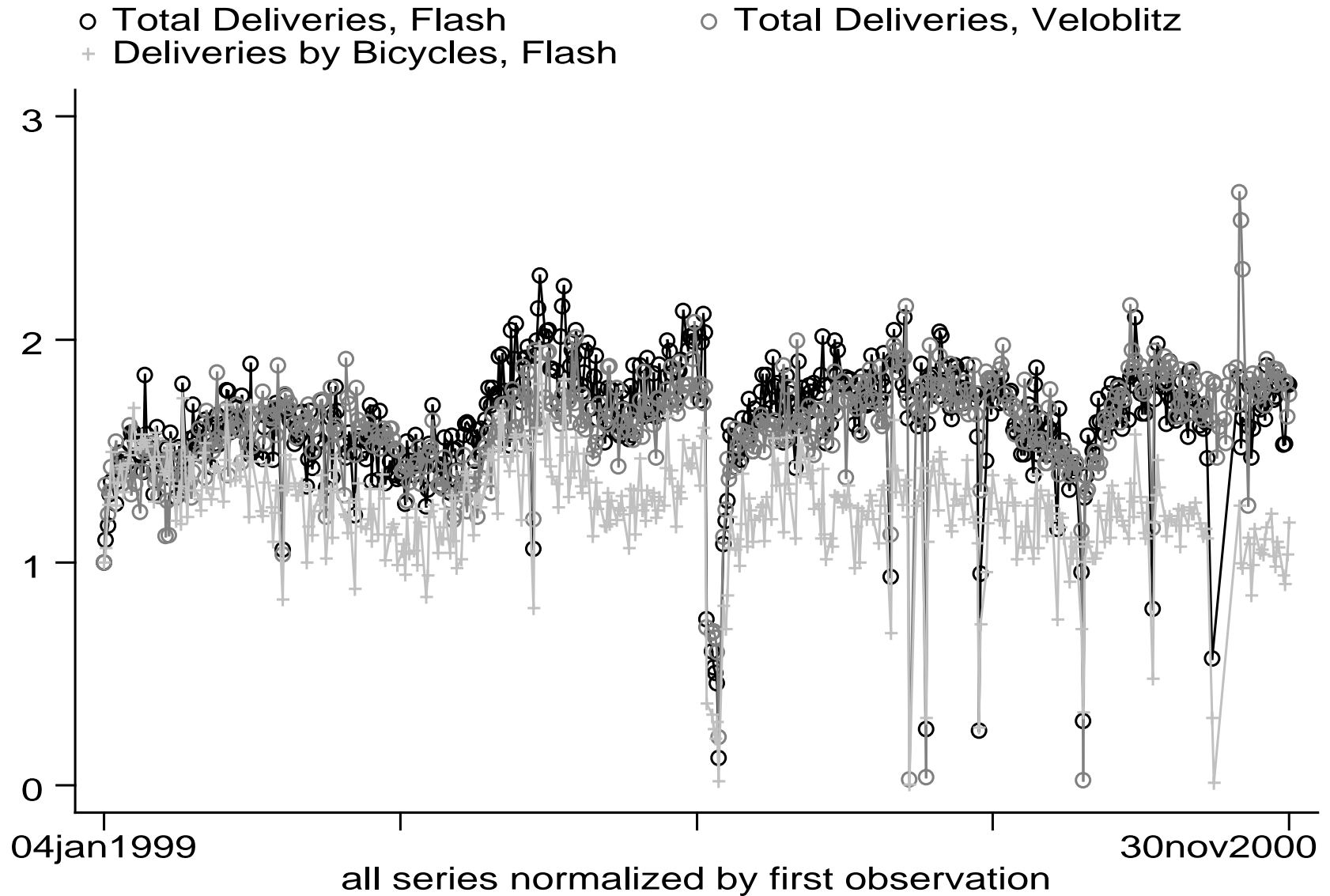
- Notes:
- a. \*, \*\*, \*\*\* denotes significance at the 10, 5, and 1 percent level, respectively
  - b. z-statistics in parentheses.
  - c. Coefficients on the composition of shifts and the constant term are omitted.
  - d. DV indicates dummy variable.

TABLE 4: EXPLAINING THE DIRECT TREATMENT EFFECT ON EFFORT  
 DEPENDENT VARIABLE IN ALL MODELS: LOG(# DELIVERIES PER SHIFT)

OLS REGRESSIONS

|   | Model (1)<br><br><b>Interaction with:<br/>"need the<br/>additional money<br/>urgently"</b><br><br><i>N</i> =21,603 | Model (2)<br><br><b>Interaction with:<br/>Non-Durable<br/>Consumption<br/>Share</b><br><br><i>N</i> =21,519 | Model (3)<br><br><b>Interaction with:<br/>Time Above<br/>Anaerobic Level</b><br><br><i>N</i> =21,236 |
|---|--|---|--|
| <b>Direct Treatment Effect alone</b><br>Direct Treatment Effect | -0.055**<br>(2.279)  | -0.063**<br>(2.482)   | -0.051**<br>(-2.04)  |
| <b>Interactions</b><br>Direct Treatment Effect                  | -0.01<br>(-0.21)   | 0.048<br>(1.01)   | -0.02<br>(0.585)   |
| Interaction Term  | -0.02<br>(-1.426)  | -0.189***<br>(-2.72)  | -0.157*<br>(-1.886)  |
| Test for joint significance                                     | significant**  | significant**   | significant**  |
| <b>Interaction alone</b><br>Interaction Term                    | -0.02***<br>(-2.918)   | -0.129***<br>(-3.541)   | -0.185***<br>(-2.716)  |

- Notes:
- a. \*, \*\*, \*\*\* denotes significance at the 10, 5, and 1 percent level, respectively.
  - b. z-statistics in parenthesis.
  - c. same controls as in Table 3, column (1).
  - d. Tests for joint significance are F-tests testing whether both coefficients equal 0.
  - c. same controls as in Table 1, column (1).
  - d. Tests for joint significance are F-tests testing whether both coefficients equal 0.



**Figure 1: The Demand For Messenger Services**

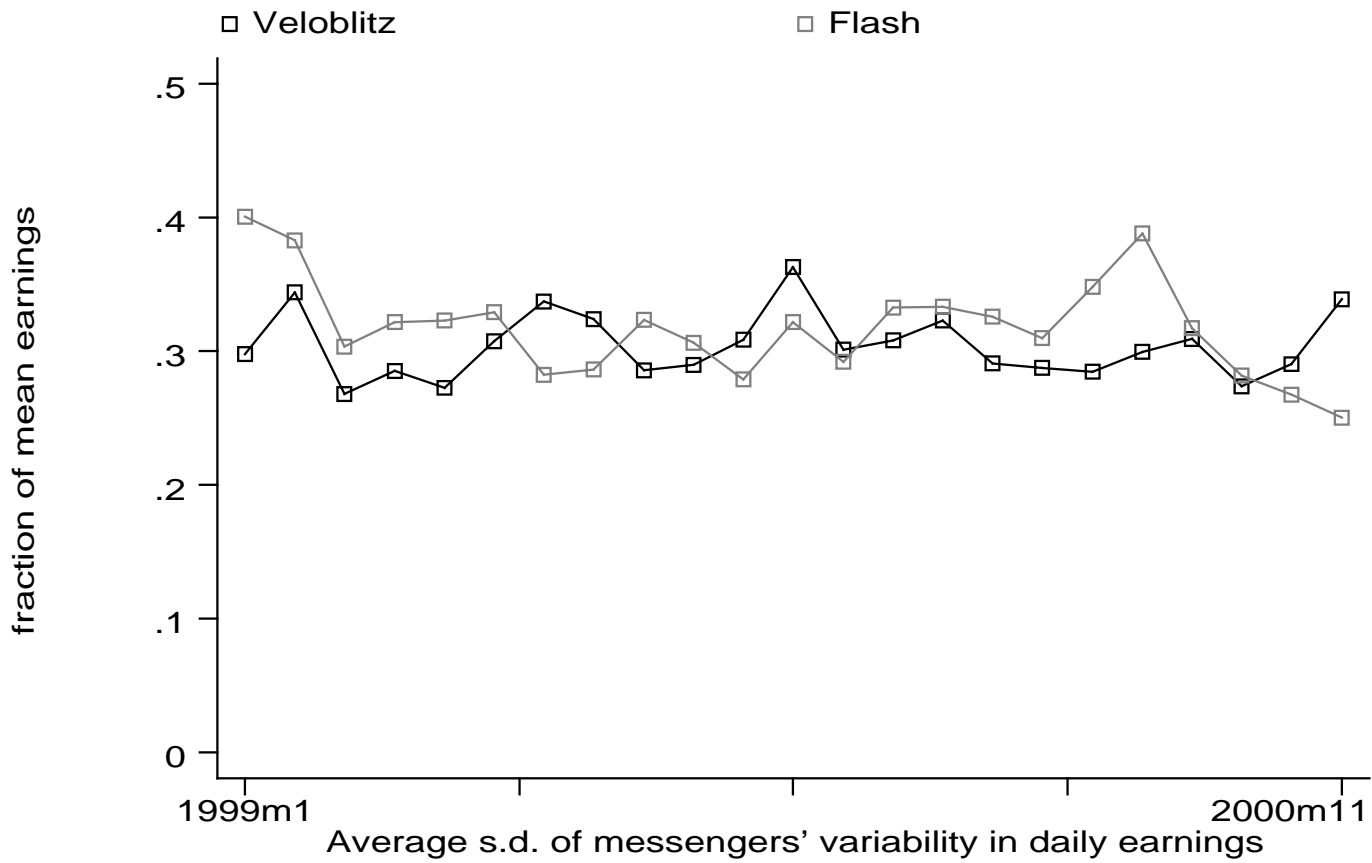


Figure 2: Volatility of Earnings per shift

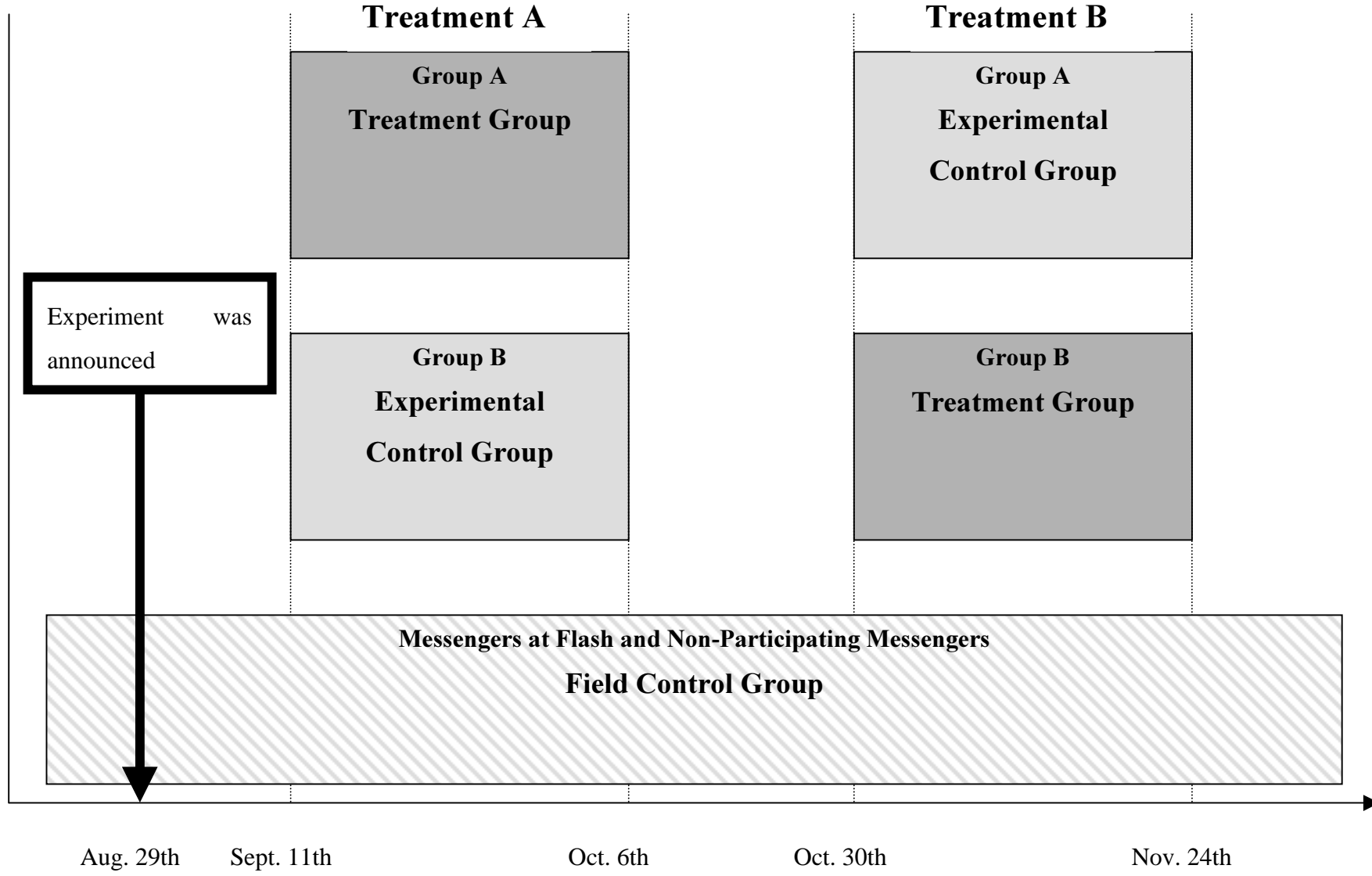
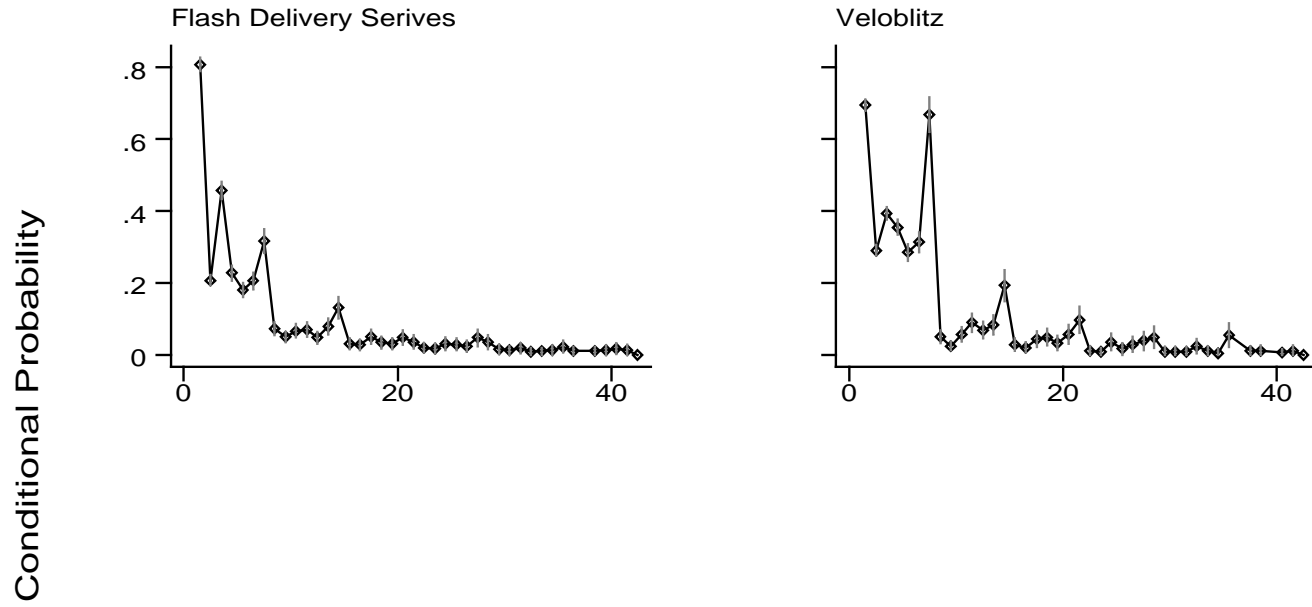


FIGURE 3: THE TIMING OF EVENTS



Pr(work today | hasn't worked T days); full sample.  
**Figure 4: Working Habits**

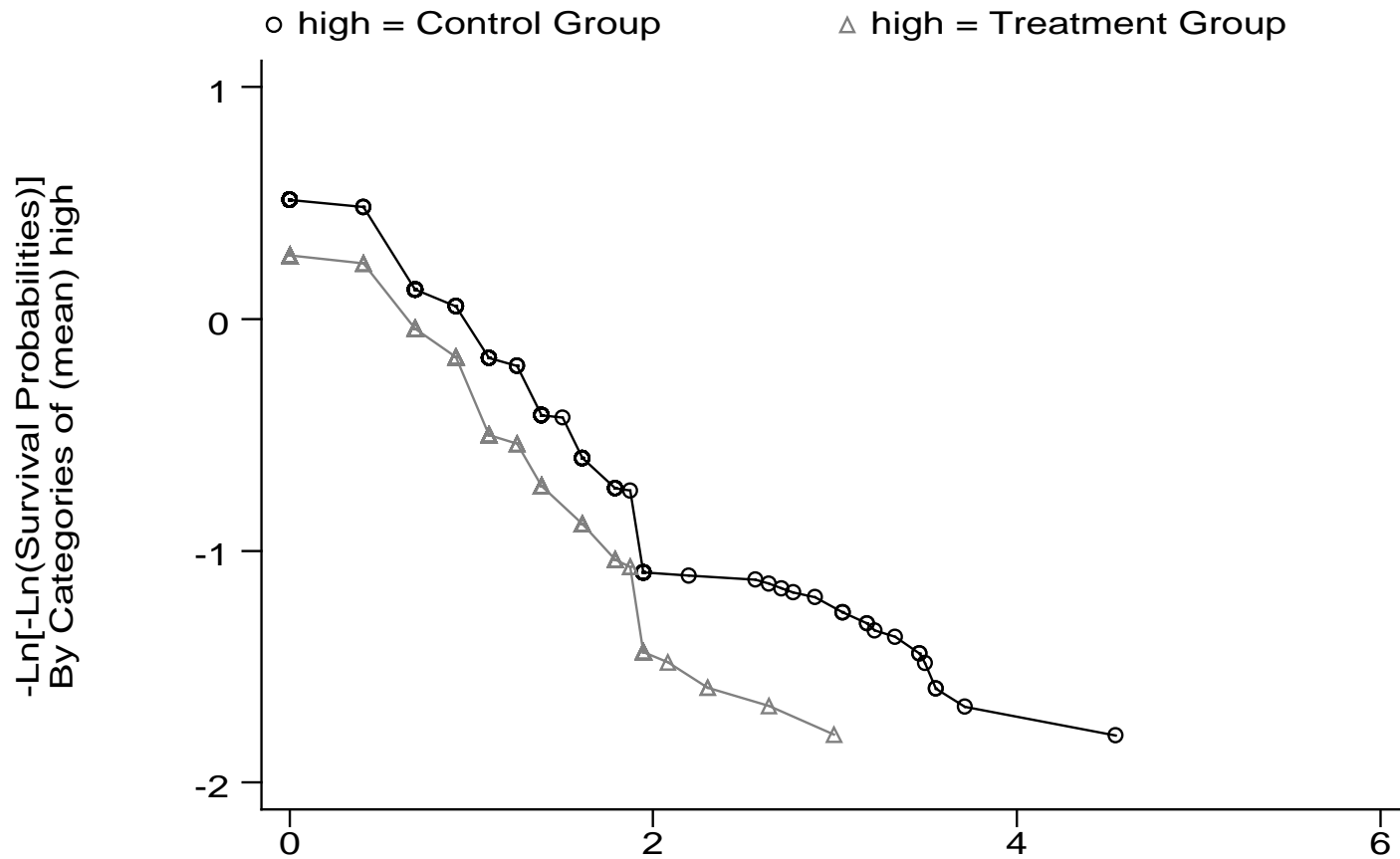


Figure 5: The Direct Treatment Effect of Shifts