Are preferences for work reference dependent or time nonseparable? New experimental evidence

Sam Cosaert, Mathieu Lefebvre, Ludivine Martin

Are preferences for work reference dependent or time nonseparable? New experimental evidence*

Sam Cosaert† Mathieu Lefebvre‡ Ludivine Martin§

June 7, 2022

Abstract

Tests of labor supply models often rely on wages. However, wage variation alone generally cannot disentangle the classical time separable model and its extensions: reference dependent preferences (income targeting) and time nonseparable preferences (disutility spillovers; timing-specific preferences). We set up a novel laboratory experiment in which individuals choose their working time. We vary, independently, wages, historical income paths, and cumulative past work. We also vary the timing of experimental sessions. Statistical tests and stochastic revealed preference methods cannot reject the classical model in favor of income targeting or disutility spillovers, but the data suggest that labor supply varies by time-of-the-day.

Keywords: Working time, lab experiment, time separable model, income targeting, disutility spillovers, timing-specific preferences, revealed preferences.

JEL classification: C91, D12, J22, J31.

1 Introduction

Labor supply can be seen as the outcome of a trade-off between income (consumption) and leisure. Real wages increase the opportunity cost of leisure and this can promote labor

*We are grateful to the editor and two anonymous referees. Their suggestions helped us to improve the paper and, in particular, to decompose 'time nonseparable' preferences into *disutility spillovers* from the real-effort task itself and *time-specific preferences for work* associated with the timing of experimental sessions. Furthermore, we thank LISER-LAB manager Francesco Fallucchi for running the experimental sessions in Luxembourg. This work was supported by the French National Research Agency Grant ANR-17-EURE-0020, and by the Excellence Initiative of Aix-Marseille University - A*MIDEX.

†University of Antwerp. Prinsstraat 13, 2000 Antwerp, Belgium. E-mail: sam.cosaert@uantwerpen.be

‡Aix Marseille Univ, CNRS, AMSE, Marseille, France. E-mail: mathieu-julien.lefebvre@univ-amu.fr

§Luxembourg Institute of Socio-Economic Research (LISER) and Centre de Recherche en Économie et Management (CREM). Porte des Sciences 11, L-4366 Esch-sur-Alzette, Luxembourg. E-mail: ludivine.martin@liser.lu

supply through substitution. Substitution effects are well-documented in the literature (Oettinger, 1999). However, over the years, the literature has also uncovered several types of heterogeneity in the preferences for work. These range from *reference dependent preferences* (Camerer et al., 1997; Fehr and Goette, 2007; Crawford and Meng, 2011) to *time nonseparable preferences* in the form of disutility spillovers (Farber, 2005; Fehr and Goette, 2007) and timing-specific preferences (Chen et al., 2019). A major challenge for empirical analyses is that wages are not necessarily orthogonal to these reference points, disutility spillovers, or aspects of timing. Higher wages can shift a worker's income target, promote participation in additional shifts and thereby increase cumulative past work, and finally affect the timing of work. All this may, in turn, influence the individual's preferences for work. The interaction between wages on one hand and possible preference covariates on the other hand partially explains the wide range of estimates of wage elasticities across the literature. A simple wage elasticity reflects the compound effect of substitution, targeting, tiredness, etc.; it is therefore a biased estimator of 'pure' substitution effects.

In this paper, we present a novel real-effort lab experiment in which subjects choose their working time. We vary, *independently*, budget constraint parameters, the income path, cumulative working time, and the timing of experimental sessions. The first objective is to isolate 'pure' budget constraint effects of wages from correlations between wages and the possible preference covariates. The second objective is to detect relevant factors that co-vary with the preferences for work. The final objective is to investigate which part of the variation in observed time choices is associated with (budget constraint) wage effects and which part can be explained by various forms of preference heterogeneity.

Reference dependent and time nonseparable preferences. The classical, time separable, model assumes that the distribution of preferences for work is stable and independent of historical income paths or the structure and timing of shifts. The literature has come up with distinct generalizations of this model: *reference dependent* and *time nonseparable* preferences. Each extension relaxes the restrictive structure on preference heterogeneity in a different way. Theories of reference dependent preferences, on one hand, state that workers are loss averse around some reference point.[1] In particular, their willingness to work changes discretely if earnings exceed the target income. We focus here on 'backward-looking' targets shaped by an individual's historical income path. Theories of time nonseparable preferences, on the other hand, posit that preferences for

[1]Since the seminal work of Kahneman and Tversky (1979), there has been increasing evidence that individual preferences depend not only on outcomes but also on how much these outcomes diverge from some reference point. See DellaVigna et al. (2017) for a list of recent works on reference dependency in different settings. The formation of reference points may depend on historical income paths, rational expectations, the status quo, and social comparisons. Given our focus on path-dependent explanations of labor supply, we follow the literature on *backward-looking* reference points (Bowman et al., 1999; Genesove and Mayer, 2001; Goette et al., 2004; Post et al., 2008; DellaVigna et al., 2017).

work depend on time itself. We make a further distinction between time nonseparability in the form of disutility spillovers between periods and nonseparability in the form of timing-specific preferences. According to the first class of theories, labor supply generates intertemporal spillovers through accumulated fatigue or boredom associated with the task itself.[2] *Cumulative working time* from past periods increases the present disutility from work and decreases the willingness to supply labor. According to the second class of theories, the willingness to work depends on the timing of that work.[3] *Cumulative time in the day* (clock hour) is thus another potential predictor of preferences for work.

Different labor supply models have different policy implications. The time separable model with strong substitution effects suggests that firms can simply offer higher wages to increase labor supply. The aforementioned extensions show that employers have other instruments at their disposal, beyond wages, to predict and enhance willingness to work (overtime) among their employees. If, on one hand, preferences are reference dependent and reference points depend on income paths, firms can anticipate employees' willingness to work on the basis of their former earnings and salaries. If, on the other hand, preferences are time nonseparable, firms can improve worker motivation by optimizing the duration and the temporal pattern of work episodes.[4] More generally, if there are correlations between the wage on one hand and reference points, cumulative work, or the timing of work on the other hand, it may be very difficult to test or predict labor supply responses. Any relationship between the wage and labor supply can stem directly from budget constraint effects but also indirectly from preferences. An original feature of the design of our study is that we let wages vary *while we observe* (and can thus control for) possible preference covariates.

Experimental design. Identification of the determinants of labor supply has been difficult in practice because individuals had little control over their working hours in the short run.[5] Moreover, real-life wages are determined by an interaction of labor supply and demand, and demand shocks are often unobservable to the econometrician (Oettinger,

[2]We refer to recent studies in emergency medical services (Brachet et al., 2012), munition factories (Pencavel, 2014), and call centers (Collewet and Sauermann, 2017), which showed the impact of the (exogenous) duration of work on the worker's performance.

[3]In an experiment similar to ours (i.e., with a real-effort task and exit choice), Hogarth and Villeval (2014) found that the timing of sessions affects the likelihood that subjects exit the game early. Chen et al. (2019) recently documented time-varying reservation wages among Uber drivers. A first (economic) explanation is that the opportunity cost of work depends on its scheduling. The timing of work is a disamenity when it impedes social life (Hamermesh, 1999). A second (biological) explanation is based on circadian rhythms: Kahneman et al. (2004) showed that tiredness perceptions increase steeply from noon onward. Cardinali (2008) described how adolescents' cognitive functions peak towards the evening.

[4]Recently, Baucells and Zhao (2019) proposed a fatigue-disutility-model to compute the optimal temporal profile of effort and breaks.

[5]A series of papers focused on special environments in which workers effectively *choose* their labor supply: taxi drivers (Camerer et al., 1997; Farber, 2005, 2008; Crawford and Meng, 2011), stadium vendors (Oettinger, 1999), and bicycle messengers (Fehr and Goette, 2007).

1999). We circumvent both issues in an original real-effort lab experiment.

The main part of the experiment consisted of four 'rounds' in which subjects did a tedious, repetitive task. In the last round, after a default work episode had ended, subjects could *choose* to stay in the lab to work additional minutes *or* to leave the room and exit the experiment. This working time choice will be our main outcome variable. Subjects earned the sum of wage W per minute of (overtime) work and a fixed payment Y. In this experiment, we controlled not only budget constraint parameters but also cumulative past working time and the historical income path. To this end, we introduced three rounds preceding the final one. In these rounds, subjects were asked to work a *given* amount of time. This exogenously determined the cumulative working time M leading up to the subject's final choice. To construct the income path, we separated rounds by implementing small breaks between them. We let subjects work one round at a time; we also clearly specified the compensation scheme at the level of each round. Since working time was fixed in the initial rounds, subjects earned a *pre-determined* income per round; we denote this earnings level I. The aim of this systematic exposure to I was to establish it as target income. At the end of the experiment, subjects finally received I from rounds 1, 2 or 3 (each with probability 1/4) or their pay-off Y plus the product of W and the labor supply choice from round 4 (with probability 1/4). Thus each round was independent and equally important in terms of the final payment. This is in line with the typical incentive structure used in lab experiments for revealed preference testing, see infra.

At this point, it is worth mentioning two further features of our experiment that facilitate identification of reference dependent and time nonseparable preferences. The first distinguishing feature is the *compensation scheme*. In most other experiments, the object of choice was effort provision, and the compensation scheme consisted of piece rate rewards per unit of output. In our setting, the object of choice was working time, and subjects received a wage per unit of time worked. There was however a minimal output requirement per minute to impose a strictly positive cost of effort. We normalized the cost of effort by scaling the output requirement to the individual's productivity in a first part of the experiment. This mitigated the influence of productivity differences.[6]

The second distinguishing feature is the *decision environment*. In a previous experiment, Abeler et al. (2011) tested the effects of expectations-based, reference dependent preferences in a real-effort lab environment.[7] The authors allowed participants to stop

[6]With relatively small sample sizes, productivity differences may bias the comparison of reference dependent preferences and time nonseparable preferences. High productivity individuals would be more likely to hit the (exogenous) target income before tiring, whereas low productivity individuals would be more likely to suffer from the (exogenous) duration of the experiment.

[7]Abeler et al. (2011) manipulated 'expectations-based' income reference points in the lab by varying the amount of a predetermined payment. Subjects received either their accumulated piece rate earnings from a tedious and repetitive task, *or* the predetermined payment. In line with Kőszegi and Rabin (2006, 2009)'s theory of reference dependent preferences, the reference point was formed by an individual's probabilistic beliefs about outcomes. Abeler et al. (2011) found that subjects stopped working when their accumulated piece rate earnings were relatively close to the fixed payment.

working at any time during the experiment. In our experiment, by contrast, subjects submitted their working time choice once and committed to the chosen duration by construction. This has a theoretical and a practical advantage. From a theoretical perspective, it circumvents issues of dynamic inconsistency (present bias) in effort choices in the lab (Augenblick et al., 2015). We embed our labor supply models in a simple 'static' framework instead of a dynamic set-up with stopping probabilities. From a practical perspective, all subjects in a given session submitted their working time choice simultaneously and independently of each other. This facilitated the experimental protocol and ruled out social comparisons and peer effects in the lab.

Treatments and tests. We compare the working time choices from round 4 across seven treatments. Treatment manipulations are variations of budget constraint parameters (W, Y) and, independently, the income path I and cumulative work M. This is one of the few studies that vary, experimentally, the income path or cumulative working time. Furthermore, to our knowledge, this is the only study that compares independent variation of the income path *and* cumulative working time. Treatments $\{W20\}$, $\{W50\}$, and $\{W80\}$ differ in budget constraint parameters only. We use $\{W20\}$ and $\{W50\}$ for our main analysis and $\{W80\}$ to test the robustness of our revealed preference tests. The next treatments $\{W20I\}$ and $\{W50I\}$ have the same budget constraint parameters (and the same cumulative working time across rounds 1 to 3) as $\{W20\}$ and $\{W50\}$ respectively, but income paths are higher. The final pair of treatments $\{W20M\}$ and $\{W50M\}$ have the same budget constraint parameters (and income path) as $\{W20\}$ and $\{W50\}$, but use longer default work episodes in rounds 1 to 3. We finally also vary the timing of experimental sessions per treatment. We study differences in leisure choices from round 4 (i.e., residual time *not* working in the lab) across subsamples characterized by different treatment conditions and/or subsamples collected at different times.

We first use parametric tests to determine the significance of differences in leisure between $\{W20\}$ and $\{W50\}$, between $\{W20\}$ and $\{W20I\}$ (respectively $\{W50\}$ and $\{W50I\}$), and finally between $\{W20\}$ and $\{W20M\}$ (respectively $\{W50\}$ and $\{W50M\}$). We find that differences in leisure between $\{W20\}$ and $\{W50\}$ are statistically significant, and in line with standard substitution effects. However, the data also suggest that the demand for leisure varies across the day. In a subsequent multivariate analysis, we pool observations from different treatments/sessions and estimate several specifications of leisure choices. Explanatory variables include age, gender, budget constraint parameters, the income path, cumulative working time, and time-of-the-day. Budget constraint wage effects are again significant. Similarly, time-of-the-day still correlates with the leisure choice, conditional on all other variables.

We then use revealed preference[8] methods for a structural nonparametric analysis

[8]'Empirical' revealed preference theory, in the spirit of Afriat (1967), Diewert (1973), and Varian

of the data. After all, while the statistical model behind the parametric exercise may be clear, its behavioral foundations are less clear. The revealed preference approach allows us to apply basic rationality conditions—consistency with utility maximization—at the individual level. We adopt Cherchye et al. (2018)'s test of rational behavior with two normal goods (leisure and income). We then follow Cherchye et al. (2019)'s *stochastic* revealed preference procedure to study the *distribution* of leisure choices for counterfactual budget constraint parameters. We find that the classical labor supply model, with heterogeneous preferences *independent* of income paths and cumulative work, produces reasonable bounds on labor supply responses. While time-of-the-day shifts some of these bounds considerably, we cannot separately identify leisure quantiles by time-of-the-day.

Structure of the paper. Section 2 defines the labor supply models. Section 3 presents the experimental design and treatments with independent variation of budget constraint parameters, the income path, and cumulative working time. We also discuss the timing of our experimental sessions. Section 4 consists of a parametric part with standard statistical tests and a nonparametric part with revealed preference analyses. Section 5 concludes.

2 Comparison of labor supply models

Consider an individual who chooses labor h and earns income y. Her choices are constrained by a linear budget:

$$y = Y + W \times h \tag{1}$$

where Y denotes a fixed payment and W the wage per unit of work. The budget constraint describes all the combinations of earnings y and working time h available to the individual.[9] Individuals are aware of all the budget constraint parameters. There is no uncertainty about the fixed payment or the wage. Correspondingly, in section 4, we offer direct tests of labor supply models, rather than tests of the 'double' hypothesis that some model *and* expected utility theory hold simultaneously.[10] Furthermore, we observe the historical income path I, cumulative past working time M, and the timing of the

(1982), is well-suited to analyze data from lab experiments. Revealed preference principles have been successfully applied to study individual rationality of children (e.g. Harbaugh et al., 2001), altruism and systematic preferences for 'giving' (e.g. Andreoni and Miller, 2002; Fisman et al., 2007; Cox et al., 2008), and inequality aversion (e.g. Bruyneel et al., 2017). It can be meaningfully applied to relatively small data sets, and avoids ad hoc restrictions on the form of utility functions.

[9] A distinguishing feature of our design is that the (total) budget is the sum of two terms: fixed income Y and labor income $W \times h$. This is different from the design of Abeler et al. (2011), in which subjects received Y *or* $W \times h$ with equal probability. Abeler et al. (2011) used variation of Y to shift the subjects' beliefs (income expectations) and to influence their expectations-based reference points.

[10] It is worth noting that expected utility theory is subject to several critiques, especially in lab experiments with relatively small earnings (Rabin, 2000).

work τ. In our experimental construction outlined below, individuals systematically earn $y_{-1} = y_{-2} = y_{-3} = I$ in three prior episodes. Actual working time in these initial episodes is $h_{-1} = h_{-2} = h_{-3} = H$. Let $M = 3 \times H$ denote cumulative working time up to the last episode. Parameter τ indicates the time-of-the-day of the experimental session.

We posit that individuals value leisure l (i.e., the complement of h) and income y. Each individual maximizes a utility function

$$U(l, y|\nu) \tag{2}$$

subject to (1). Let ν denote a vector of preference factors. These factors fully determine the shape of the instantaneous utility function for leisure and income. We impose no parametric structure on U, so each realization of ν may be associated with a completely new utility function. We also drop the common separability assumption between leisure and income. In practice, ν is unobserved. There are however possible correlations between ν on one hand and characteristics of the choice environment *and/or* of individuals (age, gender) on the other hand. We allow variation of ν both within and between individuals. In this framework, we can easily introduce the labor supply models under consideration. The main difference between the models in our study lies in the *correlation* between the (unobserved) preference vectors ν and the (observable) preference covariates I, M, τ. Depending on the empirical set-up, these covariates may be attributes of the choice environment and/or of the individual herself.[11]

We assume that, *conditional on I, M, and τ*, the distribution of preferences for work is independent of budget constraint parameters (W, Y): $F(\nu|W, Y, I, M, \tau) = F(\nu|I, M, \tau)$. We do not consider theories of price-dependent preferences in this paper. The labor supply models under consideration impose different assumptions on the conditional distribution $F(\nu|I, M, \tau)$. *The classical time separable model* assumes that $F(\nu|I, M, \tau) = F(\nu)$. In words, the distribution of preferences for work is independent of the income path, cumulative working time, and the timing of sessions. This structure is fundamentally restrictive and encompasses assumptions that rule out reference dependent preferences, assumptions that eliminate disutility spillovers, and finally assumptions that exclude timing-specific preferences. Several relaxations are possible. *Theories of reference dependent preferences* admit that the marginal rates of substitution between contemporaneous leisure and income depend on income targets.[12] We focus on 'backward-looking' targets based on the stream of former incomes. This implies that the distribution of preferences for work is conditional on the historical income path. We thus have $F(\nu|I, M, \tau) \neq F(\nu|M, \tau)$.

[11] Hence we take into account that timing τ can be an attribute of *both* the choice environment (scheduling of sessions) *and* the willingness of individuals to participate at certain times.

[12] Consider for instance the function $U(l, y) = \begin{pmatrix} \lambda(y - I) + G(l) \text{ if } y \geq I \\ \gamma\lambda(y - I) + G(l) \text{ if } y < I \end{pmatrix}$ with $G(l)$ the utility from leisure and $\gamma > 1$. The 'shape' of this utility function clearly depends on I. Each unique I is thus associated with a different preference vector ν and a different function $U(l, y|\nu)$.

Next, *theories of time nonseparable preferences* allow that the current tastes for leisure and income vary with the *number* and/or *timing* of work shifts. Applied to our set-up, preferences may depend on cumulative working time in the lab—disutility spillovers[13]—or cumulative time (clock hour) in the day—timing-specific preferences. Theories of disutility spillovers reject the notion that the distribution of preferences for work is independent of cumulative working time, i.e., $F(\nu|I, M, \tau) \neq F(\nu|I, \tau)$. Alternatively, timing-specific preferences reject that the distribution of preferences is independent of the timing of sessions, i.e., $F(\nu|I, M, \tau) \neq F(\nu|I, M)$.

classical time separable preferences:	$F(\nu\|I, M, \tau) = F(\nu)$
- preferences *not* reference dependent:	$F(\nu\|I, M, \tau) = F(\nu\|M, \tau)$
- *no* disutility spillovers:	$F(\nu\|I, M, \tau) = F(\nu\|I, \tau)$
- *no* timing-specific preferences:	$F(\nu\|I, M, \tau) = F(\nu\|I, M)$

In summary, the restrictive structure that the standard time separable model imposes on the distribution of preferences can be decomposed in assumptions that rule out reference dependent preferences, assumptions that rule out disutility spillovers, and finally assumptions that exclude time-specific preferences. This framework allows us to formulate alternative hypotheses, consistent with more flexible theories of preferences:

- Hypothesis **RD** (reference dependent preferences): The observed distribution of leisure changes with I, ceteris paribus.
- Hypothesis **DS** (disutility spillovers): The observed distribution of leisure changes with M, ceteris paribus.
- Hypothesis **TP** (timing-specific preferences): The observed distribution of leisure changes with τ, ceteris paribus.

The first hypothesis **RD** can be tested by letting I vary while keeping W, M, and τ constant. Any effect of the income path on the leisure distribution must stem from preferences ν, because the historical income path does not enter the current budget constraint (1). By construction then, any shift in the distribution of observed leisure provides empirical support for theories of reference dependent preferences. The next hypothesis **DS** can be verified by letting M vary conditional on W, I, and τ. Again, M can only affect the leisure distribution insofar it affects preferences ν; it does not enter the last period budget constraint. Then any shift in the distribution of observed leisure empirically supports theories of disutility spillovers. The final hypothesis **TP** could be

[13] Hotz et al. (1988) developed a theoretical model with intertemporal disutility spillovers. Consider for instance the utility function $U(l - M, y)$. The marginal utilities of this function can depend on the level of M. Each unique M is thus associated with a different preference vector ν and ultimately a different utility function.

tested by letting τ vary conditional on W, I, and M. However, hypothesis **TP** differs from **RD** and **DS** in the sense that **TP** came up *only after* our first round of data collection. Hence we cannot interpret tests of this hypothesis in a pure statistical way; we will nonetheless study the patterns in the data consistent with **TP**.

We will use the framework and hypotheses from this section mainly to achieve the second goal of this study: to detect the attributes that co-vary with instantaneous preferences for work.

3 Experimental design

We set up a lab experiment with a real-effort task. In this environment, we observed working time choices in various treatment conditions.[14] This will help us disentangle the labor supply models under consideration.

Subjects worked on a tedious task: counting the number of 1s in a series of tables. Each table consisted of 50 randomly ordered 0s and 1s. This task did not require any prior knowledge and performance was easily measurable. Furthermore, there was little learning possibility and effort was costly. The experiment involved two stages: an individual productivity elicitation part and a main part. Subjects were informed about the second part only after having completed the first.

Productivity elicitation part. During the first part of the experiment, subjects had five minutes to process as many tables as possible. In order to elicit individual productivity, subjects were offered a pure piece rate compensation scheme. For each correctly processed table, they received 50 tokens. The number of correctly processed tables was then used to design a feasible contractual effort in the second part of the experiment. Subjects were not aware of this.[15] Figure B.2 in Appendix B presents a screenshot of the productivity elicitation part.

Main part. In the second part of the experiment, the task was again to count randomly ordered 1s in a series of tables, but this time the compensation scheme was different. Subjects received a wage W per minute worked. The wage was conditional on having achieved a minimal output threshold in that minute. The threshold was expressed in terms of correctly processed tables. This design feature imposed a strictly positive cost of effort in each minute of work. Moreover, following Marchegiani et al. (2016), we normalized the cost of effort between subjects by scaling the output requirement per minute to each subject's individual productivity. In practice, the number of tables each

[14] Appendix A presents our instructions and Appendix B presents screenshots of the experiment.

[15] We use this information in Appendix D to demonstrate the robustness of our findings with respect to individual productivity differences.

subject was expected to process was fixed at 90% of the output from the productivity elicitation part.[16] Failure to meet this threshold in a given minute resulted in a wage loss for that minute.

The main part was further divided in four 'rounds' or work episodes.[17] The role of the first three rounds was to exogenously determine the income path and cumulative working time. In each of these rounds, subjects worked a *given* amount of time H. Cumulative working time was simply the sum of all these durations (i.e., $M = 3 \times H$). We operationalized mandatory work by means of short five-to-ten minute work episodes. Although longer episodes may generate stronger disutility spillovers, we wanted to keep the total duration of the experiment below 90 minutes, in line with the majority of studies in the lab.[18] To construct the income path, we let subjects work round-by-round. We separated rounds by small breaks during which subjects could access the internet, read magazines, or use their mobile phones. We reported earnings for each round separately. To reinforce the construction of target income, we designed the compensation scheme so that subjects systematically earned income I in each round 1 to 3. This amount of earnings also appeared explicitly in an example in the instructions. The aim of this repeated exposure to a given income target was to establish it as 'backward-looking' reference point for income in the subsequent decision problem. One caveat is that the systematic repetition of a given earnings level in early rounds may not suffice for subjects to 'internalize' the income path as reference point. Their true target income may be robust to experimental manipulations. Nonetheless, it is unlikely that an individual has just one reference point that applies universally to all her choices regardless of the decision-making context. In our experimental set-up, the income path in early rounds offered a salient and straightforward point of comparison for the working time choice in the final round. This is to some extent confirmed by the subjects' textual responses collected at the end of the experiment. We refer to Appendix D for more details.

In the fourth round, subjects could finally *choose* their working time. This last round started with a default work episode of five minutes. After the default episode, we offered subjects the opportunity to stay in the lab to earn more money. They could extend the duration of work by any number between 0 (*Stop Working*) and 15 minutes (*Working 15 additional minutes*). While the limit at 15 minutes may seem restrictive, this option increased the duration of round 4 fourfold, i.e., from 5 minutes of default work to a total of 20 minutes. Figure B.5 displays the corresponding choice menu. Subjects submitted

[16] Marchegiani et al. (2016) showed that this choice of threshold makes it possible for every subject to meet the output requirements. The 10% discount implies for instance that if the subject correctly counted 50 tables in five minutes in the productivity elicitation part, she must count at least $(0.9 \times 50)/5 = 9$ tables per minute in the main part.

[17] We informed subjects that they would receive their pay-off from just one round. The objective of this random lottery incentive scheme was to avoid that subject motivation wavered after a few rounds.

[18] In the field, it would be possible to significantly expand these time intervals, but it would be challenging to deal with typically unobserved intertemporal variation in the *demand* for labor.

their working time choice once and subsequently committed to the chosen duration by construction.[19] Those who stopped work immediately earned a fixed income Y. Each additional minute was rewarded at the same wage rate W as in the first three rounds.[20] Seen together, parameters (W, Y) and the corresponding budget constraint from section 2 describe all combinations of income and leisure that were available to subjects in round 4. Notice that this budget set was independent of predetermined parameters I and M, so that I and M can affect labor supply decisions only via preferences. To enhance the external validity of the experiment, we allowed subjects to leave the lab after they had stopped working. The opportunity cost of work was thus personal free time outside the lab.

Treatment conditions. We varied, across treatments, budget constraint parameters, the income path, and cumulative working time. Table 1 lists the seven treatments $\{W20\}$, $\{W50\}$, $\{W20I\}$, $\{W50I\}$, $\{W20M\}$, $\{W50M\}$, and $\{W80\}$. For ease of exposition, we named treatments after the wage level of the treatment $\{W\cdot\}$ and we added a letter at the end of the expression to denote the potential taste shifter that was raised. Note however that I in $\{W20I\}$ and M in $\{W20M\}$ exceed the corresponding values in $\{W50I\}$ and $\{W50M\}$. We included treatments with 'intermediate' parameter values to provide a more nuanced and complete overview of the effects of potential taste shifters.

Treatments	BC parameters (W, Y)	Income path I	Cumul work M
$\{W20\}$	(20, 1200)	900	15
$\{W50\}$	(50, 1100)	900	15
$\{W20I\}$	(20, 1200)	1500	15
$\{W50I\}$	(50, 1100)	1300	15
$\{W20M\}$	(20, 1200)	900	30
$\{W50M\}$	(50, 1100)	900	24
$\{W80\}$	(80, 900)	900	15

Table 1 – Overview of treatment conditions with budget constraint parameters (in tokens), the income path (in tokens), and cumulative working time (in minutes)

The first dimension in which treatments differ is the budget constraint. $\{W20\}$ offered the lowest wage of $W = 20$ tokens per minute (with $Y = 1200$) while $\{W50\}$ offered a wage of $W = 50$ tokens per minute (with $Y = 1100$). Price variation is needed to implement the revealed preference tests of section 4.2. To further enhance the power of

[19] Our choice menu implies a dissociation between the working time choice and the work itself. By fixing the timing of decision making, we rule out complicated interactions between the effect of cumulative working time on one hand and dynamic inconsistencies on the other.

[20] We kept the wage rate constant across rounds in order not to trigger a reciprocal response in round 4. The fixed income consisted of a lump sum and the earnings from five minutes of default work.

these tests, we introduced a third budget regime in $\{W80\}$. Treatment $\{W80\}$ changed the budget constraint to $(W, Y) = (80, 900)$. We refer to Appendix C.2 for more details.

The second dimension in which treatments differ is the income path. The income path determines the benchmark against which subjects can compare earnings, but it does not change their choice set or the hypothetical maximal pay-off from round 4. Treatment $\{W20I\}$ raised the income path from $I = 900$ in $\{W20\}$ to $I = 1500$. Subjects in $\{W20I\}$ could earn 1500 tokens per work episode in rounds 1 to 3. In round 4, subjects had to work the maximal amount of time if they wanted to avoid 'losses' compared to the income path, given $(W, Y) = (20, 1200)$. Similarly, treatment $\{W50I\}$ raised the income path from $I = 900$ in $\{W50\}$ to $I = 1300$. In round 4, subjects had to work at least four minutes to stay on the income path, given $(W, Y) = (50, 1100)$. $\{W20I\}$ and $\{W50I\}$ were in sharp contrast to all other treatments, in which subjects could not lose vis-a-vis the income path even if they stopped work immediately, because $Y \geq I = 900$.

The final dimension in which treatments differ is cumulative working time. Treatment $\{W20M\}$ increased cumulative working time in the first three rounds by 15 minutes in total, from $M = 15$ in $\{W20\}$ to $M = 30$. As such, by the time subjects in $\{W20M\}$ submitted their final work choice, subjects in $\{W20\}$ had already ended all their (default and optional) work. Treatment $\{W50M\}$ was similar to this, but increased cumulative working time by only 9 minutes from $M = 15$ in $\{W50\}$ to $M = 24$.

An important feature of our design is *independent* variation of the income path and cumulative working time. We vary the income path separately from cumulative working time (and separately from the budget constraint parameters in round 4) by changing, across treatments, the size of a fixed payment in rounds 1 to 3. The size of this fixed payment is simply the difference between I and $W \times H$. In addition, as we will explain next, we vary time-of-the-day separately from treatment conditions by implementing experimental sessions of the same treatment (i.e., a given combination of (W, Y), I, and M) at different times of the day.

The timing of sessions. We organized 31 sessions in total. Theories of timing-specific preferences state that the preferences for work cannot be seen separately from the timing of the work. The distribution of preferences for work could be very different between sessions. To limit variation in the timing dimension, we planned *no* sessions over the weekend, in examination periods, during Christmas holidays, or in the months August-September. However, the start hour of experimental sessions varied between 10h00 in the morning and 16h00 in the afternoon. Time-of-the-day is perhaps one of the most important aspects of timing, in line with the literature on circadian rhythms and tiredness patterns across the day. We therefore also include time-of-the-day as a possible preference covariate, alongside income paths and cumulative working time.[21]

[21] We thank the editor and referees for pointing this out.

In principle, any dependency between time-of-the-day and the aforementioned treatment conditions makes it difficult to separate the effect of cumulative time in the day τ on labor supply (i.e., timing-specific preferences) from the effects of (W, Y), I, or M. Labor supply differences attributed to treatment effects may actually stem from timing-specific preferences, and vice versa. One possible solution is to randomize treatments within a session. Subjects in the same session would then face different treatment conditions. This would clearly facilitate controlling for timing. The downside of such procedure is that it would introduce another potential confound: peer effects in the lab. Suppose that, within a session, some subjects were enrolled in $\{W20\}$ and others in $\{W20M\}$. Subjects in $\{W20M\}$, who are typically 15 minutes behind, would observe their peers in $\{W20\}$ leave the lab. They would then learn the full distribution of time choices made by individuals in $\{W20\}$, before submitting their own choice. By implementing just one treatment per session, we made sure that the choices of all subjects were mutually independent.

We took alternative steps to mitigate possible dependency between treatments and the timing of sessions. First, we organized several experimental sessions per treatment, at different times between 10h00 and 16h00. The distribution of these experimental sessions over the day was moreover stable across the various treatment parameters. Figure 1 shows the times at which experimental sessions started. Notice the relatively uniform spread of start hours per treatment condition. Second, we will also control for timing in the following empirical analyses. For these reasons, differences in time-of-the-day cannot explain systematic differences in average leisure between treatments. One can still distinguish, for instance, between preference variation associated with cumulative work (disutility spillovers) and preference variation associated with time-of-the-day (timing-specific preferences) in our set-up.

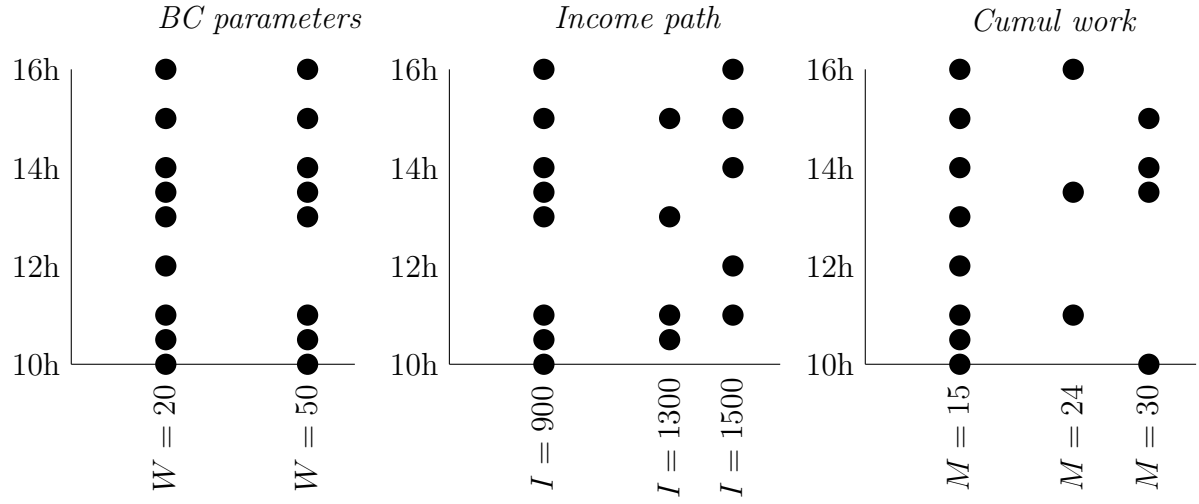

Figure 1 – Distribution of start hours of experimental sessions, by budget constraint parameters, the income path, and cumulative working time

Yet, these concrete steps do not rule out the possibility that individuals self-selected

into sessions of their preferred timing. There is a fundamental difference between variation of the treatment conditions and variation of the timing of sessions. Treatment conditions were unknown to subjects at the moment of their registration for the experiment. We did not disclose the nature of the task before subjects agreed to participate. All participants were given the same general indication of the maximal duration of the experiment (i.e., 90 minutes) prior to registration and independent of the treatment assignment. Ceteris paribus, this reduces correlations between individual characteristics and treatment conditions. By contrast, subjects could self-select into sessions of their preferred timing. This can produce correlations between individual characteristics and the timing of sessions. It is possible, for instance, that participants who chose to join mid-day sessions are less hard-working than participants who joined sessions in the morning or later in the afternoon.

For this reason, the present study cannot answer the question whether time-of-the-day co-varies with *within-individual* preferences for work or with differences in the preferences for work *between individuals*. This is in sharp contrast to preference variation induced by (exogenous) parameters I or M, which can in fact be interpreted as evidence of *within-individual* heterogeneity. We leave the distinction between these explanations for future research. In this study, we can only verify in a general way whether the distribution of preference factors ν (responsible for labor supply variation within/between individuals) depends on time-of-the-day. If this is the case, preferences for work cannot be seen separately from the timing of shifts.

Protocol of the experiment. A total of 330 individuals participated to (one of) the treatments laid out in Table 1. We conducted a first experiment in Luxembourg in 2017-2018 with 213 participants and a follow-up in Strasbourg in 2021 with 117 participants.[22] All Luxembourgish participants came from a list of experimental subjects maintained at the Laboratory for Experimentation in Social Sciences at the Luxembourg Institute of Socio-Economic Research (LISER-LAB). French participants were recruited from a list of experimental subjects maintained at BETA, University of Strasbourg. We used ORSEE software for all recruitments (Greiner, 2015). Subjects had an average age of 24 years and 55 percent were female, in both locations. They had diverse backgrounds but 28% of them were studying business management or economics.

The experiment was computerized using the ECONPLAY platform.[23] Upon arrival in the lab, each subject was assigned to a computer at random. The instructions were

[22]The objective of the follow-up was to obtain a better (more uniform) spread of the timing of sessions per treatment across the day. Apart from the scheduling of sessions (and the language of the instructions) there were no further differences in experimental design between both locations. Controlling for time-of-the-day, we detect no significant differences in labor supply between the Luxembourg and Strasbourg respondents. We nonetheless control for location/data collection phase in the regressions that follow.

[23]For more information on this software, we refer to http://www.econplay.fr

read aloud by the experimenter. Questions were answered in private before the start of the experiment. Throughout the experiment, we expressed income in tokens with a conversion rate of 100 tokens to €1.

Furthermore, we used a series of lottery choices as in Abeler et al. (2011) to elicit a subject's loss aversion. Each decision was a binary choice between a fixed payment of zero and a lottery that resulted *either* in a loss *or* in a gain. The gain was always 600 tokens, the loss increased from -200 (in Lottery 1) to -700 (in Lottery 6). Figure B.1 provides an overview of the six choice problems.

Finally, total earnings consisted of a show-up fee (€5), the payoff from one of six lotteries randomly selected, the payoff from the productivity elicitation part, and the earnings from the main part. In the main part, we informed subjects that they would receive their income from just one of the four rounds. This random lottery incentive scheme avoided that subject motivation wavered as more money was earned towards the end of the experiment.[24] After answering a demographic questionnaire, subjects left the room and received their earnings in a separate room, in private. Average earnings was €20 (standard deviation = €5).

4 Results

To separate budget constraint wage effects from possible preference covariates—and to detect the latter—we now exploit independent variation of our treatment conditions (budget constraint parameters, income paths, cumulative work) and the timing of the experimental sessions. Section 4.1 uses statistical methods to study differences in round 4's time choices. Section 4.2 applies revealed preference principles to these time choices to nonparametrically investigate properties of the underlying preference distribution.

Before presenting our results, two remarks are to be made. First, for ease of exposition, we report the subjects' choices not as working time h but in terms of its complement leisure $l = 15 - h$. This transformation leaves all our conclusions unaffected, but it streamlines the exposition of findings from the statistical and revealed preference tests. Second, for the following analyses, we will focus on the time choices in our main treatments $\{W20\}$, $\{W50\}$, $\{W20I\}$, $\{W50I\}$, $\{W20M\}$, or $\{W50M\}$. We use treatment $\{W80\}$ to test the robustness of our revealed preference results, in Appendix C.2.

4.1 Statistical analysis of time choices

The average leisure choice across all treatments was four minutes. This corresponds to an average loss of 149 tokens. There was also considerable heterogeneity across the sample.

[24] Cubitt et al. (1998) found no systematic differences between responses in 'random lottery' designs and responses in 'single choice' designs. Moreover, our construction is in line with the typical incentive system used in lab experiments for revealed preference applications (Cherchye et al., 2012).

About half of the subjects chose *some* leisure (i.e., work strictly less than 15 minutes) and five percent chose to stop work immediately (i.e., enjoy 15 minutes of leisure).

The following three subsections contain the main results from the parametric analysis. The first set of results reveals the magnitude and nature of 'pure' budget constraint wage effects, controlling for different sources of preference heterogeneity. The second set of results then indicates which attributes effectively co-vary with preferences for work. The final results shed light on which part of the observed leisure variation is due to budget constraint effects and which part is due to preference heterogeneity. This corresponds to the three research objectives we formulated at the start of this paper.

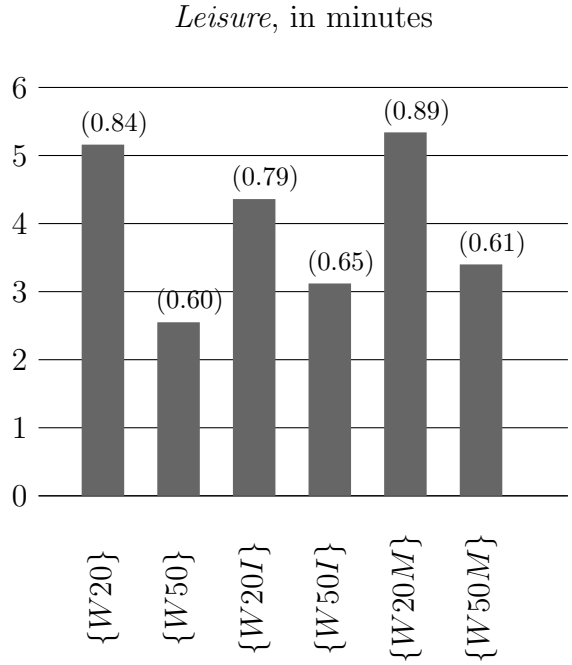

Figure 2 – Histograms of average leisure by treatment. Standard errors of means between brackets.

Analysis of 'budget constraint' wage effects. Figure 2 presents a histogram of average leisure by treatment. The main observation from this figure is that subjects selected systematically more leisure in treatments $\{W20\}$, $\{W20I\}$, and $\{W20M\}$ compared to treatments $\{W50\}$, $\{W50I\}$, and $\{W50M\}$. This can be explained by a standard substitution effect. We study the effect of budget constraint parameters on leisure by means of a two-sample t-test. Table 2 contains the results, discussed below, of several two-sample t-tests of differences in average leisure. The first test compares average leisure in $\{W20\}$ with average leisure in $\{W50\}$. These treatments are characterized by the same income path and the same cumulative working time, which allows us to isolate the effects of wage parameters. We find that subjects chose more leisure when wages were low, i.e., in treatment $\{W20\}$. The difference is statistically significant at the 1% level. We thus find

Two-sample t-tests	# obs	mean	s.e.m.	t-stat	p value
By BC parameters					
$\{W20\}$	50	5.16	0.84	2.45	0.01
$\{W50\}$	42	2.55	0.60		
By Income path					
$\{W20\}$	50	5.16	0.84	0.69	0.24
$\{W20I\}$	45	4.36	0.79		
$\{W50\}$	42	2.55	0.60	-0.63	0.74
$\{W50I\}$	51	3.12	0.65		
By Cumulative work					
$\{W20\}$	50	5.16	0.84	-0.15	0.44
$\{W20M\}$	44	5.34	0.89		
$\{W50\}$	42	2.55	0.60	-0.99	0.16
$\{W50M\}$	52	3.40	0.61		
By Time-of-the-day					
$\{\tau = 0 \vert W = 20\}$	47	3.77	0.80	-1.77	0.04
$\{\tau = 1 \vert W = 20\}$	92	5.57	0.60		
$\{\tau = 0 \vert W = 50\}$	58	2.41	0.53	-1.46	0.07
$\{\tau = 1 \vert W = 50\}$	87	3.48	0.48		

Table 2 – Two-sample t-tests of differences in average leisure (1) between treatments that vary in budget constraint parameters, (2) between treatments that vary in the income path, (3) between treatments that vary in cumulative working time, and (4) between sessions organized in different periods of the day

relatively strong budget constraint (substitution) effects in the sample.

Analysis of preference covariates. In a next step, we investigate which of the variables under consideration co-vary with preferences for work.

Figure 2 shows that the impact of the income path is small. Subjects on the higher income path in $\{W50I\}$ chose slightly *more* leisure compared to subjects in $\{W50\}$, *against* the predictions of income targeting. The second pair of t-tests in Table 2 offer formal statistical evidence. The tests examine the impact of income paths, conditional on M and (W, Y). Hypothesis **RD** states that leisure should vary between $\{W20I\}$ and $\{W20\}$. Loss aversion, a typical building block of models of reference dependent preferences, moreover implies that the demand for income increases (and the demand for

leisure decreases) in $\{W20I\}$. However, we cannot reject the null hypothesis that leisure is equal in $\{W20\}$ and $\{W20I\}$ (or in $\{W50\}$ and $\{W50I\}$).

An alternative possibility is that reference dependency only affects the decisions *closest* to the reference point. This would be the case if utility is discontinuous at the reference point (Rees-Jones, 2018). To investigate this possibility, one could first compute the probability of choosing 15 minutes of work—needed to stay on the income path in $\{W20I\}$—versus 14 or 13 minutes. In $\{W20I\}$, 24 out of 45 subjects (53%) chose precisely 15 minutes whereas no one chose 13 or 14 minutes. However, in $\{W20\}$ with much lower target, still 24 out of 50 subjects (48%) chose 15 minutes. Only two (4%) chose 13 or 14 minutes. There is bunching at 15 minutes but it seems unrelated to I. Similarly, one could compute the probability of choosing 5 or 4 minutes of work—needed to stay on the income path in $\{W50I\}$—versus 3 or 2 minutes. In $\{W50I\}$, 7 out of 51 subjects (14%) chose 4 or 5 minutes and just one (2%) chose 2 minutes. In $\{W50\}$, without higher target, 3 out of 42 subjects (7%) chose 4 or 5 minutes but no one chose 2 or 3 minutes. Again, there is some bunching at 5 minutes but it appears to be independent of I. Unfortunately, the number of observations around the respective thresholds is too small to pursue the analysis further.[25]

Another observation from Figure 2 is that subjects selected more leisure in treatments where 'prior' work episodes were long. The difference in average leisure between treatments $\{W50M\}$ and $\{W50\}$ is about one minute on average. The third set of t-tests in Table 2 show the effect of cumulative working time, conditional on I and (W, Y). Hypothesis **DS** states that leisure should vary between $\{W20M\}$ and $\{W20\}$. Spillovers typically come in the form of increasing fatigue or boredom associated with the task. This implies that the demand for leisure increases (and the demand for income decreases) in $\{W20M\}$. While subjects indeed chose slightly more leisure in $\{W20M\}$ compared to $\{W20\}$ (and clearly more leisure in $\{W50M\}$ compared to $\{W50\}$), this is not sufficient to reject the null hypotheses in favor of **DS**. Differences in average leisure are not statistically significant.

Finally, Figure 3 plots average leisure in function of the start hour of experimental sessions. We find an inverse U-shaped relationship between average leisure and time-of-the-day. The demand for leisure rises quickly between 10h30 and 12h00, reaches a final peak around 15h00, and then declines fast until 16h00. Since hypothesis **TP** was

[25] McCrary (2008) formalized a test of discontinuity at some cut-off in the density function of a variable. The method first constructs a 'finely binned' histogram and then applies local linear regression, on either side of the cut-off, to explain bin *heights* on the basis of bin *midpoints*. This method is not directly applicable here because our histogram can have at most 16 bins (0 minutes, 1 minute, ..., 15 minutes) which is clearly insufficient for local linear regressions. Alternatively, Camacho and Conover (2011) simply compared frequencies close to the threshold, without binning. To assess the magnitude of a discontinuity, the authors compared it against the discontinuity in an alternative distribution without manipulation/treatments. Our approach is similar, but in Camacho and Conover (2011) the number of observations around the cut-off (112) is considerably larger.

suggested only *after* the first round of data collection, we cannot test it in a pure statistical sense. The following analysis does not formally test **TP** but it does express an interesting data pattern. In line with Figure 3, we set $\tau = 1$ if sessions started around midday (i.e., between 11h00 and 15h00) and $\tau = 0$ otherwise. This will streamline our parametric and nonparametric analyses. Alternative definitions of the midday indicator (e.g., $\tau = 1$ if sessions started between 12h00 and 15h00) do not alter the following findings. We moreover condition on (W, Y) because we found earlier that budget constraint parameters affect leisure choices considerably. There is a sufficient number of data points (minimum 47 observations) at each intersection of budget parameters and the midday indicator. The final set of results in Table 2 indicates that subjects in midday sessions chose more leisure than subjects in the morning and late afternoon sessions.

This warrants further investigation of time-of-the-day confounds in economic studies. While it would be interesting to also *uncover the specific mechanisms* behind these patterns, we must limit our analysis to *documenting leisure variation by time-of-the-day*. The reason is that subjects could self-select in sessions of their preferred timing. This makes it difficult to identify true within-individual variation in preferences for work by time-of-the-day. In addition, there may be other (unobserved) confounds, such as variation in the value of outside options between sessions, that underlie differences by time-of-the-day. We elaborate more on this in the conclusion. We nonetheless position our finding in the literature on timing effects. It partly concurs with observed within-day variation of reservation wages for work. Chen et al. (2019) found that Uber drivers—who can easily adapt their schedules—are more likely to work in the (early) evening and less likely to work at 14h00 or 15h00. This, however, may be a specific characteristic of people who self-select as Uber drivers. This raises the question whether the patterns can be generalized. Phenomena like 'post-lunch dips' and 'afternoon naps' offer some anecdotal evidence. More formally, Kahneman et al. (2004) used the daily reconstruction method (DRM) to reconstruct tiredness patterns across the day, in a sample of employed women from Texas. The authors indeed found a rapid increase of tiredness in the early afternoon but also, inconsistent with Figure 3, low levels of tiredness at noon. Two remarks put the latter result in perspective. First, a recent follow-up by Anusic et al. (2017) has applied Kahneman et al. (2004)'s DRM to a representative sample from the German Socio-Economic Panel (GSOEP). This revealed that tiredness is W-shaped, with a small *peak* at midday, consistent with Figure 3. Second, from a methodological perspective, the tiredness patterns recovered by DRM cannot be seen separately from the types of activities—working, socializing, sleeping, ...—that take place at each time. Perhaps the most fundamental evidence comes from sleep studies (Lovato and Lack, 2010). This literature suggests that a (secondary) period of sleepiness and napping occurs generally between 13h00 and 16h00.[26]

[26]This so called afternoon slump can lead to a drop in performance in simple experimental tasks

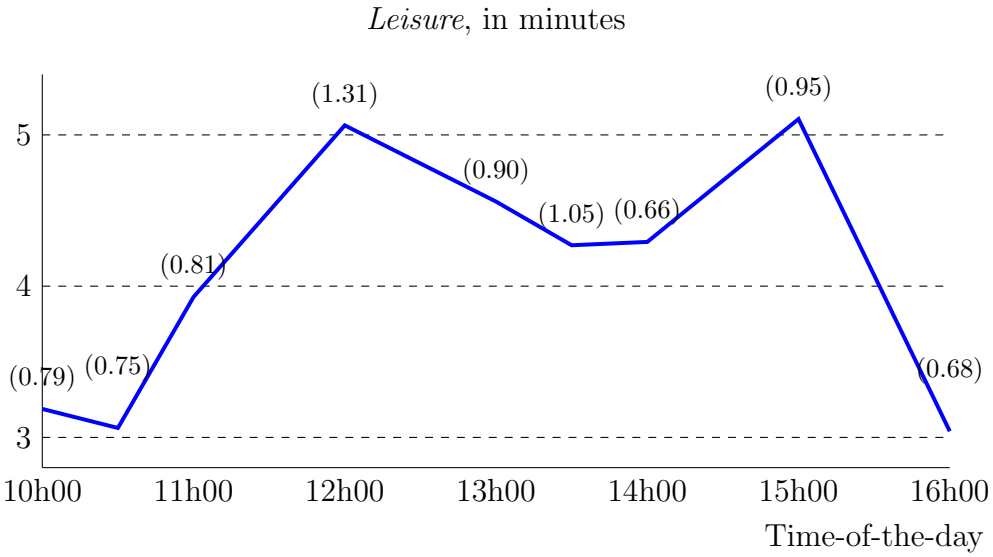

Figure 3 – Plot of predicted leisure in function of time-of-the-day at which sessions started. Leisure prediction based on a fractional polynomial regression of degree 5. Standard errors of means between brackets.

Multivariate analysis of leisure choices. We learned that the effect of budget constraint parameters (W, Y) outweighs the effects of treatment conditions I and M. But the data also suggest that the demand for leisure varies across the day. Preferences for work may not be independent of the timing of that work. This raises the question whether the budget parameters remain important predictors of the overtime choice if we condition on all other variables, including time-of-the-day. We turn to a multivariate analysis.

We predict leisure based on age, gender, a location dummy (Luxembourg or Strasbourg), budget constraint parameters, the income path, cumulative working time, and the midday dummy. The regressions use 284 observations from treatments $\{W20\}$, $\{W50\}$, $\{W20I\}$, $\{W50I\}$, $\{W20M\}$, and $\{W50M\}$. We consider several specifications. The first four control for *either* budget constraint parameters, the income path, cumulative working time, *or* time-of-the-day. The fifth controls for all variables jointly. Table 3 presents the results. In line with standard substitution effects, setting $W = 20$ (instead of $W = 50$) has a clear positive impact on leisure. The income path and cumulative working time do not seem to affect leisure choices, while the demand for leisure is stronger in midday sessions. We obtain similar results if we replace the midday dummy with linear and quadratic terms for the start hour of sessions. In a final specification, we also include an interaction term $(W = 20) \times (\tau = 1)$**.** The interaction is not significant; the results remain unchanged. This shows that, while leisure may depend on time-of-the-day, the wage

(Blake, 1967) and lower cognitive performance scores (Wertz et al., 2006). We refer to Figure D.1 in Appendix D, which shows a similar dip in (mean) performance in the productivity elicitation task at midday. Bessone et al. (2021) recently showed that short afternoon naps, between 13h30 and 14h00, improved productivity, psychological well-being, and cognition.

effect appears to be independent of time-of-the-day. Wages can in principle compensate for unfavorable timing conditions.

	(1) Leisure b/se	(2) Leisure b/se	(3) Leisure b/se	(4) Leisure b/se	(5) Leisure b/se	(6) Leisure b/se
female	-0.731	-0.643	-0.637	-0.811	-0.880	-0.901
	(0.61)	(0.63)	(0.62)	(0.62)	(0.61)	(0.61)
age	0.032	0.039	0.030	0.051	0.034	0.032
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Luxembourg	-0.066	-0.029	-0.018	-0.812	-0.779	-0.771
	(0.63)	(0.65)	(0.64)	(0.69)	(0.68)	(0.68)
$(W = 20)$	1.920***				1.853***	1.835***
	(0.60)				(0.61)	(0.62)
IncomePath		-0.000			-0.001	-0.001
		(0.00)			(0.00)	(0.00)
CumulWork			0.059		0.015	0.016
			(0.05)		(0.06)	(0.06)
Midday				1.945***	1.871***	1.832***
				(0.68)	(0.68)	(0.69)
$(W = 20) \times$ Midday						0.704
						(1.24)
constant	2.715	3.793	2.512	2.502	2.838	4.760*
	(1.74)	(2.31)	(1.92)	(1.76)	(2.80)	(2.87)
R-sqr	0.041	0.006	0.010	0.034	0.069	0.070

* p<0.1, ** p<0.05, *** p<0.01

Table 3 – Regressions of leisure based on age, gender, budget constraint parameters, the income path, cumulative working time, and timing (Midday) and location (Luxembourg) of sessions

Table 3 shows that budget constraint parameters and time-of-the-day explain some of the variation in leisure choices.[27] We conduct two robustness checks in Appendix D. We first investigate the mediating role of individual performance from the productivity elicitation task. Some individuals are more productive than others, and this may be correlated with disutility spillovers and fatigue. We then study the mediating role of loss aversion, derived from individual lottery choices. Subjects characterized by high degrees of loss aversion may react stronger to reference points. The new regressions, in Table D.1 of the appendix, confirm the results from the main analysis.

[27]Some individuals not always achieved their contractual minimal output requirement per minute. The wage associated with these minutes was then subtracted from the individual's earnings. Mistakes in early rounds may have influenced the individual's working time choice in round 4. To control for this, we constructed a new variable *PropFailed* which captures the proportion of minutes in which the output requirement was *not* achieved. *PropFailed* is just four percent on average, and zero for the majority of participants. Still, there is some variation across individuals. We repeated all our analyses for subsamples with $PropFailed \leq 0.11$ (mean + one standard deviation) and $PropFailed \leq 0.04$ (mean), and this confirms our main findings. Results are available upon request.

4.2 Revealed preference analysis of time choices

The parametric analysis of section 4.1 *first* discovered important substitution effects, *then* suggested time-of-the-day as a possible preference covariate, and *finally* showed that this preference covariate and the budget constraint parameters jointly explain some of the observed variation in time choices. The latter two findings shed light on preference variation in the sample. Preferences and utility functions are intrinsically microeconomic concepts.

Yet, the aforementioned results rely on a particular statistical model, and the microeconomic foundations of the statistical analysis are less clear. Lewbel (2001), for instance, showed that basic rationality conditions (consistency with utility maximization) at the individual level do not automatically translate to similar conditions on the statistical or econometric demand functions. Furthermore, regressions of leisure choices based on budget constraint parameters may produce biased results. A too restrictive specification of the relationship between leisure and wages may disadvantage explanations based on budget constraint parameters in favor of theories of preference variation—which are inherently flexible. But inflexible error terms and implicit 'representative agent' assumptions may also underestimate the true level of preference heterogeneity in the sample.

For all these reasons, we will now analyze our data by means of nonparametric revealed preference procedures. The revealed preference approach is independent of parametric form restrictions on utility functions. Throughout the following revealed preference analyses, we maintain just two assumptions: subjects are rational (each individual maximizes *some* utility function) and leisure $15-h$ and income y are normal goods (expansion paths are increasing with total available resources $Y + W \times 15$).[28]

In a preliminary step, we test the consistency of several subsamples with the Joint Normality Axiom of Revealed Preference ($JNARP$). When $JNARP$ is violated, the data could *not* be generated by *only one* utility function. We refer the interested reader to Appendix C for technical details. The number of $JNARP$ violations between $\{W20I\}$ and $\{W50\}$ (with a change in I) is comparable to the number of violations between $\{W20\}$ and $\{W50\}$. Variation of I does not affect the probability that preferences differ. The number of violations between $\{W20\}$ and $\{W50M\}$ (with a change in M) is slightly larger. Variation of M increases the probability that preferences differ. The number of violations increases further when we compare subsamples collected in different periods of the day. Seen together, this suggests that preferences depend on time, both cumulative working time M and time-of-the-day τ. But the violations between $\{W20\}$ and $\{W50\}$ demonstrate more generally that *one* utility function does not suffice to describe the

[28]The normality assumption is compatible with the time separable model and moreover common for goods such as leisure and income. It helps us make meaningful comparisons across choices from different budget lines. In Appendix C.2, we study the normality assumption in more detail. Our seventh treatment $\{W80\}$ increases wages to $W = 80$. This increases the degree of wage variation between treatments and enables us to compare the results of nonparametric tests with normality ($JNARP$) and without ($WARP$).

empirical distribution of leisure choices, even if one controls for the historical income path, cumulative working time, and the timing of experimental sessions. Representative agent assumptions are unrealistic even in simple real-effort tasks. One should consider a general distribution of preferences for work that can vary between/within individuals (as we do in theory section 2).

We therefore analyze wage effects across a general distribution of unobserved preferences for work. We moreover admit that this distribution depends on time-of-the-day, because our earlier parametric and nonparametric analyses showed that τ is a possible preference covariate. The only remaining assumption is that, conditional on time-of-the-day τ, the distribution of preferences for work is independent of the budget constraint parameters; thus $F(\nu|W,Y,\tau) = F(\nu|\tau)$.[29] We investigate if the substitution effects found in section 4.1 survive the introduction of these general forms of preference heterogeneity within and between individuals.

We aim at predicting the distribution of leisure $l(W_0, Y_0, \tau_0, \nu)$ for counterfactual budget constraint parameters (W_0, Y_0) and with an unobserved distribution of preferences for work ν that can depend on τ_0. We study three hypothetical budget lines. We fix the counterfactual wage to $W_0 = 25$, 35, or 45, and the lump sum payment becomes $Y_0 = 1675 - 15 \times W_0$. This construction keeps the total available income fixed as we vary W_0.[30] The empirical challenge is that we observe neither preference vectors ν (or their dependency on τ) nor counterfactual demands $l(W_0, Y_0, \tau_0, \nu)$. Indeed, none of our subjects actually faced budget parameters $W_0 = 25$, 35, or 45 in the experiment. We therefore adopt a *stochastic* revealed preference procedure. Applied to our empirical setup, stochastic revealed preference theory produces bounds on $l(W_0, Y_0, \tau_0, \nu)$ by taking as inputs the empirical distributions $l_{\{s\}}(\nu) = l(W_{\{s\}}, Y_{\{s\}}, \tau_{\{s\}}, \nu)$ from subsamples $\{s\}$ where $\tau_{\{s\}} = \tau_0$. We follow a recent procedure of Cherchye et al. (2019) to construct bounds on the cumulative distribution $\Pr(l(W_0, Y_0, \tau_0, \nu) \leq l_0)$ for each l_0. Let $lb(l_0)$ and $ub(l_0)$ denote lower and upper bounds on these probabilities:

$$lb(l_0) \leq \Pr(l(W_0, Y_0, \tau_0, \nu) \leq l_0) \leq ub(l_0).$$

To construct $lb(l_0)$, the procedure uses the empirical leisure distribution in subsamples $\{s\}$ where $W_{\{s\}} = 20 < W_0$. In our setting, $lb(l_0)$ is the proportion of bundles (l_i, y_i) that lie on budget line $(W, Y) = (20, 1200)$ and so that $l_i \leq l_0$ *and* $y_i \geq y_0$. Subjects who chose $l_i \leq l_0$ and $y_i \geq y_0$ from the actual budget line will necessarily choose less than l_0 from the counterfactual one, otherwise *JNARP* is violated. To construct $ub(l_0)$, the procedure

[29] When subjects registered for this experiment, they were unaware of the treatment parameters of their session. One may therefore expect the distribution of preferences to be reasonably similar across all our treatments, conditional on time-of-the-day.

[30] Notice furthermore that this total available income is fixed at the median of treatments $\{W50\}, \{W50I\}$, i.e., 1850 tokens, and treatments $\{W20\}, \{W20I\}$, i.e., 1500 tokens.

uses the leisure distribution in subsamples $\{s\}$ where $W_{\{s\}} = 50 > W_0$. In our setting, $ub(l_0)$ is the proportion of bundles (l_j, y_j) that lie on budget line $(W, Y) = (50, 1100)$ and so that $l_j \leq l_0$ *or* $y_j \geq y_0$. This construction has a similar interpretation as before: subjects who chose $l_j > l_0$ and $y_j < y_0$ from the actual budget line cannot choose less than l_0 from the counterfactual one, otherwise *JNARP* is violated.

Table 4 reports $lb(l_0)$ and $ub(l_0)$ for different levels of $l_0 = 0, 1, \ldots, 15$. For each counterfactual budget line, we obtain lower and upper bounds on the probability that leisure will be smaller than or equal to l_0. For instance, when $W_0 = 45$, less than 60% of the subjects working around midday choose no leisure. Between 79% and 95% of them choose less than (or equal to) ten minutes leisure.

	Non-midday ($\tau_0 = 0$)						Midday ($\tau_0 = 1$)					
	$W_0 = 25$		$W_0 = 35$		$W_0 = 45$		$W_0 = 25$		$W_0 = 35$		$W_0 = 45$	
l_0	Lb	Ub	Lb	Ub	Lb	Ub	Lb	Ub	Lb	Ub	Lb	Ub
0	0	0.69	0	0.69	0	0.69	0	0.6	0	0.6	0	0.6
1	0	0.69	0	0.69	0	0.69	0	0.61	0	0.61	0	0.61
2	0	0.69	0	0.69	0	0.84	0	0.61	0	0.61	0	0.72
3	0	0.84	0	0.84	0	0.84	0	0.72	0	0.72	0	0.75
4	0	0.84	0	0.84	0.6	0.86	0	0.72	0	0.75	0.42	0.76
5	0	0.84	0.6	0.86	0.62	0.88	0	0.75	0.42	0.76	0.45	0.78
6	0	0.84	0.6	0.86	0.66	0.88	0	0.75	0.42	0.76	0.47	0.78
7	0.6	0.86	0.62	0.88	0.77	0.9	0.42	0.76	0.46	0.78	0.61	0.82
8	0.6	0.88	0.77	0.9	0.77	0.95	0.42	0.78	0.61	0.82	0.64	0.93
9	0.62	0.9	0.77	0.9	0.77	0.97	0.45	0.82	0.61	0.82	0.64	0.94
10	0.62	0.95	0.77	0.95	0.83	0.97	0.46	0.93	0.64	0.93	0.79	0.95
11	0.77	0.97	0.83	0.97	0.85	0.97	0.61	0.94	0.79	0.94	0.8	0.98
12	0.77	0.97	0.85	0.97	0.85	0.98	0.61	0.95	0.83	0.95	0.83	1
13	0.77	0.97	0.89	0.97	0.89	1	0.61	0.98	0.84	0.98	0.84	1
14	0.77	0.98	0.89	0.98	0.89	1	0.64	1	0.85	1	0.85	1
15	0.83	1	1	1	1	1	0.79	1	1	1	1	1

Table 4 – Lower ($lb(l_0)$) and upper bounds ($ub(l_0)$) on the proportion of individuals who choose l_0 or fewer minutes of leisure, for counterfactual budget constraint parameters $W_0 = 25, 35, 45$ and $Y_0 = 1675 - 15 \times W_0$

We finally use the results from Table 4 to recover *quantiles* $l(\pi)$ of the leisure distribution. We define $l(\pi)$ so that

$$\pi = \Pr(l(W_0, Y_0, \tau_0, \nu) \leq l(\pi)).$$

To compute bounds on these quantiles, we use the probabilities $lb(l_0)$ and $ub(l_0)$ from

Table 4. We define $l^-(\pi)$ and $l^+(\pi)$ as follows[31]

$$\begin{aligned} \pi &= ub(l^-(\pi)); \\ \pi &= lb(l^+(\pi)). \end{aligned}$$

Then $l^-(\pi)$ and $l^+(\pi)$ are theoretical lower and upper bounds for leisure quantiles $l(\pi)$; we refer to Cherchye et al. (2019) for the formal argument. Our estimates in Table 5 confirm that there is substantial unobserved preference heterogeneity in the sample. Suppose that the wage is $W_0 = 35$. The median leisure choice will be situated between 0 and 5 minutes (0 and 8 minutes around midday) and the $90th$ percentile between 9 and 15 minutes. These bounds do not overlap so the $50th$ and $90th$ percentiles of the leisure distribution are separately identified.

Generally speaking, leisure quantiles are higher in midday sessions and decreasing in W_0. Focusing for instance on the midpoint of each interval $[l^-(\pi), l^+(\pi)]$, the third leisure quartile increases from 4 minutes on average ($W_0 = 45$) to 5 ($W_0 = 35$) and eventually 6.5 minutes ($W_0 = 25$) *in morning and late afternoon sessions*, and then further from 6.5 minutes ($W_0 = 45$) to 7.5 ($W_0 = 35$) and ultimately 10.5 minutes ($W_0 = 25$) *around midday*. This clearly demonstrates the substitution effects and time-of-the-day operating across the distribution of leisure choices. Admittedly, for a given quantile, there is much overlap between the bounds of different counterfactual budget constraints, and between the bounds for midday and non-midday sessions. We cannot separately identify leisure quantiles *per* budget line or time-of-the-day. However, the main takeaway from this stochastic revealed preference analysis is that the same patterns (i.e., with substitution effects) emerge even if we maximally allow for unobserved heterogeneity in the preferences for work across observations.

5 Conclusion

Insight in the determinants of labor supply is crucial for employers and policy makers. The literature has produced several models to describe and explain variation in labor supply within and between individuals. The classical 'time separable' model assumes that the distribution of preferences for work is independent of income targets, fatigue spillovers, and the timing of work. In this paper, we study popular relaxations (Fehr and Goette, 2007) of this structure: reference dependent preferences (income targeting) and time nonseparable preferences (disutility spillovers or timing-specific preferences).

This general framework admits that preferences for work co-vary with historical income paths, cumulative past working time, or the timing of shifts. Each of these attributes

[31] In practice, we recover the highest level of $l^-(\pi)$ so that $\pi \geq ub(l^-(\pi))$ and the lowest level of $l^+(\pi)$ so that $\pi \leq lb(l^+(\pi))$.

	Non-midday ($\tau_0 = 0$)			Midday ($\tau_0 = 1$)		
	$W_0 = 25$	$W_0 = 35$	$W_0 = 45$	$W_0 = 25$	$W_0 = 35$	$W_0 = 45$
Leisure						
$l(0.5)$	[0,7]	[0,5]	[0,4]	[0,11]	[0,8]	[0,7]
$l(0.55)$	[0,7]	[0,5]	[0,4]	[0,11]	[0,8]	[0,7]
$l(0.6)$	[0,9]	[0,7]	[0,5]	[0,11]	[0,8]	[0,7]
$l(0.65)$	[0,11]	[0,8]	[0,6]	[2,15]	[2,11]	[1,10]
$l(0.7)$	[2,11]	[2,8]	[1,7]	[2,15]	[2,11]	[1,10]
$l(0.75)$	[2,11]	[2,8]	[1,7]	[6,15]	[4,11]	[3,10]
$l(0.8)$	[2,15]	[2,11]	[1,10]	[8,15]	[7,12]	[6,11]
$l(0.85)$	[6,15]	[4,12]	[3,11]	[9,15]	[9,15]	[7,15]
$l(0.9)$	[9,15]	[9,15]	[7,15]	[9,15]	[9,15]	[7,15]
$l(0.95)$	[10,15]	[10,15]	[8,15]	[11,15]	[11,15]	[9,15]
$l(1)$	[15,15]	[15,15]	[15,15]	[15,15]	[15,15]	[15,15]

Table 5 – Lower ($l^-(\pi)$) and upper bounds ($l^+(\pi)$) on leisure quantiles, for counterfactual budget constraint parameters $W_0 = 25, 35, 45$ and $Y_0 = 1675 - 15 \times W_0$

may, in turn, co-vary with wages. In reality, wages change workers' income targets, incentivize work in *more* shifts—inducing fatigue spillovers—and incentivize work in *different* shifts—changing the timing of work. The aggregate wage effect is therefore a compound of 'pure' substitution and income effects on one hand and correlations with preference covariates on the other hand. A wage elasticity may say very little about pure substitution effects, let alone inform firms and policy makers about willingness-to-work predictors.

We set up a novel real-effort lab experiment in which individuals choose their working time. The experiment varies, independently, budget constraint parameters, the income path, cumulative working time, and the timing of sessions. This breaks up correlations between budget constraint parameters and possible preference covariates. To construct the income path, we rewarded subjects systematically with an exogenous maximal income per work episode in initial rounds of the experiment. To construct cumulative working time, we controlled the total duration of default work in the initial rounds.

The main objective of this research was three-fold. First, we aimed at identifying 'pure' budget constraint wage effects, i.e., by fixing the income path, cumulative working time, and the timing of sessions. In our statistical analysis, we detect significant differences in leisure between treatments with distinct budget constraint parameters. Low wages lead to less labor supply. The second objective was to detect the preference covariates. Revealed preference tests (and statistical tests) do not reject the classical model in favor of income targeting (and disutility spillovers) but the data do show different leisure choices across the day. The covariance of labor supply and time-of-the-day is remarkable since each subject in the experiment agreed to participate at that time in the first place. The choice to participate in a given shift is thus not a sufficient condition for high levels of labor supply; timing-specific preferences may still influence the intensive margin choice

within that shift. So any event that changes the timing of shifts—e.g., rainy moments for taxi drivers—may also change the length of that shift. Our final objective was to study which part of labor supply variation can be explained by pure budget constraint effects, and which part can be explained by preference variation. Multivariate regressions show that wages are important predictors of leisure choices, also after controlling for time-of-the-day. Stochastic revealed preference methods confirm that the classical model *with* preference heterogeneity, *albeit unrelated to income targets and cumulative past work*, produces reasonable bounds on the leisure distribution. Timing affects these bounds, but not to the extent that leisure quantiles are separately identified by time-of-the-day.

We finish the paper by a methodological comment and a recommendation. We studied individual behavior across several sessions; an experimental session was characterized by one treatment condition but also by its start hour, calendar date, location, etc. Our randomization of treatments *across* (rather than *within*) sessions can be problematic (Athey and Imbens, 2017); for instance, hard-working individuals may self-select into particular experimental sessions. Selection bias compromises conditional independence assumptions and obfuscates interpretation of differences between treatments (Duflo et al., 2007). More generally, differences in behavior between experimental sessions may stem from numerous (observed and unobserved) factors, beyond treatments, that vary between the sessions. The behavior of subjects in our study partially reflects this: time-of-the-day is a possible confound. While we suppressed correlations between the initial treatment conditions and the start hour of sessions—via our second round of data collection—and controlled for location and individual productivity, this does not eliminate all possible confounds. The value of experimentation lies precisely in the true randomization of treatments, which obviates the need for (ad hoc) corrections afterwards (Czibor et al., 2019). Randomization further improves credibility and transparency of the research. Investing in the appropriate physical space to conduct experiments is thus important.

Our main recommendation for future research is an experimental study that randomizes treatment conditions *within* a session. Subjects choose their working time under *different* treatment conditions (e.g., wages, income path, cumulative work) but roughly at the same time, on the same day, in the same university, etc. It may be necessary to put the subjects in different rooms—to avoid that subjects with $M = 30$ monitor the working time choices of those with $M = 15$. Furthermore, researchers interested in the time-of-the-day effect itself may invite subjects to the lab in a different way. Instead of communicating the start hour of sessions prior to registration, they could communicate only the date of the session and ask participants to leave their schedule open on that day. Next, the experimenter could randomize the start hour of experimental sessions within that day. In such empirical set-up, labor supply variation by time-of-the-day would indicate timing-specific preferences *at the individual level* rather than heterogeneity in the preferences for work *between individuals* who choose to participate at different times.

References

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011). Reference points and effort provision. *American Economic Review 101*, 470–492.

Afriat, S. N. (1967). The construction of utility functions from expenditure data. *International Economic Review 8*, 67–77.

Andreoni, J. and J. Miller (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica 70*, 737–753.

Anusic, I., R. E. Lucas, and M. B. Donnellan (2017). The validity of the day reconstruction method in the german socio-economic panel study. *Social Indicators Research 130*, 213–232.

Athey, S. and G. Imbens (2017). The econometrics of randomized experiments. In *Handbook of Field Experiments*, pp. 73–140. Elsevier.

Augenblick, N., M. Niederle, and C. Sprenger (2015). Working over time: Dynamic inconsistency in real effort tasks. *Quarterly Journal of Economics 130*, 1067–1115.

Baucells, M. and L. Zhao (2019). It is time to get some rest. *Management Science 65*, 1717–1734.

Bessone, P., G. Rao, F. Schilbach, H. Schofield, and M. Toma (2021). The economic consequences of increasing sleep among the urban poor. *Quarterly Journal of Economics 136*, 1887–1941.

Blake, M. J. F. (1967). Time of day effects on performance in a range of tasks. *Psychonomic Science 9*, 349–350.

Bowman, D., D. Minehart, and M. Rabin (1999). Loss aversion in a consumption–savings model. *Journal of Economic Behavior and Organization 38*, 155–178.

Brachet, T., G. David, and A. M. Drechsler (2012). The effect of shift structure on performance. *American Economic Journal: Applied Economics 4*, 219–246.

Bruyneel, S., L. Cherchye, S. Cosaert, B. De Rock, and S. Dewitte (2017). Measuring the willingness-to-pay for others consumption: An application to joint decisions of children. *Quantitative Economics 8*, 1037–1082.

Camacho, A. and E. Conover (2011). Manipulation of social program eligibility. *American Economic Journal: Economic Policy 3*, 41–65.

Camerer, C., L. Babcock, G. Loewenstein, and R. H. Thaler (1997). Labor supply of New York City cabdrivers: One day at a time. *Quarterly Journal of Economics 112*, 407–441.

Cardinali, D. P. (2008). Chronoeducation: How the biological clock influences the learning process. In *The Educated Brain*, pp. 110–126. Cambridge University Press.

Chen, M. K., J. A. Chevalier, P. E. Rossi, and E. Oehlsen (2019). The value of flexible work: Evidence from uber drivers. *Journal of Political Economy 127*, 2735–2794.

Cherchye, L., T. Demuynck, and B. De Rock (2012). Nash-bargained consumption decisions: A revealed preference analysis. *Economic Journal 123*, 195–235.

Cherchye, L., T. Demuynck, and B. De Rock (2018). Normality of demand in a two-goods setting. *Journal of Economic Theory 173*, 361–382.

Cherchye, L., T. Demuynck, and B. De Rock (2019). Bounding counterfactual demand with unobserved heterogeneity and endogenous expenditures. *Journal of Econometrics 211*, 483–506.

Collewet, M. and J. Sauermann (2017). Working hours and productivity. *Labour Economics 47*, 96–106.

Cox, J. C., D. Friedman, and V. Sadiraj (2008). Revealed altruism. *Econometrica 76*, 31–69.

Crawford, V. P. and J. Meng (2011). New York City cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *American Economic Review 101*, 1912–1932.

Cubitt, R., C. Starmer, and R. Sugden (1998). On the validity of the random lottery incentive system. *Experimental Economics 1*, 115–131.

Czibor, E., D. Jimenez-Gomez, and J. A. List (2019). The dozen things experimental economists should do (more of). *Southern Economic Journal 86*, 371–432.

DellaVigna, S., A. Lindner, B. Reizer, and J. F. Schmieder (2017). Reference-dependent job search: Evidence from Hungary. *Quarterly Journal of Economics 132*, 1969–2018.

Diewert, W. E. (1973). Afriat and revealed preference theory. *Review of Economic Studies 40*, 419–425.

Duflo, E., R. Glennerster, and M. Kremer (2007). Using randomization in development economics research: A toolkit. In *Handbook of Development Economics*, pp. 3895–3962. Elsevier.

Farber, H. S. (2005). Is tomorrow another day? The labor supply of New York City cabdrivers. *Journal of Political Economy 113*, 46–82.

Farber, H. S. (2008). Reference-dependent preferences and labor supply: The case of New York City taxi drivers. *American Economic Review 98*, 1069–1082.

Fehr, E. and L. Goette (2007). Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review 97*, 298–317.

Fisman, R., S. Kariv, and D. Markovits (2007). Individual preferences for giving. *American Economic Review 97*, 1858–1876.

Gächter, S., A. Herrmann, and E. J. Johnson (2007). Individual-level loss aversion in riskless and risky choices. *University of Nottingham Centre for Decision Research and Experimental. Economics Discussion Paper*.

Genesove, D. and C. Mayer (2001). Loss aversion and seller behavior: Evidence from the housing market. *Quarterly Journal of Economics 116*, 1233–1260.

Goette, L., D. Huffman, and E. Fehr (2004). Loss aversion and labor supply. *Journal of the European Economic Association 2*, 216–228.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association 1*, 114–125.

Hamermesh, D. S. (1999). The timing of work over time. *Economic Journal 109*, 37–66.

Harbaugh, W. T., K. Krause, and T. R. Berry (2001). GARP for kids: On the development of rational choice behavior. *American Economic Review 91*, 1539–1545.

Hogarth, R. M. and M. C. Villeval (2014). Ambiguous incentives and the persistence of effort: Experimental evidence. *Journal of Economic Behavior & Organization 100*, 1–19.

Hotz, J., F. Kydland, and G. Sedlacek (1988). Intertemporal preferences and labor supply. *Econometrica 56*, 335–360.

Kahneman, D., A. B. Krueger, D. A. Schkade, N. Schwarz, and A. A. Stone (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science 306*, 1776–1780.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*, 263–292.

Kőszegi, B. and M. Rabin (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics 121*, 1133–1165.

Kőszegi, B. and M. Rabin (2009). Reference-dependent consumption plans. *American Economic Review 99*, 906–936.

Lewbel, A. (2001). Demand systems with and without errors. *American Economic Review 91*, 611–618.

Lovato, N. and L. Lack (2010). The effects of napping on cognitive functioning. In *Progress in Brain Research*, pp. 155–166. Elsevier.

Marchegiani, L., T. Reggiani, and R. Matteo (2016). Loss averse agents and lenient supervisors in performance appraisal. *Journal of Economic Behavior and Organization 131*, 183–197.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*, 698–714.

Oettinger, G. S. (1999). An empirical analysis of the daily labor supply of stadium vendors. *Journal of Political Economy 107*, 360–392.

Pencavel, J. (2014). The productivity of working hours. *Economic Journal 125*, 2052–76.

Post, T., M. J. van den Assem, G. Baltussen, and R. H. Thaler (2008). Deal or no deal? Decision making under risk in a large-payoff game show. *American Economic Review 98*, 38–71.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica 68*, 1281–1292.

Rees-Jones, A. (2018). Quantifying Loss-Averse Tax Manipulation. *Review of Economic Studies 85*, 1251–1278.

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica 50*, 945–973.

Wertz, A. T., J. M. Ronda, C. A. Czeisler, and K. P. Wright (2006). Effects of sleep inertia on cognition. *JAMA 295*, 159.

A Instructions in $\{W50\}$

Thank you for participating in this experiment on decision making. In this experiment, your earnings depend on your decisions. In addition to the amount that you will earn you will receive a show-up fee of 5 euros. All your decisions are anonymous. You will never enter your name on the computer. You will indicate your choice on the computer to which you are sitting.

From now we ask you not to talk. If you have a question please raise your hand and an experimenter will meet you in private. It is forbidden to communicate with another participant during the experiment. If you violate this rule you will be disqualified from this experiment and of any potential payment.

The experiment consists of three successive games. Please start by reading the instructions for the first part carefully. You will receive the instructions for the two other parts of the experiment before starting them. In each part, your earnings are counted in tokens. Your earnings in cash from each part will be calculated according to the following conversion rate:

100 tokens = 1 €

At the end of the experiment, the total amount of your earnings in euros will be paid in cash.

GAME 1. During this game, you will make six successive decisions. Each decision is a choice between a fixed payment of zero and a lottery. The lotteries always involve a **50/50 chance** of winning 600 tokens or losing a given amount of tokens. All the six lotteries are displayed below and the loss varies from -200 to -700 tokens. There is no initial endowment and should you experience losses they will be withdrawn from your further earnings in the experiment.

For each lottery, you are asked to indicate whether you would prefer to play the lottery or to obtain zero token by ticking the preferred option. Here is the screen on which you will have to make your decision.

Figure B.1 here

> Example: at decision 1, you have to choose between receiving nothing and playing a lottery that gives you 600 tokens with a probability of 50% and makes you lose 200 tokens with a probability of 50%. The other five decisions are similar.

After you have made your six decisions, you must confirm this set of decisions by clicking the Validate-button. Once you have clicked this button, you can no longer change your decisions. At the end of the experiment, one decision will be selected randomly and be paid. Each decision has an equal chance of being chosen for your earnings.

GAME 2. In this game, your task is to count ones in a series of tables of zeros and ones during five minutes. The figure shows the work screen you will see later.

Figure B.2 here

On the screen you will find a table containing zeros and ones. You have to enter the number of ones into the box on the right side of the screen. After you have entered the number, click the OK-button. If you enter the correct result, a new table will be generated. If your input is wrong, you have two additional trials to enter the correct number into the table. You therefore have a total of three trials to solve each table. If you fail three times, a new table is automatically generated.

If you entered the correct number of ones **you receive 50 tokens for each table** you solved correctly.

> Example: you solve three tables correctly and you miscount one table, your earnings are $3 \times 50 = 150$ tokens.

You have 5 minutes to complete as many tables as you can. The number of correctly counted tables is displayed at the bottom of the screen. The remaining time, expressed in second, is displayed in the upper right-hand corner of the screen. The earnings of this part of the experiment will be paid to you at the end of the experiment.

GAME 3. This game consists of 4 successive rounds. At the end of the experiment one round will be randomly drawn for determining your earnings.

Each round is divided in two periods of time. During the first three rounds, rules are the same. The fourth round is different.

Rounds 1–3 During round 1 to 3, each round is divided into two periods: a working time period of 5 minutes and a leisure time period of 5 minutes.

During the working time, you have to accomplish an individual task. The task is once again to count ones in a series of tables, but new rules are now in effect. Your task is to solve a given number of tables per minute. Indeed at the very start of this game you will be assigned a goal, expressed as a number of tables, which you have to achieve every minute. This goal will be displayed on the screen before you start the game and will remain the same all along the game (see the screenshot below).

At the start of a round, you receive a fixed endowment of 650 tokens for taking part to this round. In addition to this endowment, you will be paid a wage per minute of work of 50 tokens conditionally on achieving the goal per minute you have been assigned. That means that for each minute in which you solve the number of required tables you receive a wage of 50 tokens. The goal per minute, the payment per minute and the endowment are specified on the initial screen (see below) and on the screen during the round.

Figure B.3 here

Your earnings during a round of the game is made of two components: a fixed endowment and a variable income depending on your ability to achieve the goal. Here, you are no more paid by tables solved but by minute.

> Example: Assume that the goal you are assigned is to solve 7 tables per minute. In round 1, if for each minute of the five minutes of the working time you solve 7 tables, you will be paid the fixed endowment of 650 tokens plus 5 $\times$ 50 tokens for a total earning of 900 tokens. If you fail to reach your goal for only one minute out of the five, then you will be paid $650 + 4 \times 50 = 850$ tokens for this round.

During the working time, you will face the same screenshot as the one you find below. As before, there is a table in which you have to count ones and report the correct number in the box on the right. Above the box, the minute in which you are playing (out of the five in total) and the remaining time within this minute is displayed. Under the box, you find the number of tables that you have already solved in the given minute. Be careful that the timer is expressed in seconds remaining in one minute. As before, you are offered three trials to correctly count the number of ones before a new table is generated. When a minute is elapsed, a new table is automatically generated.

Figure B.4 here

When the working time is over, you are offered an automatic period of leisure time of five minutes for you to rest. During this leisure time, you can access the internet or simply rest and wait for the next round to start. Again the remaining time is displayed on the screen.

Once the leisure time is over a new round will automatically start. The rules are exactly the same and the goal to achieve in terms of number of tables, the payment per minute and the fixed endowment are the same in each round. In total, you will play three rounds like that.

Round 4 After having played three rounds of this game, a last round will start. On the contrary to the three first rounds, the fixed endowment you receive at the start of the round will be 850 tokens instead of 650 tokens. Moreover, you will now have the opportunity to work during your leisure time and the total leisure time available is 15 minutes. Thus after the working time, you can decide to use as many minutes as available from your leisure time to solve tables at the given wage per minute.

Practically, once the regular working time is finished, you have to indicate how many additional minutes you want to work in addition to those already worked. You cannot work more minutes than those available in your leisure time but you can however decide to work less than the total time available. The screen below displays the choice you will have to make, i.e. deciding not to work during your leisure time or to work 1 minute or 2 minutes or 3 minutes... until at most 15 minutes.

Figure B.5 here

Your earnings for each additional minute is the same as previously, that is you receive a wage per minute of 50 tokens if you achieve the goal you are assigned. However, there is no additional endowment.

> Example: If you decide to work 2 additional minutes, you will receive $2 \times 50 = 100$ tokens in addition to your earnings from the regular working time.

Figure B.6 here

To summarize, during round 4, you solve tables during the regular working time and then you can decide to work more minutes or to stop. If you decide not to work anymore, game 3 is finished.

END OF THE EXPERIMENT. After completion of game 3 you will have to answer a few demographic questions and then you receive your payment and you may leave the lab. Each of you may leave independently of the other participants. Thus the time you stay in the lab depends on your decision to work or not during the leisure time in round 4. There is no obligation for you to wait until all of the other participants will have completed their tasks.

One round out of the four you played in Game 3 will be randomly picked and will be included in your total earnings of this experiment. Your earnings will be determined by the use of a virtual lottery. Thus your total earnings are the sum of your earnings in Game 1, Game 2 and one round of Game 3.

B Experiment's screenshots

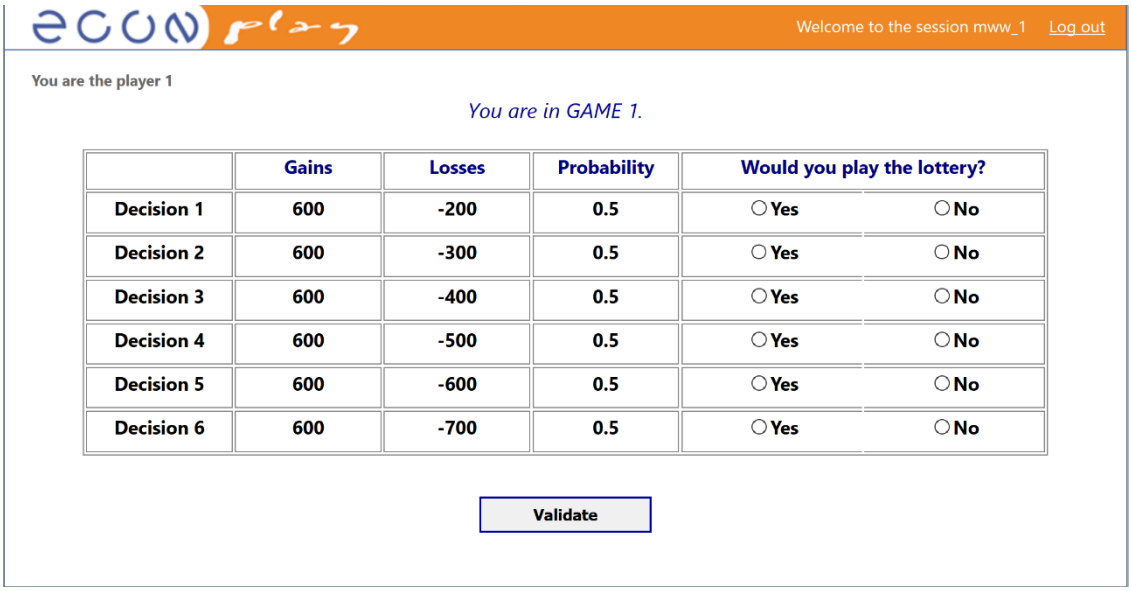

Figure B.1 – Screenshot of lotteries

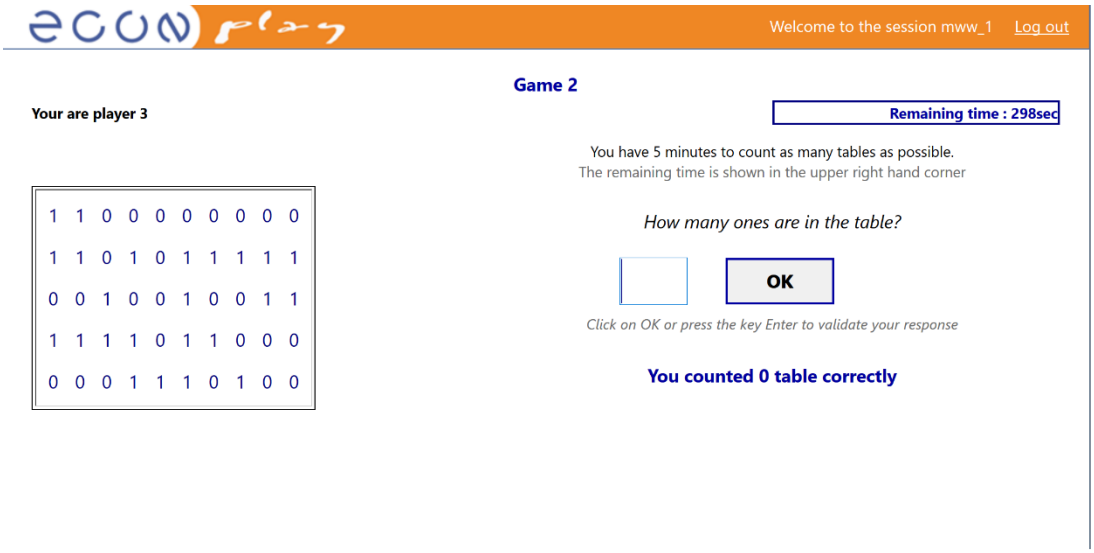

Figure B.2 – Screenshot of individual productivity elicitation

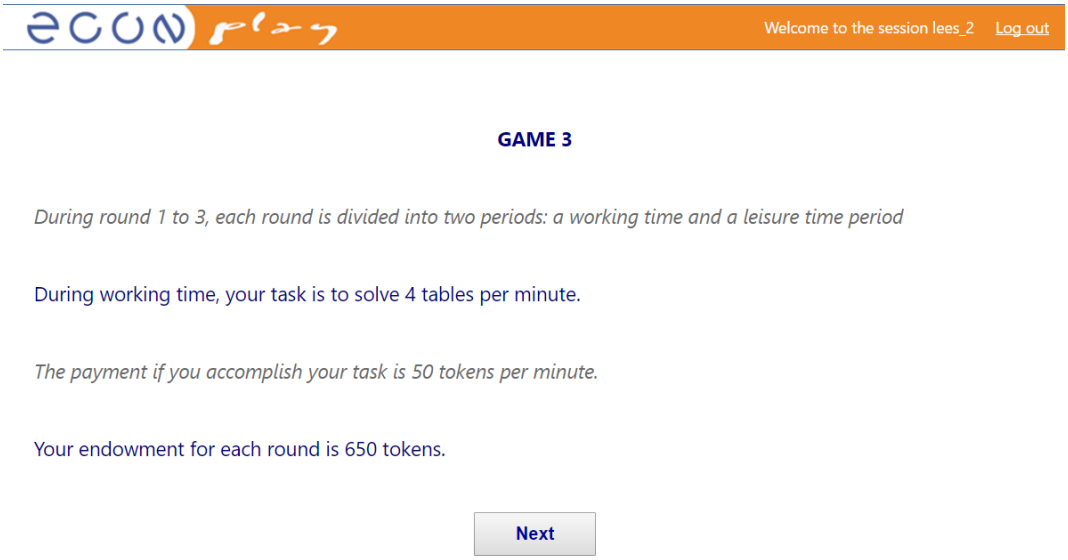

Figure B.3 – Screenshot of instructions in rounds 1 to 3

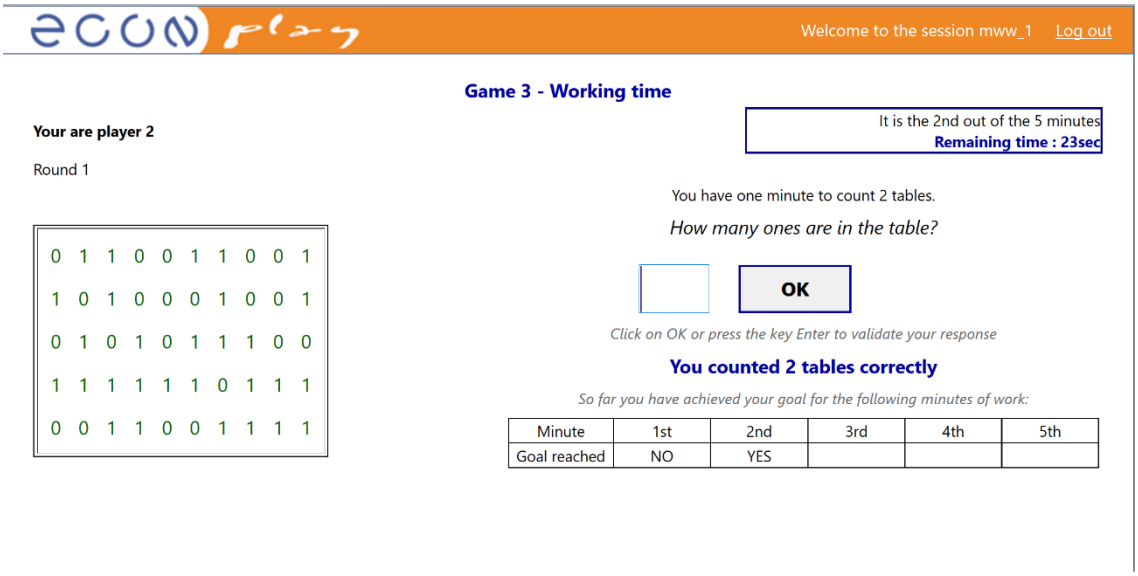

Figure B.4 – Screenshot of round 1

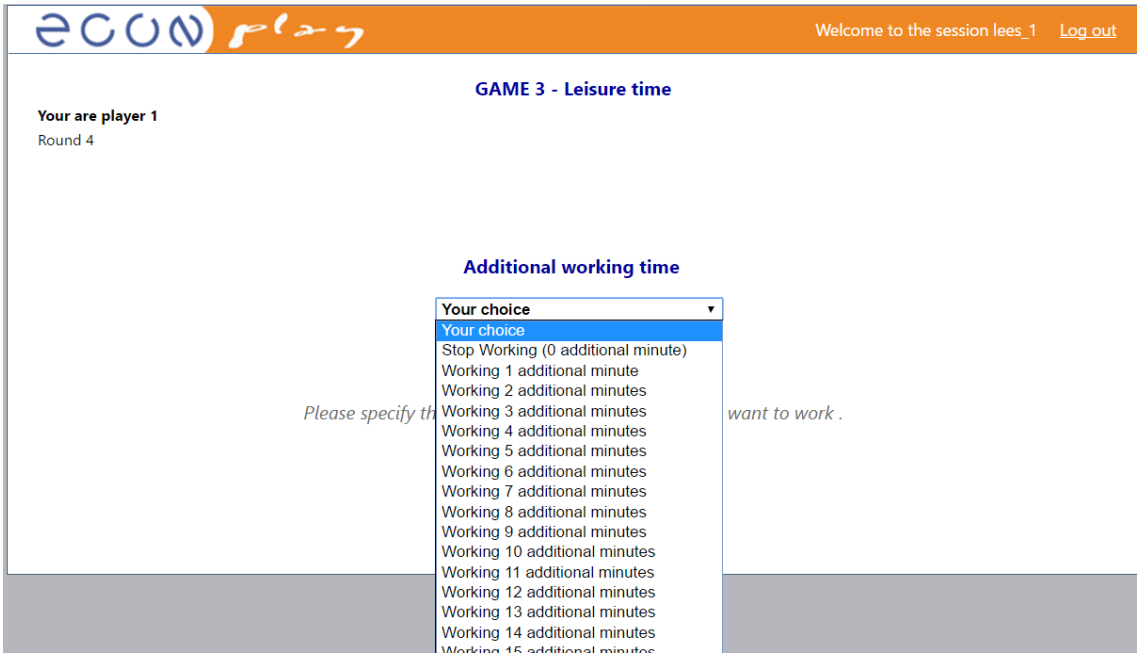

Figure B.5 – Screenshot of working time choice in round 4

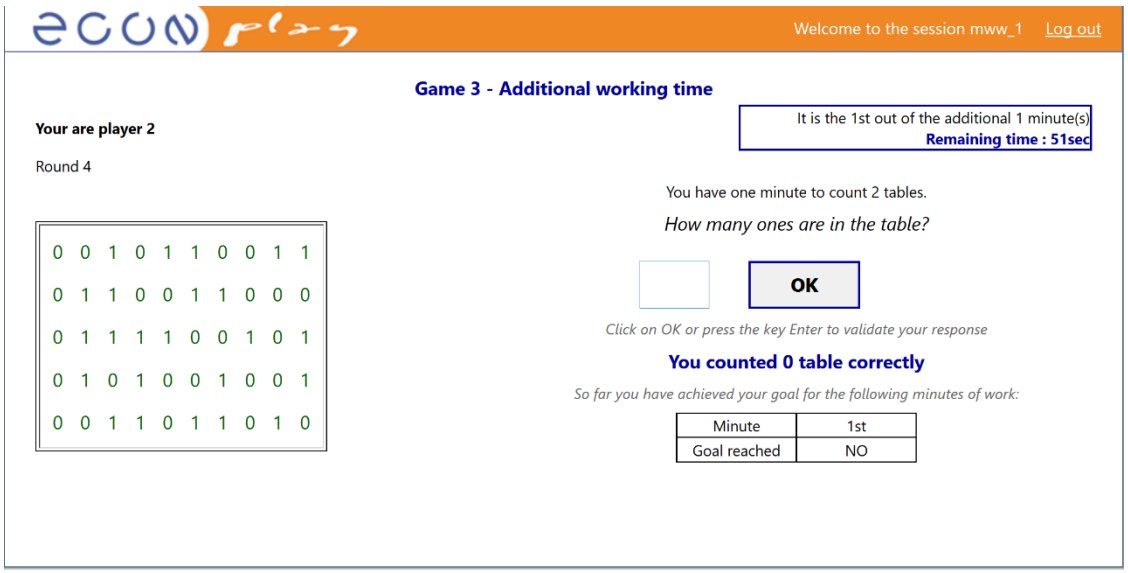

Figure B.6 – Screenshot of round 4

C Deterministic revealed preference tests

C.1 Nonparametric tests of preference covariates

In this more 'deterministic' revealed preference exercise, we start from the premise that the preferences for work should be (relatively) stable across subsamples with the same historical income path, the same cumulative working time, and the same timing of the sessions. This is a useful starting point for the analysis because the different extensions of the classical time separable model suggest that variation of the income path, cumulative work, or timing *shifts* the tastes for leisure and income. Following this logic, variation of leisure is informative, not if it is 'significant' in the usual statistical sense, but if it cannot be explained by *just one* utility function. Taking utility maximization and normality of leisure and income as given, we use Cherchye et al. (2018)'s Joint Normality Axiom of Revealed Preference (*JNARP*) to test if *one* utility function is sufficient to describe the data. As long as there exists at least one (stable) utility function that explains a set of leisure observations, we cannot say with certainty that preferences for work differ between the observations. By contrast, the more violations of *JNARP* we detect, the more we can reject the notion of preference homogeneity in the data.

Consider a pair of bundles: (l_i, y_i) chosen in subsample $\{s\}$ and (l_j, y_j) chosen in subsample $\{v\}$. Subsamples are characterized by treatment conditions (budget constraint parameters (W, Y), the income path I, or cumulative working time M) and/or by time-of-the-day τ. Denote the parameters associated with subsample $\{s\}$ by $(W_{\{s\}}, Y_{\{s\}})$, $I_{\{s\}}$, $M_{\{s\}}$, and $\tau_{\{s\}}$. Suppose that the wage in subsample $\{s\}$ was lower than the one in $\{v\}$. If leisure increased between i and j, then income should have increased too. If this condition is violated, then there is no utility function that explains i and j simultaneously. Definition 1 consists of simple combinatorial restrictions that are easy to implement.

Definition 1 (JNARP). *Let $W_{\{s\}} < W_{\{v\}}$. Consider a pair of bundles (l_i, y_i) from subsample $\{s\}$ and (l_j, y_j) from subsample $\{v\}$. The bundles satisfy the Joint Normality Axiom of Revealed Preference if*

$$\begin{aligned} l_i \leq l_j \quad &\textit{implies} \quad y_i \leq y_j \\ l_i < l_j \quad &\textit{implies} \quad y_i < y_j \end{aligned}$$

We measure deviations from preference homogeneity by the number of *JNARP* violations[32] in pairwise comparisons of bundles (l_i, y_i) from subsample $\{s\}$ and bundles (l_j, y_j)

[32] Admittedly, there are other ways to quantify preference heterogeneity in the nonparametric literature. One way is to 'split' the sample in the minimal number of partitions so that all subsets are internally consistent. However, in a setting with just two budget lines, two subsets trivially suffice because this eliminates all the price variation in the data. Another way is to use standard 'goodness-of-fit' criteria like the Afriat or Houtman and Maks index. However, these indices are harder to compute and are more commonly used to capture deviations from rationality.

from subsample $\{v\}$. The number of violations is always situated between zero and the product of the number of subjects per subsample.[33] The ability of revealed preference tests to pick up preference variation generally also depends on the budget constraint parameters used in the experiment. Discriminatory power measures the 'strength' of a revealed preference test, i.e., the likelihood that simulated *random* data will violate the conditions. Definition 1 demonstrates that wage variation is strictly required to detect *JNARP* violations. We designed the experiment so that it can easily pick up these violations. First, wages increase by 150 percent between $W = 20$ and $W = 50$.[34] A second important feature of the design is that treatments $\{W20I\}$ and $\{W50M\}$ differ from $\{W20\}$ and $\{W50\}$ in the income path or cumulative working time but *not* in the budget constraint parameters for round 4. We apply *JNARP* separately to subsamples based on treatments $\{W20\}$ and $\{W50\}$; subsamples based on $\{W20I\}$ and $\{W50\}$; and finally subsamples based on $\{W20\}$ and $\{W50M\}$. All these tests are characterized by the same set of prices ($W_{\{s\}} = 20$ and $W_{\{v\}} = 50$) and the same pair of budget lines. Thus, by construction, the chance that any pair of random bundles (from two budget lines) violates *JNARP* is *identical* across the following comparisons.

The first three comparisons in Table C.1 (top panel) focus on subjects who participated in midday sessions. This limits heterogeneity in time-of-the-day. We first apply *JNARP* to subsamples from $\{W20\}$ and $\{W50\}$. We find 67 violations of *JNARP*, which corresponds to 11 percent of the total number of pairwise comparisons between observations from $\{W20\}$ and $\{W50\}$. Given that this analysis kept experimental taste shifters and timing constant, we take this result as our benchmark. We then apply *JNARP* to treatments $\{W20I\}$ and $\{W50\}$. Treatment $\{W20I\}$ raises the income path from $I_{\{W20\}} = I_{\{W50\}} = 900$ to $I_{\{W20I\}} = 1500$ tokens. We find 78 violations of *JNARP* or, equivalently, 13 percent of the pairwise tests reject preference homogeneity. These numbers are relatively close to the benchmark, which kept the income path fixed. Thus we cannot reject the (seemingly restrictive) assumption that $F(\nu|I, M, \tau) = F(\nu|M, \tau)$. We finally apply *JNARP* to treatments $\{W20\}$ and $\{W50M\}$. Treatment $\{W50M\}$ increases cumulative working time from $M_{\{W20\}} = M_{\{W50\}} = 15$ to $M_{\{W50M\}} = 24$ minutes. We find 166 violations; *JNARP* is violated in 18 percent of the tests. This reflects a 50% increase relative to the benchmark results. Variation in the amount of cumulative work leads to more severe rejections of preference homogeneity. This suggests that $F(\nu|I, M, \tau) \neq F(\nu|I, \tau)$.

The final three analyses in Table C.1 (bottom panel) compare individuals who joined midday sessions with those who joined morning or late afternoon sessions. We compare subsamples characterized by high wages but midday work (i.e., sessions of treatments

[33] For instance, 31 subjects participated to $\{W20\}$ and 19 subjects participated to $\{W50\}$ in midday sessions. The hypothetical maximal number of violations is $31 \times 19 = 589$.

[34] To better balance earnings between treatments, we let fixed income Y decrease correspondingly.

$\{W50\}$, $\{W50I\}$, or $\{W50M\}$ implemented between 11h00 and 15h00) with subsamples characterized by low wages but work outside midday (i.e., sessions of treatments $\{W20\}$, $\{W20I\}$, or $\{W20M\}$ implemented before 11h00 or after 15h00). The percentage of $JNARP$ violations increases from 11% to 15% (conditional on I and M), from 13% to 14% (conditional on M), and finally from 18% to 23% (conditional on I). We infer from this that $F(\nu|I, M, \tau) \neq F(\nu|I, M)$. The main takeaway from Table C.1 is that the combined effect of variation in cumulative working time (disutility spillovers) and variation in time-of-the-day (timing-specific preferences) *doubles* the percentage of tests in which preference homogeneity is rejected, from 11% to 23%. This makes a compelling case for theories of time nonseparable preferences.

Limited variation in time-of-the-day				
Subsample 1:	treatment	$\{W20\}$	$\{W20I\}$	$\{W20\}$
	timing	$\tau = 1$	$\tau = 1$	$\tau = 1$
Subsample 2:	treatment	$\{W50\}$	$\{W50\}$	$\{W50M\}$
	timing	$\tau = 1$	$\tau = 1$	$\tau = 1$
Total nr of comparisons		589	589	930
Nr of $JNARP$ violations		67	78	166
% of $JNARP$ violations		11.38	13.24	17.85
Maximal variation in time-of-the-day				
Subsample 1:	treatment	$\{W20\}$	$\{W20I\}$	$\{W20\}$
	timing	$\tau = 0$	$\tau = 0$	$\tau = 0$
Subsample 2:	treatment	$\{W50\}$	$\{W50\}$	$\{W50M\}$
	timing	$\tau = 1$	$\tau = 1$	$\tau = 1$
Total nr of comparisons		361	266	570
Nr of $JNARP$ violations		55	39	133
% of $JNARP$ violations		15.24	14.66	23.33

Table C.1 – Frequency of pairwise $JNARP$ violations between observations from different subsamples. Subsamples are partitions of the data with specific treatment conditions and a specific timing of sessions

C.2 Normality: $WARP$ versus $JNARP$

WARP. In a two-goods setting, $JNARP$ is necessary and sufficient for the existence of continuous and $WARP$ consistent demand functions **that are also normal**. Below we

define the Weak Axiom of Revealed Preference (*WARP*) for finite data sets.

Definition 2 (WARP)**.** *Consider a pair of bundles* (l_i, y_i) *from subsample* $\{s\}$ *and* (l_j, y_j) *from subsample* $\{v\}$*. The bundles satisfy the Weak Axiom of Revealed Preference if*

$$W_{\{s\}} \times (l_i - l_j) + (y_i - y_j) \geq 0 \text{ implies } W_{\{v\}} \times (l_j - l_i) + (y_j - y_i) < 0.$$

Assume towards a contradiction that $W_{\{s\}} \times (l_i - l_j) + (y_i - y_j) \geq 0$ and $W_{\{v\}} \times (l_j - l_i) + (y_j - y_i) \geq 0$. Agent i chose bundle (l_i, y_i) while (l_j, y_j) was affordable; thus (strictly) preferring (l_i, y_i) over (l_j, y_j). At the same time, agent j selected (l_j, y_j) while (l_i, y_i) was affordable; thus (strictly) preferring (l_j, y_j) over (l_i, y_i). We conclude from this that i and j have distinct preferences. *WARP* consistency is a necessary requirement for the existence of a utility function and a corresponding demand function that rationalize the data.[35]

Figure C.1 presents the budget lines associated with $\{W20\}$, $\{W50\}$, and $\{W80\}$ graphically. Every line has a slope $-W_{\{s\}}$ and an intercept $(15, Y_{\{s\}})$ with the vertical axis on the right. Crossing budget lines make it possible to detect violations of *WARP*.

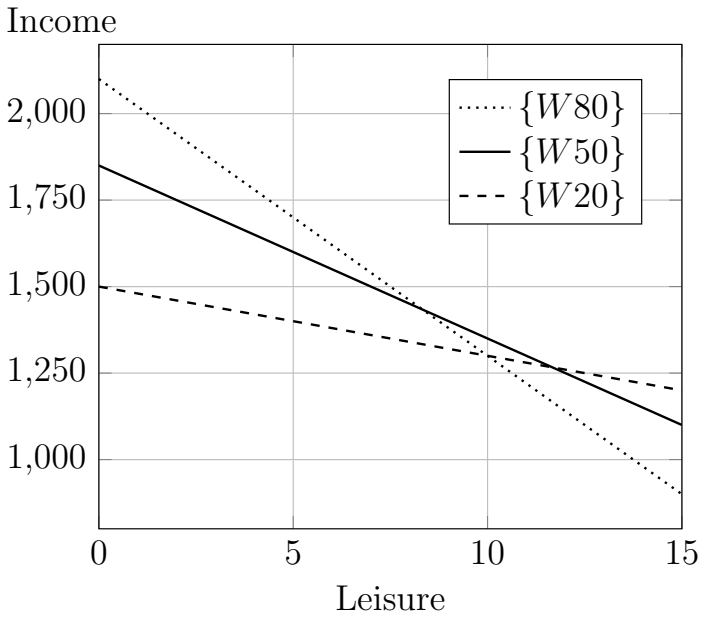

Figure C.1 – Graphical representation of budget lines associated with $(W, Y) = (20, 1200)$; associated with $(W, Y) = (50, 1100)$; and associated with $(W, Y) = (80, 900)$

Normality. In addition to *WARP* consistency, we also assumed that the demands for leisure and income are *normal*. The normality assumption improves the empirical bite of

[35] In a setting with just two goods, the conditions are moreover sufficient for the existence of a utility function that rationalizes the data.

the revealed preference conditions, but it could be violated in practice. One may worry that this assumption is (partly) responsible for the lack of support for theories of reference dependent preferences.

In the following exercise, we verify that this lack of support still holds *without* the normality assumption. To this end, we use the time choices in subsamples based on treatment $\{W80\}$. $\{W80\}$ is similar to $\{W20\}$ and $\{W50\}$ with respect to the income path and cumulative working time, but wages in $\{W80\}$ are higher. In an analogous way as before, we compute the number of pairwise violations between subsamples based on $\{W80\}$ and $\{W20\}$, between subsamples based on $\{W80\}$ and $\{W20I\}$, between subsamples based on $\{W80\}$ and $\{W50\}$, and finally between subsamples based on $\{W80\}$ and $\{W50I\}$. We do the analysis with the normality assumption ($JNARP$) and without ($WARP$). The results are in Table C.2. As expected, the number of $JNARP$ violations is higher than the number of $WARP$ violations. Still, many pairs of individuals violate $WARP$. The main takeaway is that the results of $WARP$ and $JNARP$ all go in the same direction: the percentage of violations is not very sensitive to variation in I. This confirms our earlier result that there is little empirical evidence of income targeting in our set-up.

	Treatments			
	$\{W80\}$ $\{W20\}$	$\{W80\}$ $\{W20I\}$	$\{W80\}$ $\{W50\}$	$\{W80\}$ $\{W50I\}$
Total nr of comparisons	682	682	418	836
Nr of $WARP$ violations	92	112	64	120
% of $WARP$ violations	13.49	16.42	15.31	14.35
Nr of $JNARP$ violations	110	128	139	273
% of $JNARP$ violations	16.13	18.77	33.25	32.66

Table C.2 – Frequency of pairwise $WARP$ and $JNARP$ violations between observations from different treatments. $\{W80\}$ is characterized by $(W, Y) = (80, 900)$, $I = 900$, and $M = 5$

D Additional results

In this appendix, we first discuss the robustness of our findings with respect to individual productivity and loss aversion in more detail. Table D.1 contains the full set of regression results. We then perform textual analysis of the subjects' stated preferences.

Robustness. First, some individuals may be more productive than others, e.g., they are better at the experimental task. Intertemporal disutility spillovers and fatigue may be affected not only by (exogenous) cumulative working time or time-of-the-day but also by the level of effort subjects exerted per minute. To mitigate the latter source of variation, we normalized the effort requirement per minute by the subject's initial productivity in the first part of the experiment. Denote the number of correctly counted tables in the productivity elicitation exercise by *InProd*. We first plot *InProd* by time-of-the-day in Figure D.1. The patterns are less outspoken compared to Figure 3 from the main text but we do see a post-lunch, early-afternoon dip in productivity. To assess the effectiveness of our normalization procedure, we test whether the subject's working time choice in round 4 is independent of her initial productivity *InProd*. We include *InProd* as an additional control variable in our leisure regressions. This does not change our earlier findings; moreover, the coefficients of *InProd* are all insignificant.

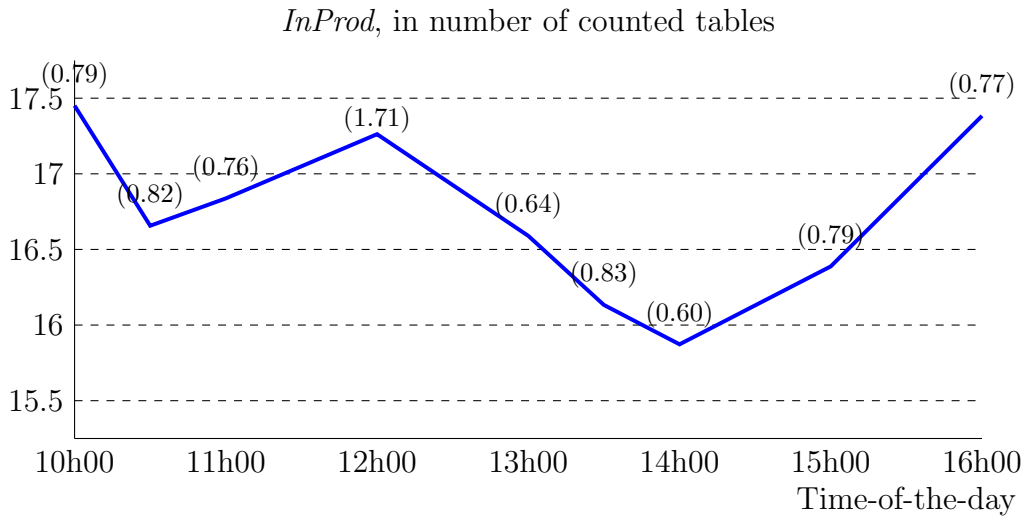

Figure D.1 – Plot of predicted productivity in function of time-of-the-day at which sessions started. Productivity prediction based on a fractional polynomial regression of degree 5. Standard errors of means between brackets.

Second, we study the relationship between loss aversion and labor supply. As indicated earlier, loss aversion is a crucial element in models of reference dependent preferences. In theory, it makes individuals work more when earnings are below the reference point, and less when earnings are above the reference point. Subjects characterized by higher degrees of loss aversion can be expected to respond stronger to variation in reference

points. To check this, we elicited each individual's degree of loss aversion from a series of six lottery choices (see above). The least risky Lottery was rejected by only 5% of the subjects; the most risky Lottery was rejected by 84%. Following Fehr and Goette (2007), Gächter et al. (2007), and Abeler et al. (2011), we define each subject's *LossAversion* in terms of the number of lotteries she rejected. We repeat our earlier multivariate analyses with *LossAversion* as an additional regressor. We find that budget constraint parameters and the midday indicator, but not the income path or cumulative working time, vary with leisure. The coefficients of *LossAversion* are all insignificant. Overall, the estimates confirm that time nonseparable preferences outweigh reference dependent preferences in our data.

	(1) Leisure b/se	(2) Leisure b/se	(3) Leisure b/se	(4) Leisure b/se	(5) Leisure b/se	(6) Leisure b/se	(7) Leisure b/se	(8) Leisure b/se	(9) Leisure b/se	(10) Leisure b/se
female	-0.740	-0.654	-0.647	-0.817	-0.883	-0.715	-0.628	-0.623	-0.803	-0.871
	(0.62)	(0.63)	(0.63)	(0.62)	(0.61)	(0.63)	(0.64)	(0.64)	(0.63)	(0.62)
age	0.030	0.036	0.028	0.050	0.033	0.032	0.039	0.029	0.051	0.033
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Luxembourg	-0.069	-0.033	-0.021	-0.812	-0.779	-0.069	-0.031	-0.020	-0.813	-0.780
	(0.64)	(0.65)	(0.65)	(0.69)	(0.68)	(0.64)	(0.65)	(0.65)	(0.69)	(0.69)
inprod	-0.014	-0.018	-0.015	-0.010	-0.004					
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)					
$W = 20$	1.917***				1.852***	1.920***				1.853***
	(0.60)				(0.61)	(0.60)				(0.61)
IncomePath		-0.000			-0.001		-0.000			-0.001
		(0.00)			(0.00)		(0.00)			(0.00)
CumulWork			0.058		0.015			0.059		0.015
			(0.05)		(0.06)			(0.05)		(0.06)
Midday				1.940***	1.869***				1.944***	1.870***
				(0.68)	(0.68)				(0.68)	(0.68)
lossaversion						-0.033	-0.032	-0.029	-0.016	-0.019
						(0.24)	(0.25)	(0.25)	(0.24)	(0.24)
constant	3.021	4.161	2.843	2.723	2.931	2.805	3.882	2.593	2.548	2.893
	(2.19)	(2.66)	(2.37)	(2.22)	(3.10)	(1.86)	(2.42)	(2.04)	(1.89)	(2.89)
R-sqr	0.041	0.007	0.011	0.034	0.069	0.041	0.006	0.010	0.034	0.069

* p<0.1, ** p<0.05, *** p<0.01

Table D.1 – Regressions of leisure based on age, gender, budget constraint parameters, the income path, cumulative working time, timing (Midday) and location (Luxembourg) of sessions, *InProd*, and/or *LossAversion*. Regressions (1)–(5) control for *InProd*; Regressions (6)–(10) control for *LossAversion*

Textual analysis of subjects' motivation. We end this section with an exploratory overview of the subjects' *stated* preferences. After round 4, we asked the following two questions: 'Would you have liked to work more?' and 'Why/Why not?'. The first question had only two options: *yes* or *no*. The correlation between this response and the observed leisure choice is negative (less than -0.5) and statistically significant. More importantly, we notice that virtually all subjects who chose leisure $l < 15$ answered *yes* while all subjects who chose leisure $l = 15$ answered *no*. This means that the answer to the first question matches the subject's labor supply choice at the extensive margin (stay in the lab versus exit immediately). In this case, follow-up question 'Why/Why not?' reveals

the motivation underlying the subject's extensive margin choice. This offers insight into the individual reasons for working—at least some minutes—overtime. Moreover, given the strong correlation between the leisure choice at the intensive and extensive margins, the second question has the potential to reveal general information about the subject's attitude towards staying longer in the lab.

The aim is to complement our formal analysis with qualitative information on the individual-specific motivations behind time choices. These data are unstructured by nature and although no hypotheses were formulated regarding the content of the answers, some meaningful patterns/insights can be retrieved from it. By applying text mining tools, it is possible to establish functional relationships between the text and the subjects; we then categorize words or sentences with respect to our treatment categories. To do so, we coded all subjects' responses and applied text mining tools for lemmatization and word counting.[36] This allows us to compare the (relative) frequency of each keyword in different treatments.

We first find that the words 'tokens' and 'gains' appear more frequently in $\{W80\}$ compared to $\{W20\}$. Luxembourgish subjects in $\{W80\}$ expressed for instance their desire 'to gain more tokens and have an another chance', 'to win more tokens', or '[to] increase the chances of gain'. We also find that keywords 'count' and 'too' appear more frequently in $\{W20M\}$. 'count' was sometimes used in combination with 'distracting' ('I find it distracti[ng] to count the ones in a table'). 'too' was used in combination with time ('too much time', but also 'too boring' and 'too little money for extra time and effort'). This supports the idea that motivation is especially low when wages are low and when time factors are not favorable. The textual analysis does not reveal large differences in keywords between treatments with different income paths. Individuals' stated motivations are therefore not inconsistent with the results from our choice experiment. Here is the list of keywords as they appear in subjects' responses:

- in $\{W20M\}$: **count** and **too** (good to count which makes me faster than before / only one round of the fourth game will count. I did not want to waste too much time on something which may not count in the end / to motivate myself and check my ability in reading and counting / Because its too much time for less earning / Too boring and too little money for extra time and effort gains)

- in $\{W80\}$: **gain** and **token (**to gain more tokens and have an another chance / I do have the time and want to increase the chances of gain resp. decrease the chance of loss / for the chance of gaining more / Because even with the small chances to get the money from it, I wanted to reach 1000 tokens by playing 2 minutes more. / chance of getting more tokens / Because I really liked the game and also wanted to win more tokens.)

[36] This was done using the Stata's command txttool.

The textual analysis also allows us to assess the validity of our operationalization of the reference income as 'income path' in the experiment. As we explain in the introduction, to our knowledge, no studies have attempted to implement income reference points in a way similar to ours. The question is then whether subjects effectively internalized the gains in the first three periods (the historical income path) as possible point of comparison. The textual analysis of answers to the question *why* subjects stayed in the lab offers comforting albeit still anecdotal evidence. Several individuals referred in their stated motivations to their gains from initial rounds of the experiment. This confirms that the generated historical income path is, in principle, a salient and relevant point of comparison. Some subjects even explicitly stated their willingness to reach the same gains as in initial rounds. This is particularly true in treatments $\{W20I\}$ and $\{W50I\}$. Here are some examples of subjects' answers:

- To get at the end the same potential amount I was able to get in the first three rounds
- TO EQUALIZE THE RESULTS TO OTHER ROUNDS
- to make some extra bucks and equilibrate the winnings in each round