

# Pessimism and Overcommitment\*

Claes Ek<sup>†</sup>

Margaret Samahita<sup>‡</sup>

August 20, 2020

## Abstract

Commitment devices are often proposed as solutions to self-control problems, but little attention has been given to the possibility of excess demand for commitment. We provide evidence for overcommitment in a laboratory experiment. Subjects face a tedious productivity task and a tempting option to surf the internet. After stating their willingness-to-pay for a commitment device that removes the option to surf, subjects are allocated commitment with some probability, thus allowing us to observe the behavior of subjects who demand commitment but still face temptation. We find that a significant share of subjects overestimate their commitment demand, as indicated by their willingness-to-pay, when compared to their realized material and psychological costs from facing the temptation. This overcommitment appears driven by pessimism in predicted performance under temptation. Although we find that undercommitment reduces welfare more than overcommitment, our results nevertheless suggest a need to consider the possibility that commitment could be harmful for some.

*JEL classification:* C91, D03, D91.

*Keywords:* Commitment device, pessimism, self-control, willpower.

---

\*We thank Giuseppe Attanasi, Pol Campos-Mercade, Christina Gravert, Olof Johansson-Stenman, Leonhard Lades, Lisa Norrgren, Séverine Toussaert and participants at the SABE ECR Workshop (Dublin), UCD Behavioural Science and Policy seminar, University of Gothenburg Behavioural and Experimental Economics seminar and the ESA Junior Faculty webinar for helpful comments and suggestions. We also thank Miloš Fišar and the team at Masaryk University Experimental Economics Laboratory (MUEEL) for their assistance during the experimental sessions. Financial support from the Torsten Söderberg Foundation [grant number E25/18] is gratefully acknowledged. Our analysis plan is pre-registered at <https://osf.io/rn8uc/>. In accordance with the Swedish Ethical Review Act (SFS 2003:460), which applies in the country where the authors' institutions were located, no IRB approval was required since no sensitive personal data was processed and the experiment did not involve a risk of harming subjects physically or psychologically. Informed consent was obtained for experimentation with human subjects. Declarations of interest: none.

<sup>†</sup>Department of Economics, University of Gothenburg. E-mail: [claes.ek@economics.gu.se](mailto:claes.ek@economics.gu.se)

<sup>‡</sup>School of Economics, University College Dublin. E-mail: [m.samahita@gmail.com](mailto:m.samahita@gmail.com)

# 1 Introduction

Commitment devices have often been proposed as a solution to self-control problems in a variety of settings.<sup>1</sup> The existing literature has largely focused on the underdemand of commitment, i.e. overoptimistic decision makers not demanding sufficient levels of commitment to completely eliminate the self-control problem (DellaVigna and Malmendier, 2006; Heidhues and Köszegi, 2009; John, 2020; Bai et al., forth.); the policy implication being that commitment take-up should increase (for example, Acland and Chow, 2018; Sadoff and Samek, 2019). However, recent experimental research suggests that, possibly driven by errors or demand effects (Carrera et al., 2019), commitment demand is often present among those who do not seem to need it. Royer et al. (2015) find that agents with a high frequency of pre-study gym visits are relatively more likely to demand commitment to go to the gym. Similarly, in a home grocery delivery program, Sadoff et al. (2020) show that commitment demand is higher amongst agents who are less likely to exchange healthy items for unhealthy items. This implies that “overcommitment” could be a real problem among pessimistic decision makers, and suggests that increased commitment take-up may sometimes be detrimental.

However, although such findings are suggestive, they cannot be taken as direct proof of overcommitment. These studies use either costless (Sadoff et al., 2020) or refundable commitment (Royer et al., 2015) and do not determine whether observed commitment demand is truly excessive and thus irrational, leading to an *ex-post* utility loss to the decision maker. In this paper, therefore, we present a laboratory experiment designed to provide exactly such clear evidence. Our design allows us to estimate, within a single sample, the existence and magnitude of both excess and insufficient commitment demand. We are also able to directly compare the severity of each error, and thus infer whether or not overcommitment is an economically significant phenomenon on par with undercommitment.

In our experiment, the commitment device removes a tempting option to surf the internet during a tedious productivity task. Subjects state their willingness-to-pay (WTP) for this non-refundable commitment device, and commitment is then allocated with some

---

<sup>1</sup>For example, smoking cessation (Giné et al., 2010), grocery shopping (Schwartz et al., 2014; Sadoff et al., 2020), gym attendance (Milkman et al., 2013; Royer et al., 2015; Carrera et al., 2019), alcohol consumption (Schilbach, 2019), gaming (Acland and Chow, 2018), savings (Ashraf et al., 2006; Beshears et al., 2015; John, 2020) and work tasks (Kaur et al., 2010; Augenblick et al., 2015; Toussaert, 2018; Houser et al., 2018).

probability, thus allowing us in an incentive-compatible way to observe the behavior of subjects who demand commitment but have to face temptation. We also elicit subjective beliefs regarding expected productivity, as well as the *ex-ante* expected and *ex-post* actual experienced difficulty of resisting temptation. These measures allow us to decompose subjects' valuation of the commitment device into expected material loss from being exposed to temptation and any non-material 'psychological costs', such as the mental burden of maintaining self-control while facing temptation.

The take-up rate for commitment in our experiment, as measured by the proportion stating positive WTP, is modest at 27%. This is in line with previous studies where, with only a few exceptions (Schilbach, 2019; Beshears et al., 2015), agents who demand commitment are typically in the minority.<sup>2</sup> For example, the share of participants demanding commitment is 28% in Ashraf et al. (2006) and 25% in Acland and Chow (2018). Since the allocation of commitment incorporates a random implementation rule, all but 12 of our 289 participants still have to face temptation. However, of these 277 exposed subjects, only four actually succumb and surf the internet.

We find that 22% of exposed subjects overestimate their demand for commitment when compared to their actual material loss from facing the temptation. By the same measure, and perhaps surprisingly, only around 6% act according to standard accounts of naïve decision makers and underestimate their demand for commitment. Moreover, this fraction is significantly lower than the above proportion of overestimating subjects. We obtain similar results when we comparing against both material and psychological costs of temptation: 15% of subjects seem to state a WTP higher than seems justified from the sum of these costs. By the same measure, only 7% of subjects underestimate the material and psychological costs of facing temptation and understate their WTP. On the other hand, the summed welfare loss from undercommitment is about double that from overcommitment. Thus, although overcommitment is the more widespread phenomenon in our sample, undercommitment is relatively more severe, seemingly vindicating the traditional focus on the latter bias in the literature.

We also explore drivers of overcommitment, framing our results in relation to models

---

<sup>2</sup>Given the extent of self-control problems in domains such as personal finance and health, this level of commitment take-up may be consistent with underdemand by naïve present-biased agents. On the other hand, Laibson (2015) shows that in the presence of commitment cost, commitment demand may be irrational even for these agents – suggesting the observed level of take-up, while low, may in fact be excessive and consistent with overcommitment.

often used to rationalize commitment demand (see Bryan et al. (2010) for a review), specifically models of present-biased preferences (Strotz, 1955; Laibson, 1997; O'Donoghue and Rabin, 1999), random indulgence (Chatterjee and Krishna, 2009; Dekel and Lipman, 2012), and costly self-control (Gul and Pesendorfer, 2001). In general, it is not clear from the literature whether subjects (over)commit because they expect they may succumb to temptation if exposed (as in the former pair of models), or because they expect disutility simply from having to exercise self-control (as in the latter). For example, in the recent lab experiment by Toussaert (2018), up to a third of subjects preferred removing a tempting option from the full choice set, but also preferred the full choice set to the tempting option by itself, suggesting they expected not to pick the tempting option when exposed.

In our data, overcommitment appears driven mainly by considerations related to costly self-control: WTP overestimators tend to be too pessimistic about the cost of exercising self-control in the face of temptation, but not the success thereof, as captured by the low predicted likelihood of succumbing. The fact that overcommitment is systematically related to pessimism also indicates that it is not simply driven by random error on the part of subjects. This contrasts with, for example, the field experiment of Carrera et al. (2019). Among members of a fitness facility, the authors find that about half of those who demand commitment for more gym visits also demand it for fewer visits, suggesting commitment demand is driven by errors and demand effects. While the setting considered by Carrera et al. (2019) involves uncertainty about the future, we use a transparent laboratory setting where subjects can familiarize themselves with the task before making choices.

We describe the experimental setting in Section 2 and derive our hypotheses in Section 3. The results are presented in Section 4 and Section 5 concludes.

## 2 Experimental Design

The experiment was conducted at Masaryk University Experimental Economics Laboratory (MUEEL) in Brno, Czech Republic during the period 27-30 May 2019 and programmed using z-Tree (Fischbacher, 2007). Participants were recruited from the laboratory subject pool consisting of students at Masaryk University, using hroot (Bock et al., 2014). In total we ran 12 sessions with 289 subjects. Each session lasted around 2 hours and average

earnings were CZK 707 per subject, including CZK 100 participation fee.<sup>3</sup>

The experiment consists of two stages and follows the sequence shown in Figure 1. In Stage 1, subjects complete an attention task without temptation (Task 1). At the start of Stage 2, subjects learn that they will complete the same task again but with temptation (Task 2). They are told about the possibility to purchase a commitment device that removes the temptation. We elicit their WTP for this commitment device, using the incentive-compatible Becker-DeGroot-Marschak (BDM) mechanism, and various beliefs about productivity in the second attention task. Subsequently subjects complete the second attention task, followed by an exit survey. Subjects are informed before Task 1 that one of the two tasks will be randomly chosen for payment. This randomization is performed after Task 2. The use of within-subject design is motivated by wanting all subjects to indicate their WTP for the commitment device with all information about the attention task available to them, including their own experience of it. Instructions and screenshots are included in Online Appendix F and G.

## 2.1 The attention task, temptation, and commitment device

We use an attention task similar to that used in Toussaert (2018): for a period of up to 30 minutes, subjects are asked to pay attention to their computer screen where a four-digit number increments every third second. At five random (subject-specific) times, they are prompted to enter the last number they saw, after which the number is reinitialized; the last prompt always occurs after the 25-minute mark, ending the task. Subjects cannot access the internet after having secured all 5 answers even if they end before 30 minutes. Each correct answer is worth 120 tokens per question, where 1 token corresponds to CZK 1. The potential earnings from this task are set to be relatively high to induce subjects to be interested in completing the task.

During this attention task, subjects are not allowed to do any other activity, including checking their phone or studying. All personal belongings have to be put away and subjects caught doing something else forfeit their experimental payment. In our sessions, the experimenter or lab manager walks around the room at random times to check that no phone is in sight. Therefore, we are highly confident of subjects' expectation of the enforcement of the rules, and thus of their compliance during the task.

---

<sup>3</sup>CZK 1 corresponded to EUR 0.039 at the time of the experiment. The Czech minimum wage is CZK 13,000, or EUR 500, per month.

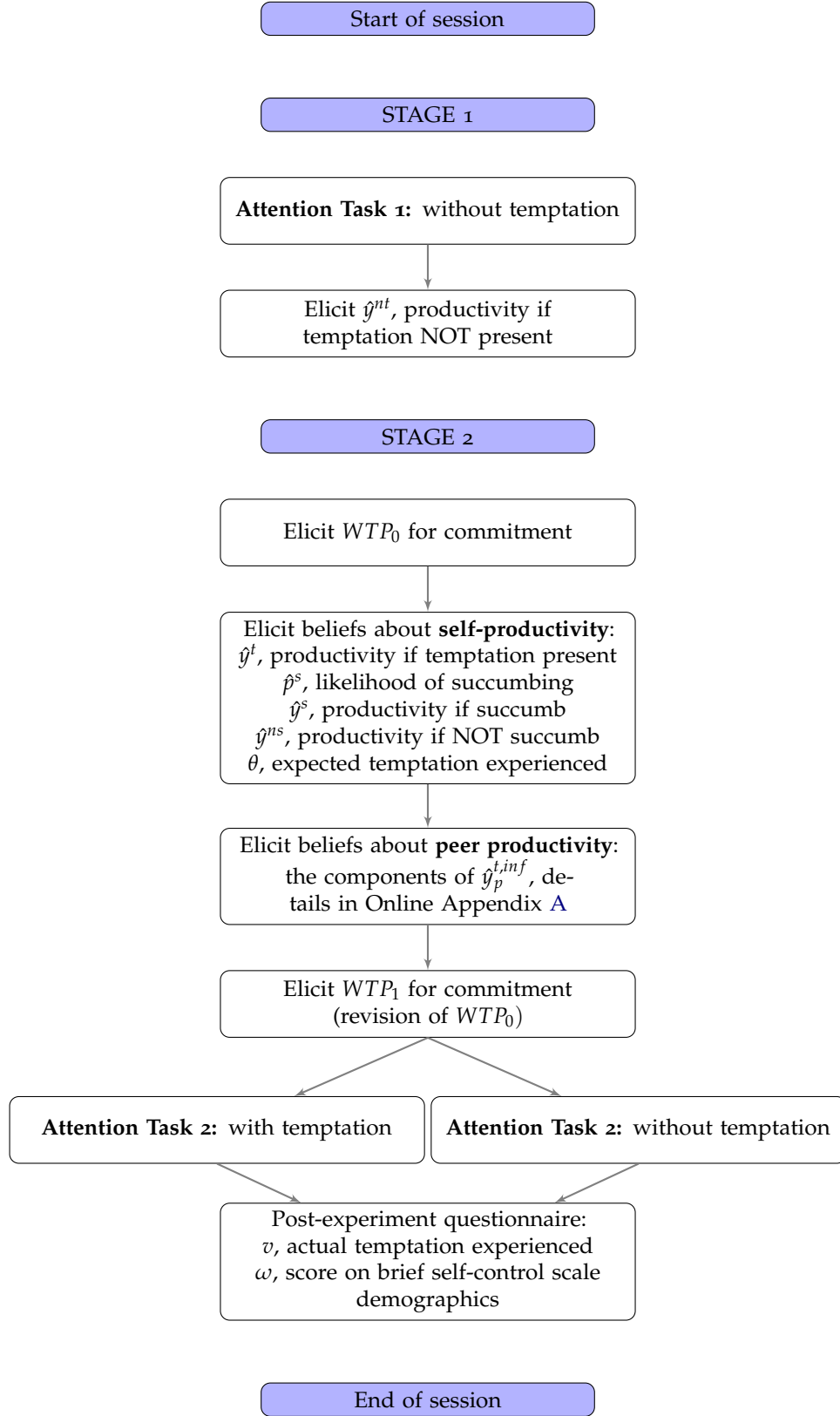


Figure 1: Experiment timeline

In Stage 1 subjects complete the attention task as described above (Task 1) and are given feedback about their performance. At the start of Stage 2, subjects are informed that they will do the attention task again (Task 2), but this time there will be an additional button on the screen which allows internet access. Subjects are shown a screenshot of the temptation in the paper instructions. Clicking the internet access button means that the subject surfs the internet for the remainder of the period instead of continuing with the attention task. The subject will forfeit the chance to earn any more money from the attention task, but will retain any money earned from correct answers up until the point of clicking the internet access button. Subjects are thus aware that to get the highest possible monetary payoff they would have to exercise willpower to overcome the temptation. This temptation, as also used in [Houser et al. \(2018\)](#) and [Bonein and Denant-Boèmont \(2015\)](#), is highly familiar to subjects, which should imply more accurate WTP estimates. It also has immediate appeal given the tedious attention task and lack of other distraction, and should be perceived to be bad since subjects choosing this option forfeit the possibility to earn more money from the attention task. As revealed in subjects' feedback elicited at the end of the experiment, they appear to consider the internet as a temptation to be avoided.

We then offer subjects a commitment device: the possibility of paying to remove the internet access button which would guarantee participation in Task 2 for the whole 30-minute period. WTP for the commitment device is elicited using the BDM method. Subjects state a price between 0 and 100 tokens representing the maximum price they are willing to pay to remove the option to surf. The computer will then simulate a coin toss. If Heads comes up, the internet button continues to be present regardless of the subject's WTP. Only if Tails comes up will the WTP be taken into account. The computer will draw a random number  $R$  between 0 and 100 and if this number is less than or equal to the stated price then the internet option will be removed and the subject pays the amount  $R$ . Hence the probability of getting the commitment device to remove the option to surf is equal to  $WTP/200$  and increases linearly with subject WTP, up to a value of 0.5.<sup>4</sup> Subjects are told that the possibility of successfully removing the internet button increases the higher their stated price. In particular, subjects are informed that this possibility is maximized (but not guaranteed) if they state a price of 100, and that they should enter a

---

<sup>4</sup>The use of a random implementation rule is motivated by the following objectives: to ensure incentive-compatibility, to observe the performance under temptation of subjects who have a positive demand for commitment, and to maximize the number of subjects who in fact face temptation. See other uses in [Karlan and Zinman \(2009\)](#), [Augenblick et al. \(2015\)](#) and [Toussaert \(2018\)](#).



price of 0 if they are not willing to pay anything to remove the button.

WTP is elicited twice. The initial measure, denoted  $WTP_0$ , is not used in our main analysis. The second elicitation (denoted  $WTP_1$ ) occurs after subjects have been asked to reflect on their own productivity (as will be described in the next section), which may promote more accurate preferences.  $WTP_1$  will later be compared against various measures of subjects' productivity in the attention task.<sup>5</sup>

In designing the task, we carefully adjusted the level of difficulty to make the temptation suitably appealing. In particular, using a more difficult task would not necessarily make commitment seem more desirable. Indeed, beyond a certain point, subjects may simply give up and start perceiving temptation as a good: we found this to be the case in a pilot session where the four-digit number was made to update randomly every second. In any case, if the task is too easy (or too difficult), this should shift both actual *and* expected performance under temptation, thus having no net effect on the likelihood of WTP overestimation, which is the focus of our experiment.

## 2.2 Measures of subjects' productivity beliefs

The subject's actual productivity is measured as  $y_1$  and  $y_2$ , the number of correct answers in Tasks 1 and 2 respectively. We also elicit unincentivized subject beliefs about their productivity. Directly after the conclusion of Task 1 and after subjects are told about their performance  $y_1$ , we ask subjects how many questions they expect to answer correctly if they were to redo Task 1, where there was *no* temptation ( $\hat{y}^{nt}$ );<sup>6</sup> this is followed by the elicitation of  $WTP_0$ . Next, we ask how many questions they expect to answer correctly if the temptation is present in Task 2 ( $\hat{y}^t$ ). We also ask subjects to state the percentage likelihood that they would succumb to temptation, if present ( $\hat{p}^s$ ), as well as how many questions they expect would be answered correctly if they succumb ( $\hat{y}^s$ ) or do *not* succumb

---

<sup>5</sup>The distribution of  $WTP_1$  is not significantly different from that of  $WTP_0$ , as confirmed by a Kolmogorov-Smirnov test. Our results are robust to using  $WTP_0$  instead of  $WTP_1$ . In Appendix B we use both  $WTP_0$  and  $WTP_1$  to demonstrate the robustness of our results to potential measurement errors using the approach in [Gillen et al. \(2019\)](#).

<sup>6</sup>We consistently apply 'hat notation' to all belief variables elicited in the experiment.



$(\hat{y}^{ns})$ .<sup>7,8</sup> These measures allow us to check for the presence and source of misestimation of performance. We do not incentivize these measures of self-productivity to prevent subjects hedging with their stated beliefs against adverse performance in Task 2.<sup>9</sup>

After the above variables have been measured, we repeat the WTP elicitation, framing it as an option to revise the amount stated previously. We explain that this second measure ( $WTP_1$ , henceforth  $WTP$ ) will be the one that determines whether the internet button is present in Task 2.

### 2.3 Other variables

To better understand subjects' estimation of the psychological cost of temptation, we elicit a measure of expected temptation strength before Task 2 with the question "How tempted do you think you would be by internet access?". Subjects respond on a scale from 1 to 4 (not at all tempted, not that tempted, quite tempted, very tempted), yielding  $\theta$ . Subjects' actual experience of the temptation is elicited in the post-experiment questionnaire, where those who did face temptation are asked to respond whether they think ignoring the internet button was easier, more difficult, or neither easier nor more difficult than expected. We derive the variable  $v$ , representing actual temptation, and set it equal to 1 for those who respond that ignoring internet access was easier than expected, that is, who overestimated the psychological cost of temptation. We set  $v$  equal to -1 for whom the temptation was more difficult than expected to ignore, and 0 otherwise.

In the post-experiment questionnaire we also elicit subjects' perception of their willpower using the brief self-control measure (Tangney et al., 2004). This question-

---

<sup>7</sup>There are some potential concerns here. Although we ask subjects to state a single expected value, it may be that they instead consider a distribution of cases, across which they are risk-averse; or report some other statistic, such as a modal outcome; or round values up or down. Such 'behind-the-scenes' considerations may imply that their behavior will appear less rational than it actually is. In this paper, we make the assumption that subjects do state a single expected value when asked to do so, and that this expected value can be treated as certain conditional on the outcome of the coin toss and the BDM mechanism (and, for  $\hat{y}^s$  and  $\hat{y}^{ns}$  in Online Appendix C, additionally whether or not the subject succumbs to temptation).

<sup>8</sup>These values allow us to construct a second, *inferred* measure of expected productivity under temptation,  $\hat{y}^{t,inf} = \hat{p}^s(\hat{y}^s) + (1 - \hat{p}^s)(\hat{y}^{ns})$ . However, since this measure is not different from  $\hat{y}^t$  on average among subjects who face temptation (4.61 vs. 4.56,  $p = 0.0696$ ), we do not use it in our analysis.

<sup>9</sup>In order to obtain a payment-contingent measure of productivity, we also elicit and incentivize subjects' belief of the performance of a similar peer who faces temptation ( $\hat{y}_p^{t,inf}$ ). See, for example, Gächter and Renner (2010) who find that incentivizing beliefs significantly improves accuracy. These measures are described in more detail in Online Appendix A, which also shows that results are similar when we use them in place of  $\hat{y}^t$ .

naire consists of 13 statements, to each of which the subject indicates their agreement on a five-point scale, reverse-coded where necessary (Cronbach's  $\alpha = 0.84$ ). Some example statements include: "I am good at resisting temptation" and "I often act without thinking through all the alternatives". These values are aggregated to give  $\omega$ , the perceived general level of willpower. We collect data on demographic variables such as age, gender, degree program (1 for Bachelor, 2 Master, 3 PhD), major, and GPA. Finally, we include an optional question asking subjects to comment on their choice of WTP in order to better understand their motivation.

### 3 Hypotheses

Our aim is to investigate whether a substantial share of subjects overstate their WTP to remove temptation. WTP can be decomposed as the sum of expected material losses due to productivity reduction, either from succumbing or purely from being exposed to temptation (e.g. if devoting cognitive resources to self-control reduces productivity in the task); and non-material psychological costs from facing temptation.

The decomposition of WTP can be shown within a simple expected-utility model. Recall that each task is paid with probability  $1/2$  and that the BDM outcome is only used with probability  $1/2$ . Conditional on the BDM being used, with probability  $WTP/100$  the random number  $R$  drawn by the computer is no larger than the stated price, in which case the internet button is removed and the subject pays the amount  $R$ . Note that conditional on the BDM being used,  $R$  is paid regardless of the task selected for payment. Although this is not made explicit, this interpretation should be clear from the instructions which specify that the randomization is performed across Task 1 and Task 2, while the BDM falls *after* Task 1 and *before* Task 2. In any case, we check for the possibility that subjects assume that  $R$  is paid only if Task 2 is chosen for payment. We explore this possibility in Online Appendix D.1, deriving modified predictions and showing that our main results are similar in the alternative setting.

We assume that Bernoulli utility takes as argument the sum of state-dependent costs and benefits, so  $u = u(x + PC)$ , with wealth  $x$  and psychological costs  $PC$ . We also treat the random number  $R$  as a continuous variable in  $[0, WTP]$ , for maximization purposes. Then, for some fixed (possibly accurate) expectations on earnings with and without

temptation ( $y^t, y^{nt}$ ), subjects are taken to maximize expected utility as

$$\begin{aligned}
U(WTP) = & \frac{1}{2} \left[ \frac{1}{2} \cdot u(100 + 120y_1 - PC) \right. \\
& + \frac{1}{2} \left( \frac{100 - WTP}{100} \cdot u(100 + 120y_1 - PC) + \frac{1}{100} \int_0^{WTP} u(100 + 120y_1 - R) dR \right) \Big] \\
& + \frac{1}{2} \left[ \frac{1}{2} \cdot u(100 + 120y^t - PC) \right. \\
& + \frac{1}{2} \left( \frac{100 - WTP}{100} \cdot u(100 + 120y^t - PC) + \frac{1}{100} \int_0^{WTP} u(100 + 120y^{nt} - R) dR \right) \Big] \quad (1)
\end{aligned}$$

The last line, for example, is the case when Task 2 is used for payment (with probability  $1/2$ ) and the BDM outcome is taken into account (with probability  $1/2$ ). Then, with probability  $(100 - WTP)/100$ , the subject does not get the commitment and payoff is given by the show-up fee of 100, plus 120 times the number of correct answers when exposed to temptation, less any psychological cost from the temptation. The final integral term means that with probability  $WTP/100$  the subject does get the commitment and payoff is given by the show-up fee, plus 120 times the number of correct answers when not exposed to temptation, less the payment of  $R$ , the random number drawn.

We solve for maximum WTP under this assumption of risk neutrality, which is reasonable given the small lab payments at stake ([Rabin, 2000](#)). Nevertheless, we also provide a robustness check assuming risk aversion in Online Appendix C. In any case, risk neutrality yields

$$WTP = 60(y^{nt} - y^t) + PC \quad (2)$$

The first term (expected material loss) is the subject's expected productivity without temptation less their expected productivity with temptation. The model is silent on whether any positive difference  $y^{nt} - y^t$  arises from succumbing to temptation or from performing worse due to having to exercise more self-control when the internet button is present, even if it is never clicked. For example, in an exploratory analysis of the effort task of [Toussaert \(2018\)](#), subjects who were exposed to temptation despite preferring to commit to a temptation-free choice set were found less productive, suggesting that self-control is indeed costly. In any case, these values  $y^{nt}$  and  $y^t$  are elicited in the experiment, as noted in Section 2.2.

The second term (expected psychological cost) may reflect (i) the effort of resisting

temptation, (ii) the (option) value of surfing the internet, and (iii) any self-image loss should the subject succumb to temptation. We do not measure the latter two components, although we note that if a subject places relatively great value on surfing the internet (the second component),  $PC$  may in principle be negative, lowering WTP compared to material losses.

The analysis proceeds with the following steps.

1. We begin by comparing WTP with **actual** performance in the attention tasks. As suggested above, subjects might overstate WTP relative to actual material loss because they (i) inaccurately estimate future material payoffs and/or (ii) anticipate (correctly or not) non-material psychological costs of facing temptation. Although the latter explanation remains consistent with ‘correct’ demand for commitment as given by equation (2), examining whether commitment exceeds material losses is arguably interesting in its own right.

Note that, for the sub-sample of subjects who obtain commitment, we are unable to infer realized material losses because we never observe how these subjects perform when exposed to temptation. Thus, these subjects cannot be used to compare WTP with material losses. By contrast, for the sub-sample of subjects who do face temptation in Task 2, we may infer material losses by comparing those subjects’ Task 2 performance with that in Task 1, where they did not face temptation. Thus, performance in Task 1 is effectively used as the counterfactual. Although we cannot check the validity of this approach directly, we will attempt to provide supporting evidence by checking whether performance differed across Task 1 and 2 among those who were never exposed to temptation, i.e. obtained commitment, to exclude concerns about learning or time effects.

We then classify each subject exposed to temptation according to whether they exhibit  $WTP > 60(y_1 - y_2)$  or  $WTP < 60(y_1 - y_2)$  or neither, and test the following null hypothesis. All tests of proportions are based on the standard normal approximation of binomial parameters and assuming that the share that can be attributed to subject confusion or demand effects is not more than 10%, an arbitrary but arguably conservative threshold picked given a lack of existing studies measuring such drivers in this setting.<sup>10</sup>

---

<sup>10</sup>De Quidt et al. (2018) suggest that typical demand effects are generally modest. The average demand

**Hypothesis 1.** *Among the subjects who face temptation in Task 2, no more than 10% have  $WTP > 60(y_1 - y_2)$ .*

If this hypothesis is rejected, we conclude that a ‘substantial’ share of subjects overestimate WTP compared to material losses.

2. Next, we compare WTP with **expected** performance in the attention tasks, as given by  $\hat{y}^t$ . We test whether or not WTP is overstated relative to the expected material loss, as captured by the difference between expected number of correct answers without temptation less the expected number when exposed to temptation:

**Hypothesis 2.** *Among the subjects who face temptation in Task 2, no more than 10% have  $WTP > 60(\hat{y}^{nt} - \hat{y}^t)$ .*

3. As noted, a WTP higher than expected material losses may reflect expected psychological costs as well as ‘true’ overestimation (of both material losses and psychological costs). Because the actual  $PC$  is unknown, we cannot test directly whether stated WTP is larger than the entire realized right-hand side of equation (2). However, an indirect test is possible. Starting from equation (2) and denoting expected quantities by subscript  $e$  and actual values (i.e. accurate expectations) by subscript  $a$ , true overestimation in the expected-utility model with risk neutrality would be characterized by

$$\begin{aligned} WTP(\cdot_e) > WTP(\cdot_a) &\iff 60(y_e^{nt} - y_e^t) + PC_e > 60(y_a^{nt} - y_a^t) + PC_a \\ &\iff 60((y_e^{nt} - y_e^t) - (y_a^{nt} - y_a^t)) > -(PC_e - PC_a) \end{aligned} \quad (3)$$

Thus the overestimation of material losses needs to exceed any *underestimation* of psychological costs. Our approach is thus to ask the subjects whether they think resisting temptation was easier than expected, yielding  $v$ . If  $v = 1$ , our interpretation

---

effect in that paper (0.13 SD) would *ex-post* only affect 5.8% of our subjects. Also, our setting involves little uncertainty regarding the task, and additionally WTP is elicited twice, which should minimize any confusion. Nevertheless, it is conceivable that confusion and experimenter demand would affect a larger share of subjects than 10%. Yet even then, a result that, say, 15% of subjects overstate WTP seems notable, especially since experimenter demand effects are not unlike the impact of a nudge designed to increase commitment take-up.

is that  $PC_e > PC_a$ ,<sup>11</sup> and the RHS of the inequality is negative. This in turn implies that a sufficient condition for overestimation is that the LHS is greater than or equal to 0. Furthermore, for subjects who find that ignoring the temptation was just as easy or difficult as expected ( $v = 0$ ), the RHS of the above inequality is taken to be zero, implying that a sufficient condition for overestimation is that the LHS is strictly greater than 0. Note that because these conditions are not necessary, they imply a lower bound on the number of overestimators.

Testing the pair of conditions on the sub-sample of subjects who are exposed to temptation, we set  $y_a^t = y_2$ ; however, we also need to choose an appropriate counterfactual  $y_a^{nt}$ . As in testing Hypothesis 1, we suggest to use  $y_a^{nt} = y_1$ , in which case the LHS gives  $(y_e^{nt} - y_e^t) - (y_1 - y_2)$ . Additionally, given that subjects cannot state negative WTP,  $WTP(\cdot_e) > 0$  is a necessary condition for strict overestimation.

**Hypothesis 3.** *Among the subjects who face temptation in Task 2, no more than 10% have  $(\hat{y}^{nt} - \hat{y}^t) - (y_1 - y_2) \geq 0$  and  $v \geq 0$ , with at least one strict inequality, and  $WTP > 0$ .*

Rejection of this hypothesis suggests that a substantial share of subjects have overestimated WTP. However, this holds only if it is valid to use  $y_1$  as counterfactual; recall that we evaluate this assumption by testing for time effects among the subjects who obtain commitment. Only if that test is not rejected while Hypothesis 3 is rejected, do we conclude that a substantial share of subjects have overestimated WTP.<sup>12</sup>

---

<sup>11</sup>This admittedly misses any misestimation of the net disutility of succumbing: self-image loss, which may be potentially large, less any utility from being able to surf. We ignore the possibility of misestimation in these dimensions for several reasons. First, the realized value of succumbing can be elicited only from subjects that do succumb, and as will be seen later, very few subjects in fact do so. Second, we expect selection bias in terms of who succumbs, making any attempt at extrapolation to the entire population problematic. Third, regarding the value of surfing the internet, this is a common activity for our subjects and they should have little trouble estimating its appeal.

<sup>12</sup>Complete results of the analysis performed according to our original pre-analysis plan are available at <https://osf.io/4dgm7/>. The current paper deviates from the original plan mainly in the following ways. First, we only report results using  $\hat{y}^t$ , instead of all three measures  $\hat{y}^t, \hat{y}^{inf,t}, \hat{y}_p^{inf,t}$ , for clarity. All results are robust to using all three measures. Second, we use  $y_1$  instead of  $\hat{y}^{nt}$  as the counterfactual in Hypothesis 3 since  $y_1$  is a more intuitive choice consistent with Hypothesis 1. Our results are robust to using  $\hat{y}^{nt}$  as a counterfactual in both Hypotheses 1 and 3. Third, the analyses from Section 4.2 onward, investigating the drivers of commitment demand, are not pre-registered. Nevertheless, we believe the topics covered are important to study.

## 4 Results

### 4.1 Main results

Summary statistics from the experiment are presented in Table 1. The majority of subjects have little willingness to pay for the commitment device, consistent with previous studies such as Augenblick et al. (2015). 211 subjects (73%) state  $WTP = 0$ . The average WTP is 6.94 tokens, or 25.73 for those with positive WTP. The distribution of positive WTP is given in Figure 2. In total, 12 subjects are successful in getting the temptation removed after stating a price greater than the random number drawn and given the BDM outcome is used.<sup>13</sup> Overall, subjects are successful in resisting temptation even without the commitment device, as only 4 subjects decide to access the internet. This figure is much smaller than, for example, the 18.4% who succumbed in Toussaert (2018) – it appears that subjects are sufficiently incentivized to complete the tedious (but easy) task: 87% of subjects get all 10 correct answers in Task 1 and Task 2 combined.

We start by comparing subjects' WTP for removing temptation with their *actual* productivity loss when exposed to temptation. As discussed in Section 3, we do so for the 277 subjects who do not obtain commitment. This is a selected sample that tends to exclude subjects with very high WTP and thus high probability of getting commitment. We address concerns about selection bias in two ways. First, we use just the half of the (exposed) sample where the coin flip decides against commitment, regardless of WTP. Second, we apply frequency weights with respect to WTP to the 277 subjects to account for the 12 “missing” subjects. Both methods do not change our conclusions.

As described in Hypothesis 1 in Section 3, we classify subjects as *material loss* (ML) overestimators, ML accurate estimators or ML underestimators according to whether their WTP is above, equal to, or below what would maximize utility when only material loss is considered (to distinguish it from the overestimator as defined in Hypothesis 3, which also takes into account psychological cost, and which will be used as a dummy variable in later analysis). An ML accurate estimator would state  $WTP = 60(y_1 - y_2)$ .<sup>14</sup> Table 2 summarizes these classifications. Around 72.2% of subjects accurately estimate their WTP

---

<sup>13</sup>Due to a coding error, it was possible to get the commitment device despite bidding 0 if the computer also drew the number 0. In our sample, a single subject received commitment in this way.

<sup>14</sup>ML accurate estimators include 11 subjects who state WTP equal to 0 but who have  $y_1 < y_2$ . According to equation (2) these subjects' WTP should have been negative but the experiment only allows for values greater than or equal to zero.



Table 1: Summary statistics.

Variable	Mean	SD	Min	Max	N
WTP, final stated maximum price (in tokens) for removing internet button	6.94	18.11	0	100	289
<i>Actual self-productivity</i>					
$y_1$ , number of correct answers in Task 1	4.92	0.38	0	5	289
$y_2$ , number of correct answers in Task 2, temptation NOT present	4.75	0.45	4	5	12
$y_2$ , number of correct answers in Task 2, temptation present	4.88	0.54	0	5	277
<i>Beliefs about self-productivity in Task 2</i>					
$\hat{y}^{nt}$ , predicted self-productivity if temptation NOT present	4.72	0.62	0	5	289
$\hat{y}^t$ , predicted self-productivity if temptation present	4.60	0.72	0	5	289
$\hat{y}^{ns}$ , predicted self-productivity if subject does NOT succumb	4.69	0.59	2	5	289
$\hat{y}^s$ , predicted self-productivity if subject succumbs	3.27	1.25	0	5	289
$\hat{p}^s$ , predicted self-likelihood of succumbing to temptation	0.09	0.15	0	1	289
<i>Psychological measures</i>					
$\theta$ , how tempted subject expects to be	1.76	0.76	1	4	289
$v$ , ignoring temptation was easier than expected	0.47	0.58	-1	1	277
$\omega$ , score on brief self-control scale (Tangney et al., 2004)	38.61	8.64	19	64	289
<i>Subject characteristics</i>					
Age	23.03	2.71	18	36	289
Male	0.52	0.50	0	1	289
Degree	1.53	0.56	1	3	289
Econ major	0.37	0.48	0	1	289
GPA	1.99	0.52	1	4	289

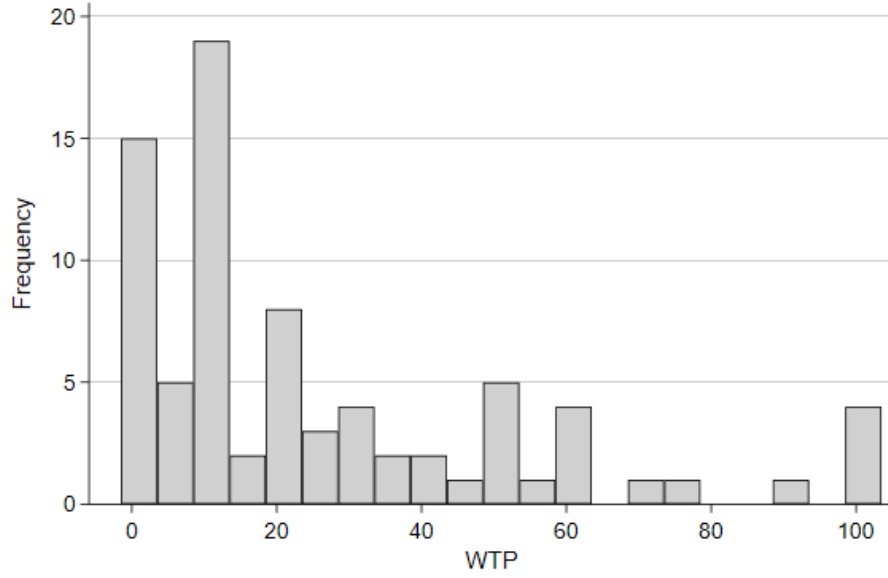


Figure 2: WTP for commitment device, for the 27% of subjects stating  $WTP > 0$ . Each bin represents an interval of length 5, i.e. the first bin is  $0 < WTP < 5$ , the second  $5 \leq WTP < 10$  and so on.

for the commitment device, the vast majority unwilling to pay and not incurring material loss from facing temptation.

As shown in Table 2, around 22.0% of subjects are ML overestimators, greater than the 10% attributed to confusion ( $p < 0.0001$ ), thus rejecting Hypothesis 1. As seen in the column  $y_1 - y_2$ , these ML overestimators on average even perform better in Task 2, thus making their positive WTP for commitment device (on average 18.67) seemingly larger than justified. Interestingly, and in contrast to the classical narrative of underdemand for commitment devices, there are more ML overestimators than ML underestimators (two-sample test of proportion,  $p < 0.0001$ ). Fewer than 6% of subjects underestimate their need for the commitment device: they are unwilling to pay (or have low WTP) and yet, when facing temptation, perform worse by nearly 2 questions.

We are able to calculate the welfare loss for ML overestimators by first substituting each subject's stated WTP,  $y_1$  and  $y_2$  into the expected utility function (1), and then comparing the result with the expected utility when WTP is set to its optimal value  $60(y_1 - y_2)$  (or 0 if negative). Summing over all 61 ML overestimators yields a total welfare loss of 156.0. Repeating this procedure with the 16 ML underestimators yields a total welfare loss of 346.4. Hence, welfare loss from underestimators is more than double that from ML

Table 2: Classification of subjects who face temptation.

Classification	N	Average WTP	$y_1 - y_2$	Frequency
ML overestimator	61	18.67	-0.07	22.02%
$WTP > 0$	61	18.67	-0.07	
ML accurate estimator	200	0	-0.06	72.20%
$WTP = 0$	200	0	-0.06	
ML underestimator	16	8	1.5	5.78%
$WTP > 0$	6	21.33	1	
$WTP = 0$	10	0	1.8	
Total	277	4.57	0.03	100%

overestimators. Note that we are only able to measure utility with respect to material loss since we do not have a direct measure of psychological cost. If this measure had been taken into account, the optimal WTP would likely be higher and thus welfare loss from overestimators (underestimators) would be lower (higher).

Given that we have used Task 1 performance as the counterfactual for how a subject would have performed without temptation in Task 2, we check if this is indeed the case for those who do manage to obtain commitment. Of those 12 subjects, 10 perform equally well in Task 1 and Task 2, and 2 subjects obtain one fewer correct answer in Task 2 compared to Task 1; averages are  $\bar{y}_1 = 4.92$  and  $\bar{y}_2 = 4.75$ , respectively. Given the small sample, any test of such differences is likely to be underpowered; nevertheless, we use the Wilcoxon matched-pairs signed-ranks test and find that individual differences  $y_1 - y_2$  are not significantly different from zero (exact  $p = 0.5000$ ).

In summary, we state our first result:

**Result 1.** *A substantial share of subjects overestimate WTP compared to actual material losses.*

It is possible that some subjects overstate WTP compared to realized material losses because they have mispredicted those losses. However, if subjects also state WTP higher than their *expected* material losses, it appears some other factor must be in play as well. Thus, we next compare WTP with subjects' *expected* material losses as elicited in the experiment,  $60(\hat{y}^{nt} - \hat{y}^t)$ . Among subjects who face temptation, 17.7% have WTP greater than this quantity ( $p < 0.0001$ ). Note that, since this test does not consider realized outcomes and thus does not require a counterfactual, it may also be performed on the full sample of subjects, including those who do not face temptation. Then, the proportion of

subjects who overestimate their WTP is 20.1% ( $p < 0.0001$ ).

In any case, we conclude the following:

**Result 2.** *A substantial share of subjects overestimate WTP compared to expected material losses.*

A possible explanation for such a discrepancy between WTP and expected material losses is that subjects expect to experience psychological discomfort from temptation, and seek to avoid it by stating a higher WTP. We next check if WTP is still overestimated even when accounting for subject's expectation of the psychological cost of temptation. As derived from equation (3) above, a sufficient condition, given  $WTP > 0$ , is that the subject's overestimation of material losses exceeds their underestimation of psychological cost, i.e. that  $(\hat{y}^{nt} - \hat{y}^t) - (y_1 - y_2) \geq 0$  and  $v \geq 0$ , with at least one strict inequality. The proportion of subjects who overestimate WTP using this measure is 14.8%, with  $p = 0.0039$ .<sup>15</sup> In contrast, the proportion of undercommitters for whom  $(\hat{y}^{nt} - \hat{y}^t) - (y_1 - y_2) \leq 0$  and  $v \leq 0$ , with at least one strict inequality, is 7.2%, significantly lower than the proportion of overcommitters (two-sample test of proportion,  $p = 0.0044$ ). Note that this definition of overestimation is again based on  $y_1$  as the counterfactual. We therefore conclude that:

**Result 3.** *A substantial share of subjects overestimate WTP compared to expected material losses and psychological temptation costs.*

## 4.2 Mechanisms and drivers

In this section, we exploit individual heterogeneity to explore potential explanations for commitment demand. We will relate our findings to existing commitment models, specifically:

- **Costly Self-Control** model (Gul and Pesendorfer, 2001): here, subjects do not expect to succumb, instead demanding commitment to avoid the expected cost of exercising self-control while resisting temptation. Note that the material component of such expected self-control costs is  $\hat{y}^{nt} - \hat{y}^{ns}$ , and commitment demand should therefore increase with this quantity.

---

<sup>15</sup>In some cases (14.8% of those facing temptation), stated WTP and expected material losses are such that Equation 1 would imply a negative  $PC_e$ . However, this is assumed to be due to a random error as the average implied  $PC$  is not significantly different from 0 ( $p = 0.3830$ ).

- **Random Indulgence** model (Chatterjee and Krishna, 2009; Dekel and Lipman, 2012): demand commitment since the agent expects to succumb with positive probability. Commitment demand increases with the expected likelihood of succumbing, as captured by  $\hat{p}^s$ .<sup>16</sup>

Clearly, while our experiment was not conceived as a sharp test of these theories, our belief variables do provide some circumstantial evidence as to their plausibility. Also, it seems plausible that the same basic mechanisms that drive commitment demand generally (as captured by the raw WTP values) may also drive *overcommitment*. For example, a subject who buys commitment because they expect to succumb to temptation (random indulgence) may also e.g. overstate WTP when overestimating  $\hat{p}^s$ . Thus, the above models will be used as potential explanations for overcommitment as well.

#### 4.2.1 Overcommitment

Our analysis of overcommitment relies on comparisons between the 14.8% of subjects who overestimate their demand for commitment, as defined in Hypothesis 3, and the remaining 85.2% who do not, in the group of 277 subjects who face temptation. Recall that, for a subject to be classified as an overestimator, we require  $(\hat{y}^{nt} - \hat{y}^t) - (y_1 - y_2) \geq 0$ . From this condition, however, it is not clear which LHS term(s) are the main mechanism for overcommitment. For example, overestimators might do equally well with and without temptation ( $y_1 = y_2$ ) but believe they do worse when exposed ( $\hat{y}^{nt} > \hat{y}^t$ ); or, they might think they will do equally well while in fact performing *better* when exposed.

Our data points towards the former, i.e. that WTP overestimation is driven by expected rather than actual material loss. Panel (a) of Figure 3 shows that actual productivity loss for overestimators and non-overestimators is nearly identical, with the majority in both groups performing equally well in Task 2 relative to Task 1. The distribution is not significantly different (Kolmogorov-Smirnov test,  $D = 0.0467$  and  $p = 1.000$ ). By contrast, in panel (b), more overestimators expect positive productivity loss when facing temptation, while the vast majority of non-overestimators do not expect any productivity difference. The difference in distribution is significant (Kolmogorov-Smirnov

---

<sup>16</sup>WTP being driven by fear of succumbing is also consistent with models of present-biased preferences. However, present-biased subjects do not exercise self-control when exposed to temptation and should therefore succumb with certainty (see, e.g., Krusell et al., 2010). By contrast, we find that exceedingly few exposed subjects fail to resist, and no subject expects to succumb with probability 1. Thus, present-biased preferences would seem to be a marginal phenomenon in our data.

test,  $D = 0.2801$ ,  $p = 0.008$ ) and, interestingly, is driven specifically by beliefs about productivity when facing temptation  $\hat{y}^t$ , with overestimators being more pessimistic about their performance (Kolmogorov-Smirnov test,  $D = 0.2866$ ,  $p = 0.006$ ). Overestimators and non-overestimators do not differ in their belief of performance without temptation  $\hat{y}^{nt}$  (Kolmogorov-Smirnov test,  $D = 0.0180$ ,  $p = 1.000$ ).

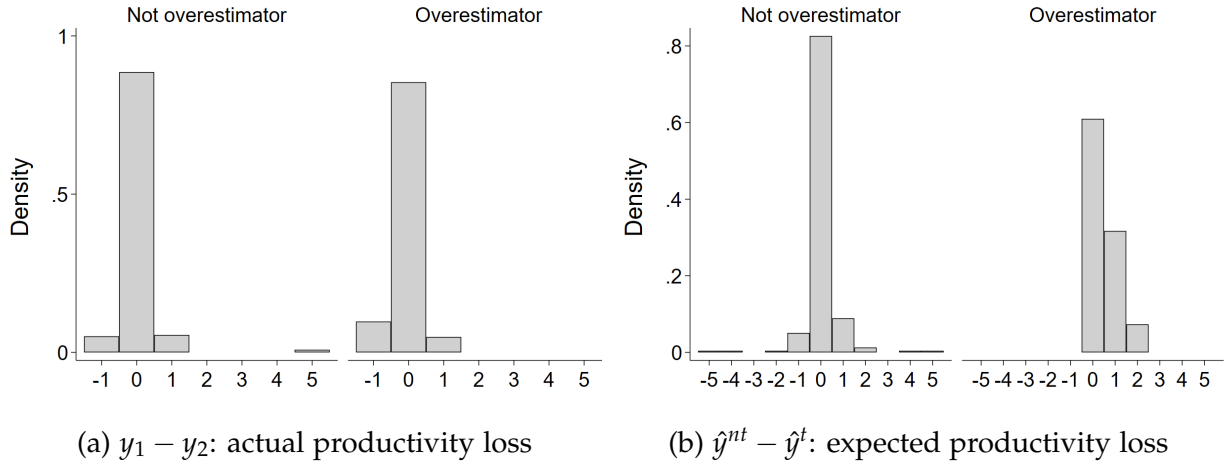


Figure 3: Distribution of realized outcomes and beliefs about productivity

Thus, overcommitment appears to be driven by pessimism: overestimators expect to do worse when exposed to temptation. Is this because they expect to succumb, or because they expect to do worse even if they manage to resist? In other words, is overcommitment consistent with the random indulgence or the costly self-control account? In fact, both mechanisms appear to be in play in our data. Regarding the self-control explanation, the distributions of self-control cost,  $\hat{y}^{nt} - \hat{y}^{ns}$ , for overestimators and non-overestimators are plotted in panel (a) of Figure 4. Consistent with the self-control explanation, overestimators expect a higher self-control cost (Kolmogorov-Smirnov test,  $D = 0.2323$ ,  $p = 0.046$ ). When comparing to actual outcomes, these pessimistic beliefs appear irrational: overestimators who do not succumb believe they would get 0.58 fewer correct answers than they in fact do, while non-overestimators who do not succumb only mispredict by 0.19. This difference in  $\hat{y}^{ns} - y_2$  between the two groups is significant ( $p = 0.0004$ ).

However, the random indulgence account also finds some support: panel (b) of Figure 4 shows that those who overestimate commitment demand also predict an overall higher likelihood of succumbing  $\hat{p}^s$  (Kolmogorov-Smirnov test,  $D = 0.2813$ ,  $p = 0.008$ ). Again, this appears irrational: as mentioned earlier, very few (only 4) subjects succumb to

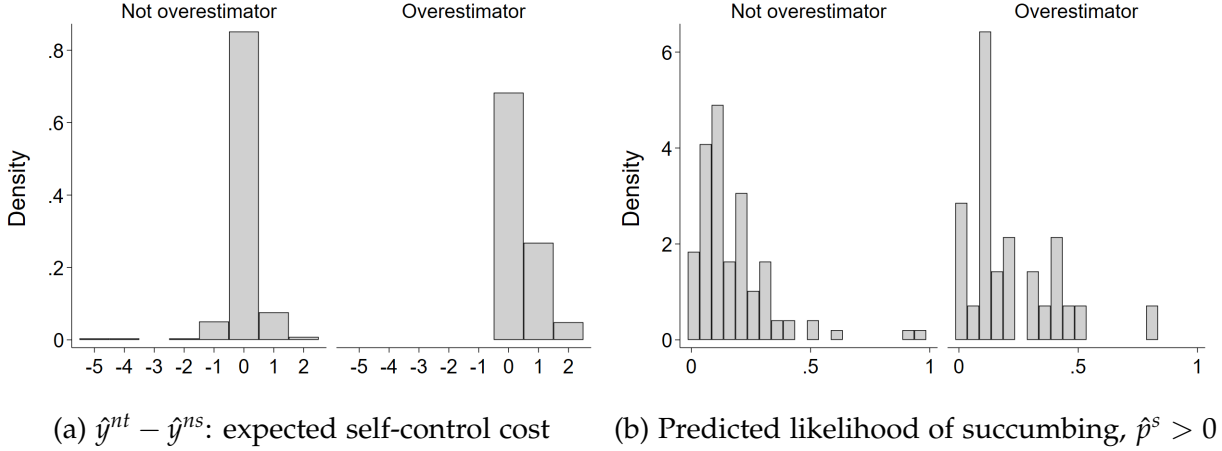


Figure 4: Beliefs about temptation

internet surfing, 1 overestimator and 3 non-overestimators. Thus, the actual probability of succumbing is lower than predicted, for both overestimators (2.44% vs 14.54%,  $p = 0.0019$ ) and non-overestimators (1.27% vs 7.12%,  $p < 0.0001$ ), though less so for the latter group.

To evaluate the relative strength of these two types of motivation, we turn to a regression framework which controls for the various components of misestimation ( $y_1$ ,  $y_2$ ,  $\hat{y}^{nt}$ ,  $\hat{y}^t$ , and the separate components  $\hat{y}^{ns}$ ,  $\hat{y}^s$ ,  $\hat{p}^s$ ). We also include  $\theta$ , how tempted subjects expect to be, as well as a set of subject-specific demographic variables. The dependent variable in these regressions, *overestimator*, derives from Hypothesis 3 above and is a dummy which equals 1 if the subject states  $WTP > 0$  and satisfies  $((\hat{y}^{nt} - \hat{y}^t) - (y_1 - y_2) \geq 0$  and  $v = 1$ ) or  $((\hat{y}^{nt} - \hat{y}^t) - (y_1 - y_2) > 0$  and  $v = 0$ ); it is 0 otherwise.  $\mathbf{X}$  is a set of subject-specific demographic variables.

The results are shown in Table 3. Column (1) confirms our earlier finding that the likelihood of a subject overestimating their demand for commitment increases with pessimism in the predicted productivity under temptation,  $\hat{y}^t$ . Expecting to get one fewer question increases the likelihood of overcommitment by around 7%. Next, we decompose  $\hat{y}^t$  into  $\hat{y}^{ns}$ ,  $\hat{y}^s$  and  $\hat{p}^s$  while retaining (column 2) or dropping (column 3) the original measure. In both cases, we again find that overcommitment is driven by expected lower performance when resisting temptation. Expecting to answer one fewer question, when the subject faces temptation but does not succumb, is associated with a 11-12% higher likelihood of overestimating demand for commitment, in line with the self-control account. Consistent with this, expecting to perform better *without* temptation is also



Table 3: Marginal effects from logistic regressions of WTP overestimation.

	(1)	(2)	(3)	(4)	(5)	(6)
$y_1$	-0.0313 (0.0630)	0.00638 (0.0609)	0.0106 (0.0615)			
$y_2$	0.0840 (0.0758)	0.0804 (0.0706)	0.0833 (0.0714)			
$\hat{y}^{nt}$	0.0573 (0.0455)	0.104** (0.0528)	0.104* (0.0533)			
$\hat{y}^t$	-0.0726** (0.0292)	-0.0233 (0.0305)				
$\hat{y}^{ns}$		-0.105** (0.0428)	-0.122*** (0.0380)			
$\hat{y}^s$		0.0172 (0.0177)	0.0141 (0.0171)			
$\hat{y}^{nt} - \hat{y}^{ns}$				0.135*** (0.0372)		0.123*** (0.0377)
$\hat{p}^s$		0.237* (0.130)	0.251* (0.130)		0.293** (0.132)	0.195 (0.130)
$\theta$	0.0208 (0.0276)	0.00182 (0.0283)	0.00531 (0.0280)	0.0202 (0.0264)	0.0141 (0.0299)	0.00409 (0.0290)
$\omega$	-0.00190 (0.00250)	-0.00255 (0.00238)	-0.00251 (0.00240)	-0.00199 (0.00244)	-0.00188 (0.00255)	-0.00212 (0.00243)
Demographics	X	X	X	X	X	X
N	277	277	277	277	277	277

**Notes:** Dependent variable *overestimator*: dummy variable which equals 1 if subject states  $WTP > 0$  and satisfies both  $(\hat{y}^{nt} - \hat{y}^t) - (y_1 - y_2) \geq 0$  and  $v \geq 0$  with one strict inequality, and 0 otherwise.  $y_1$  number of correct answers out of 5 in Task 1.  $y_2$  number of correct answers out of 5 in Task 2.  $\hat{y}^{nt}$  predicted number of correct answers out of 5 in Task 2 if temptation is NOT present.  $\hat{y}^t$  predicted number of correct answers out of 5 in Task 2 if temptation is present.  $\hat{y}^{ns}$  predicted number of correct answers out of 5 in Task 2 if subject does NOT succumb.  $\hat{y}^s$  predicted number of correct answers out of 5 in Task 2 if subject succumbs.  $\hat{p}^s$  predicted likelihood of succumbing to temptation, percentage point.  $\theta$  how much subject expects to be tempted on a scale from 1 to 4.  $\omega$  score on the brief self-control scale which ranges from 13 to 65 (Tangney et al., 2004). *Demographics* variables include Age, Male, Degree, Econ major and GPA. Standard errors in parentheses. \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

associated with a higher likelihood of overestimation. The coefficient for the probability of succumbing, on the other hand, is only marginally significant.

In columns (4)-(6), to further disentangle the two potential motivations for commitment, we regress overestimation on self-control cost ( $\hat{y}^{nt} - \hat{y}^{ns}$ ) and likelihood of succumbing  $\hat{p}^s$  separately and then together. Column (4) again confirms that expected self-control cost increases the likelihood of overestimation. One extra question answered incorrectly increases the likelihood of overestimation by 13.5%. In column (5), we see that the likelihood of overestimation also increases with probability of succumbing. Expecting to succumb by an extra 10 percentage points increases the likelihood of overestimation by 2.9%. However, when controlling for self-control cost in column (6), likelihood of succumbing is no longer significant, while the effect of self-control cost is stable at 12.3%.

Overall, our results point to subjects' pessimism in expected productivity when resisting temptation as the driver of overcommitment, consistent with models of costly self-control.

#### 4.2.2 Commitment demand in general

To investigate whether costly self-control or fear of random indulgence drive commitment demand more generally, Figure 5 plots expected self-control cost and likelihood of succumbing against WTP, for all 289 subjects in our sample, with trendline and 95% confidence intervals. Panel (a) shows that self-control cost does not explain WTP. The correlation between the two measures is 0.03 ( $p = 0.6142$ ). Panel (b), however, shows a strong relationship between WTP and expected likelihood of succumbing. The correlation is 0.39 ( $p < 0.0001$ ), indicating the presence of random-indulgence agents in our sample.

This is confirmed in regressions of WTP on various components of material loss and psychological cost in columns (1)-(3) of Table 4. In column (1), the expected strength of temptation  $\theta$  is highly significant, which seems unsurprising since it captures the effect of temptation on performance both when the subject resists and succumbs. However, in columns (2) and (3), when controlling for the likelihood of succumbing,  $\theta$  is no longer significant, but  $\hat{p}^s$  is. Finally, when directly testing costly self-control against random indulgence in columns (4)-(6),  $\hat{p}^s$  remains consistently highly significant, with a 10 percentage point increase in the likelihood of succumbing increasing WTP by 10 tokens, while self-control cost is not significant. Overall, consistent with random-indulgence models, subjects appear to state a high WTP not because of expected costs from exercising

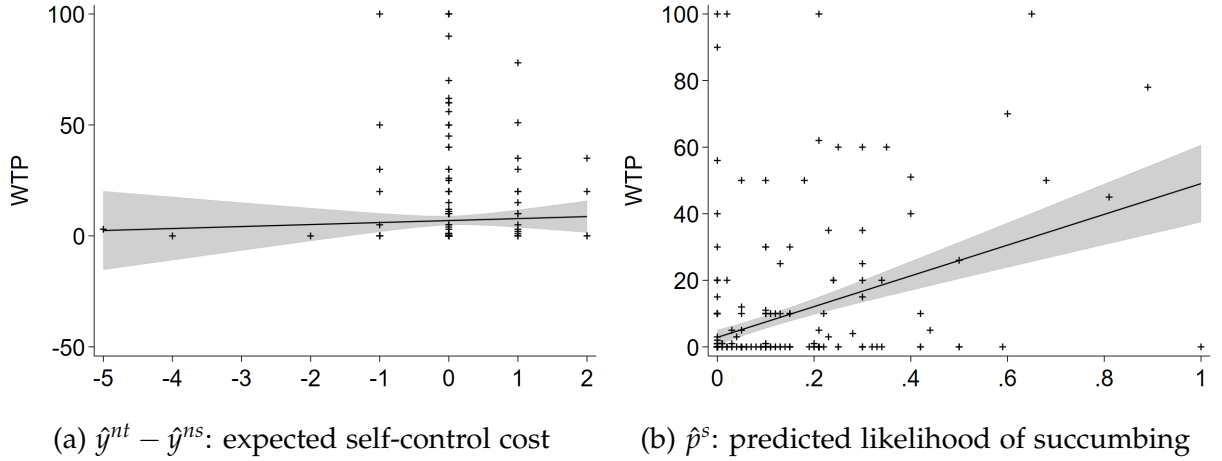


Figure 5: Beliefs about temptation

self-control, but because they anticipate succumbing to temptation.<sup>17</sup>

This result seems somewhat puzzling given our earlier finding that *overestimation* is driven mainly by costly self-control: it seems reasonable to expect any mechanism that drives overestimation to also drive WTP in general. To see more clearly what is going on, in Figure 6 we again plot WTP against self-control cost and likelihood of succumbing, this time split by overestimator status for the 277 subjects exposed to temptation. In the figure, the slope of the lines indicates whether each relevant variable drives WTP; whether it also drives overestimation may be deduced by comparing across the two subgraphs of each panel. For example, in panel (a), self-control costs clearly differ between overestimators and non-overestimators, and indeed are always nonnegative within the former group (0.00 vs 0.37,  $t$ -test,  $p = 0.0003$ ). In panel (b), average  $\hat{p}^s$  is also larger for overestimators (7.12% vs 14.54%,  $t$ -test,  $p = 0.0023$ ), though less clearly so, consistent with our results in Section 4.2.1.

As for WTP, we first note that it clearly increases with  $\hat{p}^s$  for both groups, though the slope is greater for overestimators. By contrast, in panel (a), WTP is not significantly increasing in expected self-control costs for non-overestimators. It is also not significant for overestimators ( $p = 0.3269$ ), although we note that the small sample ( $n = 41$ ), along with the limited variation in expected self-control cost within this group make drawing strong

<sup>17</sup>Given that demand for commitment appears to be driven by the fear of randomly succumbing to the temptation, overcommitment may be rationalised by risk aversion. In Online Appendix C we repeat our main analysis using a risk-averse utility function. We show that WTP correctly values subjects' expectations of their material loss, however these expectations are overly pessimistic. As a result, WTP is still overestimated relative to the actual material and psychological costs experienced.

Table 4: Tobit regressions of WTP.

	(1)	(2)	(3)	(4)	(5)	(6)
$y_1$	-5.868 (8.773)	3.904 (8.975)	4.144 (8.877)			
$y_2$	-3.764 (7.111)	-2.048 (6.965)	-2.077 (6.958)			
$\hat{y}^{nt}$	-2.967 (5.684)	-3.814 (5.740)	-3.867 (5.732)			
$\hat{y}^t$	-6.098 (4.928)	-1.048 (5.910)				
$\hat{y}^{ns}$		-1.409 (7.357)	-2.127 (6.148)			
$\hat{y}^s$		0.342 (3.047)	0.199 (2.934)			
$\hat{y}^{nt} - \hat{y}^{ns}$				0.960 (5.312)		-1.855 (4.847)
$\hat{p}^s$		99.51*** (22.96)	100.1*** (22.76)		101.0*** (21.37)	101.9*** (21.48)
$\theta$	13.02*** (4.762)	4.416 (4.849)	4.493 (4.830)	15.25*** (4.681)	4.272 (4.790)	4.359 (4.787)
$\omega$	0.381 (0.425)	0.0981 (0.409)	0.0997 (0.409)	0.316 (0.430)	0.0879 (0.399)	0.0921 (0.399)
Demographics	X	X	X	X	X	X
$N$	289	289	289	289	289	289

**Notes:** Dependent variable WTP: willingness-to-pay to remove internet access, from 0 to 100.  $y_1$  number of correct answers out of 5 in Task 1.  $y_2$  number of correct answers out of 5 in Task 2.  $\hat{y}^{nt}$  predicted number of correct answers out of 5 in Task 2 if temptation is NOT present.  $\hat{y}^t$  predicted number of correct answers out of 5 in Task 2 if temptation is present.  $\hat{y}^{ns}$  predicted number of correct answers out of 5 in Task 2 if subject does NOT succumb.  $\hat{y}^s$  predicted number of correct answers out of 5 in Task 2 if subject succumbs.  $\hat{p}^s$  predicted likelihood of succumbing to temptation, percentage point.  $\theta$  how much subject expects to be tempted on a scale from 1 to 4.  $\omega$  score on the brief self-control scale which ranges from 13 to 65 (Tangney et al., 2004). *Demographics* variables include Age, Male, Degree, Econ major and GPA. Standard errors in parentheses. \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

conclusions difficult. Thus, we cannot obviously rule out that expected self-control costs may drive *both* WTP and overcommitment specifically among the group of overestimators who are also more pessimistic about those costs. In any case, for everyone else, WTP seems clearly responsive only to  $\hat{p}^S$ .

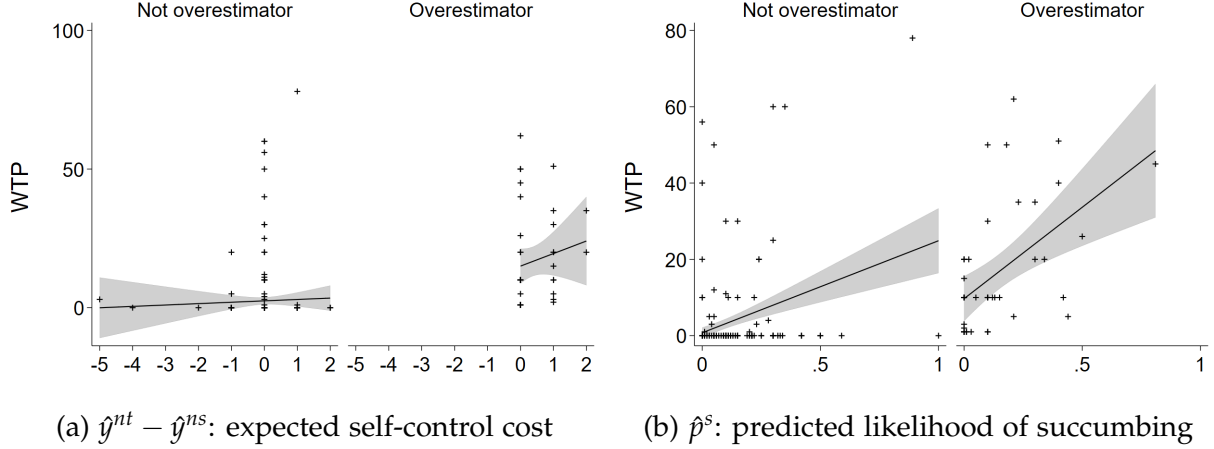


Figure 6: Beliefs about temptation

Overall, our findings contrast with those of [Toussaert \(2018\)](#), where subjects' WTP appears to be driven by how interesting the temptation is rather than the likelihood of succumbing, which in fact has a (marginally significant) negative effect on WTP. One explanation for this difference is that the temptation used in that study, reading a personal story from one subject, appeals more to subjects' curiosity, while the internet temptation used in our study is a familiar good for the subjects. The chance of succumbing, on the other hand, may be more important in our study as internet is more likely to be consumed mindlessly.

## 5 Concluding Remarks

When facing self-control problems, people may misestimate future outcomes or their own preferences, leading to suboptimal commitment demand. Previous research has focused on overoptimistic beliefs resulting in underdemand for commitment. With a few exceptions, excess commitment demand has so far been understudied in the literature. We present the first clear evidence of this possibility in the lab and make the following contributions.

First, we are the first to rationalize the demand for commitment by quantifying the different components of temptation costs. We show that a significant share of subjects overdemand a commitment device with real and non-refundable costs, thus incurring non-negligible welfare losses. This is true when we compare WTP with material loss from temptation, but also when we take into account psychological cost as well – which has so far eluded existing studies using field settings. Although it is conceivable that these estimates are specific to our setting, we use a design where, if anything, subjects should be unwilling to pay: internet access is easily available outside the lab and should have less immediate appeal. Additionally, shortly before the productivity task with temptation, subjects did the exact same task without temptation. Hence they should have a good idea about the difficulty of the task – yet even so they still overestimate their WTP.

Second, we are able to calculate the welfare loss due to excessive or insufficient commitment demand. Results indicate that, although overcommitment is the more widespread error in our sample, total welfare loss from undercommitment is around twice that from excess commitment. Although this suggests that the existing literature is right to focus on undercommitment, policy should consider the risk that commitment may yield lower utility than facing temptation and thus be harmful.

Third, we investigate the drivers of overcommitment. We find evidence that overdemand for the commitment device is systematically driven by subjects' pessimism in material loss, predicting lower-than-realized performance due to costly exercise of self-control in resisting the temptation. This contrasts with recent findings in [Carrera et al. \(2019\)](#), where commitment demand appears driven by subject errors and demand effects in a setting with high uncertainty about the future.

The above conclusions do come with some caveats; mainly, that other motives besides material losses and psychological costs of temptation may drive behavior. For example, if an agent derives signals about what kind of person he is from his actions, then succumbing to temptation may lead to psychological self-image costs beyond those considered in this paper. Similarly, obtaining commitment may have additional value as a signal of being capable of sophisticated reasoning. It is conceivable that such motivations lead to higher WTP and even that subjects construct motivated beliefs about subsequent material losses to justify their decision. However, most subjects do resist temptation, and we would expect that achieving such an outcome would provide *higher* self-image, self-confidence, or signalling value than obtained when the agent achieves the same goal

using the commitment device. Moreover, facing temptation also allows learning to occur (see, e.g., [Ali, 2011](#)). If so, a high WTP for commitment should appear even less rational than it does in our analysis. Further research should explore the implications of these issues for welfare and public policy.

We see our findings as a first step towards showing the existence of excess demand for commitment. The empirical applications and their policy implications will need to be investigated in future research.

## References

- Acland, D. and Chow, V. (2018). Self-control and demand for commitment in online game playing: Evidence from a field experiment. *Journal of the Economic Science Association*, 4(1):46–62.
- Ali, S. N. (2011). Learning self-control. *The Quarterly Journal of Economics*, 126(2):857–893.
- Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica*, 76(3):583–618.
- Ashraf, N., Karlan, D., and Yin, W. (2006). Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *The Quarterly Journal of Economics*, 121(2):635–672.
- Augenblick, N., Niederle, M., and Sprenger, C. (2015). Working over time: Dynamic inconsistency in real effort tasks. *The Quarterly Journal of Economics*, 130(3):1067–1115.
- Bai, L., Handel, B., Miguel, E., and Rao, G. (forth.). Self-control and demand for preventive health: Evidence from hypertension in India. *The Review of Economics and Statistics*.
- Beshears, J., Choi, J. J., Harris, C., Laibson, D., Madrian, B. C., and Sakong, J. (2015). Self control and commitment: Can decreasing the liquidity of a savings account increase deposits? NBER Working Paper 21474.
- Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, 19(5):547–556.



- Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.
- Bonein, A. and Denant-Boèmont, L. (2015). Self-control, commitment and peer pressure: A laboratory experiment. *Experimental Economics*, 18(4):543–568.
- Bryan, G., Karlan, D., and Nelson, S. (2010). Commitment devices. *Annual Review of Economics*, 2(1):671–698.
- Carrera, M., Royer, H., Stehr, M., Sydnor, J., and Taubinsky, D. (2019). How are preferences for commitment revealed? NBER Working Paper 26161.
- Chatterjee, K. and Krishna, R. V. (2009). A “dual self” representation for stochastic temptation. *American Economic Journal: Microeconomics*, 1(2):148–67.
- De Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302.
- Dekel, E. and Lipman, B. L. (2012). Costly self-control and random self-indulgence. *Econometrica*, 80(3):1271–1302.
- DellaVigna, S. and Malmendier, U. (2006). Paying not to go to the gym. *American Economic Review*, 96(3):694–719.
- Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.
- Gächter, S. and Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, 13(3):364–377.
- Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the Caltech Cohort Study. *Journal of Political Economy*, 127(4):1826–1863.
- Giné, X., Karlan, D., and Zinman, J. (2010). Put your money where your butt is: A commitment contract for smoking cessation. *American Economic Journal: Applied Economics*, 2(4):213–35.

- Gul, F. and Pesendorfer, W. (2001). Temptation and self-control. *Econometrica*, 69(6):1403–1435.
- Harrison, G. W., Lau, M. I., and Rutström, E. E. (2007). Estimating risk attitudes in Denmark: A field experiment. *Scandinavian Journal of Economics*, 109(2):341–368.
- Heidhues, P. and Köszegi, B. (2009). Futile attempts at self-control. *Journal of the European Economic Association*, 7(2-3):423–434.
- Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655.
- Houser, D., Schunk, D., Winter, J., and Xiao, E. (2018). Temptation and commitment in the laboratory. *Games and Economic Behavior*, 107:329–344.
- John, A. (2020). When commitment fails: Evidence from a field experiment. *Management Science*, 66(2):503–1004.
- Karlan, D. and Zinman, J. (2009). Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica*, 77(6):1993–2008.
- Kaur, S., Kremer, M., and Mullainathan, S. (2010). Self-control and the development of work arrangements. *American Economic Review*, 100(2):624–28.
- Krusell, P., Kuruşçu, B., and Smith Jr, A. A. (2010). Temptation and taxation. *Econometrica*, 78(6):2063–2084.
- Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478.
- Laibson, D. (2015). Why don't present-biased agents make commitments? *American Economic Review*, 105(5):267–72.
- Milkman, K. L., Minson, J. A., and Volpp, K. G. (2013). Holding the hunger games hostage at the gym: An evaluation of temptation bundling. *Management Science*, 60(2):283–299.
- O'Donoghue, T. and Rabin, M. (1999). Doing it now or later. *American Economic Review*, 89(1):103–124.

- Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, 68(5):1281–1292.
- Royer, H., Stehr, M., and Sydnor, J. (2015). Incentives, commitments, and habit formation in exercise: Evidence from a field experiment with workers at a fortune-500 company. *American Economic Journal: Applied Economics*, 7(3):51–84.
- Sadoff, S. and Samek, A. (2019). Can interventions affect commitment demand? a field experiment on food choice. *Journal of Economic Behavior & Organization*, 158:90–109.
- Sadoff, S., Samek, A., and Sprenger, C. (2020). Dynamic inconsistency in food choice: Experimental evidence from two food deserts. *The Review of Economic Studies*, 87(4):1954–1988.
- Schilbach, F. (2019). Alcohol and self-control: A field experiment in india. *American Economic Review*, 109(4):1290–1322.
- Schwartz, J., Mochon, D., Wyper, L., Maroba, J., Patel, D., and Ariely, D. (2014). Healthier by precommitment. *Psychological Science*, 25(2):538–546.
- Strotz, R. (1955). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3):165–180.
- Tangney, J. P., Baumeister, R. F., and Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, 72(2):271–324.
- Toussaert, S. (2018). Eliciting temptation and self-control through menu choices: A lab experiment. *Econometrica*, 86(3):859–889.

# Appendices

## A Results using beliefs about a similar peer

The peer belief measures are elicited according to the following procedure. Each subject is matched with another participant whose WTP is closest to the subject's own WTP. Then, subjects estimate how many questions the peer will answer correctly if the peer faces temptation and succumbs ( $\hat{y}_p^s$ ) or does not succumb to temptation ( $\hat{y}_p^{ns}$ ). The subject is paid 20 tokens if their answer matches the actual outcome of the chosen peer. Finally, the subject is matched to a group of five participants whose WTP values are closest to the subject's own WTP and who face temptation in Task 2. The subject is asked to estimate how many of these five participants would succumb to temptation and press the internet button ( $\hat{n}_p$ ). Again, subjects have the possibility of earning 20 tokens for a correct answer. The peer measure of expected productivity under temptation, which is inferred from the above quantities, is then  $\hat{y}_p^{t,inf} = (\hat{n}_p/5)(\hat{y}_p^s) + (1 - \hat{n}_p/5)(\hat{y}_p^{ns})$ . We use this measure to test the following hypotheses:

**Hypothesis A.2.** *Among the subjects who face temptation in Task 2, no more than 10% have  $WTP > 60(\hat{y}^{nt} - \hat{y}_p^{t,inf})$ .*

**Hypothesis A.3.** *Among the subjects who face temptation in Task 2, no more than 10% have  $(\hat{y}^{nt} - \hat{y}_p^{t,inf}) - (y_1 - y_2) \geq 0$  and  $v \geq 0$ , with at least one strict inequality, and  $WTP > 0$ .*

We first note that self-beliefs are not equal to peer beliefs: average  $\hat{y}^t$  is slightly higher than average  $\hat{y}_p^{t,inf}$  for subjects who face temptation (4.61 vs 4.52,  $p = 0.0263$ ) and also for all subjects (4.60 vs 4.49,  $p = 0.0107$ ). This is also confirmed in a Kolmogorov-Smirnov test ( $D = 0.02202$ ,  $p < 0.001$  for subjects who face temptation, and  $D = 0.2284$ ,  $p < 0.001$  for all subjects). Subjects are more optimistic when stating beliefs about their own performance rather than a similar peer. However the two variables are highly correlated (46%,  $p < 0.0001$ ). As with self-beliefs, subjects generally underestimate the number of correct answers in Task 2: actual performance  $\bar{y}_2$  is higher than average  $\hat{y}_p^{t,inf}$  (4.88 vs 4.52,  $p < 0.0001$ ).

Using the incentivized peer measure to test Hypothesis A.2, we find that 12.6% of subjects overstate their WTP ( $p = 0.0715$ ). (Repeating the test for the entire subject sample,

the proportion is 14.5%,  $p = 0.0051$ .) Testing Hypothesis A.3, the proportion of subjects who overestimate WTP is 16.6%, with  $p = 0.0001$ . Again, note that this result uses  $y_1$  as a valid counterfactual for Task 2 performance in the committed group. Hence, we conclude that our main results are robust to using the peer measure in place of the self-measure.<sup>18</sup>

## B Accounting for possible measurement error in WTP

Given that WTP is measured twice in our experiment, we can use the instrumental-variable approach suggested by [Gillen et al. \(2019\)](#) to correct for measurement error, assuming that these errors are independent across the two elicitation procedures. Essentially, one WTP variable is used as an instrument for the other WTP variable and vice versa. The predicted values from these two first-stage regressions are then pooled and used in place of the WTP values we used in the original analysis. We describe this robustness check in this section.

Figure [A.1](#) plots the distribution of  $WTP_0$  and  $WTP_1$  with frequency weights, showing the bunching at zero. Given the truncation of WTP at zero, we do not expect that measurement errors are independent for those subjects stating  $WTP_0 = 0$  and  $WTP_1 = 0$ . In the first-stage regressions, we therefore impose the restriction that any zero WTP should be predicted by zero. In any case, we are primarily concerned with WTP overestimation among subjects with positive WTP, which is where we would expect measurement errors to be independent.

Using the predicted values of  $WTP_0$  and  $WTP_1$  to test Hypotheses 1, 2 and 3, the proportions of subjects who overestimate WTP are, respectively, 22.02%, 17.33%, and 14.98%. These are the same or very similar to our original results of 22.02%, 17.69% and 14.80%. We therefore conclude that our results are robust to measurement errors in WTP elicitation.

---

<sup>18</sup>All results are robust to using  $\hat{y}_p^{t,inf} \pm 0.5$  to account for possible rounding in  $\hat{n}_p$ . As noted above, all results are also robust to correction for multiple hypotheses testing when results for both self- and peer measures are combined.

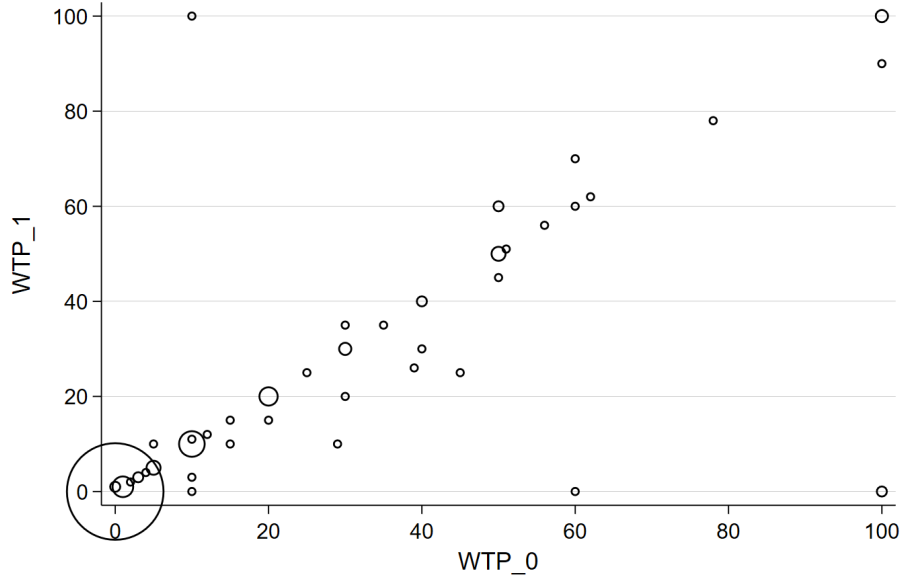


Figure A.1: Scatterplot of  $WTP_0$  and  $WTP_1$ , with frequency weights.

## C Assuming aversion to uncertainty in the lab

Given that subjects' demand for commitment appears to stem from the fear of randomly indulging their preference for surfing the internet, the temptation may have been perceived as a 'risk' or uncertainty which subjects would like to avoid. We note it is unclear that such uncertainty aversion is the same as 'risk aversion' in the traditional sense given the small stakes in the lab, over which subjects should be risk-neutral ([Rabin, 2000](#)). Nevertheless, what looks like an overstated WTP for commitment compared to the optimal WTP of a risk-neutral agent may become rationalizable or even understated when compared to the optimal choice under uncertainty aversion. In the absence of a better way to parametrize such aversion to uncertain lab payments, in this section we check whether WTP remains overstated if subjects are assumed to be (strongly) risk-averse with CRRA utility.

Recall that, in Section 3, subjects maximize a utility function which only captures risk aversion in the BDM 'lottery' and misses the second 'lottery' faced by the subject: the possibility of earning a lower amount if she succumbs to temptation. We therefore change

the temptation-related utility terms to obtain

$$\begin{aligned}
U(WTP) = & \frac{1}{2} \left[ \frac{1}{2} \cdot u(100 + 120y_1 - PC) \right. \\
& + \frac{1}{2} \left( \frac{100 - WTP}{100} \cdot u(100 + 120y_1 - PC) + \frac{1}{100} \int_0^{WTP} u(100 + 120y_1 - R) dR \right) \\
& + \frac{1}{2} \left[ \left( 1 - \frac{WTP}{200} \right) (p^s u(100 + 120y^s - PC) + (1 - p^s) u(100 + 120y^{ns} - PC)) \right. \\
& \left. \left. + \frac{1}{2} \left( \frac{1}{100} \int_0^{WTP} u(100 + 120y^{nt} - R) dR \right) \right] \right]
\end{aligned}$$

assuming that  $PC$  is the same regardless of whether the subject succumbs or not – there is, for example, no self-image loss or guilt from succumbing, and nor is there utility from internet surfing. The solution under risk neutrality, denoted  $WTP_{RN}$ , is unchanged:

$$WTP_{RN} = 60(y^{nt} - y^t) + PC$$

where  $y^t = p^s y^s + (1 - p^s) y^{ns}$ .

Assume now that the subject is risk-averse and has CRRA utility function defined as:

$$u(x) = \begin{cases} \frac{x^{1-\eta}-1}{1-\eta} & \eta \neq 1 \\ \ln(x) & \eta = 1 \end{cases}$$

No closed-form solution for WTP then exists, but we may derive the first-order condition

$$\begin{aligned}
\frac{dU}{dWTP} = & \frac{1}{200(1-\eta)} \left\{ (100 + 120y_1 - WTP)^{1-\eta} + (100 + 120y^{nt} - WTP)^{1-\eta} \right. \\
& - (100 + 120y_1 - PC)^{1-\eta} - p^s (100 + 120y^s - PC)^{1-\eta} \\
& \left. - (1 - p^s) (100 + 120y^{ns} - PC)^{1-\eta} \right\} = 0
\end{aligned} \tag{A.1}$$

To show the robustness of our results under risk aversion, our strategy is the following. We seek to calculate the optimal WTP for the risk-averse agent, denoted  $WTP_{RA}$ , and show that there are still a significant number of subjects who overestimate WTP. We obtain values for  $WTP_{RA}$  using numerical simulations of (A.1) with the relevant  $y_1$ ,  $y^{nt}$ ,  $y^s$ ,  $y^{ns}$  and  $p^s$  values inserted for each individual subject.  $\eta$ , the coefficient of relative risk



aversion, has been estimated in different studies to be around 1.<sup>19</sup> To be conservative, we present results for several values of  $\eta$  up to  $\eta = 3$ , though as will be shown our results do not change drastically.

We start by asking whether risk-averse subjects overestimate their WTP when only considering actual material loss (corresponding to Hypothesis 1 in the risk-neutral case). In equation (A.1),  $y^{nt}$  is thus interpreted as the *actual* number of correct answers when the subject is not exposed to temptation; as per the risk-neutral case above, we use  $y_1$  as the counterfactual.  $p^s$  is obtained using the percentage of subjects who succumb out of all subjects exposed to temptation, this equals 1.44%. For subjects who do not succumb,  $y^{ns} = y_2$ , while  $y^s$ , the counterfactual had they succumbed, is obtained using the average productivity of subjects who do succumb, which is  $y^s = 2$ . In the same way, for subjects who succumb,  $y^s = y_2$  while the counterfactual  $y^{ns} = 4.93$ , the average productivity for those who do not succumb. Comparing the resulting  $WTP_{RA}$  with the WTP stated by each subject, the proportion of ML overestimators under different values of  $\eta$  are given in the first row of Table A.1. Around 17% of subjects are still considered to be ML overestimators under conventional levels of risk aversion, stating WTP greater than what should be optimal when considering the actual material loss. A much higher number of subjects are now underdemanders of commitment (79% under  $\eta = 1$  or 1.5), though the magnitude of underestimation compared to material loss is much smaller (around 7 tokens under  $\eta = 1$  or 1.5) compared to the risk-neutral case (82 tokens as calculated using Table 2). Nevertheless, our first result is robust to assuming CRRA with  $\eta \leq 3$ .

Table A.1: Proportion of overestimators under risk aversion.

Relative to	$\eta = 0.5$	$\eta = 1$	$\eta = 1.5$	$\eta = 2$	$\eta = 3$
(1) Actual material loss	17.33%	16.97%	16.97%	15.52%	15.52%
(2) Expected material loss	11.55%	11.55%	11.55%	11.19%	9.75%
(3) Actual material loss and psychological cost	14.08%	14.08%	14.08%	14.08%	12.64%

We next turn to subjects' WTP considering expected material loss (corresponding to Hypothesis 2). We proceed as above, except that we now use each subject's predictions of their own performance  $\hat{y}^{nt}$ ,  $\hat{y}^s$ ,  $\hat{y}^{ns}$  and  $\hat{p}^s$ . As shown in the second row of Table A.1,

<sup>19</sup>For example, in one of the most widely cited lab experiments on risk aversion, [Holt and Laury \(2002\)](#) find that almost all subjects have  $\eta \leq 1.37$ . In a field experiment in Denmark, [Harrison et al. \(2007\)](#) find the mean  $\eta$  to be 0.67. The estimate is 0.74 in [Andersen et al. \(2008\)](#), who also estimate a population standard deviation for  $\eta$  of 0.056.

we find a lower number of risk-averse subjects overestimate their WTP, compared to the case with actual material loss above. The proportion of overestimators is less than 12% for all values of  $\eta$  and no longer significant. Putting aside psychological cost, many subjects appear to underestimate their performance relative to how well they actually do in the face of temptation and their WTP is a relatively accurate reflection of this pessimism.

Finally, we check whether WTP is still overestimated by risk-averse subjects when allowing for psychological costs of temptation. Since we do not know the actual  $PC$  faced by each subject, our strategy is analogous to the test of Hypotheses 3 under risk neutrality. First, we note that optimal WTP is strictly increasing in  $PC$  under any degree of risk aversion; the proof is given in Online Appendix D.2. Given this fact, we may proceed as follows.

For all subjects with  $WTP > 0$  and  $v \geq 0$ , and for all expected  $PC$  values consistent with  $0 < WTP < 100$ , we plug in appropriate outcome variables in (A.1) to calculate what the WTP should have been for a risk-averse subject based on *actual* material losses. We then repeat the exercise for the subject's *expected* material loss; denote these two (sets of) WTP values  $WTP_a$  and  $WTP_e$  for actual and expected WTP, respectively. Now, suppose for some particular expected psychological cost  $PC_e$ ,  $WTP_e \geq WTP_a$  while  $v \geq 0$  (implying  $PC_e \geq PC_a$ ), with at least one of these two inequalities strict. Since WTP is increasing in  $PC$ , we then have  $WTP_e(PC_e) \geq WTP_a(PC_e) \geq WTP_a(PC_a)$ , again with at least one strict inequality. Thus, WTP has been strictly overestimated in relation to both actual material losses and actual psychological costs. To be conservative, we classify as overestimators those subjects who have  $v \geq 0$  and  $WTP_e \geq WTP_a$ , with at least one strict inequality, for *all* values of  $PC_e$  consistent with  $0 < WTP_e < 100$ .<sup>20</sup>

As shown in row (3) of Table A.1, we find that such subjects make up between about 13-14% of all subjects who face temptation ( $p < 0.1$  for all values of  $\eta$ ,  $p < 0.05$  for conventional values of  $\eta \leq 2$ ).<sup>21</sup> Hence, our result of overestimation relative to both actual material and psychological costs is also robust to assuming CRRA with  $\eta \leq 3$ .

Overall, even assuming a very strong degree of risk aversion, our conclusion that a significant share of subjects overstate their demand for the commitment device is

---

<sup>20</sup>In principle, since both stated WTP and all parameters related to expected material losses are known, we might use them in (A.1) to solve for a single implied value of  $PC_e$ . The reason why we do not check whether  $WTP_e \geq WTP_a$  only at this implied  $PC_e$  is because, as in Footnote 15, it is sometimes negative, which we interpret as there being some random error in subjects' WTP responses.

<sup>21</sup>Using  $\hat{y}^{nt}$  as our counterfactual in place of  $y_1$  yields stronger results: the proportion of overestimators are 16.3% under  $\eta \leq 2$  and 14.4% when  $\eta = 3$ .

unchanged. The subjects in question appear to state a higher WTP than motivated either by actual material losses, or when including actual psychological costs.

## D Proofs

### D.1 That results are robust to assuming WTP is paid conditional on Task 2 being chosen for payment

In this subsection we show that our analysis is robust to assuming that WTP is paid only conditional on Task 2 being chosen for payment. In this case subjects maximize expected utility, with equal probabilities of either Task 1 or Task 2 being paid, as

$$\begin{aligned}
U(WTP) = & \frac{1}{2} \left[ \frac{1}{2} \cdot u(100 + 120y_1 - PC) \right. \\
& + \frac{1}{2} \left( \frac{100 - WTP}{100} \cdot u(100 + 120y_1 - PC) + \frac{1}{100} \int_0^{WTP} u(100 + 120y_1 - 0) dR \right) \Big] \\
& + \frac{1}{2} \left[ \frac{1}{2} \cdot u(100 + 120y^t - PC) \right. \\
& + \frac{1}{2} \left( \frac{100 - WTP}{100} \cdot u(100 + 120y^t - PC) + \frac{1}{100} \int_0^{WTP} u(100 + 120y^{nt} - R) dR \right) \Big]
\end{aligned}$$

and the solution under risk neutrality is given by

$$WTP = 120(y^{nt} - y^t) + 2PC$$

Clearly, the analysis for Hypothesis 3, and thus Hypothesis 4, is equivalent. For Hypothesis 1, the proportion of overestimator using the new WTP threshold given above is 21.7% ( $p < 0.0001$ ). For Hypothesis 2, the proportion of overestimators using  $\hat{y}^t$  is 17.3% ( $p < 0.0001$ ).

## D.2 That WTP under risk aversion is increasing in psychological cost

Assuming CRRA with  $\eta > 1$ , the first-order condition is restated below:

$$\begin{aligned} \frac{dU}{dWTP} = \frac{1}{200(1-\eta)} \Big\{ & (100 + 120y_1 - WTP)^{1-\eta} + (100 + 120y^{nt} - WTP)^{1-\eta} \\ & - (100 + 120y_1 - PC)^{1-\eta} - p^s (100 + 120y^s - PC)^{1-\eta} \\ & - (1 - p^s) (100 + 120y^{ns} - PC)^{1-\eta} \Big\} = 0 \end{aligned}$$

The second derivative is

$$\begin{aligned} \frac{d^2U}{dWTP^2} = \frac{1}{200} \Big[ & -\frac{1}{(100 + 120y_1 - WTP)^\eta} - \frac{1}{(100 + 120y^{nt} - WTP)^\eta} \Big] \\ & < 0 \end{aligned}$$

The partial derivative of the first-order condition with respect to  $PC$  is

$$\begin{aligned} \frac{\partial^2 U}{\partial WTP \partial PC} = \frac{1}{200} \Big[ & \frac{1}{(100 + 120y_1 - PC)^\eta} + \frac{p^s}{(100 + 120y^s - PC)^\eta} + \frac{1 - p^s}{(100 + 120y^{ns} - PC)^\eta} \Big] \\ & > 0 \end{aligned}$$

Using the implicit function theorem,

$$\frac{dWTP}{dPC} = -\frac{\frac{\partial^2 U}{\partial WTP \partial PC}}{\frac{d^2 U}{dWTP^2}} > 0$$

Hence, WTP is strictly increasing in  $PC$ . The proof for  $0 < \eta < 1$  and  $\eta = 1$  is similar and is left to the reader.

## E Power calculations

Our power calculations are based on the concept of minimum detectable effect (MDE) (see, e.g., [Bloom, 1995](#)). Using 90% power and 5% level of statistical significance, the MDE is the smallest effect that, if true, has a 90% chance of producing an estimate that is

significant at the 5% level. The MDE is calculated to be

$$\begin{aligned} MDE &= (t_{\alpha/2} + t_{\beta})\sigma \\ \hat{\theta} - \theta &= 1.645\sigma + 1.282\sigma \\ \hat{\theta} - \theta &= 2.927\sigma \end{aligned}$$

where  $\sigma$  is the standard error of the estimator  $\hat{\theta}$ . We use the score test for binomial proportion which is based on the null standard error, yielding

$$\begin{aligned} \hat{\theta} - \theta &= 2.927\sqrt{\frac{\theta(1-\theta)}{n}} \\ \hat{\theta} &= 0.1 + 2.927\sqrt{\frac{0.1(0.9)}{289}} \\ &= 0.1517 \end{aligned}$$

with the null  $\theta = 0.1$  (attributable to subject confusion) and a sample size of 289 ( $\hat{\theta} = 0.1528$  for  $n = 277$ , for tests using the sample who face temptation). This means that if the true proportion of WTP overestimators is at least 15.17%, our experimental design will detect this with 90% probability at a 5% level of statistical significance.

For the purpose of our hypothesis tests, the critical value above which the hypothesis would be rejected would thus be  $\theta + t_{\alpha/2}\sigma = 12.90\%$  for tests with  $n = 289$  and 12.97% for tests with  $n = 277$ .

## F Instructions

Begin on next page.

## General instructions

You are about to participate in an experiment on decision-making. Before we start, please make sure your phones are on silent and put away all personal belongings.

The experiment will take place through your computer terminals. Please do not talk or try to communicate with other participants during the session. If you have any question, please raise your hand and the experimenter will approach you to answer it.

This experiment consists of 2 stages plus a short questionnaire at the end. The whole session will last up to 2 hours. After the session, you will receive your experimental payment. This payment consists of a **participation fee of 100 CZK** plus your **experiment earnings**. Your experiment earnings will depend on your own decisions, on the decision of another participant, and on chance. It is therefore important to think about each of your decisions carefully.

During the experiment, your payoff will be denominated in experimental tokens that will be converted to CZK at the end at the following rate:

$$1 \text{ CZK} = 1 \text{ token}$$

You are about to begin with Stage 1. Out of Stage 1 and Stage 2, only one will be used for payment. Which stage is chosen will be determined by a random draw at the end of the experiment.

We will now read together the instructions for Stage 1.

## Stage 1

During Stage 1, your main task will be to focus attentively on a four-digit number that will appear on your computer screen for a period of up to 30 minutes. This number will increment every 3 seconds. At random times during the 30 minute period, you will be prompted to **enter the last number you saw on your screen**. The number will be reinitialized after every prompt and you will receive a total of 5 prompts during the period. You will earn 120 tokens per correct answer, should this stage be chosen for payment.

Besides performing the attention task, no other activity will be allowed (including checking your phone, surfing the internet, studying...). If you are caught doing something else, you will not be paid for your participation in the experiment.

Are there any questions at this point?

You will now practice the attention task for 1 minute on the computer before moving on to the real task.

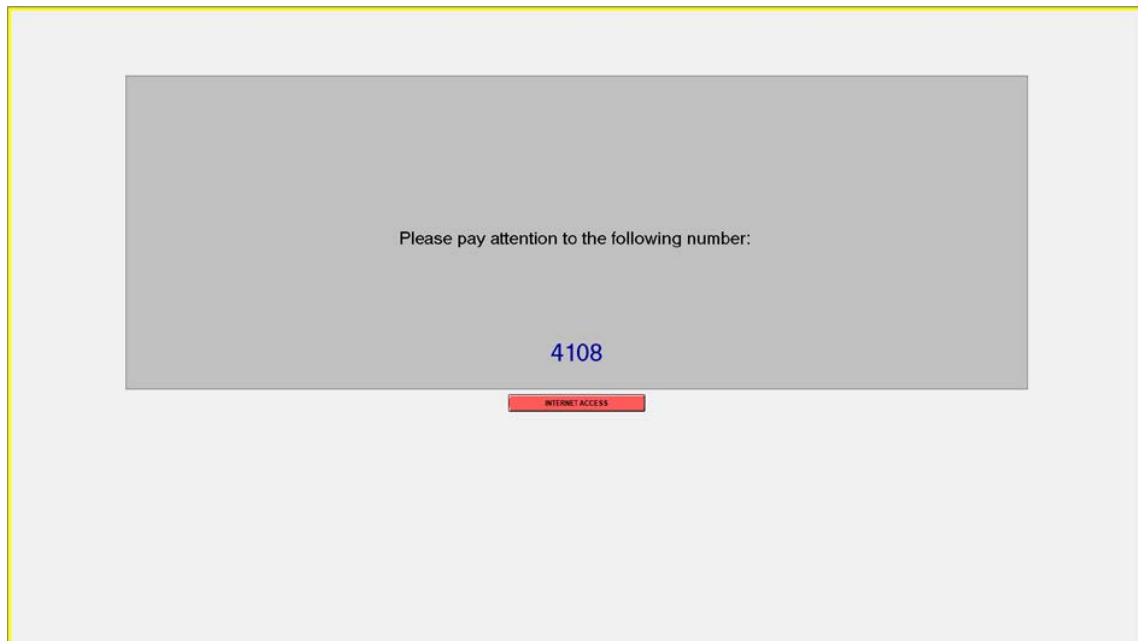
[After stage 1]

In Stage 2, you will repeat the same attention task with a small modification. We will explain this modification in more detail soon, and afterwards you will be asked to answer some questions regarding the new attention task.

### The modification

The attention task in Stage 2 is similar to the one you completed in Stage 1: your main task is to focus attentively on a four-digit number that will appear on your computer screen for a period of up to 30 minutes. There will again be 5 prompts to enter the last number you saw on the screen, and you will again earn 120 tokens per correct answer.

**However, below the four digits, you will now see a button labeled “Internet Access”, as shown in the screenshots below.** You can click on this button at any point during the attention task.





What was the number you just saw?

Please enter this number in the following box:

Submit

INTERNET ACCESS

The screenshot shows a light gray rectangular area with a yellow border. Inside, the text 'What was the number you just saw?' is centered. Below it, 'Please enter this number in the following box:' is centered. Underneath is a small, empty, light blue rectangular input box. Below the input box is a small gray button labeled 'Submit'. At the bottom is a red rectangular button labeled 'INTERNET ACCESS'.

If you click this button, you will be able to surf the internet for the remainder of the 30-minute period instead of continuing with the attention task. Once you click the button, you will not be able to return to the task, so clicking it means that you forfeit the chance to earn more money through answering any future prompts correctly. You will, however, earn money from all correct answers up to the point of clicking the button.

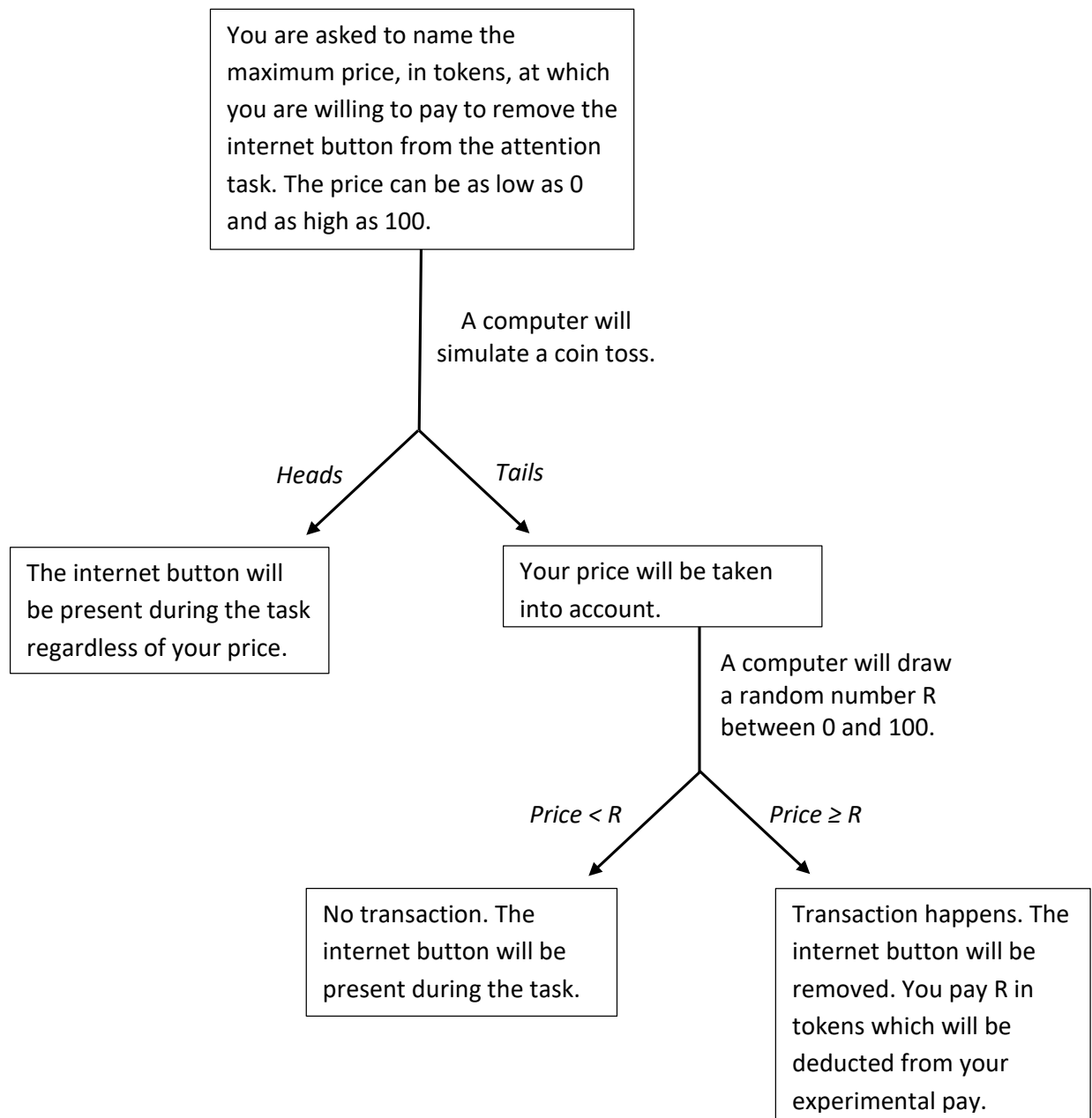
If you do not click the button, you will simply continue with the attention task. The button will continue to be present for the remainder of the 30-minute period.

#### Removing the “Internet Access” button

If you would like to remove the “Internet Access” button, for example because you think you might be able to concentrate better without it, you have the possibility to do so. We will now describe how this works.

You can pay tokens to remove the “Internet Access” button. Removing the button means that there will be no possibility for internet access during the attention task and you will therefore be participating in the attention task for the whole 30 minutes. The screen displayed will exactly be the same as in Stage 1, without the “Internet Access” button.

Starting from the top, determining whether to remove the button or not involves the following steps:



As you can see, the possibility of successfully removing the internet button increases the higher your stated price. Note in particular that:

- Your chance of removing the internet button is maximized (but is not guaranteed) if you state a price of 100.
- If you are not willing to pay anything to remove the internet button, you should enter a price of 0.

Your decision is final and cannot be changed.

**No matter if you pay to remove the internet button, keep the button but never click it, or keep the button and click it, you will spend the same amount of time (up to 30 minutes) in Stage 2.**

You will now have a practice round to ensure you understand how the process of removing the internet button works. When you have finished the practice round, you will be asked to state your **actual** price for removing the button, as well as answer a few questions about the new attention task.

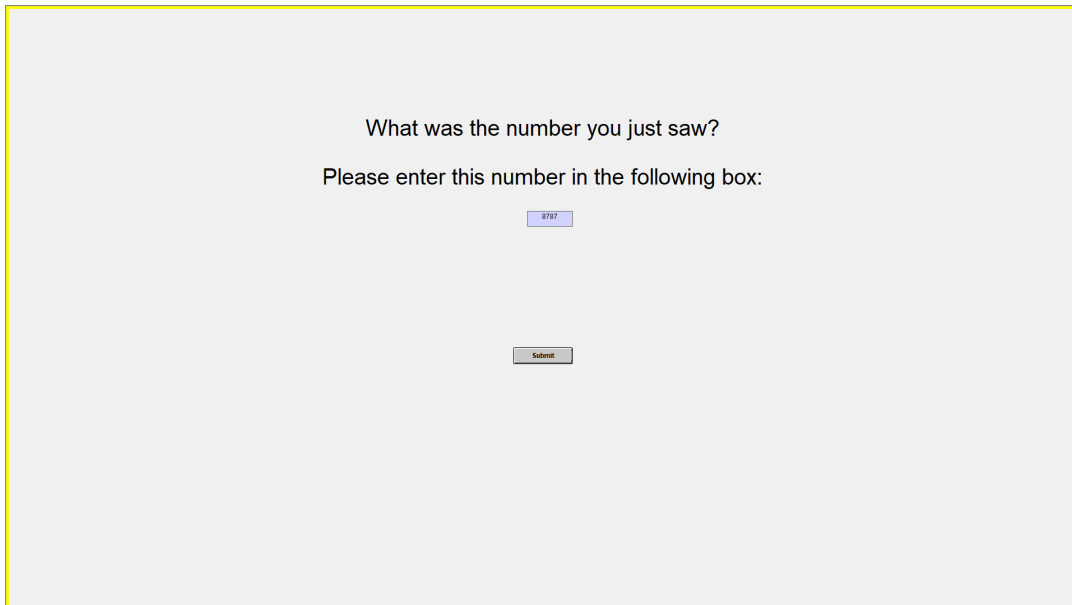
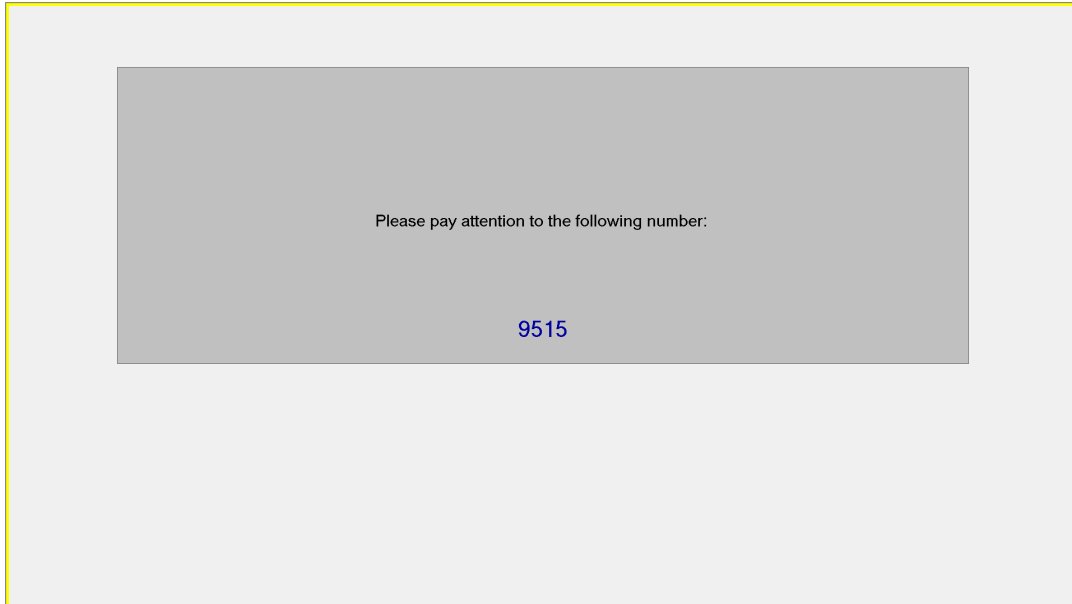
Finally, the computer will toss the coin and (if TAILS comes up) draw the random price to determine the outcome of the transaction. You will be informed about whether the “Internet Access” button will be present or not, and subsequently you will start the attention task in Stage 2.

## **Stage 2**

On-screen instructions.

## G Screenshots

### G.1 Attention Task 1 and elicitation of $\hat{y}^{nt}$



Number of correct answers from Stage 1

4

OK

Suppose that now you are to repeat the same task again.

This second time round, how many correct answers would you expect to get out of 5?

Submit

## G.2 WTP and beliefs elicitations

### PRACTICE ROUND

**Maximum price for removing "Internet Access" button**

State a test price, in tokens, you are willing to pay to remove the button to surf the internet.  
Please enter a number between 0 and 100 (inclusive)

Submit

The coin toss resulted in  
**HEADS**

Therefore, your test price will NOT be taken into account.

The outcome for the next stage would have been  
**Attention task WITH internet access**

END of PRACTICE ROUND

OK

Before we move on to the actual transaction round,  
please take a moment to consider what you would prefer for the upcoming attention task.

- ☐ I prefer to have the option of internet access in the upcoming attention task.  
☐ I prefer to do the upcoming attention task without the option of internet access

Think about what your choice implies for your price to remove the button in the next screen.

Submit

## ACTUAL ROUND

### Maximum price for removing "Internet Access" button

What is the highest price, in tokens, you are willing to pay to remove the button to surf the internet?  
Please enter a number between 0 and 100 (inclusive)

50

Submit

Before going on to Stage 2, please answer the following questions.

Continue

You will soon learn if the "Internet Access" button is removed or not in Stage 2.  
Suppose the button is NOT removed, and you continue to have the option for internet access for the rest of the attention task.

In this situation, how many correct answers do you expect to get out of 5?

3

Submit



You will soon learn if the "Internet Access" button is removed or not in Stage 2.  
Suppose the button is NOT removed, and you continue to have the option for internet access for the rest of the attention task.

How likely (in %) would you say you are to press the button?



Suppose that you DO press the "Internet Access" button.  
In this situation, how many correct answers do you then expect to get out of 5, prior to pressing the button?

3

Suppose that you DO NOT press the "Internet Access" button.  
In this situation, how many correct answers do you then expect to get out of 5?

3

Submit

How tempted do you think you would be by internet access?

- ☐ Not at all tempted
- ☐ Not that tempted
- ☐ Quite tempted
- ☐ Very tempted

Submit

You have previously stated a price of 55 tokens for removing internet access.  
You have since answered questions and thought more about the potential outcomes in Stage 2.  
Now you have a chance to revise your stated price, if you would like to do so.  
Would you like to revise your stated price?  
If yes, please enter a new price between 0 and 100 (inclusive).  
If not, simply enter the same price of 55.

Submit

Before continuing with Stage 2, we will match you with another participant in this room  
who has stated a price which is the same as (or closest to) your price, 3 tokens, for removing internet access.  
Please answer the following questions about what you expect that other participant would do in Stage 2.  
You will be paid 20 tokens if your answer matches the actual outcome of the matched participant.

OK

Suppose the "Internet Access" button IS removed for this person in Stage 2.

How many correct answers do you expect this person to get out of 5?

4

Submit

Suppose the "Internet Access" button is NOT removed in Stage 2, and this person continues to have the option for internet access for the rest of the attention task.

Suppose that this person DOES press the "Internet Access" button.

How many correct answers do you expect this person to get out of 5, prior to pressing the button?

Suppose that this person DOES NOT press the "Internet Access" button.

How many correct answers do you expect this person to get out of 5?

Submit

Just for this question, we will create a group of participants similar to you, consisting of 5 people each of whom:

- i) stated a price which is the same as (or closest to) your price, AND
- ii) continues to have the "Internet Access" button for the rest of the attention task.

How many of these 5 participants would you say are likely to press the button?

You will be paid an extra 20 tokens if your answer is correct.

Submit

The coin toss resulted in

**HEADS**

Therefore, your price will NOT be taken into account.

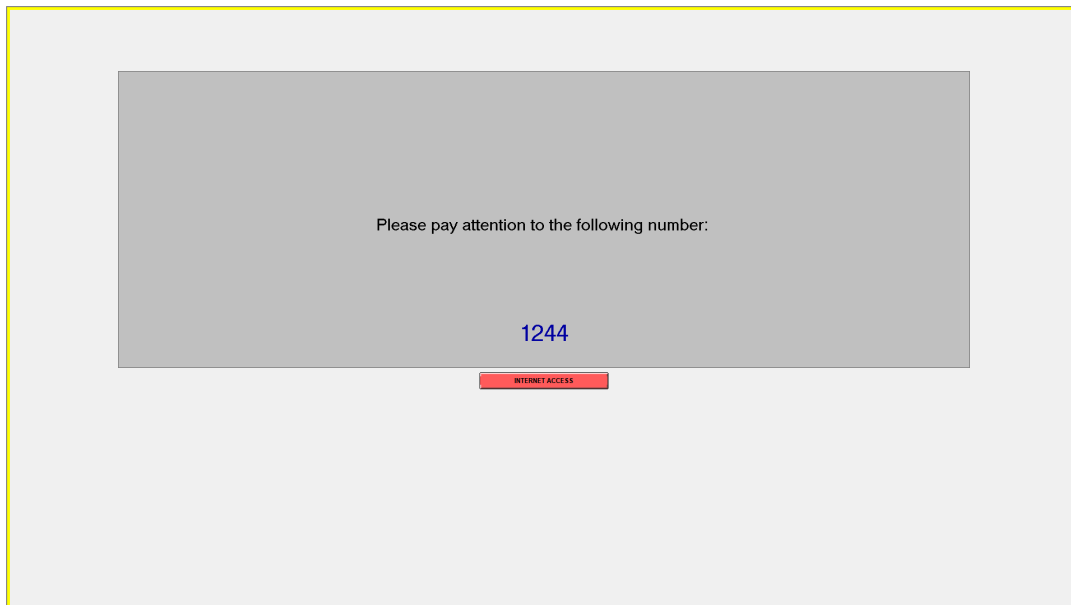
The outcome for the next stage is

**Attention task WITH internet access**

Click OK to start Stage 2

OK

### G.3 Attention Task 2 and post-experiment questionnaire

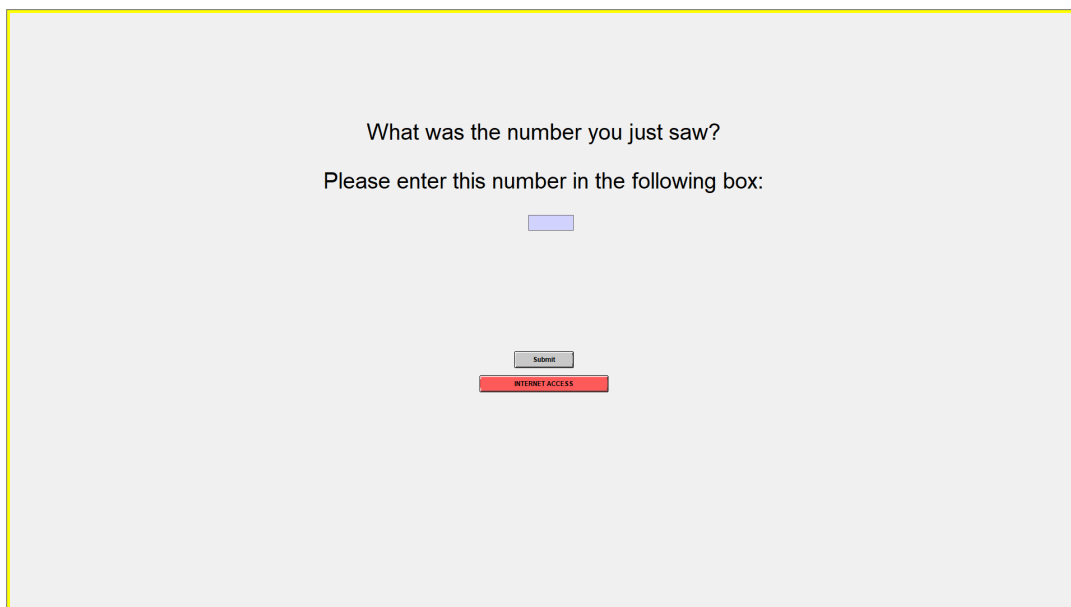


Please pay attention to the following number:

1244

INTERNET ACCESS

This screenshot shows a gray rectangular box centered on a light gray background. Inside the box, the text "Please pay attention to the following number:" is displayed in a small black font. Below this text, the number "1244" is shown in a larger, blue font. At the bottom center of the box, there is a small red rectangular button with the text "INTERNET ACCESS" in white capital letters.



What was the number you just saw?

Please enter this number in the following box:

Submit

INTERNET ACCESS

This screenshot shows a light gray background. At the top, the text "What was the number you just saw?" is centered. Below it, the text "Please enter this number in the following box:" is centered. Underneath this text is a small, empty blue rectangular input box. Further down, centered, is a small gray rectangular button with the text "Submit" in black. At the bottom center, there is a small red rectangular button with the text "INTERNET ACCESS" in white capital letters.

### END QUESTIONNAIRE

What is your age (in years)?

What is your gender? ☐ Male  
☐ Female

What is your degree program? ☐ Bachelor  
☐ Master  
☐ PhD

What is your field of study? ☐ Economics  
☐ Business  
☐ Computer Science  
☐ Law  
☐ Social Science  
☐ Language  
☐ Humanities  
☐ Education  
☐ Science  
☐ Medicine  
☐ Other

If other, please specify

What is your average study grade?

Each of the previous university exams you attempted yields a grade A-F, where A corresponds to value of 1, B is 1.5, C is 2, D is 2.5, E is 3, and Fail is 4.  
Your average study grade is the average of these values for all exams attempted. If you do not know your average study grade, please give an estimate.

Submit

### Which of the following best applies to you?

- ☐ I was not interested in internet access at all because I did not care about it  
☐ I was not interested in internet access at all because I was concentrating on the number  
☐ At first I was not interested in internet access, but as time passed, I got bored and started thinking about it  
☐ At first I thought a lot about internet access, but as time passed, I managed to start focusing more on the number  
☐ I kept thinking about internet access and this prevented me from staying focused on the number  
☐ I chose to surf the internet as soon as possible

**Do you think the difficulty of ignoring the "Internet Access" button and concentrating on the attention task was higher or lower than expected (when you chose your price for removing that button)?**

- ☐ Ignoring the internet button and concentrating on the attention task was more difficult than expected  
☐ Ignoring the internet button and concentrating on the attention task was neither easier nor more difficult than expected  
☐ Ignoring the internet button and concentrating on the attention task was easier than expected

Submit

Using a 5-point scale, please indicate how much each of the following statements reflects how you typically are.

<i>I am good at resisting temptation.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>I have a hard time breaking bad habits.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>I am lazy.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>I say inappropriate things.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>I refuse things that are bad for me.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>I wish I had more self-discipline.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>I do certain things that are bad for me, if they are fun.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>Pleasure and fun sometimes keep me from getting work done.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>I have trouble concentrating.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>I am able to work effectively toward long-term goals.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>People would say that I have iron self-discipline.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>Sometimes I can't stop myself from doing something, even if I know it is wrong.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH
<i>I often act without thinking through all the alternatives.</i>	NOT AT ALL    ○ ○ ○ ○ ○    VERY MUCH

Submit

Your final stated price for removing internet access was 3 tokens.  
We are interested to hear your motivations. What made you decide on this price?

Submit

Stage chosen for payment  
**Stage 1**

Number of correct answers in that Stage  
**4**

Earnings from correct answers  
**480 tokens**

Cost of removing internet access  
**-0 tokens**

Earnings from correctly answering question about other participant  
**+0 tokens**

Show-up fee  
**+100 tokens**

Total earnings for today  
**580 tokens**

OK