Revealing temptation through menu choice: field evidence

Séverine Toussaert
University of Oxford

This version: February 2019

Abstract

In a field study with participants in a weight loss challenge, I use the menu choice approach of Gul and Pesendorfer (2001) to explore the extent to which preference for smaller menus may "reveal" temptation. Focusing on the temptation to eat unhealthy, I elicit participants' preferences over a set of lunch reimbursement options ("the menus"), which differed in the range of foods covered. I extract information from the entire ordering to develop measures of temptation allowing to study its *source*, *strength* and *structure*, and validate those measures with survey data. Finally, I test whether temptation measured through menu choice predicts other behaviors that could be symptomatic of self-control problems, such as take-up of, and performance on, a goal setting contract. I find that choices to restrict the coverage are very common and generally target the foods rated as most tempting and unhealthy. Furthermore, the structure of commitment choices appears largely consistent with the restrictions imposed by the theory. Finally, those who revealed their temptation through menu choice were more likely to take up the contract and less likely to achieve their goals. The elicitation of menu preferences thus offers a promising venue for measuring self-control problems.

JEL Classification Numbers: C93, D03, I12.

Key Words: Commitment, temptation, flexibility, self-control, experiment.

Acknowledgments: I feel greatly indebted to my advisors Guillaume Fréchette and David Cesarini, and to my former colleague Sevgi Yüksel, for their support and continuous feedback in the development of this project. I thank Ned Augenblick, Yves Le Yaouanq, and Pietro Ortoleva for their many insightful comments. I also thank seminar participants at the University of Alicante, Amsterdam, Bonn, Harvard, Lyon, Maastricht, NYU, NYU Shanghai, PSE, SUFE, UCL, Warwick and Zürich, as well as participants at the 2014 North American ESA Conference and the 2018 CESifo Conference in behavioral economics, for their useful comments. Finally, I would like to acknowledge the enthusiasm and interest of the manager of the NYU wellness services, Christine Gould, without whom this project would have never seen light.
Contact: severine.toussaert@economics.ox.ac.uk

1. Introduction

Commitment behavior is at the heart of economic theories of temptation (Laibson (1997), O'Donoghue and Rabin (1999), Gul and Pesendorfer (2001), Fudenberg and Levine (2006)). While standard decision makers always weakly prefer larger choice sets, individuals aware of their self-control problems may choose to constrain their future choices at a cost to themselves. This behavioral implication of temptation has been widely tested, both in the lab and in the field, by studying the take-up of a commitment device. Examples of such devices are illiquid accounts designed to increase savings (Ashraf et al. (2006), John (2019), Beshears et al. (2015)), deadlines to reduce procrastination on an effort task (Ariely and Wertenbroch (2002), Bisin and Hyndman (2018)), or deposit contracts with financial penalties for smoking or failing to exercise (Giné et al. (2010), Royer et al. (2015)). Despite the wealth of studies, a recurring finding in this literature is that commitment take-up is low, usually in the range of 10% to 35%.[1] In turn, the weak evidence of commitment demand has motivated researchers to reassess the robustness of its theoretical foundations to various assumptions about agents' uncertainty, their naiveté, or the price of commitment (Laibson (2015)).

One challenge in interpreting the literature findings is that most experimental designs greatly constrain the ways individuals can express their preference for commitment, making it hard to infer temptation from commitment behavior. First, commitment opportunities are usually limited to a single instance or, at best, to several instances of the same type of commitment decision. In particular, virtually no study examines commitment demand across different decision environments.[2] Second, individuals are typically constrained in terms of the option they can commit to. For instance, in consumption-savings problems, the standard assumption is that individuals are tempted to overconsume now (and thus undersave); however, as evidenced in Ameriks et al. (2007), some individuals may need commitments to dissave, especially towards the end of the life cycle. Finally, commitment rarely presents different degrees of flexibility.[3] For instance, an individual might prefer to keep his options open if

[1] For instance, only 11% of smokers took up the deposit contract proposed by Giné et al. (2010) and 28% of bank customers adopted the commitment savings product of Ashraf et al. (2006). Similarly, Royer et al. (2015) report a 12% enrollment rate in their exercise commitment contract, while Kaur et al. (2015) report a 35% take-up rate of their work contract. In the lab, Augenblick et al. (2015) found that less than 10% of subjects were willing to pay even \$0.25 to constrain their effort choices. Two notable exceptions are Milkman et al. (2014) who found that 61% of students were willing to pay to access audio novels only when exercising at the gym, and Schilbach (2018) who found that over 50% of his study participants demanded commitment to increase their sobriety.

[2] Schilbach (2018) studies the impact of sobriety incentives and commitment savings on overall savings and finds that they are substitute. However, the design does not allow to explicitly measure demand for commitment savings. In addition, the account with a commitment feature also included a savings bonus (10-20% matching contribution rate); thus, it is not obvious how deposits in this account should be interpreted.

[3] One exception is Beshears et al. (2015) who implemented one treatment in which they offered subjects

offered the choice between a (apple) and b (brownie), but to eliminate b from his choice set if offered a third alternative c (chocolate), thus choosing $\{a, c\}$ over $\{a, b, c\}$. In this paper, I enrich the dataset of commitment choices to study the extent to which temptation can be revealed through agents' decisions to constrain their choice sets.

To this purpose, I conduct a field study with individuals expected to face real temptation dilemmas: participants in a weight loss challenge. My methodological approach is grounded in a large revealed preference literature, which relies on the observation of preferences over menus to elicit temptation (Gul and Pesendorfer (2001), Dekel et al. (2009); see Lipman and Pesendorfer (2013) for a review). In this literature, commitment is defined as a preference for a smaller menu and an agent reveals his temptation for an option if he strictly prefers a menu that excludes it. Following this literature, I construct a dataset of menu choice to study participants' commitment to eating healthy through a lunch reimbursement program. The "menus" were various reimbursement options that differed only in the range of foods included in the coverage. Instead of asking participants to select one option, I elicit their preference ordering over the entire set and implement their preferences using an incentive-compatible procedure. This rich dataset allows me to take a more agnostic and comprehensive approach to measuring temptation. First, I allow for any degree of commitment, from commitment to a single food category to complete flexibility. Second, I allow commitment to target any food category, thus letting individuals reveal their own temptations.

Extracting information from the entire ordering, I build various measures of temptation to analyze its *source*, its *strength* and its *structure*. I study the source and strength of temptation by (i) looking at participants' top option(s), and by (ii) constructing an index of "Global Temptation", which computes the number of times a participant preferred to exclude food category x from the coverage when comparing two nested options M and $M\backslash\{x\}$. I study the structure of temptation by testing the consistency of commitment choices with the seminal model of Gul and Pesendorfer (2001) (henceforth GP) and its main axiom, Set Betweenness. Although simple and intuitive, Dekel et al. (2009) argue that Set Betweenness might be too restrictive to capture the multiple facets of temptation and propose two relaxations: one allowing temptation to have a cumulative effect, the other acknowledging the stochastic nature of temptation. I construct falsification tests to study the compatibility of Set Betweenness and its relaxations with the individual preference orderings.

Finally, I assess the validity of the menu choice measures I developed along two dimensions. First, using detailed survey data on participants' food perceptions and self-reported consumption, I test whether those behavioral measures are in line with participants' assessments of what foods are tempting and unhealthy. Second, I test whether temptation revealed through menu choice can predict other behaviors during the challenge that could result from

the choice between two different commitment savings products that varied in their degree of illiquidity.

self-control problems. In particular, I test whether a preference for removing unhealthy foods from the coverage predicts take-up of, and performance on, a goal setting contract designed to encourage exercise and a more active participation in the challenge. Relatedly, I study whether menu choice is predictive of three measures of engagement: challenge completion, study completion, and likelihood of claiming reimbursement.

I find that the elicitation of menu preferences offers a promising venue for measuring temptation in real life settings. First, while the largest coverage maximized reimbursement, only a third of participants strictly preferred to be reimbursed for all food categories. Choices to restrict the reimbursement coverage appear far from random: nearly half of participants systematically ranked higher those options that removed the most unhealthy foods from the coverage. Second, most decisions to exclude unhealthy foods are structurally consistent with temptation as modeled by Gul and Pesendorfer (2001), suggesting that temptation takes a simple form. Third, this behavioral expression of temptation coincides with what participants perceive to be tempting and unhealthy foods. Fourth, those who revealed temptation through their menu preferences were more likely to enter the goal setting contract, less likely to reach their goals, and overall, less likely to stay engaged until the end.

While the menu choice literature on temptation is large, so far very limited attempts have been made to bring the various models to the data. In Toussaert (2018), I rely on the elicitation of menu preferences to identify "self-control types" i.e., agents who may demand commitment despite expecting to resist temptation. The two papers complement each other in several respects. On the one hand, Toussaert (2018) is a lab experiment in which I observe both preferences over a restricted set of menus, and actual choices from menus. On the other hand, the present paper is a field study in which the environment of menu choice is much richer, but actual food selections are not observed. While Toussaert (2018) focuses on the identification of costly self-control, this study is concerned with the expression of temptation and the extent to which it can be revealed from menu choice only.

The rest of this paper proceeds as follows. Section 2 describes the weight loss challenge and study design. Section 3 studies temptation through menu preferences in the lunch reimbursement program. Section 4 relates menu preferences to survey evidence on food perceptions and to behaviors possibly linked to self-control problems. Section 5 provides a summary and discussion of the results. Supplementary material is available in a detailed Online Appendix (henceforth OA).

2. Study Design

2.1. Description of the weight loss challenge

The study was conducted with participants in an eight-week weight loss challenge organized every year since 2011 by the wellness services of New York University (NYU). Only NYU faculty and staff members were eligible to participate. The data concerns the 2014 edition, which took place over the months of March and April. The basic rules of the challenge were inspired by the American TV show "The Biggest Loser": the contestant who loses the highest percentage of body mass over the challenge period wins a prize. Contestants were required to participate in an initial weigh-in on Week 1 of the challenge, which determined their reference weight. Three follow-up weigh-ins were scheduled during the challenge but only the final weigh-in on Week 8 was required in order to be considered for the prize. All weigh-ins were conducted at a private gym club near NYU. To encourage participants to attend the follow-up weigh-ins, the top loser between any two weigh-ins received a small prize. The final winners (one male and one female) were the participants who lost the highest percentage of their reference weight at the end of the challenge.

In addition to the weigh-ins, participants could sign up for many activities and events to help them stay on track during the challenge. First, participants were offered free access to the gym facilities during the first month of the challenge. Interested participants received a gym badge during the first weigh-in, which they had to scan every time they wished to access the facilities. Second, participants could sign up for any of four wellness events (two nutrition seminars and two exercise classes) scheduled by the wellness services. Participants received regular reminders about those events, which were advertised at the start of the challenge. More details about the challenge rules and procedures can be found in OA-D.1.

2.2. Recruitment procedures and structure of the study

Participants were recruited for the experiment at the first weigh-in. The experiment was advertised as an online study on improving health through exercise and nutrition conducted in collaboration with the wellness services. Participants were told that for completing a two-part survey, they could receive a $20 gift card as well as the chance to be reimbursed for their lunch meals during the month of April. Interested participants were asked to complete an online consent form, the link to which was printed on flyers and on a weight booklet (see OA-D.2 and -D.3). The study was composed of two online surveys, one completed during the first week of the challenge and one after the final week. Eligibility for the full incentives was contingent on having completed the entire study ($10 for partial completion). Overall, 195

participants were present at the first weigh-in and 117 signed up for the study. Among those 117 participants, 113 completed the first survey and 87 also completed the second survey.

TABLE 1: Timeline and content of the study

Online Survey	Completion Period	Survey Content
Survey 1 $N = 113$	March 4th through March 11th, 2014	**Part 1**: Basic socio-demographics Questions about participation **Part 2**: Goal setting contract **Part 3**: Reimbursement program
Survey 2 $N = 87$	April 29th through May 6th, 2014	Feedback questions about challenge and study Intertemporal choice tasks Self-control measures of Ameriks et al. (2007)

Table 1 summarizes the structure of the study. Survey 1 had three main components. The first part gathered data on basic socio-demographics, past participation in the challenge, expectations of weight loss and physical activity, and perceived obstacles to success. The second part studied commitment demand through goal setting by offering participants to receive their study payment only if they achieved self-set attendance goals (see Section 4.2). The third part of the survey studied participants' commitment demand to eat healthy through a lunch reimbursement program offered during the second month of the challenge. Participants were asked to rank various reimbursement options, which systematically differed in the range of foods covered by the reimbursement. The elicitation and implementation procedures are discussed in Section 3.1. After having submitted their ranking, participants answered a set of questions related to their food habits and were asked to rate food items in terms of their healthiness and attractiveness (see Section 4.1).

The aim of Survey 2 was threefold. The first objective was to gather information about respondents' evaluation of their participation in the challenge and to identify barriers to attendance and success. The second aim was to understand respondents' Survey 1 decisions to constrain their future choices through goal setting and the reimbursement program. The final objective was to assess whether commitment decisions in the challenge correlate with standard measures of self-control problems used in the literature. I focused on two such measures: (*i*) present bias over time-dated monetary rewards measured through a Multiple Price List mechanism; (*ii*) the self-control measures developed by Ameriks et al. (2007), which rely on survey answers to an hypothetical intertemporal consumption problem; see OA-C for an analysis of these measures and their relationship to commitment decisions.[4]

[4]The link between commitment demand and those two standard measures of self-control problems is at

2.3. Sample characteristics

Appendix Table 4 presents summary statistics about the subject pool. Surveyed participants were 79% female and more educated than the general population. About 30% had previously participated in the challenge. On average, participants entered the challenge with the goal of losing 14.3 lbs (1.8 lbs per week). These ambitious numbers do not vary significantly with gender or prior participation in the program. Although individual-level data on weight or BMI is not available, casual observation suggests that the large majority of participants were overweight at the start of the challenge. Looking at aggregate-level data, the average starting weight was 204.0 lbs for male participants (min: 146.6 - max: 253.6) and 172.3 lbs for female participants (min: 117.2 - max: 312.2). These averages are above the US national average of 195.5 lbs for males and 166.2 lbs for females, and much beyond the ideal body weight of 157.1 lbs for males and 119.3 lbs for females.[5] Furthermore, more than 80% of Survey 1 respondents reported having attempted at least one diet over the last 10 years, and among those, about 20% had attempted at least 10 diets. The population under study is therefore greatly concerned with losing weight and struggling to achieve this goal.

3. Commitment in the reimbursement program

3.1. Description of the program and preference elicitation

I used the approach of menu choice to study commitment to eating healthy in the context of a lunch reimbursement program. Respondents who completed the first survey could enter a lottery with a 10% chance of being reimbursed for all their lunch meals over the second month of the challenge (04/01/14 - 04/28/14). The reimbursement was capped at 20 meals and $15 per meal. In order to claim reimbursement, participants were required to provide itemized receipts with their name or credit number on it. As a way to encourage all respondents to participate in the program, the winners were announced only after the challenge was over. More information on the logistical aspects of the program and reimbursement claims can be found in Section 4.2.3 and OA-B.4.4.

After receiving the above details about the program, participants were asked for their preferences over various reimbursement options, which differed in what food categories they covered. More precisely, the reimbursement could cover one, two, or all three categories

best weak; however, attrition between the two surveys does not allow to draw any strong conclusions.

[5]Source: Anthropometric Reference Data for Children and Adults: United States, 2007-2010. The Ideal Body Weight (IBW) was computed using the formula IBW_M = 50kg + 2.3*(height(in)-60) for males and IBW_F = 45.5kg + 2.3*(height(in)-60) for females, for the US average height of 69.3in for males and 63.8in for females.

listed in Table 2. The instructions emphasized that the options only differed in the range of foods covered by the reimbursement; otherwise, the terms of reimbursement were identical. Without being more explicit, it was also mentioned that being reimbursed only for a certain category of foods could provide an incentive to eat more of those foods and less of the foods not covered by the reimbursement (see instructions in OA-E for more details).

TABLE 2: Lunch Categories

Green Category G	- salads (regular, kale, quinoa), soups (veggie, noodle) - natural fruits and low-fat yogurt - water (spring or sparkling)
Orange Category O	- sandwiches (bagels, wraps, baguette, club, hot sandwiches) - cereal bars, fruit bars or trail mix - fruit juice
Red Category R	- burgers, pizzas or fried foods (French fries, chicken wings, barbecue) - pastries (cookies, cakes, muffins, donuts, croissants) - soda

I elicited participants' weak preference ordering $\succeq$ over the set of reimbursement options $\mathcal{M} := \{G, O, R, GO, GR, OR, GOR\}$. Formally, each option can be seen as a menu of reimbursable foods; for instance, G is the singleton menu $\{G\}$ that commits a respondent to be reimbursed only for green foods, while GOR is the most flexible menu $\{G, O, R\}$ that allows to be reimbursed for all lunch categories. The full preference ordering $\succeq$ over $\mathcal{M}$ was elicited using an incentive compatible procedure. Respondents were asked to assign a rank number between 1 and 7 to each of the seven reimbursement options and could express indifferences by assigning the same rank to several options. To elicit a truthful report of the entire ordering, participants were told that their reimbursement option would be determined through a lottery assigning higher odds to higher ranked options.[6] To incentivize a respondent to truthfully express an indifference, the ranking procedure made it easier to report indifferences than strict rankings: participants first selected all the options they wished to assign rank 1 and then proceeded iteratively to assign all other ranks until the list of options was empty. The seven options appeared listed in a random order so as to

[6]The exact odds were (0.35, 0.3, 0.2, 0.1, 0.03, 0.02, 0) where $0.35 = \mathbb{P}\{\text{rank 1}\}$ and $0 = \mathbb{P}\{\text{rank 7}\}$. Options assigned the same rank received in expectation the same chances of being selected. For instance, if three options were assigned rank 1, then each option was equally likely to be drawn with probability 0.35, 0.3 or 0.2. Information about the specific odds was accessible to participants.

control for order effects. Participants were individually informed of their assigned option after having completed Survey 1.

3.2. Identification of temptation through menu preferences

3.2.1 Temptation and Self-Control à la GP

With such a dataset of menu preferences, one can construct a fairly rich language to talk about temptation. The idea of using menu choice to elicit temptation was introduced by Gul and Pesendorfer (2001) and investigated in several extensions (Dekel et al. (2009), Stovall (2010), Kopylov (2012); see Lipman and Pesendorfer (2013) for a review). The basic framework is a two-period expected utility model in which the decision maker (henceforth, DM) first chooses from a set of menus and then makes a choice from the selected menu at some later (unmodeled) stage. The primitive of those models is a preference relation $\succeq$ defined on a set $\mathcal{M}$ of menus (of lotteries). Besides standard assumptions in this environment, GP imposed on $\succeq$ a new behavioral axiom allowing to capture temptation.[7] This axiom called *Set Betweenness* states that for any two menus M and M',

$$M \succeq M' \text{ implies } M \succeq M \cup M' \succeq M'$$

This relaxation of the standard framework allows to capture the behavior of a tempted DM. To illustrate, consider an individual who would ex ante prefer to eat green foods rather than the more unhealthy red foods, $\{G\} \succ \{R\}$. A standard DM free of temptation (STD) evaluates a menu by its best element(s) and is unaffected by the presence of dominated options, implying $\{G\} \sim \{G, R\} \succ \{R\}$. On the other hand, a DM who is tempted by unhealthy foods would prefer to restrict his access to R than to leave himself the choice between G and R in Period 2, $\{G\} \succ \{G, R\}$. More generally, say that option x is a *temptation* in menu M if $M \backslash \{x\} \succ M$. In addition, say that x is *globally tempting* in $\mathcal{M}$ if $M \backslash \{x\} \succ M$ for all $M \in \mathcal{M}$ such that $x \in M$. For instance, R is a global temptation for an individual if he would prefer to exclude red foods from any reimbursement option.

Set Betweenness allows to capture the behavior of a DM who may prefer to restrict his choice set despite expecting to resist temptation with positive probability. This situation corresponds to the Strict Set Betweenness (SSB) ordering $\{G\} \succ \{G, R\} \succ \{R\}$. In the

[7] Regarding the other assumptions, $\succeq$ is required to be a complete and transitive relation that satisfies the expected utility axioms of continuity and independence adapted to a menu choice setting. Those axioms are not tested in the experiment. Completeness and transitivity follow directly from the rank ordering procedure, while continuity and independence are treated as maintained assumptions. For that reason, I will not discuss models that depart from GP by relaxing the independence axiom; see Fudenberg and Levine (2006, 2012) or Noor and Takeoka (2010, 2015) for some examples.

GP model, SSB reveals costly self-control: confronted with the choice, the DM expects to resist R by exerting self-control, thus preferring $\{G, R\}$ to $\{R\}$; however, self-control comes at a cost, which makes $\{G\}$ better than $\{G, R\}$. An alternative interpretation of SSB is that of "random indulgence" (Dekel and Lipman (2012)): the DM expects to succumb to R with some probability $p > 0$ and to have no temptation otherwise. In either case, we will say that R is revealed to be a resistible temptation for the DM.[8] When self-control costs are prohibitively high, the DM may expect to give in to R if offered the choice, thus assigning the same value to $\{R\}$ and $\{G, R\}$ i.e., $\{G\} \succ \{G, R\} \sim \{R\}$. In this case, R is revealed to be an overwhelming temptation for the DM. More generally, say that the DM (i) has *Self-Control* at two sets $(M,\ M')$ if $M \succ M'$ implies $M \succ M \cup M' \succ M'$ (SC) and (ii) has no *No Self-Control* if $M \succ M'$ implies $M \succ M \cup M' \sim M'$ (NSC). GP show that in the standard lottery framework of menu choice, imposing Set Betweenness on $\succeq$ leads to the following self-control representation

$$V_{GP}(M) = \max_{x \in M} [u(x) + v(x)] - \max_{y \in M} v(y)$$

The *commitment* utility u measures the normative preferences of the agent i.e., when committed to a singleton choice so that temptation concerns are absent. The *temptation* utility v measures the temptation value of an alternative and $c(x, M) = \max_{y \in M} v(y) - v(x)$ can be interpreted as the self-control cost of choosing x over the most tempting alternative in M. When choosing from M, the DM chooses as if he maximized the compromise utility $u + v$.

3.2.2 Multidimensionality of Temptation

Although simple and intuitive, the Set Betweenness axiom has strong implications for the structure of temptation: the cost of self-control only depends on the most tempting alternative in the menu, and not on the number of potential temptations. In some sense, Set Betweenness reduces temptation to a one-dimensional object. As argued by Dekel et al. (2009), this simplification rules out many behaviors, which may plausibly reflect temptation. To illustrate their point, consider a DM for whom $\{G\} \succ \{G, O\} \succeq \{G, R\} \succ \{G, O, R\}$. Despite violating Set Betweenness, there are at least two reasons why such an ordering might be natural for a dieting agent. First, resisting two temptations simultaneously might be strictly harder than resisting only one at a time; that is, the DM might suffer from cumulative temptation effects when facing both O and R. Second, the agent might be uncertain as to what option in O or R will be most tempting at the time of choice i.e, the source of temptation

[8] Dekel and Lipman (2012) show that costly self-control and random indulgence cannot be distinguished on the basis of menu choice only. In fact, for every costly self-control representation à la GP, there is an equivalent random indulgence (i.e., random Strotz) representation. Separating these two classes of models would require a much richer dataset of choices *between* menus and *from* menus. As actual food choices were not observed in this experiment, both classes of models offer a plausible interpretation of the SSB ordering.

might be stochastic. To study *cumulative* and *stochastic temptations* (henceforth, CT and ST), Dekel et al. (2009) suggested the following two relaxations of Set Betweenness:

Positive Set Betweenness (CT): $M \succeq M'$ implies $M \succeq M \cup M'$

Weak Set Betweenness (ST): If $\{x\} \succeq \{y\}$ for all $x \in M$, $y \in M'$ then $M \succeq M \cup M' \succeq M'$

The first relaxation extends the GP model to allow for a multidimensional cost of self-control:

$$V_{CT}(M) = \max_{x \in M} [u(x) + \sum_{j \in J} v_j(x)] - \sum_{j \in J} \max_{y \in M} v_j(y)$$

where $c_j(x, M) = \max_{y \in M} v_j(y) - v_j(x)$ can be interpreted as the self-control cost suffered along temptation dimension j (e.g., sugar) and the total cost of choosing x from M is given by $C(x, M) = \sum_{j \in J} c_j(x, M)$. As can be readily seen, the GP case corresponds to $|J| = 1$.

The second relaxation allows for a representation with a stochastic cost of self-control:

$$V_{ST}(M) = \sum_{v \in V} p(v) \{\max_{x \in M} [u(x) + v(x)] - \max_{y \in M} v(y)\}$$

where the DM must pay a self-control cost $c_v(x, M) = \max_{y \in M} v(y) - v(x)$ with probability $p(v)$ in order to choose x from M. The GP representation is obtained when $|V| = 1$.

These two different ways of extending the GP model were axiomatized by Dekel et al. (2009) and Stovall (2010).[9] Clearly, PSB and WSB are two independent axioms in the sense that neither implies the other. For instance, the ordering $\{G\} \succ \{R\} \succ \{G, R\}$ satisfies PSB but violates WSB. Therefore, both classes of models capture a different aspect of the multidimensional nature of temptation. At the same time, one can show that the two models cannot always be separated if only choices between (but not within) menus are observed (see Section 7.3 of Lipman and Pesendorfer (2013) for an example). It is thus important to understand how easily they can be distinguished in real-choice settings with limited data, and how well each can rationalize observed behaviors.

3.2.3 Commitment without Temptation, Temptation without Commitment

Although they allow for a wider range of temptation-driven behaviors, the above weakenings of Set Betweenness might be too permissive to solely capture temptation concerns i.e., without introducing other potential motives for commitment such as guilt or regret. For instance,

[9] More precisely, Dekel et al. (2009) show that a cumulative temptation representation is a finite additive EU representation that satisfies PSB (see their Theorem 3). Stovall (2010) shows that a pair of conditions equivalent to WSB given his other axioms characterizes stochastic temptation representations for the case in which $|V| < \infty$, thus proving an early conjecture of Dekel et al. (2009). Finally, Dekel and Lipman (2012) generalize the result proved by Stovall (2010) to the case where $|V| = \infty$, assuming $\succeq$ satisfies an additional continuity requirement.

the ordering $\{G\} \succ \{R\} \succ \{G, R\}$ might reflect the agent's aversion to guilt when choosing from R while the choice to act virtuously (by selecting something from G) was also available (see Kopylov (2012)). In the analysis, I therefore study the extent to which Set Betweenness violations of the form $M \succ M' \succ M \cup M'$ indeed reflect temptation concerns. To acknowledge the fact that motives other than temptation might explain those violations, I use the generic term "Global Commitment"(GC) when referring to the ordering $M \succ M' \succ M \cup M'$.

Finally, while the models discussed here differ in their assumptions on the structure of commitment choices (and thus the shape of temptation), they all make the stark assumption that the normative preference u is known and one-dimensional.[10] Behaviorally, this means that the DM can never exhibit a strict *Preference for Flexibility* ($FLEX$) $M \cup M' \succ M \succ M'$ at any two choice sets M and M'. As a result, none of the above models can capture a DM who may have a preference for diversity or may feel uncertain about his future tastes, an idea originally motivated by Kreps (1979). In practice, individuals often have to balance a desire to restrict their access to temptations with a need to accommodate uncertainty and preserve choice variety. This tension between commitment and flexibility might justify orderings of the form $\{G, R\} \sim \{G, O\} \succ \{G\} \succ \{G, O, R\}$ through which the DM reveals that he would rather not have both O and R available, but that G alone is too restrictive.

In summary, the link between commitment and temptation is not straightforward: on the one hand, commitment might reveal other motives than temptation; on the other hand, temptation might not always translate into a commitment choice. As a result, data from a single binary comparison is unlikely to give an accurate picture of the prevalence of temptation concerns. Below I exploit data on the entire ordering to extract this information.

3.3. Findings

In this section, I analyze individual rank orderings to study the extent to which preferences in the reimbursement program exhibit temptation. Besides assessing its prevalence, I study three properties of temptation based on the discussion from the previous section:

Source: When participants prefer a restricted coverage, what food categories tend to be excluded? This question about the source of temptation can be asked because the elicitation procedure imposed a priori no restrictions on when commitment demand should arise.

Strength: What lunch categories appear as "robust temptations" in the sense of being classified as globally tempting? More generally, comparing two nested options, how often would

[10] Identification in models of menu choice becomes extremely challenging if one allows the DM to simultaneously experience uncertainty about his normative preferences and uncertainty about his temptation preferences; see Dekel et al. (2009) and Stovall (2018) for a discussion of this identification problem.

a participant prefer the option that excludes a given food category from the coverage?

Structure: When commitment is strictly preferred, what form does it take? This question refers to the consistency of commitment choices with Set Betweenness and, more generally, to the compatibility of preference orderings with the temptation models presented above.

To provide a reference point, the findings below will be compared against a random benchmark. Overall, 82.3% of respondents gave a strict ordering of the seven options (thus completing their ranking in 7 steps), while the remaining expressed at least one indifference.[11] To preserve the structure of indifferences observed in the dataset, the benchmark consists of 1,000 random permutations of the ranks that each of the 113 participants assigned to the seven options. The observed preference patterns will be compared to the pooled distribution of these 1,000 permutations ($N = 113,000$).

3.3.1 Source and strength of temptation

Figure 1 shows the mean rank (1-7) assigned to each reimbursement option; the full rank distributions are reported in OA-A.1. Option GO is by far the most popular option followed by GOR and G; these three options have a significantly lower mean rank than the other four options ($p < 0.001$, two-sided t-tests) and the benchmark. As would be expected from individuals trying to lose weight, OR and R are the least popular options and perform significantly worse relative to the benchmark. Options O and GR appear to be more neutral options, with respectively 60% and 72% of respondents assigning them rank 3, 4 or 5.

Table 3 shows the proportion of times G, GO and GOR were strictly preferred to any of the other six options. Almost one third of respondents expressed a strict preference for GO, while 15% of respondents strictly preferred to restrict their choice to G. Therefore, close to 50% of people expressed a desire for commitment to eating healthy by excluding R from the reimbursement coverage, while keeping G. At the same time, about 32% of respondents selected GOR as their unique rank 1 option. Together, these findings highlight a tension between commitment and flexibility, and suggest that commitment take-up may be limited if it requires the individual to greatly restrict his choice set. Removing GO from the set of options, only 32% of respondents would strictly prefer to commit to G, a take-up rate for commitment closer to previous findings in the literature, where commitment rarely presents different levels of flexibility (see OA-A.1).

[11] See OA-A.1 for more details. The percentage of strict orderings generated in this study is somewhat higher than the one observed in Toussaert (2018) (between 53% and 75% depending on the classification adopted). These different percentages could be due to several reasons such as differences in the elicitation procedure, the number of menus to rank (7 vs. 3), or specific features of the menu items.

FIGURE 1: Mean rank assigned to each reimbursement option

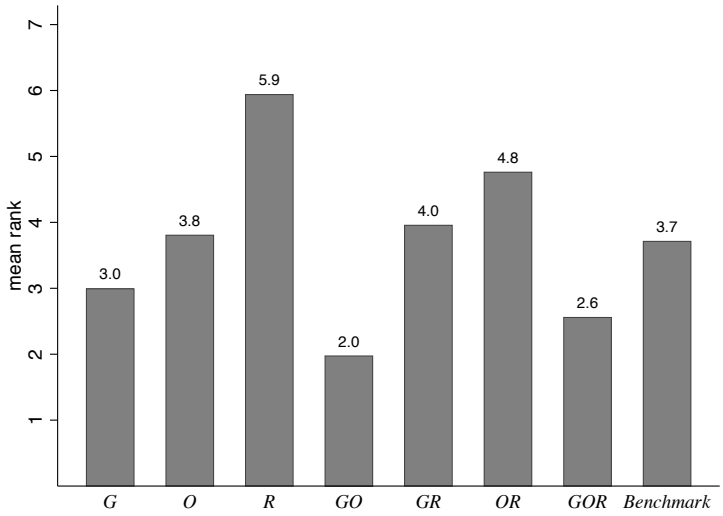


TABLE 3: Distribution of top choices

Top option	Actual sample % (N)	Benchmark %	p-value
G	15.0 (17)	12.3	0.388
GO	32.7 (37)	12.3	< 0.001
GOR	31.9 (36)	12.3	< 0.001
Other	20.4 (23)	63.1	< 0.001
Total	100 (113)	100.0	

Notes: "Other" refers to participants who either (i) had their unique top in the set $\{O, R, GR, OR\}$ (respectively 3, 1, 1, and 2 respondents) or (ii) assigned rank 1 to several options (16 respondents). The p-values are the result of a two-sided binomial test that the observed frequency is equal to the benchmark frequency.

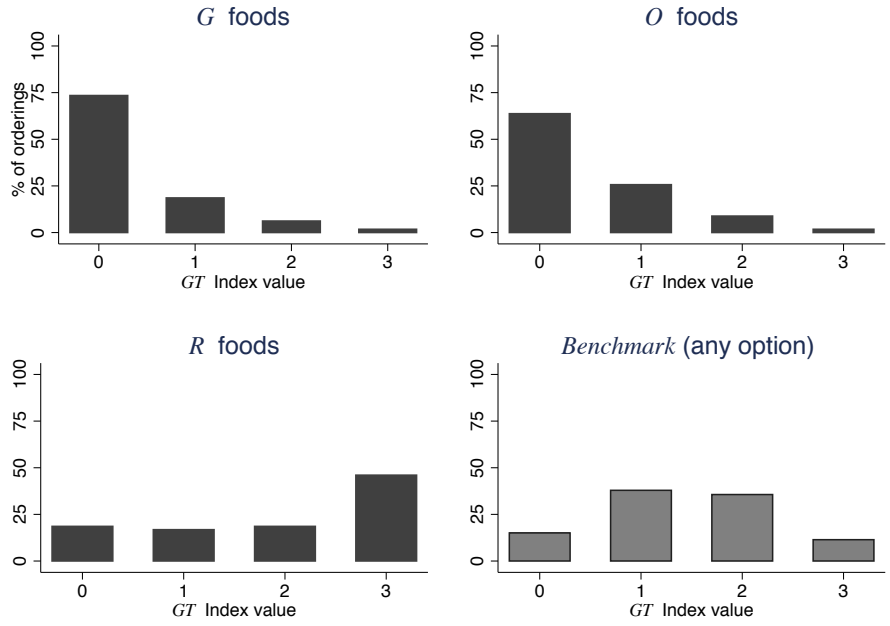
To study the strength of temptation, I now define the following Global Temptation Index for food category $x \in \{G, O, R\}$

$$GT_{-x} = \sum_{\mathcal{M}_x} \mathbb{1}_{\{M \setminus \{x\} \succ M\}}$$

where $\mathcal{M}_x = \{M \in \mathcal{M} \mid x \in M \text{ and } M \neq \{x\}\}$. For instance, $\mathcal{M}_R = \{GR, OR, GOR\}$ and $GT_{-R} \in \{0, 1, 2, 3\}$ measures the number of times a given respondent strictly preferred to eliminate R from a reimbursement option that includes it. Therefore, GT_{-x} measures

the "strength" or "robustness" of an individual's temptation for x; in particular, $GT_{-x} = 3$ implies that x is globally tempting in $\mathcal{M}$. Figure 2 shows the distribution of the Global Temptation Index for each lunch category as well as the distribution that would be observed under the benchmark.

FIGURE 2: Distribution of the Global Temptation Index for G, O and R



Red foods appear to be tempting for a large majority of participants and globally tempting for 46% of respondents, a proportion four times higher than under the random benchmark. Unsurprisingly, green foods are not revealed to be a temptation ($GT_{-G} = 0$ for 73% of respondents). More surprisingly, O is also rarely revealed to be tempting ($GT_{-O} = 0$ for 64% of respondents), and the distributions of the GT_{-O} and GT_{-G} indices are not significantly different from each other ($D = 0.097$, $p = 0.658$). Below I show that temptation by x as measured by participants' preference for a coverage that excludes x is consistent with their subjective assessment of how tempting x is. Yet, very partial information on the preference ordering over $\mathcal{M}$ - such as information on the choice between G and GOR - would not be sufficient to infer the source of temptation and its strength.

3.3.2 Structure of temptation

I now study the structure of temptation by assessing the consistency of commitment choices with the Set Betweenness axiom and its relaxations. To this end, I look at all comparisons

between any two non-nested reimbursement options, making 9 pairwise comparisons in total.[12] Following the discussion of Section 3.2, I consider five categories of menu preferences depending on how the individual ranks the union of any two non-nested menus M and M':

1- Standard (STD): $M \succ M'$ implies $M \sim M \cup M' \succ M'$

2- Flexibility-loving ($FLEX$): $M \succ M'$ implies $M \cup M' \succ M \succ M'$

3- No Self-Control (NSC): $M \succ M'$ implies $M \succ M \cup M' \sim M'$

4- Self-Control (SC): $M \succ M'$ implies $M \succ M \cup M' \succ M'$

5- Global Commitment (GC): $M \succ M'$ implies $M \succ M' \succ M \cup M'$

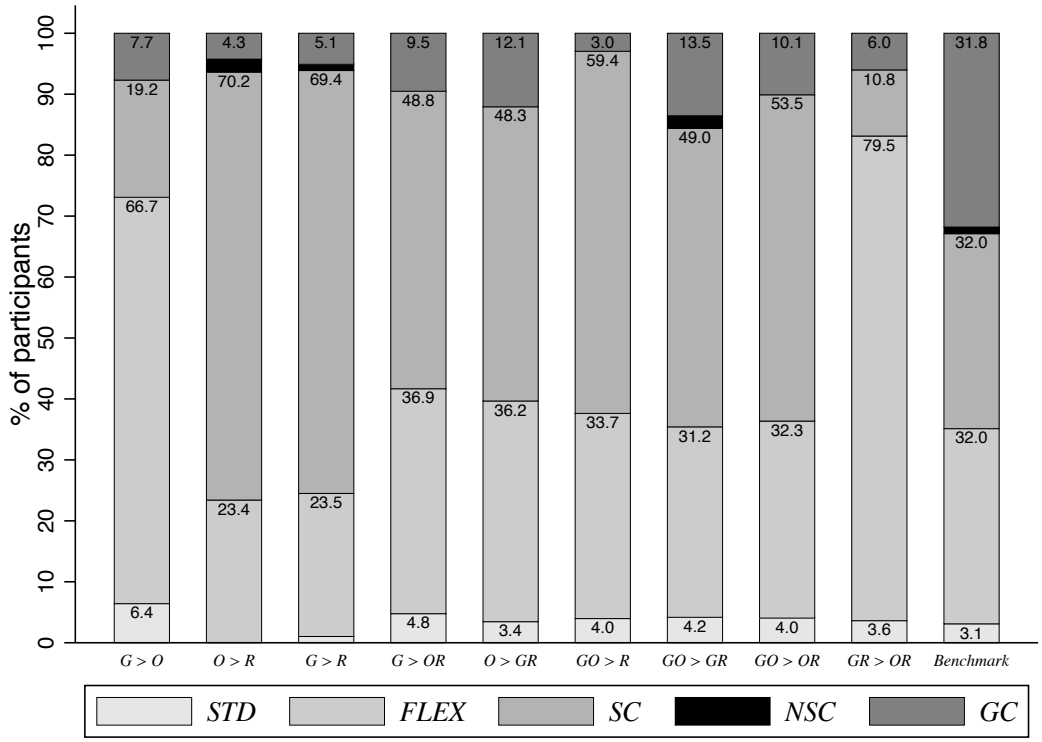
Remember that while 1, 3 and 4 are consistent with Set Betweenness, 2 and 5 are not. Furthermore, only 3, 4 and 5 are consistent with temptation. Figure 3 shows the proportion of respondents who behaved according to each of these five categories for the most frequent binary choices (see OA-A.2.2 for a complete breakdown).

There are four main findings. First, due to the low percentage of expressed indifferences, the fraction of choices satisfying STD or NSC is very low. Second, a large percentage of respondents behaved either according to $FLEX$ or SC in each of these binary choices. For instance, about 67% of those who ranked G strictly above O also ranked GO strictly above these two menus, thus favoring flexibility. On the other hand, 69% of those who ranked G strictly above R placed GR strictly in between. Third, commitment is mostly consistent with the strict form of Set Betweenness (SC), rather than with a multiple temptation interpretation (GC); in fact, GC preferences appear at a much lower frequency than under the random benchmark. Finally, whether commitment or flexibility prevails depends on the presence or absence of R in the menus being compared. When either both menus contain R or both exclude R, a majority of participants prefer flexibility. On the other hand, participants who strictly prefer a menu that does not contain R to a menu that contains R tend to favor commitment (SC or GC). Thus, commitment is essentially motivated by a willingness to exclude R from the set of options, which is consistent with R being a tempting alternative for most participants. Flexibility is instead favored when respondents face a trade-off between G and O, all else constant.[13]

[12] Comparisons between one menu and a proper subset of it are excluded from the analysis since the implications of Set Betweenness are trivial in this case.

[13] This observation also holds for the least frequent binary choices. For instance, only 18.5% of the 27 respondents who strictly preferred O to G put GO strictly in between. As another example, only 12% of the 50 participants who placed GR strictly above O also put GOR strictly in between. Putting all 9 comparisons together in a regression framework, respondents are 22-26 (15-21) percentage points more likely to exhibit SC preferences when comparing two menus M and M' such that $M \succ M'$ and $R \notin M$ ($R \in M'$); additional regressions show that it is the interaction of the two conditions $R \notin M$ and $R \in M'$ that predicts SC

FIGURE 3: Distribution of menu preferences in bilaterial comparisons



Notes: Distribution of menu types for the most frequent preferences over two options (M, M'); see OA-A.2.2 Table 6 for the least frequent preferences. For instance, the number 66.7% on the first bar means that among the respondents who ranked G strictly above O ($G \succ O$), 66.7% had the $FLEX$ ordering $GO \succ G \succ O$.

Aggregating over all binary comparisons, I now construct an index for each food category $x \in \{G, O, R\}$, which measures how frequently a given respondent exhibits a temptation for x in the form of SC, the strict version of Set Betweenness. For this purpose, let $\tilde{\mathcal{M}}_x = \{M \in \mathcal{M} \setminus \{GOR\} \mid x \in M\}$ and $\tilde{\mathcal{M}}_{\neg x} = \{M \in \mathcal{M} \setminus \{GOR\} \mid x \notin M\}$. Then the Strict Set Betweenness Index for food category x is computed as

$$SSB_{-x} = \sum_{\mathcal{P}_x} \mathbb{1}_{\{M \succ M \cup M' \succ M'\}}$$

where $\mathcal{P}_x = \{(M, M') \in \tilde{\mathcal{M}}_{\neg x} \times \tilde{\mathcal{M}}_x \mid M \not\subseteq M' \text{ and } M' \not\subseteq M\}$. For example, $\mathcal{P}_R = \{(G, R), (O, R), (G, OR), (GO, R), (O, GR), (GO, GR), (GO, OR)\}$ and $SSB_{-R} \in \{0, 1, ..., 7\}$ measures the number of times R is revealed to be a resistible temptation. The distribution of the index for each food category is presented in Appendix Figure 8, together with the

preferences. No such pattern can be found for GC preferences. See OA-A.2.2 for more details.

benchmark distribution. Consistent with the previous analysis, green and orange foods score very low on this index, with SSB_{-G} (SSB_{-O}) being equal to 0 or 1 for 93% (83%) of respondents. On the other hand, R is widely revealed to be a temptation {a la GP, with SSB_{-R} taking a value of 5 or higher for close to 40% of respondents and 83% of those for whom R is globally tempting (see OA-A.2.2 for more details).

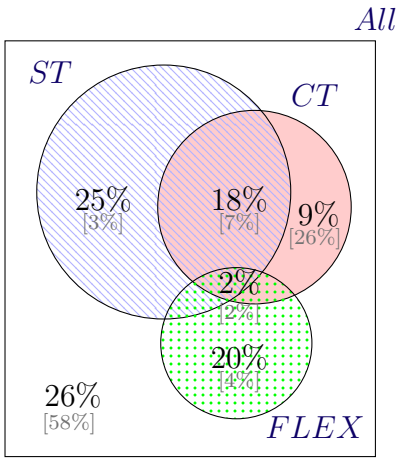
As a final step, I now test the consistency of individual preference orderings with the temptation models discussed in Section 3.2. To this end, I use data on the entire ordering to construct falsification tests of the Set Betweenness axiom (SB) and its relaxations, Weak Set Betweenness (WSB) and Positive Set Betweenness (PSB). As explained above, the SB axiom originally proposed by Gul and Pesendorfer (2001) constrains temptation to be a one-dimensional object, while the PSB and WSB axioms of Dekel et al. (2009) acknowledge the multiple facets of temptation. In particular, WSB allows for stochastic temptations (ST) that may affect the agent at different times, while PSB captures cumulative temptations (CT) that may affect the agent simultaneously. The three axioms put specific structure on commitment choices, leaving no room for the expression of preference for flexibility. To contrast commitment with flexibility, I also test the monotonicity axiom (MON) of Kreps (1979), which states that $M \subseteq M'$ implies $M' \succeq M$. In words, MON says that larger menus are always preferred to their subsets i.e., the agent prefers to maintain flexibility by keeping options open.

To count the number of violations of SB, PSB and WSB, I look at all comparisons of any two non-nested menus, implying 9 potential violations in total. For MON, I look at all comparisons of two nested menus, bringing to 12 the number of possible violations. Since some axioms are easier to violate than others, I contrast the number of violations observed in the actual sample with the number generated in the random dataset. More information on the falsification tests can be found in OA-A.2.3. The distribution of violations for each axiom is presented in Appendix Figure 9. Preferences in the actual sample appear far more consistent with SB, WSB, and $FLEX$ than preferences in the random benchmark. The opposite is true for PSB; in fact, the PSB axiom is frequently satisfied in the random sample and thus appears to be quite weak. Despite these observations, only a third of rank orderings are fully consistent with one of the four axioms (17% for benchmark). Because each axiom allows for either flexibility or commitment concerns but not both, this finding stresses the importance of integrating the two concerns into a single framework, rather than modeling them separately.[14]

[14] In turn, modeling the trade-off between commitment and flexibility requires to understand the specific conditions under which each concern might prevail. For instance, SB appears to perform quite well when comparing two singleton options, while its performance seems weaker for the other comparisons (see OA-A.2.3). By contrast, MON might perform better when the options under comparison contain many and/or complex objects, leading the DM to defer his choice rather than commit; see for instance Tversky and Shafir (1992) for empirical evidence and Pejsachowicz and Toussaert (2017) for a theoretical discussion.

To understand the overlap between models, I finally classify participants into types (ST, CT, $FLEX$) depending on whether they satisfy the corresponding axiom (WSB, PSB, MON) in all comparisons except possibly one. Allowing for one violation increases the number of participants who can be classified from one third to nearly 75%. Figure 4 shows the distribution of types. Of the 113 participants, about 20% exhibit preference for flexibility, a proportion which is five times higher than the benchmark frequency ($p < 0.001$, two-sided binomial test). However, commitment concerns overall dominate, with a majority of participants exhibiting temptation-driven preferences (52% vs. 36% for benchmark, $p = 0.001$). Interestingly, actual preferences appear far more consistent with stochastic temptation (WSB) than with cumulative temptation (PSB), contrary to benchmark preferences. There is nevertheless significant overlap between the two classes of models, with 18% of orderings consistent with both classes. Among them, nearly two thirds satisfy SB (about 11% of the total sample, $p < 0.001$ against 1% benchmark).[15]

FIGURE 4: Consistency of individual orderings with existing models



Notes: Typology of participants when allowing for at most one violation of each axiom ($N = 113$). Category ST (respectively, CT and $FLEX$) refers to participants who violate WSB (respectively, PSB and MON) no more than once. Participants at the intersection of all three categories are fully indifferent. The numbers in square brackets refer to benchmark frequencies.

[15] In OA-A.2.3, I show how the typology changes if one allows for (*i*) no violation, or (*ii*) two violations. The proportion of individuals classified as having temptation-driven preferences ($TEMPT$) is stable at around 50% when going from one to two violations; however, consistency with SB increases from 10% to 30%. Although not allowing for any violation is a strict requirement, a relaxation of this criterion may come at the cost of weakening the identification of temptation. One sign of this potential problem is the positive correlation between the SSB_{-G} score and being classified as $TEMPT$, since a higher SSB_{-G} score may arguably capture other concerns than temptation.

4. Linking Commitment to Temptation

The previous section presented evidence of commitment demand in the context of a lunch reimbursement program. While GOR was the option that maximized meal reimbursement, only 32% of participants strictly favored this option. Among the others, the modal preference was for excluding only the most unhealthy foods from the coverage. In particular, R was revealed to be a strong temptation for most participants, but not O. Finally, temptation by R presents a high level of consistency with the simple structure imposed by Gul and Pesendorfer (2001). Together these findings suggest that research on the elicitation of menu preferences may offer a promising venue for measuring temptation in real life settings. In this section, I test whether temptation revealed through menu choice is in line with participants' perceptions of what is tempting (4.1) and study its power to predict self-control problems in other choice contexts (4.2).

4.1. Survey validation of menu choice measures of temptation

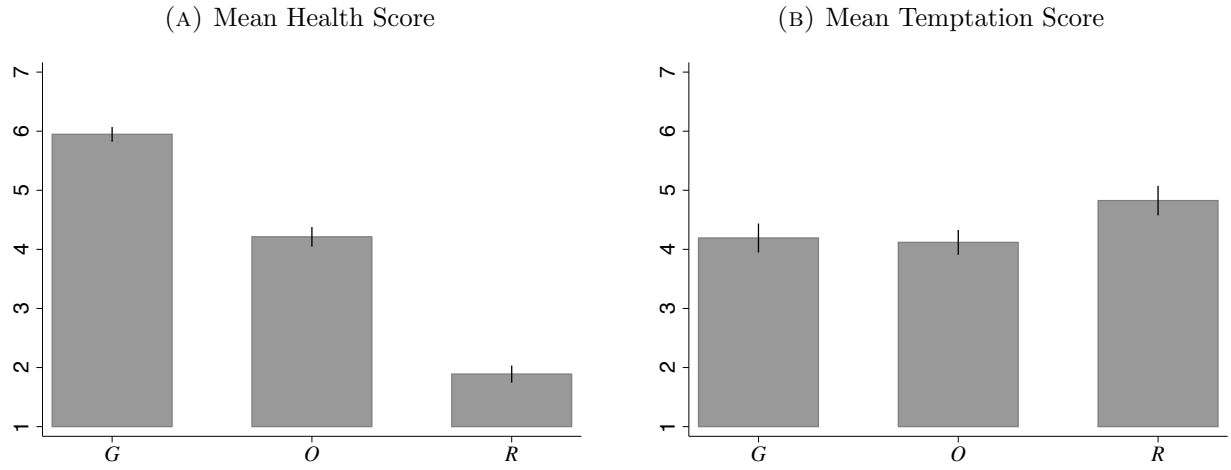
The fact that a participant chose to exclude R from the reimbursement coverage could a priori reflect other concerns than avoiding temptation. Respondents might have chosen to remove from the reimbursement the unattractive foods that are never part of their diet, and are thus irrelevant to their choice of coverage. At the same time, unfrequent consumption of red foods does not mean per se absence of temptation if a respondent usually resists alternatives he craves.[16] The question is therefore whether a significant fraction of commitment choices can be explained by food tastes rather than temptation concerns. To investigate this issue, I use data collected in Surveys 1 & 2 about subjective perceptions of the various food alternatives, consumption habits and attitudes, and respondents' comments about their favorite reimbursement coverage. All sources of data support the hypothesis that commitment to a large extent reflects the temptation to eat unhealthy.

The first source of evidence pertains to participants' subjective perception of each lunch category. After submitting their ranking, respondents were asked to rate a list of food items, each belonging to one of the three lunch categories, in terms of (i) *how healthy* and (ii) *how tempting* they considered this item (1-7 scale).[17] Each lunch category was then assigned a Health (Temptation) Score computed as the mean health (temptation) rating over all food

[16] This is a major implication of models of costly self-control: even agents who would resist temptation in the absence of restrictions might prefer to be committed ex ante. For instance, someone who always orders salads when other alternatives are available on the menu could be better off if the menu exclusively offered salads. See Toussaert (2018) for a more extensive discussion of this point.

[17] The food items were: salad, soup, yogurt and fruit (category G); cold sandwich, hot sandwich and cereal bar/trail mix (category O); burger, pizza, fried food and pastry (category R).

FIGURE 5: Perceived health and temptation value of G, O and R foods



Notes: The error bars are 95% confidence intervals obtained from linear regressions of the Health (Temptation) score on dummies for the lunch category; standard errors clustered at the subject level ($N = 113$).

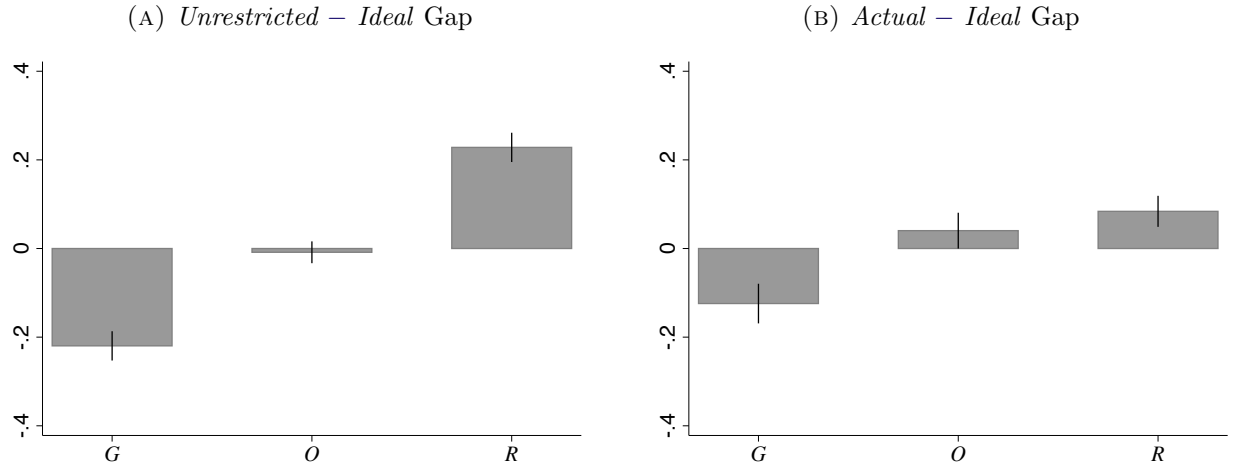
items in that category. Figure 5 shows the average Health and Temptation scores over all participants, while Appendix Table 5 presents a breakdown by menu preference (top choice and value of GT_{-R}). Respondents' ratings of the various food items are consistent with R being a tempting but unhealthy category relative to G and O. First, respondents unambiguously perceived G as the healthiest category and R as the most unhealthy, with O being moderately healthy. Second, while O was not rated as more tempting than G, respondents assigned a significantly higher temptation value to R than to both G and O. Furthermore, the Health and Temptation Scores of G, O and R do not significantly differ based on respondents' ranking of the various reimbursement options. The ratings of participants whose top choice was one of the two most popular options (GO or GOR) are very similar; in particular, R was rated as significantly more tempting than G and O in both cases ($p < 0.05$ in all comparisons).[18]

Turning to the second source of evidence, participants were asked to report on a 0-100 scale (*iv*) how often they had each food item for lunch since the beginning of the year (Survey 1); (*v*) how often they thought they should consume each item (Survey 2); and (*vi*) how

[18] As shown in Appendix Table 5, the mean Temptation Score for G (R) appears to be highest (lowest) for respondents with stronger commitment preferences (i.e., G *top* and $GT_{-R} = 3$); however, no differences are significant on any pairwise test. Because (*ii*) might measure taste rather than temptation per se, I asked a follow-up question to the 87 participants who took Survey 2, which explicitly defined tempting foods as foods that are hard to resist but that should be avoided. Differences in mean rating between R and the other two categories are more pronounced, as are the differences across menu preference categories; see OA-A.4 for more details.

often they would consume each item if unrestricted (Survey 2). In the spirit of Ameriks et al. (2007), I study temptation by measuring the extent to which actual (unrestricted) consumption deviates from ideal consumption. In models of costly self-control à la GP, one can think of ideal consumption as maximizing u, unrestricted consumption as maximizing v, and actual consumption as maximizing $u+v$. Letting $s_w(j)$ be the share in total consumption of food category $j \in \{G, O, R\}$ that maximizes utility $w \in \{u, v, u+v\}$, one can then define the *Actual* $-$ *Ideal* gap for j as $\Delta_{A-I}(j) := s_{u+v}(j) - s_u(j)$ and the *Unrestricted* $-$ *Ideal* gap as $\Delta_{U-I}(j) := s_v(j) - s_u(j)$. Both gaps should be positive for tempting foods; furthermore, $\Delta_{A-I} \leq \Delta_{U-I}$ if individuals exert self-control to resist temptation (with $\Delta_{A-I} = \Delta_{U-I}$ for overwhelming temptations). Figure 6 shows the mean *Unrestricted* $-$ *Ideal* and *Actual* $-$ *Ideal* gaps for each lunch category, while Appendix Table 5 presents a breakdown by menu preference ($N = 87$).[19]

FIGURE 6: Mean deviation from ideal consumption



Notes: The error bars are 95% confidence intervals obtained from linear regressions of the *Unrestricted* $-$ *Ideal* (*Actual* $-$ *Ideal*) gap on dummies for the lunch category; standard errors clustered at the subject level ($N = 87$).

The *Unrestricted* $-$ *Ideal* gap for R is large and positive: in the absence of restrictions, participants would on average increase their consumption share of red foods by more than 20 percentage points relative to their ideal "should" consumption ($\bar{\Delta}_{U-I}(R) = 0.23$, $t = 13.7$, $p < 0.001$). The size of the gap is similarly large across the various menu preference categories. By contrast, $\Delta_{U-I}(O) \approx 0$ for all menu preferences, suggesting that O is not a source of temptation for participants. This last observation implies that the positive gap for

[19] More information on the mean actual, ideal, and unrestricted consumption shares of each lunch category can be found in OA-A.4, both overall and by menu preference. In each of the three questions, the consumption share of food category $j \in \{G, O, R\}$ was computed as $s(j) = \frac{f(j)}{f(G)+f(O)+f(R)}$, where $f(j)$ is the average of a respondent's answers to that question for all foods belonging to category j.

red foods is nearly perfectly counterbalanced by a negative gap for green foods ($\bar{\Delta}_{U-I}(G) = -0.22$, $t = -13.2$, $p < 0.001$). As would be expected if self-control is exerted, the *Actual* − *Ideal* gap for red foods is significantly smaller than the *Unrestricted* − *Ideal* gap, but remains positive ($\bar{\Delta}_{A-I}(R) = 0.08$, $t = 7.86$, $p < 0.001$). Although not significantly different from zero in a few cases, $\Delta_{A-I}(R)$ is positive and fairly comparable across all menu preferences. On the other hand, there is some noticeable variation in $\Delta_{A-I}(G)$ across preference categories: participants with stronger commitment preferences (*G top* and $GT_{-R} \geq 2$) consume green foods in a proportion closer to their ideal consumption share than others. This variation is almost entirely due to differences in how participants distribute their consumption between the G and O categories, as the actual consumption share of red foods is fairly stable across menu preferences (see OA-A.4 for more details).

The last source of evidence concerns participants' explanations of their ranking. In Survey 2, respondents were asked to explain why they chose or did not choose to assign rank 1 to the most flexible option GOR. Out of 87 respondents, 72 provided motivated answers, which form the basis of the analysis.[20] Among those who preferred a restricted coverage, nearly half explicitly mentioned that they wanted to be motivated to eat healthier and/or discouraged to eat unhealthy foods. Another 42% explained having ranked the options based on their preferences or current consumption habits, most often specifying that they essentially eat green foods and/or rarely eat red foods. The remaining 11% mentioned a desire not to be reimbursed for unhealthy foods or to eat healthy, but without being more specific. However, for the last two categories of respondents, information about their food tastes allows to discard the hypothesis that commitment is orthogonal to temptation. Participants were asked to indicate the five food items they craved the most; unsurprisingly, all respondents who strictly preferred G or GO mentioned at least one craving belonging to category R.[21]

Together, the findings of this section support the hypothesis that commitment to a restricted coverage is largely taken as an incentive mechanism to shift consumption away from foods perceived as tempting but unhealthy.

4.2. Predictive power of menu choice measures of temptation

I now investigate whether the menu preference measures of temptation developed in Section 3.3 are predictive of several behaviors during the challenge that could reflect self-control problems. First, I examine the link between menu preferences and take-up of a goal setting contract. I then look at the relationship between menu choice and goal achievement. Finally,

[20] See respondents' comments in OA-A.5. The 15 remaining respondents either did not respond to the question, mentioned that they did not remember the ranking procedure/were not sure of their choice, or answered something unrelated to the question. The entire set of comments is available upon request.

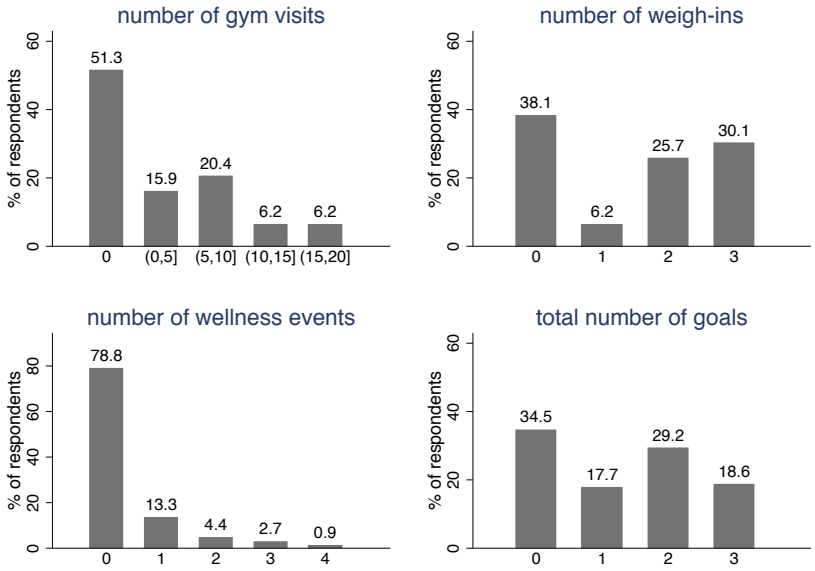
[21] The cravings they most commonly cited were pastries (8), chocolate (7), chips/fries (7) and pizza (5).

I discuss the link between participants' preferences in the reimbursement program and their likelihood of completing the challenge, completing the study, and claiming reimbursement.

4.2.1 Goal setting

Description of the contract In Survey 1, participants were offered to receive their study payment ($20 gift card) only if they achieved self-set attendance goals. The goals belonged to three categories: number of gym visits (during a four-week period), number of follow-up weigh-ins (out of three), and number of wellness events (out of four). Participants were free to set a goal in none, one, or multiple categories; in case they did not want to set a goal, they were required to enter 0 for the corresponding category. It was emphasized that setting goals could help stay on track during the challenge, but that participation was entirely voluntary (see instructions in OA-E). To make the commitment credible, participants were informed before setting their goals that their attendance would be verified at the end of the challenge. They were also asked to confirm that they understood the contract terms before validating their answers. After completing Survey 1, participants were emailed a summary of their commitment decisions to confirm or readjust their choices.[22] In addition, they received a reminder of their goals before each weigh-in. Figure 7 shows the proportion of participants who committed to one or multiple goals as well as the distribution of goals by category.

FIGURE 7: Distribution of goals



[22] Six of the 113 participants asked to adjust their goals; more information is available in OA-B.1.

Contract take-up is high: despite the absence of any financial reward for reaching their goals, about 65% of respondents committed to at least one goal, and 73% of them committed to multiple goals.[23] The selected targets were non trivial. Among the 55 respondents who set a gym attendance goal, the mean target was about 9 visits (2.25 visits per week). Attending the weigh-ins was the most popular goal: 62% (70/113) of participants chose to commit to at least one weigh-in, with about half of them setting the goal to attend all follow-up weigh-ins. On the other hand, only 21% (24/113) chose to commit to attending a wellness event.

Menu preferences and goal setting I now test whether a preference for removing unhealthy foods from the reimbursement coverage predicts take-up of the goal setting contract. In theory, the correlation in commitment demand across the two decision domains could be positive, negative or zero depending on what subjects believe about (*i*) their self-control in both domains, and (*ii*) the separate and joint efficacy of the two commitment technologies. If self-control presents some domain generality - e.g., participants face and anticipate similar struggles with food and exercise - and the two technologies are perceived as complementary tools to achieve weight loss, then one should observe a positive correlation between commitment decisions. On the other hand, the correlation might be negative if participants doubt their ability to successfully commit in both domains at the same time (for instance, because self-control is limited) or if the two commitment strategies are perceived as substitutes. Finally, the correlation might be zero if participants have different beliefs about their self-control problems in the two domains or about the efficacy of each technology. To study this relationship, I consider the following linear probability model

$$\mathbb{1}(\text{set a goal})_i = \mathbf{X}_i\beta + \mathbf{M}_i\gamma + \epsilon_i$$

where $\mathbf{X}_i$ is a vector of individual controls and $\mathbf{M}_i$ measures participant i's menu preferences in the reimbursement program. Following the analysis of Section 3.3, I consider the following menu preference measures: dummies for i's top choice (*G top*, *GO top*, *Other top*; *GOR top* is treated as the reference category), i's score on the Global Temptation Index GT_{-x} for each food category $x \in \{G, O, R\}$, and i's score on the Strict Set Betweenness Index SSB_{-x} for each category $x \in \{G, O, R\}$.[24] In the baseline regressions, $\mathbf{X}_i$ only includes a control

[23] Due to a programming error, some of the participants who responded to Survey 1 on the first day it was administered managed to move to the next survey section without having entered a response for one or several of the goal categories; non-response was recorded as 0. The error was corrected within the first 24 hours and all respondents were emailed a summary of their choices in order to confirm their decisions. However, the 47% (18/38) contract take-up recorded on Day 1 is significantly lower than the 75% (56/75) recorded for the subsequent days ($z = 2.82$, $p = 0.004$); as a result, the overall take-up rate of 65% is likely to be a lower bound estimate of the actual take-up. To account for these differences in take-up, the econometric analysis presented below controls for the survey completion date. See OA-B.1 for more details.

[24] More information about the relationship between these variables can be found in OA-A.2. I do not consider variables that measure the overall frequency of commitment choices (such as measures of consistency with PSB or WSB), since the analysis of Sections 3.3 and 4.1 identified R as the only real temptation.

for the day at which the participant completed Survey 1. Additional regressions control for a richer set of socio-demographic and challenge-related variables collected before the goal setting decisions were taken (see Appendix Table 4 for a description of these variables).

Regression results are presented in Appendix Table 6; probit regressions yield similar results. Participants who revealed a temptation for R through their menu preferences were overall more likely to take up the goal setting contract. As shown in columns (1) & (2), participants who strictly preferred GO were 19-23 percentage points more likely to commit to a goal than those who strictly preferred no restrictions; the effect of a strict preference for G is however not significant and has the wrong sign. There is also a positive and significant relationship between take-up of the goal setting contract and the strength of the preference for removing R, with a 7-10 percentage point increase in take-up for each additional point on the GT_{-R} score. Importantly, the effect of the GT_{-R} score is stronger and more precisely estimated after controlling for the corresponding scores for categories G and O (while the latter two have no predictive power). Findings are similar, although slightly weaker, if one considers the SSB_{-R} index. In other words, it is temptation by R - and not simply commitment in general - that predicts take up of the goal setting contract. Looking at the effect of the other covariates, goal setting is positively correlated with age (weakly), but negatively correlated with higher education.[25] Interestingly, there is a concave relationship between goal setting and participants' confidence that they will reach their weight loss target; this result is in line with Royer et al. (2015) who find that commitment demand is highest among those with a medium confidence in their capacity to reach their goals.[26]

4.2.2 Goal achievement

Attendance Data Data regarding attendance of the weigh-ins and the wellness events was collected through the wellness services of NYU. The gym attendance data comes from two different sources. First, the gym provided data on the number of times a participant scanned his gym badge, a requirement to access the facilities. However, the system retrieved only 69 of the 113 names.[27] Furthermore, the number of scans is only a proxy for gym

[25] Unlike prior studies that find a positive correlation between education and commitment take-up (Ashraf et al. (2006), John (2019)), note that the population of this study is highly educated (almost 60% have more than a bachelor degree); thus, the effect of education may have a very different interpretation.

[26] The regression analysis of this section focuses on the extensive margin i.e., on the decision to set a goal or not. In OA-B.2, I study both the extensive and intensive margins through a set of regressions that take the target number for each goal category as the outcome variable. The relationship between menu preferences and goal setting tends to be weaker for target levels than contract take-up. Findings are also somewhat weaker for the number of goals set; a regression analysis of the latter outcome is available upon request.

[27] Participants who were found in the system do not significantly differ in terms of their observable characteristics from those who were not, except for their lower likelihood of having participated in a prior edition of the challenge; see OA-B.4.2 for more details. All regressions control for prior participation.

attendance because participants could have checked in without exercising, as could have been the case on weigh-in days. Therefore, most respondents were also asked to report their gym attendance.[28] Among the 44 participants for whom the data could not be retrieved, 20 had set a gym attendance goal. These participants received their gift card provided they had achieved their other goals, which was only the case for 9 of them. Gym reports were collected after payments were announced, which should have lowered incentives to misreport. When data from both sources was available, the final estimate was taken to be the minimum between the self-declared report and the number recorded by the gym system. The final gym attendance data consists of 112 observations, due to one missing report.[29]

On average, participants used the gym about 5.9 times (1-2 times per week), with a higher attendance of those who committed to a specific number of gym visits compared to those who did not (7.2 vs. 4.4 visits, $t = 2.35$, $p = 0.021$). Among the goal setters, 44.4% (24/54) achieved or surpassed their exercise goal. Participants attended an average of 1.4 weigh-ins, again with a significantly higher participation of the goal setters (1.6 vs. 1.0 weigh-ins, $t = 2.84$, $p = 0.005$). Among the latter, 48.6% (34/70) achieved their attendance goal. Wellness events were much less attended: 87.6% (99/113) of participants attended none of the four events. Although 62.5% (15/24) of the goal setters attended at least one event, only 20.8% (5/24) attended all the events they had committed to.

Menu preferences and goal achievement The analysis of Section 4.2.1 showed a positive link between preference for removing R from the coverage and take-up of the goal setting contract. A natural follow-up question is whether the menu preferences of those who took up the contract also predict their likelihood of reaching the goals they set. As with goal setting, the link between goal achievement and preference for removing unhealthy foods from the reimbursement coverage is a priori unclear. On the one hand, participants who made a clear commitment to eating healthy could be more likely to reach their goals if such commitment signals (and maybe reinforces) their ex ante motivation to lose weight. On the other hand, those who made stricter commitments could be less likely to fulfill their goals if their choice of tighter constraints partly reflects overoptimistic beliefs about their self-control and motivation during the challenge. To test the relationship between goal achievement and menu preferences, I study linear probability models of the form

$$\mathbb{1}(\text{achieved goal } g)_{ij} = \mathbf{X}_i\beta + \mathbf{M}_i\gamma + \mathbf{G}_{ij}\eta + \epsilon_{ij}$$

[28] Respondents were not contacted when the system recorded zero visits or when there was no conflict with a weigh-in date.

[29] When a respondent reported a bracket rather than a single number, the self-reported estimate was chosen to be the median in this bracket rounded down to the nearest integer. In the regression analysis presented below, I assume that the participant with a missing report failed to reach his gym goal of 8 visits, a plausible assumption given that this participant failed to reach his weigh-in goal; however, findings are very similar if I instead exclude this observation from the analysis.

where g is the target number selected by individual i for goal category $j \in$ {gym visits, weigh-ins, wellness events}, $\mathbf{X}_i$ and $\mathbf{M}_i$ are as previously defined, and $\mathbf{G}_{ij}$ is a vector of controls for the goal category and corresponding target number. Appendix Table 7 presents the regression results; standard errors are clustered at the subject level. Consistent with partial naiveté, participants who preferred to exclude R from the reimbursement coverage were less likely to achieve their goals. First, those who ranked G as their top option were about 35 percentage points more likely to miss their targets; the coefficient for GO is also negative, but smaller and at best marginally significant. Thus, it appears that participants who preferred stricter commitments were more likely to miss their goals. Higher GT_{-R} and SSB_{-R} scores are also associated with a significantly higher likelihood of failing to reach a given goal; these findings are robust to controlling for the corresponding scores for O and R. Looking at the other covariates, first time participants and repeat dieters were less likely to reach the goals they set, suggesting that the observed patterns could be due to both a lack of experience with the environment and failures of self-control.[30]

4.2.3 Menu preferences and other outcomes

In OA-B.4, I investigate the relationship between preferences in the reimbursement program and three other types of outcomes: (*i*) challenge completion, (*ii*) study completion, and (*iii*) submission of lunch receipts. Although connected, each of the three outcomes measures a different dimension of participation and engagement. Below I briefly discuss how those three outcomes relate to one another and to the menu preference measures studied in this paper.

Challenge completion To complete the challenge and be eligible for the grand prize, participants were required to attend the last weigh-in. Among the 113 study participants, only 36% (41) did so.[31] In OA-B.4.1, I test whether temptation in the reimbursement program predicts likelihood of attending the final weigh-in. Since goal setters were somewhat more likely to (*i*) attend the last weigh-in, and (*ii*) prefer a restricted coverage, regressions control for whether a participant set a weigh-in goal. To focus on completion rather than overall attendance (which, besides perseverance, also captures initial motivation), regressions also control for attendance of the second and third weigh-ins. Participants with a higher GT_{-R} score were less likely to attend the final weigh-in, with a 7 percentage point decrease in the

[30] In OA-B.3, I present an analysis by goal category, looking at rates of failure for gym attendance on the one hand, and for the weigh-ins and wellness events on the other hand; the above findings do not appear to be driven by a specific goal category. I also report results from regressions looking at contract default (i.e., failure to reach all the goals set); findings appear somewhat weaker with this outcome measure.

[31] As could be expected, the rate of attendance of the weigh-ins drops over time, with 61% (69) of participants who attended the second weigh-in, and 38% (43) who attended the third weigh-in. Since the second and third weigh-ins were not mandatory in order to be part of the competition, participants who attended the last weigh-in did not necessarily attend the second and third weigh-ins.

likelihood of attendance for each additional point in the GT_{-R} score. The relationship between attendance and the SSB_{-R} score is similar, but somewhat less robust; the relationship between top choice and challenge completion is not significant.

Study completion To complete the study and receive the $20 gift card, participants had to answer a 10-15 minute follow-up survey at the end of the challenge, with a deadline of one week. Only 77% (87) of participants made the effort to complete this survey. Unsurprisingly, participants who defaulted on their goal setting contract, thus forfeiting the gift card, were less likely to respond than those who achieved all their goals and those who set no goal (51.1%, 100%, and 89.7% response rate, respectively).[32] Since those who preferred to exclude R from the coverage were less likely to achieve their goals, I test whether the propensity to complete Survey 2 can be explained by menu preferences after controlling for eligibility to receive the gift card. Regressions also control for whether the participant completed the challenge, as those who completed it were more likely to respond to Survey 2 than those who did not (97.6% vs. 65.3%, $z = 3.92$, $p < 0.001$). As shown in OA-B.4.3, an increase in the GT_{-R} score by one point is associated with a 5-7 percentage point decrease in the likelihood of completing Survey 2. Furthermore, those who ranked G as their top option were about 20 percentage points less likely to reply to Survey 2 than those who strictly preferred GOR.

Submission of lunch receipts Despite some initial interest in the reimbursement program, only 17% (19) of participants ended up submitting receipts for reimbursement. An examination of Survey 2 responses suggests that the logistical costs might have been too high.[33] As a reminder, participants were required to submit itemized receipts with their name or credit number on it during the last weigh-in (although late submissions were accepted). Participating in the program thus required a high level of organization, motivation and effort. In OA-B.4.4, I test whether participants' preferences for a restricted coverage predict their likelihood of submitting receipts. Regressions control for attendance of the last weigh-in and for the rembursement option received.[34] In line with the other outcomes, the likelihood of claiming reimbursement decreases by 6-7 percentage points for each additional point in the GT_{-R} score. Furthermore, those who ranked GO as their top choice were about 20 percentage points less likely to return their receipts (the coefficient is negative but smaller and insignificant for those who ranked G first).

[32] Participants still received a $10 payment if they only completed Survey 1 and/or missed their goals.

[33] Among the 87 Survey 2 respondents, 41% explained that they either lost receipts, forgot to ask for receipts, or ate at places that were not giving detailed receipts; 34% considered that participation in the program was not worth the effort given that only 10% of respondents would be reimbursed. Among other reasons for non participation, respondents mentioned that they usually brought their own lunch/rarely ate out (26%), or that their reimbursement option did not cover the foods they ordered (10%).

[34] As could be expected given the randomization procedure and the distribution of top choices, most participants received either option G (20%), GO (24%) or GOR (25%). See OA-A.3 for more details.

5. Discussion

A key behavioral implication of economic theories of temptation is that agents with self-control problems should be willing to demand commitment. Many experimental studies have tested this hypothesis, both in the lab and in the field, finding only limited support. However, commitment opportunities are usually restricted either to a single instance or to repeated instances of the same type of decision. As a result, only limited inferences can be drawn from existing studies about the link between temptation and commitment demand. In this paper, I gather more granular data on commitment preferences to study the extent to which temptation can be revealed through agents' decisions to restrict their choice sets.

In order to do so, I conduct a field study with individuals expected to struggle with real temptations: participants in a weight loss challenge. Following the revealed preference approach of Gul and Pesendorfer (2001), I construct an environment of menu choice to study participants' commitment to eating healthy through a lunch reimbursement program. The menus were various reimbursement options, which only differed in the range of foods included in the coverage. Instead of asking participants to select one option, I elicit their preference ordering over the entire set. By allowing for the expression of any type of preference, the elicitation procedure lets individuals reveal their own temptations.

Exploiting data on the entire ordering, I develop measures of temptation to study its *source*, its *strength* and its *structure*. I then validate those measures using survey data on participants' food perceptions and self-reported consumption. Finally, I test whether temptation revealed through menu choice can predict other behaviors during the challenge that could be symptomatic of self-control problems, such as take up of, and performance on, a goal setting contract. What can be learned about temptation from eliciting preferences over menus?

I find that menu choice offers potential for measuring temptation in real life settings. In this study, only a third of participants strictly preferred the most comprehensive coverage, thus possibly leaving money on the table. Preference for a restricted coverage appears far from random: nearly half of participants systematically preferred options that excluded the foods they perceived as most tempting and unhealthy. Furthermore, the expression of temptation generally takes the simple form assumed in Gul and Pesendorfer (2001), making temptation concerns easy to identify in the data. Finally, menu choice is predictive of other decisions during the challenge that likely reflect self-control problems.

Although the language of menus appears rich enough to deliver important insights about the nature of temptation, it is by no means exhaustive. Below I return to the interpretation of the main findings of this paper to discuss the extent to which menu choice can reveal temptation. Building on this discussion, I suggest several directions for future research.

In this study, I find that preference for a restricted coverage largely coincided with a desire to avoid the foods perceived as most tempting and unhealthy. However, it could be argued that identifying temptation solely from commitment choices is too weak of a test, for it does not take into account preferences over the singleton options (Stovall (2010)). For instance, should R be interpreted as a temptation if $G \succ GR$ but $R \succeq G$? In OA-A.2, I therefore consider a tighter definition of the Global Temptation Index, which imposes additional restrictions on preferences over $\{G, O, R\}$. The distribution of the GT_{-R} index changes very little; in particular, R remains globally tempting ($GT_{-R} = 3$) for 46% of participants. Furthermore, as I argue in OA-A.2, it is unclear whether a definition of temptation based on menu choice warrants such restrictions on preferences over singletons.[35]

A related concern is that differences in commitment choices across participants could simply reflect differences in preferences for the three food categories. For instance, those who ranked GO strictly above GOR were more likely to rank G strictly above R (98.4% vs. 71.4%, $z = 4.19$, $p < 0.001$). To test whether the predictive power of the menu choice measures presented in this paper simply comes from variation in preferences over $\{G, O, R\}$, I repeated the analysis of Section 4.2 for the subsample of 90 participants who ranked R strictly below G and O (see OA-B). For the majority of outcomes and temptation measures, the above findings remain very similar when restricting the analysis to this subsample.[36]

While the definitions provided in this paper might be considered too weak to identify temptation (i.e, there could be commitment without temptation), an alternative view is that they are in fact too strong (i.e., a lot of temptation occurs without commitment). Indeed, a key assumption for temptation to be identified from commitment behavior is that of sophistication: only individuals aware of their self-control problems should demand commitment. Yet, recent studies have documented significant levels of naiveté among individuals asked to form beliefs about their future self-control (Augenblick and Rabin (2018), Fedyk (2018)). If beliefs are highly inaccurate, then a large wedge will exist between *perceived* temptation, as revealed from menu choice, and *actual temptation*, as individuals experience it.

In the present context, there are, however, reasons to believe that participants were at least somewhat sophisticated about their temptations. First, they were recruited from a population of highly educated people who already made one commitment: they entered a challenge

[35] For instance, $GO \succ GOR \succ R \succ G$ could be a reasonable ordering for a participant who would rather avoid R if he can get enough calories/proteins by combining G and O, but would prefer R to only G.

[36] One exception is the non-robust effect of the GT_{-R} and SSB_{-R} indices in the regressions for contract take-up. Among the 23 participants excluded, 8 ranked R strictly above both G and O, 5 had the ranking $G \succ R \succ O$, 4 had the ranking $O \succ R \succ G$, 3 the ranking $G \succ O \sim R$, 2 the ranking $G \sim O \sim R$ and 1 the ranking $O \sim R \succ G$; see OA-A.1 for the distribution of preferences over singletons. While men were more likely than women to rank R above G (29.2% vs. 9.0%, $z = 2.59$, $p < 0.01$), no other baseline characteristic appears to explain a preference for R over G and/or O; an examination of respondents' comments in Survey 2 suggests that some of the decisions to give R a higher ranking were likely due to confusion/mistakes.

to lose weight. This suggests that participants were at least partially aware of the self-control problems associated with controlling their weight. Second, while sophistication might be low in unfamiliar contexts, individuals are unlikely to be completely naive about their food cravings. Thus, provided that participants (i) acknowledged some struggle with maintaining a healthy diet, and (ii) perceived a commitment value to the reimbursement program, preference for a restricted coverage provides a reasonable way to identify temptation.

While the observed commitment demand appears consistent with participants being aware of their self-control problems on the extensive margin, the high failure rates on the goal setting contract suggest a misunderstanding of the intensive margin of self-control. On the one hand, participants who preferred not to be reimbursed for unhealthy foods were *more* likely to take up the contract, suggesting they had some global understanding of their temptations. On the other hand, consistent with partial naiveté, they were *less* likely to reach the goals they set. One conjecture is that they overestimated how active they would be during the challenge and lost motivation over time. In line with this interpretation, participants with a higher GT_{-R} score were less likely to attend the last weigh-in, despite having attended the previous weigh-ins at a similar, if not higher, rate (see OA-B.4.1). To better understand the link between sophistication and commitment demand, menu choice data could be combined in future work with information on beliefs and actual choices when facing temptation.

Although beliefs about self-control are a likely candidate to explain the correlation in commitment demand across domains, a perhaps complementary explanation is signaling or experimenter demand (Exley and Naecker (2017), de Quidt et al. (2018)). Despite being free to choose their restrictions (including none), participants might have felt compelled to commit to the behaviors they thought were expected of them. However, if the primary goal was to seek social approval, those who chose to commit should have been more likely to follow through on their commitments. Instead, they were less likely to meet their goals or submit lunch receipts (despite being more likely to be assigned a healthy option). Furthermore, the methodological work of de Quidt et al. (2018) suggests that typical demand effects are likely modest in size. Future research could use a similar method to bound the effect of experimenter demand, both on point estimates (take-up rates) and correlations (take-up across domains).

Of course, given the specificity of the subject pool and decision context, it is important to test whether the findings of this paper generalize to other choice environments. It also remains to understand how menu choice measures of temptation relate to other measures of self-control problems frequently used in the literature, such as measures of present bias. Finally, more research is needed to understand how choice architecture (size of the choice set, type of items...) may influence commitment take-up and encourage behavior change.

References

Ameriks, J., A. Caplin, J. Leahy, and T. Tyler (2007): "Measuring Self-Control Problems," *American Economic Review*, 97, 966–972.

Ariely, D. and K. Wertenbroch (2002): "Procrastination, Deadlines, and Performance: Self-Control by Precommitment," *Psychological Science*, 13, 219–224.

Ashraf, N., D. Karlan, and W. Yin (2006): "Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines," *Quarterly Journal of Economics*, 121, 635–672.

Augenblick, N., M. Niederle, and C. Sprenger (2015): "Working Over Time: Dynamic Inconsistency in Real Effort Tasks*," *Quarterly Journal of Economics*.

Augenblick, N. and M. Rabin (2018): "An Experiment on Time Preference and Misprediction in Unpleasant Tasks," *Review of Economic Studies*, forthcoming.

Beshears, J., J. Choi, C. Harris, D. Laibson, B. Madrian, and J. Sakong (2015): "Self Control and Commitment: Can Decreasing the Liquidity of a Savings Account Increase Deposits?" NBER Working Paper No. 21474.

Bisin, A. and K. Hyndman (2018): "Present-Bias, Procrastination and Deadlines in a Field Experiment," Working paper.

de Quidt, J., J. Haushofer, and C. Roth (2018): "Measuring and Bounding Experimenter Demand," *American Economic Review*, 108, 3266–3302.

Dekel, E. and B. L. Lipman (2012): "Costly Self-Control and Random Self-Indulgence," *Econometrica*, 80, 1271–1302.

Dekel, E., B. L. Lipman, and A. Rustichini (2009): "Temptation-Driven Preferences," *Review of Economic Studies*, 76, 937–971.

Exley, C. L. and J. K. Naecker (2017): "Observability Increases the Demand for Commitment Devices," *Management Science*, 63, 3262–3267.

Fedyk, A. (2018): "Asymmetric Naiveté: Beliefs about Self-Control," Working paper.

Fudenberg, D. and D. K. Levine (2006): "A Dual-Self Model of Impulse Control," *American Economic Review*, 96, 1449–1476.

——— (2012): "Timing and Self-Control," *Econometrica*, 80, 1–42.

Giné, X., D. Karlan, and J. Zinman (2010): "Put Your Money Where Your Butt Is:

A Commitment Contract for Smoking Cessation," *American Economic Journal: Applied Economics*, 2, 213–35.

Gul, F. and W. Pesendorfer (2001): "Temptation and Self-Control," *Econometrica*, 69, 1403–1435.

John, A. (2019): "When Commitment Fails - Evidence from a Field Experiment," *Management Science*, forthcoming.

Kaur, S., M. Kremer, and S. Mullainathan (2015): "Self-Control at Work," *Journal of Political Economy*, 123, 1227–1277.

Kopylov, I. (2012): "Perfectionism and Choice," *Econometrica*, 80, 1819–1843.

Kreps, D. (1979): "A Representation Theorem for Preference for Flexibility," *Econometrica*, 47, 565–576.

Laibson, D. (1997): "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112, 443–478.

——— (2015): "Why Don't Present-Biased Agents Make Commitments?" *American Economic Review*, 105, 267–72.

Lipman, B. and W. Pesendorfer (2013): "Temptation," in *Advances in Economics and Econometrics: Tenth World Congress*, ed. by D. Acemoglu, M. Arellano, and e. Dekel, Eddie, Cambridge University Press, vol. 1.

Milkman, K., J. Minson, and K. Volpp (2014): "Holding the Hunger Games Hostage at the Gym: An Evaluation of Temptation Bundling," *Management Science*, 60, 283–299.

Noor, J. and N. Takeoka (2010): "Uphill self-control," *Theoretical Economics*, 5, 127–158.

——— (2015): "Menu-Dependent Self-Control," *Journal of Mathematical Economics*, 61, 1–20.

O'Donoghue, T. and M. Rabin (1999): "Doing It Now or Later," *American Economic Review*, 89, 103–124.

Pejsachowicz, L. and S. Toussaert (2017): "Choice deferral, indecisiveness and preference for flexibility," *Journal of Economic Theory*, 170, 417–425.

Royer, H., M. Stehr, and J. Sydnor (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, 51–84.

Schilbach, F. (2018): "Alcohol and Self-Control: A Field Experiment in India," *American Economic Review*, forthcoming.

Stovall, J. (2010): "Multiple Temptations," *Econometrica*, 78, 349–376, forthcoming.

——— (2018): "Temptation with uncertain normative preference," *Theoretical Economics*, 13, 145–174.

Toussaert, S. (2018): "Eliciting temptation and self-control through menu choices: a lab experiment," *Econometrica*, 86, 359–889.

Tversky, A. and E. Shafir (1992): "Choice under conflict: the dynamics of deferred decision," *Psychological Science*, 3, 358–361.

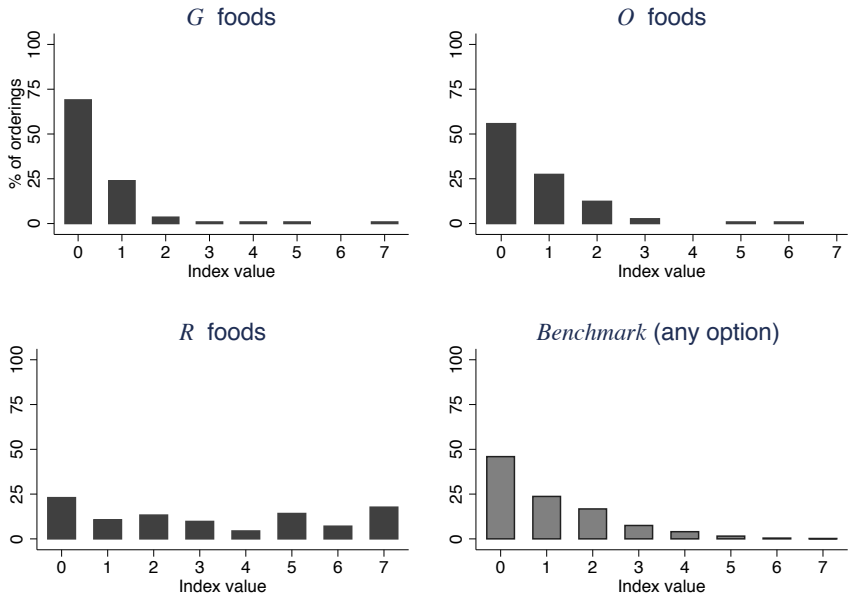
Appendix

Table 4: Descriptive Statistics

Variables	scale/unit	mean	s.d.
female	0-1	0.79	0.41
single	0-1	0.61	0.49
age	25-65	34.95	9.99
years of educ	0-9	4.93	1.60
prior participant	0-1	0.31	0.46
starting weight	lbs	179.03	40.59
weight loss goal	lbs	14.36	10.73
goal confidence	1-7	4.85	1.17
diets attempted	$n \in \{0, 1, ...\}$	4.23	4.72
Observations			113

Notes: *single* refers to participants who are single without children; *age* was constructed from a categorical variable by taking the mid-point of the age categories 21-29, 30-39, 40-49, 50-59 (rounded up to nearest integer) and setting 60+ at 65; *years of educ* is a proxy for number of years of education post high school based on highest degree obtained; *prior participant* = 1 if the individual participated in a previous edition of the challenge. The variable *starting weight* is weight (in lbs) at the first weigh-in and *weight loss goal* is targeted weight loss (in lbs) during the challenge. Finally, *goal confidence* measures on a 1-7 scale the confidence with which a participant expects to reach his weight loss goal (1=extremely uncertain; 7=extremely confident) and *diets attempted* is the number of diets attempted over the last 10 years.

FIGURE 8: Distribution of the SSB index for G, O and R foods



Notes: For each category $x \in \{G, O, R\}$, "Index value" refers to the value of the Strict Set Betweenness Index $SSB_{-x} = \sum_{\mathcal{P}_x} \mathbb{1}_{\{M \succ M \cup M' \succ M'\}} \in \{0, 1, ..., 7\}$ (see main text for a definition of $\mathcal{P}_x$).

FIGURE 9: Distribution of violations of SB, WSB, PSB and MON

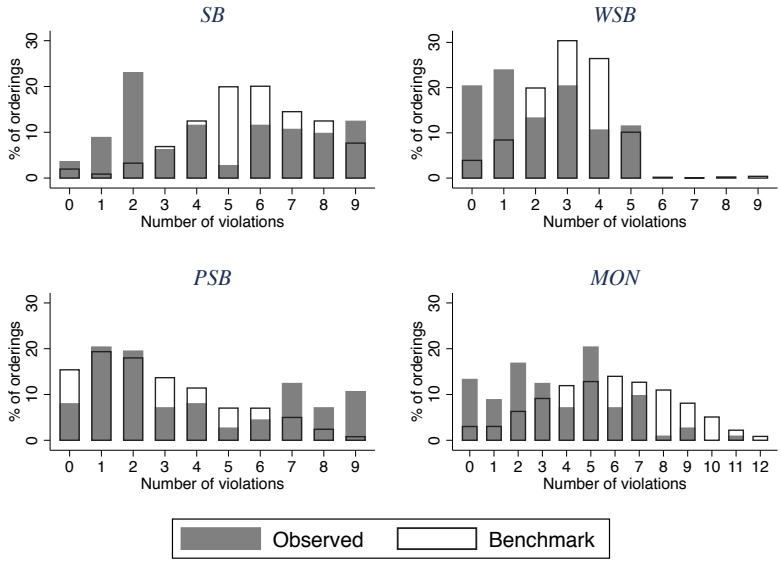


TABLE 5: Mean health, temptation and consumption value of G, O, R by menu preference

	Health Score			Temptation Score			*Unrestricted* − *Ideal* Gap			*Actual* − *Ideal* Gap		
Lunch category	G	O	R	G	O	R	G	O	R	G	O	R
Panel A: By top choice												
G top	5.88	3.90	1.96	4.41	4.22	4.46	-0.21	0.008	0.20	-0.007	-0.06	0.07
	(0.16)	(0.21)	(0.19)	(0.32)	(0.27)	(0.33)	(0.05)	(0.04)	(0.05)	(0.06)	(0.06)	(0.05)
GO top	5.90	4.35	2.05	4.12	4.02	4.74	-0.21	0.01	0.20	-0.10	0.02	0.08
	(0.11)	(0.15)	(0.13)	(0.22)	(0.19)	(0.22)	(0.03)	(0.02)	(0.03)	(0.04)	(0.04)	(0.03)
GOR top	6.08	4.25	1.86	4.12	4.30	5.08	-0.24	-0.02	0.26	-0.13	0.02	0.11
	(0.11)	(0.15)	(0.13)	(0.22)	(0.19)	(0.22)	(0.03)	(0.02)	(0.03)	(0.04)	(0.03)	(0.03)
Other	5.87	4.16	1.63	4.26	3.93	4.84	-0.21	-0.03	0.23	-0.21	0.16	0.06
	(0.14)	(0.18)	(0.16)	(0.28)	(0.24)	(0.28)	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)
F-stat	0.68	1.05	1.44	0.25	0.66	0.90	0.27	0.59	0.92	2.42*	3.89**	0.50
Panel B: By value of the Global Temptation Index for R												
$GT_{-R} = 0$	5.92	4.37	1.82	3.93	4.22	5.20	-0.23	-0.05	0.27	-0.23	0.12	0.11
	(0.15)	(0.19)	(0.17)	(0.29)	(0.25)	(0.29)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)	(0.04)
$GT_{-R} = 1$	5.93	4.09	1.89	4.12	4.00	4.99	-0.24	-0.02	0.26	-0.16	0.03	0.12
	(0.15)	(0.20)	(0.18)	(0.30)	(0.26)	(0.31)	(0.04)	(0.03)	(0.04)	(0.05)	(0.05)	(0.04)
$GT_{-R} = 2$	5.98	4.22	1.64	4.06	4.37	4.67	-0.22	0.02	0.20	-0.04	0.02	0.02
	(0.15)	(0.19)	(0.17)	(0.29)	(0.25)	(0.29)	(0.04)	(0.03)	(0.04)	(0.05)	(0.05)	(0.04)
$GT_{-R} = 3$	5.95	4.19	2.01	4.38	4.02	4.68	-0.20	0.002	0.20	-0.09	0.005	0.09
	(0.09)	(0.12)	(0.11)	(0.18)	(0.16)	(0.19)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
F-stat	0.03	0.34	1.22	0.72	0.60	0.95	0.28	1.18	1.25	3.02**	1.70	1.20
Observations	113	113	113	113	113	113	87	87	87	87	87	87

Notes: Results from linear regressions of the Health Score, Temptation Score, *Unrestricted* − *Ideal* Gap, and *Actual* − *Ideal* Gap for lunch category $m \in \{G, O, R\}$ on dummies for top choice (Panel A) and level of GT_{-R} (Panel B); standard errors are in parentheses. See Section 4.1 for the definition of each variable. Reported F-statistic corresponds to a test of the null hypothesis that all dummy coefficients are equal.
* $p < 0.10$ and ** $p < 0.05$

TABLE 6: Determinants of contract take-up

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
G top	0.031	0.049						
	(0.136)	(0.136)						
GO top	0.191*	0.232**						
	(0.109)	(0.107)						
Other top	0.057	0.039						
	(0.124)	(0.121)						
GT_{-R}			0.075**	0.099***	0.103***			
			(0.037)	(0.037)	(0.038)			
GT_{-G}					0.034			
					(0.070)			
GT_{-O}					-0.014			
					(0.070)			
SSB_{-R}						0.026	0.038**	0.050***
						(0.017)	(0.017)	(0.018)
SSB_{-G}								-0.034
								(0.041)
SSB_{-O}								-0.060
								(0.044)
female		-0.094		-0.126	-0.122		-0.119	-0.153
		(0.106)		(0.105)	(0.107)		(0.107)	(0.109)
single		0.051		0.087	0.089		0.079	0.071
		(0.091)		(0.088)	(0.089)		(0.089)	(0.089)
age		0.007		0.008*	0.007		0.008*	0.009**
		(0.005)		(0.005)	(0.005)		(0.005)	(0.005)
years of educ		-0.092***		-0.097***	-0.098***		-0.097***	-0.100***
		(0.029)		(0.029)	(0.029)		(0.029)	(0.029)
prior participant		-0.092		-0.073	-0.071		-0.082	-0.089
		(0.091)		(0.089)	(0.090)		(0.090)	(0.090)
weight loss goal		0.005		0.004	0.005		0.004	0.003
		(0.004)		(0.004)	(0.004)		(0.004)	(0.004)
goal confidence		0.421**		0.385**	0.386*		0.417**	0.421**
		(0.197)		(0.193)	(0.195)		(0.194)	(0.193)
$(goal\ confidence)^2$		-0.048**		-0.045**	-0.046**		-0.048**	-0.048**
		(0.020)		(0.020)	(0.020)		(0.020)	(0.020)
diets attempted		0.006		0.005	0.005		0.006	0.006
		(0.009)		(0.009)	(0.009)		(0.009)	(0.009)
Day 1 decision	-0.252***	-0.213**	-0.270***	-0.232***	-0.230**	-0.259***	-0.215**	-0.193**
	(0.093)	(0.088)	(0.091)	(0.085)	(0.088)	(0.092)	(0.087)	(0.088)
N	113	113	113	113	113	113	113	113
adj. R^2	0.069	0.184	0.092	0.213	0.199	0.078	0.196	0.201

Notes: Linear probability models where the dependent variable is equal to 1 if the respondent set at least one goal. *Day 1 decision* is an indicator variable equal to 1 if the respondent completed Survey 1 within the first 24 hours the survey was administered; see Section 3.3 and Table 4 for a description of the other control variables. Standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 7: Determinants of goal achievement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
G top	-0.341**	-0.353***						
	(0.143)	(0.130)						
GO top	-0.165	-0.243*						
	(0.135)	(0.133)						
Other top	-0.154	-0.112						
	(0.149)	(0.129)						
GT_{-R}			-0.098**	-0.120**	-0.116**			
			(0.048)	(0.050)	(0.050)			
GT_{-G}					0.025			
					(0.080)			
GT_{-O}					-0.044			
					(0.093)			
SSB_{-R}						-0.029	-0.036*	-0.059**
						(0.019)	(0.018)	(0.022)
SSB_{-G}								0.038
								(0.046)
SSB_{-O}								0.111
								(0.083)
female		-0.038		-0.005	-0.007		-0.000	0.044
		(0.104)		(0.112)	(0.113)		(0.121)	(0.124)
single		0.054		0.042	0.042		0.067	0.079
		(0.100)		(0.098)	(0.100)		(0.097)	(0.100)
age		-0.005		-0.008*	-0.007		-0.009*	-0.010**
		(0.005)		(0.005)	(0.005)		(0.005)	(0.004)
years of educ		0.036		0.048	0.045		0.034	0.053*
		(0.031)		(0.031)	(0.033)		(0.030)	(0.030)
prior participant		0.196*		0.178*	0.179*		0.177*	0.213**
		(0.105)		(0.103)	(0.103)		(0.105)	(0.105)
weight loss goal		-0.004		-0.003	-0.003		-0.002	-0.002
		(0.003)		(0.003)	(0.003)		(0.003)	(0.003)
goal confidence		0.053		0.050	0.049		0.041	0.014
		(0.278)		(0.274)	(0.274)		(0.256)	(0.266)
$(goal\ confidence)^2$		-0.002		-0.001	-0.001		-0.002	0.001
		(0.029)		(0.029)	(0.029)		(0.027)	(0.028)
diets attempted		-0.027**		-0.026**	-0.025**		-0.025**	-0.028**
		(0.011)		(0.011)	(0.010)		(0.011)	(0.012)
Day 1 decision	0.165	0.180	0.185	0.204*	0.211*	0.162	0.176	0.145
	(0.125)	(0.113)	(0.127)	(0.115)	(0.116)	(0.127)	(0.118)	(0.130)
N	149	149	149	149	149	149	149	149
adj. R^2	0.090	0.147	0.103	0.162	0.151	0.083	0.138	0.153

Notes: Linear probability models where the dependent variable is equal to 1 if the respondent achieved his goal for goal category j; participant with missing gym attendance assumed to have failed his gym goal. All regressions include a control for selected target numbers and dummies for goal category. *Day 1 decision* is an indicator variable equal to 1 if the respondent completed Survey 1 within the first 24 hours it was administered; see Section 3.3 and Table 4 for a description of the other controls. Standard errors in parentheses clustered at the subject level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$