Pre-Analysis Plan

How Robust is Commitment Demand? Experimental Evidence[*]

Rafael Suchy[†] Séverine Toussaert[‡]

November 11, 2024

Abstract

While commitment demand has been documented in various settings, recent evidence suggests that some of it might be due to mistakes. We propose to investigate the robustness of commitment demand in a field study involving meal choices on a food ordering platform. We elicit participants' preferences over the range of meals presented on the platform. Preferences are elicited twice using three methods: ranking, monetary valuations, and binary choices. We propose to compare the consistency of expressed preferences across methods and the stability within a method, as well as to explore drivers of the variability observed. We will benchmark our results against expert forecasts and provide guidance on method selection. This detailed pre-analysis plan outlines (i) the rationale for the study; (ii) the full experimental design; (iii) the analyses to be conducted.

JEL Codes: C80, D01, D90, D91

Keywords: Preference elicitation, measurement, robustness, commitment, field experiment

[*] This PAP document was originally submitted as a registered report to the Journal of Political Economy Microeconomics, with some differences in the experimental design (setting and timing of decisions, population of respondents and sample size). We thank Johannes Abeler, Marina Agranov, Miguel Ballester, Anna Dreber, Daniel Evans, Drew Fudenberg, Taisuke Imai, Barton Lipman, Pietro Ortoleva and Dmitry Taubinsky for inspiring discussions and helpful comments. We are extremely grateful to Stefania Merone, Katja Michlbauer and Anna Popova for their excellent research assistance. We gratefully acknowledge funding from the John Fell Fund, the OEP Fund at the University of Oxford as well as the Economic and Social Research Council (ES/V003461/). This research has been reviewed according to the ethics procedures for research involving human subjects of the University of Oxford, approval #ECONCIA21-22-10.

[†] University of Oxford; rafael.suchy@economics.ox.ac.uk

[‡] University of Oxford; severine.toussaert@economics.ox.ac.uk

Contents

1 Introduction

A central insight of economic theories of dynamic inconsistency and limited self-control is that individuals who suffer from temptation may benefit from tools or incentive mechanisms that can influence their own future behavior (Strotz, 1955; Laibson, 1997; O'Donoghue and Rabin, 1999; Gul and Pesendorfer, 2001). One of the self-control strategies most studied in the literature is the use of commitment devices, which constrain future choices by strictly removing certain alternatives from the choice set or by rendering them less attractive (Bryan et al., 2010). As documented in Table A.1, Appendix A, over 50 empirical studies in economics provide estimates of commitment take-up in a variety of settings (whether lab or field), domains (e.g., savings or exercise), and forms (e.g., financial penalties or hard restrictions).[1]

While commitment devices have received much empirical attention, demand for commitment is by no means a universal phenomenon. Indeed, panel A of Figure 1 shows that commitment take-up rates across studies display remarkable variation, ranging from 11% (Giné et al., 2010) to 93% (Casaburi and Macchiavello, 2019). This variation is in itself unsurprising given the large heterogeneity in study populations and features of the commitment devices, which likely moderate their effectiveness and ultimate appeal. To put it simply, certain commitment devices might just be better suited than others to meet the demand of those who need them. One competing and more worrying interpretation of the data, however, is that commitment decisions often differ in how clearly and transparently they are framed, which could induce some users to take commitment devices by mistake. In other words, preferences for commitment might be incorrectly measured, which could lead to substantial variation in observed take-up rates.

Speaking to these concerns, several recent studies document that individuals who restrict their choices ex ante often fail in their commitments ex post (John, 2020; Bai et al., 2021), or they may demand commitment devices without even using them (Robinson et al., 2018), suggesting they would not do it again. Clearly pointing towards measurement error, Carrera et al. (2022) find that approximately half of their study participants who take up commitment contracts for higher gym attendance also take up contracts for lower gym attendance. As the authors conclude, because commitment take-up is a coarse measure of demand for behavior change due to its binary nature, even modest stochastic valuation errors in the perception of incentives may lead to upward-biased estimates of commitment demand. Relatedly, if commitment preferences are subtle (close to indifference or indecision), estimates of commitment take-up may be highly sensitive to details of the elicitation procedure. The goal of this paper is to carefully test this conjecture in an experiment that isolates the impact of the choice of measurement.

Despite recent concerns about measurement, we indeed know surprisingly little about the impact of certain methodological choices on observed take-up rates. Looking at the literature, three main elicitation methods have been used to different degrees in order to recover estimates of commitment take-up: (i) direct choices (often binary) between a flexible option and some commitment option(s); (ii) an ordinal ranking of these options; (iii) a valuation exercise eliciting people's price for a commitment option. A cursory glance at the literature hints at the potentially non-neutral impact of the chosen method on

[1] The estimate of the number of studies varies depending on the inclusion criteria applied. Our review in this paper restricts attention to studies (i) written in economics, (ii) involving commitments with real financial consequences if broken, or irreversible choice set restrictions (no hypothetical choices), (iii) with "pure" commitment devices that come with no additional benefits. See Appendix A for more information on the construction of our sample and variable definitions.

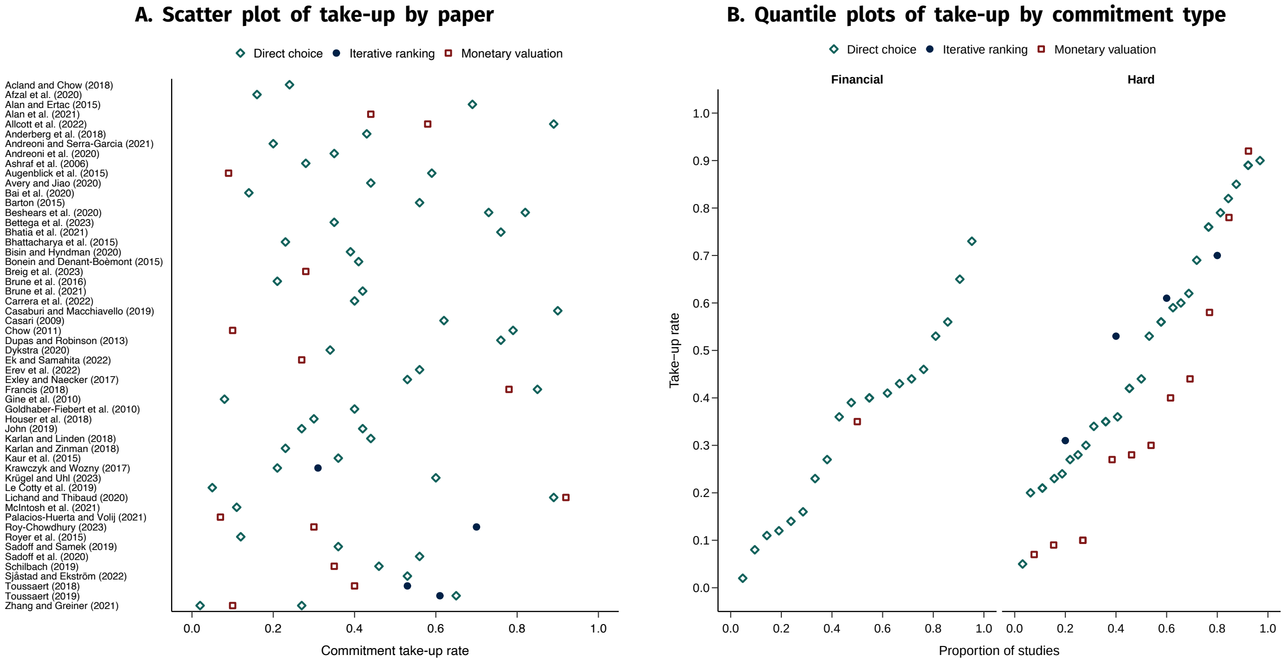


Figure 1. Illustration of take-up rates in the literature. Panel A. shows the take-up rates reported in the respective paper categorized by the elicitation method(s) used. Panel B. shows the same take-up rates ordered in quantile plots and split by the commitment type, which can be financial (i.e., entailing a penalty if broken) or hard (irreversible choice set restriction); some papers have both.

the distribution of take-ups. Panel B of Figure 1 indeed shows that take-up rates appear to be generally higher for the ranking method, intermediate for direct choice, and lowest when using valuations. At the same time, the evidence is at best suggestive given the limited and non-systematic variation in the choice of elicitation method across studies, which could be correlated with other important study characteristics relevant to take-up. In addition, panel A of Figure 1 shows that very few papers collect measures of take-up using a combination of methods and, when they do so, those measures are usually not independent, leaving little hope of exploiting within-study variation.[2] More generally, very few studies take any repeated measurements of commitment take-up even using the same method over time (see Table A.1 for a breakdown).

Using an online experiment, the purpose of this paper is to address three sets of related questions left open by the literature. First, how sensitive are estimates of commitment demand to the choice of elicitation method? Second, how reliable are these estimates within a given method? Third, what are the implications for the choice of elicitation procedure of any discrepancies between and within methods?

We examine these questions by conducting an online study as part of a larger field experiment involving meal choices on a food ordering platform that was specifically designed for this study. The meal orders take place in the context of a challenge to follow a meat- and carb-controlled diet for 5 consecutive days. Participants indicate their preferences over 7 versions of the food ordering platform

[2] For example, some studies ask participants to make a direct choice between a flexible and a commitment option and then ask for a similar decision, this time with the commitment option costing $X. If participants satisfy monotonicity in money, estimates of take-up will mechanically decrease once money is introduced.

or “menus”, which only differ in the range of meal categories available. The week after taking the survey, 50% of respondents will order 5 meals using one version of the platform selected based on their responses.

Building on existing studies on food choice (Schwartz et al., 2014; Sadoff and Samek, 2019; Sadoff et al., 2020), our customizable food ordering platform allows us to take full control over the choice architecture and implement a symmetric design in which people can express preferences for commitment and flexibility in any way they want. More specifically, we say that respondents reveal a preference for commitment if, when comparing any two platform versions, they would prefer the one displaying fewer meal categories (and conversely for preference for flexibility). Importantly, our approach entails that respondents can simultaneously express a desire for commitment when comparing two specific platforms and a desire for flexibility when comparing two others; in addition, we allow them to express indifferences. By enabling the expression of any type of preference, we avoid framing decisions in a specific direction, thus reducing the potential for demand effects.

Since our study is concerned with the robustness of commitment take-up to the choice of elicitation procedure, we repeatedly elicit respondents’ preferences over the platform in a within-subject design. In connection to the literature, preferences are elicited using three methods shown in a randomized order: a ranking procedure, a monetary valuation exercise, and binary choices. To assess the reliability of responses for a given method, each respondent completes the three elicitation exercises twice, both times in the same order. Each decision task is given an equal chance of determining the platform received.

Contrasting these methods is useful, for they impose a different amount of structure on preferences, thus varying the flexibility with which a desire for commitment can be revealed. At one end of the spectrum, eliciting binary choices is the most direct and transparent way of recovering preferences, but does not guarantee transitivity. At the other end of the spectrum, eliciting valuations not only induces a complete and transitive ordering over alternatives but also allows to quantify the strength of preferences. In between, a ranking procedure also induces a complete and transitive ordering but without the ability to make cardinal statements. Although the available evidence hints at one possible direction, which method should generate more commitment demand is a priori unclear. On the one hand, if commitment decisions are the product of mistakes, then using a more direct procedure should reduce the proportion of commitment choices. On the other hand, procedures that require respondents to think on a numerical scale and evaluate options holistically could generate more compression in evaluations, thus reducing the frequency of commitment choices. Our design allows us to carefully examine potential differences.

We begin our empirical analysis by examining whether and how much the prevalence of commitment demand changes depending on the method used, and provide bounds on commitment take-up within our setting. To better understand potential differences between methods, we then construct consistency indices, which calculate the number of times any two methods lead to the same classification of preferences (i.e., commitment, flexibility, or indifference). While observing inconsistencies between methods could be due to each method measuring preferences differently, another possibility is that respondents’ answers are simply unstable even within a given method. To test this conjecture, we next fix the method and study how the prevalence of commitment demand changes between the two elicitation rounds of this study.

Because we expect to observe at least some commitment demand and some inconsistencies or instabilities, making meaningful claims about the prevalence and robustness of commitment demand requires us to provide benchmarks against which to judge the observed frequencies. In this paper, we propose to compare our results against a purely random choice benchmark and expert predictions collected through a separate online survey. Combining all these pieces of data, we paint a rich picture of commitment decision patterns and assess their reliability.

If the use of different methods generates (potentially large) differences in commitment take-up rates, an important question for the designer is what method(s) she should select for her specific use-case. In the last section of our paper, we evaluate each method according to three performance criteria. First, we compare the three methods in terms of the plausibility of the structural assumptions they impose on preferences, such as completeness or transitivity. Second, we assess respondents' subjective views of each method and measure their preferred method through an incentivized mechanism that allows us to ask: if respondents could recommend a method for measuring their own preferences, which one would they choose? Finally, we conduct a prediction exercise to evaluate the ability of each method to pin down a respondent's preferred platform version as measured with two simple questions.

Understanding the extent to which estimates of commitment take-up may vary across measurements is important from a policy perspective. For instance, consider a company that contemplates building search filters on its website in order to allow customers to hide tempting alternatives on their screen. Because building search filters is costly, the company will only incur the development costs if it can be certain that a sufficient share of customers will make use of the filters. Without taking into account how estimates might vary across measurements, the company may end up making the wrong business development decision. Our study aims to gauge the importance of this concern and the need for experimenting with measurement.

Our work lies at the intersection of several active areas of research. First and foremost, our paper is connected to recent empirical studies that measure commitment take-up in lab and field settings (Kaur et al., 2015; Toussaert, 2018, 2019; Cheung et al., 2022). Most similar to our context, Sadoff and Samek (2019) and Sadoff et al. (2020) leverage food delivery programs to study the effect of information and experience on the prevalence of commitment demand, and its relation to dynamic inconsistency. With a few exceptions noted earlier, most studies draw inferences based on a single measurement of take-up using a single method. In contrast, we offer a systematic study of commitment take-up by observing repeated commitment decisions across multiple elicitation methods.

Second, our work speaks to a large literature that studies how using different methods affects the distribution of elicited preferences. The most famous illustration concerns the so-called "preference reversal phenomenon", which refers to the robust finding that binary choices and monetary valuations elicit opposite risk preferences (Lichtenstein and Slovic, 1971; Grether and Plott, 1979). Even when the relationship is positive, correlations in elicited preferences across methods tend to be low e.g., ranging from 0.06 to 0.37 in Frey et al. (2017) and Holzmeister and Stefan (2021). Closer to our study domain, Hascher et al. (2021) elicit preferences for food options through incentivized and unincentivized ratings as well as stated willingness to pay, finding that ratings outperform monetary valuations in predicting subsequent food choices.

Complementing this comparative literature, many studies have examined the stability of elicited preferences from repeated measurements using a given method. Chuang and Schechter (2015) document large variations in the correlation of risk (0.13 to 0.55), time (0.09 to 0.68) and social (0.12 to 0.28) preferences elicited repeatedly at different time horizons ranging from a few days to multiple years. In the domain of preferences over menus, we meld the two literature strands by examining across-method inconsistencies, within-method instabilities, and the potential interaction between the two.

Third, exploring potential sources of inconsistencies and instabilities, our work speaks to a growing literature that aims to understand the link between stochastic choice behavior or sensitivity to framing effects and the fact that individuals may have imprecise or incomplete preferences (e.g., non-transitive preferences) or face cognitive uncertainty (Loomes and Pogrebna, 2017; Agranov and Ortoleva, 2022; Costa-Gomes et al., 2022; Enke and Graeber, 2023). We contribute to this literature by proposing an incentivized measure of confidence in answers provided with a given method, which we relate to participants' stability of responses and subjective perceptions of each method.

Finally, we contribute to a burgeoning literature that seeks to understand the impact of certain degrees of freedom in research design on the distribution of experimental outcomes, and experts' awareness of these potential effects (DellaVigna and Pope, 2018; Bryan et al., 2019; Huber et al., 2023).

The rest of this pre-analysis plan document proceeds as follows: Section 2 introduces the study design, Section 3 describes our main outcomes and benchmark data, Section 4 presents our analysis of the consistency and stability of our preference measures, Section 5 assesses the relative strengths and weaknesses of each method, and Section 6 concludes with a discussion.

2 Study Design

2.1 General Description

Recruitment and Incentives. We will recruit students from the University of Oxford who have no dietary restrictions to participate in a study on the design of food ordering platforms. The study consists of a 35-min survey and a food ordering stage (the exact structure is illustrated in Figure B.1 in Section B.1). To recruit participants, we shared a "registration of interest" survey with students in three targeted Oxford colleges: Balliol, St John's, and St Anne's the week prior to the experiment.[3] We selected these three colleges based on the number of students and the feasibility to cater to our specific needs. Besides a £15 survey completion fee, 50% of participants will be randomly selected to receive free lunches for five consecutive days (5 meals worth around £25). The meals will be served during regular opening hours and prepared by the usual kitchen staff, ruling out any ambiguity regarding the quality of the meals. The selected participants will place a single order of 5 meals on Monday of the week in which they will receive the meals. Specifically, they will receive the link to the ordering platform at 8:00am on Monday and must complete their orders by 11:00am for the entire week. The short time between placing the order and receiving the first meal is a key design feature intended to induce temptation. Participants will eat their lunches between 12:00pm and 1:30pm, Monday to Friday, during standard

[3] The survey was circulated via various channels including mailing lists, text messages, social media channels, and in person-recruitment.

Table 1. Outline of the survey structure.

Survey section	Description
Section 1	Collection of socio-demographic variables and current meal habits of participants.
Section 2	Presentation of the food challenge and ordering platform with customizable menus.
Section 3	Elicitation of preferences over 7 possible menus via three methods (2 rounds): (i) iterative ranking, (ii) monetary valuations, (iii) binary choices (randomized order); Incentivized elicitation of preferred method and round.
Section 4	Debriefing questions on preferred platform and perceptions of the decision tasks

lunch hours. Orders will be placed through one of 7 possible versions of a customized food ordering platform developed specifically for this study; the assigned version will depend on respondents' answers in the survey.

Survey Structure. Participants will go through 4 survey sections summarized in Table 1. In the introductory section, participants answer basic questions about themselves (including gender, and program year) and provide information on their meal habits over the last 6 months (percentage of lunch/dinner meals that contained vegetables, carbohydrates, and meat).[4] The next three sections, described below, introduce participants to the food challenge and the food ordering platform they will use (Section 2.2), elicit their preferences for the meal options to be shown on the platform (Section 2.3), and collect follow-up data to better interpret the preferences they expressed (Section 2.4).

2.2 Food Challenge and Platform Presentation

Food Challenge and Meal Categories. Section 2 of the survey introduces respondents to a food challenge associated with the free meal orders. The goal of the challenge is to limit their consumption of animal products and/or refined carbs for 5 days in a row and instead focus on plant-based sources of protein and vegetables. Participants are explicitly told that there is no financial reward for completing the challenge and that the only prize is what they might learn by challenging themselves. To structure the challenge, we divided the meals that might be available on the platform into three mutually exclusive categories: Daily Harvest (G), Carb Powerhouse (O), and Carnivore Corner (R).[5] Table 2 summarizes the general content of each meal category, and Table B.1 in Appendix B provides examples of typical meals available under each category at one of the colleges.

Design of the Food Ordering Platform and Menu. Meal categories are combined into a set of 7 possible menus $\mathscr{M} := \{\text{GOR}, \text{GO}, \text{GR}, \text{OR}, \text{G}, \text{O}, \text{R}\}$. Since different platform versions will only differ in the number of available meal categories, we will interchangeably use "menu" or "platform" to refer to a specific combination of meal categories. For instance, a platform that represents the menu GOR can

[4] These variables will be presented as summary statistics in the appendix of the paper, overall and broken down by college.

[5] We use the letters G, O and R for the green, orange and red colors often associated with vegetables, carbohydrates and meat but these letters are not used in the actual survey.

be accessed via https://foodivery-college-shop-1.myshopify.com. Importantly, regardless of the number of meal categories available on the platform, participants must order all 5 meals from the same meal category. Thus, while having access to a larger platform allows one to postpone the choice of meal category to order from, it does not provide more variety in the combination of meals. To reduce any room for misunderstanding and ambiguity regarding the challenge, meal categories, or differences between platform versions, respondents have to answer multiple quiz questions throughout the survey, some of which require them to explicitly interact with the food ordering platform. More details about the food ordering process on the platform are available in Appendix B.

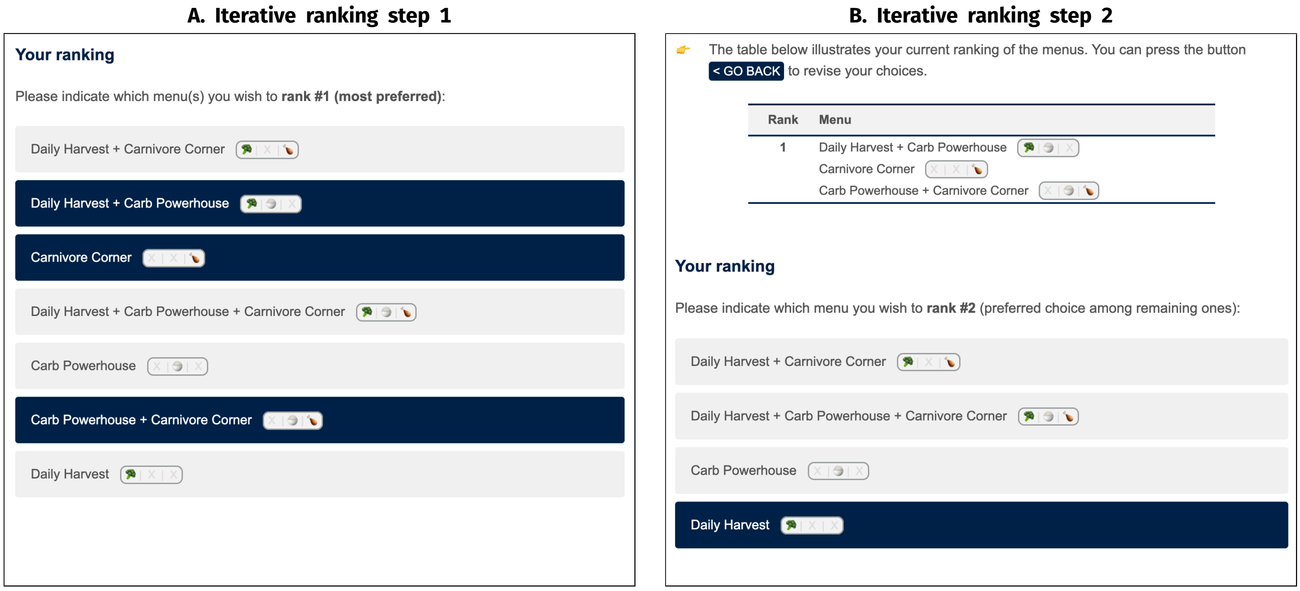
2.3 Elicitation of Preferences Over Platforms

Overall Structure and Notation. In Section 3 of the survey, we elicit respondents' preferences over $\mathscr{M}$, the set of 7 possible menus on the platform. At the start of the section, respondents are informed that their preferences will be elicited through a series of decision tasks and that their responses in one randomly selected task will determine their assigned platform (if invited to make an order). Using a within-subject design, we elicit respondents' preferences twice using three elicitation methods: iterative ranking (method R), monetary valuations (method V), and binary choices (method B). Formally, let $\succsim_{j,r}$ denote the binary relation that captures the respondent's expressed preference over two elements M, $M' \in \mathscr{M}$ under elicitation method j and round $r \in \{1,2\}$, where $j \in \mathscr{J} := \{R, V, B\}$. To control for order effects, we randomize between subjects the order of the three methods e.g., $(\succsim_{B,r}, \succsim_{V,r}, \succsim_{R,r})$, but we keep the same fixed order between the two rounds e.g., $(\succsim_{B,1}, \succsim_{V,1}, \succsim_{R,1}; \succsim_{B,2}, \succsim_{V,2}, \succsim_{R,2})$. In total, participants thus complete 6 decision tasks to provide their preferences. They are not informed of the total number of decision tasks a priori. After the first round, a screen notifies them that they will complete another round of the three decision tasks to either confirm or reconsider their preferences, with all tasks having an equal chance of determining their menu. To minimize submission errors, each decision task ends with a summary page displayed for at least 5 seconds; respondents are asked to inspect the consequences of their choices, and revise them if necessary. Once confirmed, choices can no longer be revised. Below we describe our choice of elicitation methods in detail; a discussion of alternative implementations is provided in Section D.1.3 of Appendix D.

Table 2. Overview of the meal categories.

Meal Category	Description
Daily Harvest (G)	Vegetarian-friendly meals rich in vegetables and plant-based sources of protein and carbs. *Carb-controlled. No meat or fish.*
Carb Powerhouse (O)	High-energy meals with a large portion of starchy carbs such as potatoes, pasta or rice, accompanying plant-based protein options. *No meat or fish.*
Carnivore Corner (R)	High-protein meals with a meat or fish option such as chicken, beef, pork, salmon or cod, complemented by a mix of vegetables and plant-based sources of carbs. *Carb-controlled.*

2.3.1 Iterative Ranking (R)



A. Iterative ranking step 1

Your ranking

Please indicate which menu(s) you wish to **rank #1 (most preferred)**:

- Daily Harvest + Carnivore Corner
- Daily Harvest + Carb Powerhouse
- Carnivore Corner
- Daily Harvest + Carb Powerhouse + Carnivore Corner
- Carb Powerhouse
- Carb Powerhouse + Carnivore Corner
- Daily Harvest

B. Iterative ranking step 2

The table below illustrates your current ranking of the menus. You can press the button < GO BACK to revise your choices.

Rank	Menu
1	Daily Harvest + Carb Powerhouse Carnivore Corner Carb Powerhouse + Carnivore Corner

Your ranking

Please indicate which menu you wish to **rank #2** (preferred choice among remaining ones):

- Daily Harvest + Carnivore Corner
- Daily Harvest + Carb Powerhouse + Carnivore Corner
- Carb Powerhouse
- Daily Harvest

Figure 2. Illustration of the iterative ranking method. Panel A shows the first step in the ranking procedure where the respondent assigned rank #1 to the options GO, OR, and R. Panel B shows the second step of the procedure, where G was assigned rank #2; the table at the top summarizes the choices made in Step 1.

We elicit an ordinal ranking from respondents in an iterative way. At the start of the procedure, the 7 menus are listed in a random order (with a different order between rounds).[6] Respondents first select the menu(s) they wish to designate as their rank #1 (i.e., top) option(s). After this first step, the list of menus is updated and respondents select their #2 option(s). The procedure continues until the list of remaining options is empty. Importantly, respondents can express as many indifferences as they wish by assigning the same rank to two or more menus. This procedure generates a weak order $\succsim_{R,r}$ on $\mathscr{M}$. For instance, the respondent in Figure 2 assigned rank #1 to GO, OR, and R and rank #2 to G only, which we interpret as GO $\sim_{R,r}$ OR $\sim_{R,r}$ R $\succ_{R,r}$ G. At every iteration in the ranking, a table at the top of the page summarizes the current ranking of the menus (see panel B of Figure 2).

To elicit a truthful report of the entire ordering, we rely on a probabilistic assignment procedure: the platform version that a respondent receives is determined by a lottery assigning a higher probability of selection to a better-ranked menu (and an equal probability in expectation to equally-ranked menus).[7] This procedure is incentive-compatible under the assumption that respondents' preferences respect first-order stochastic dominance on the set of lotteries over $\mathscr{M}$, with the caveat that truth-telling is only weakly

[6] By randomizing the order of the menus between rounds, we aim to minimize order effects and create some variation in respondents' perceptions when repeating the task, thus making it more difficult to mechanically repeat the pattern from the first elicitation.

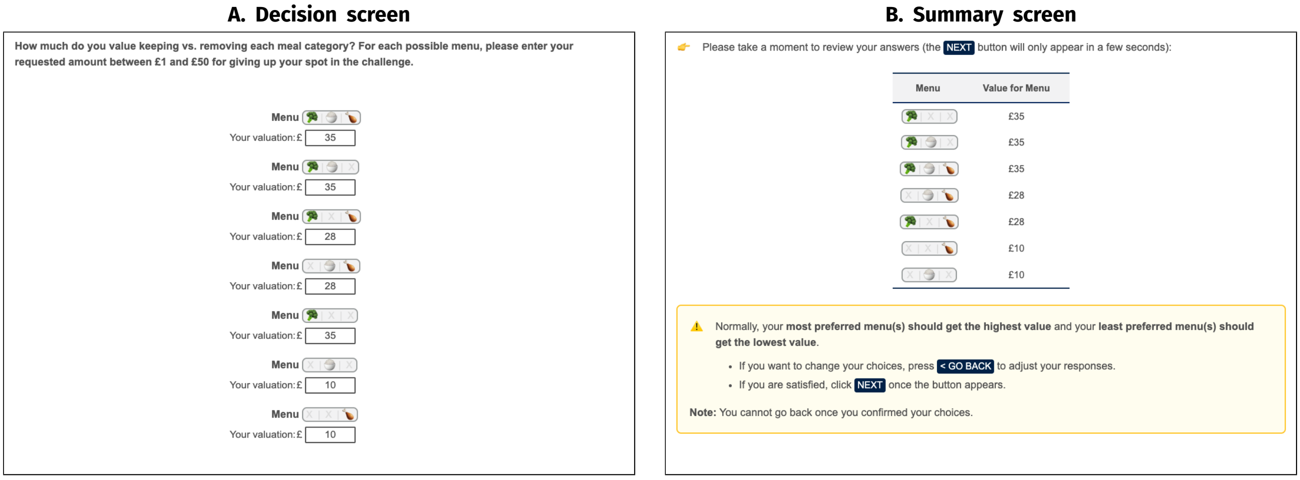
[7] More precisely, the vector of implementation probabilities is $\boldsymbol{P} = (0.35, 0.3, 0.2, 0.1, 0.03, 0.02, 0)$, where $0.35 = P(\text{best rank})$ and $0 = P(\text{worst rank})$, which is information that respondents can access by clicking on a button. For example, if GO, OR, and R are assigned rank #1, then each platform version is equally likely to be drawn with probability 0.35, 0.3 or 0.2 and the menu ranked #2, G in our example, is implemented with probability 0.1.

dominant for indifferences.[8] To provide a strict incentive to report indifferences, the iterative ranking procedure makes them easier to report. Indeed, while a respondent who wants to express a strict ranking of all 7 menus needs to take 7 steps to complete the procedure, someone who is indifferent between all menus can complete the procedure in a single step.

2.3.2 Monetary Valuations (V)

We elicit monetary valuations from respondents for each of the 7 menus using a WTA framing. Concretely, we ask participants for the minimum amount of money $v_M \in [0, 35]$ they would request (in £) to give up their spot in the challenge if offered menu M on the platform.[9] As for the iterative ranking procedure, the 7 menus are listed in a random order and respondents are asked to enter their valuation below each menu (see panel A of Figure 3).

We explain to respondents that they should request a higher amount of money for menus they like more and the same amount for menus they like equally. We then identify a strict preference for menu M over menu M' if $v_M - v_{M'} > 0$, and an indifference between the two menus if $v_M - v_{M'} = 0$. With this interpretation, the elicitation of monetary valuations $\{v_M\}_{M \in \mathscr{M}}$ induces a weak order $\succsim_{V,r}$ on $\mathscr{M}$. For instance, the respondent in Figure 3 requested £35 for GOR, GO and G, £28 for OR and GR, and £10 for O and R, implying GOR $\sim_{V,r}$ GO $\sim_{V,r}$ G $\succ_{V,r}$ OR $\sim_{V,r}$ GR $\succ_{V,r}$ O $\sim_{V,r}$ R. After making all their entries, participants proceed to a summary page that lists menus in a descending order by monetary valuation



A. Decision screen

How much do you value keeping vs. removing each meal category? For each possible menu, please enter your requested amount between £1 and £50 for giving up your spot in the challenge.

Menu 🥦 | 🍚 | 🍗
Your valuation: £ 35

Menu 🥦 | 🍚 | X
Your valuation: £ 35

Menu 🥦 | X | 🍗
Your valuation: £ 28

Menu X | 🍚 | 🍗
Your valuation: £ 28

Menu 🥦 | X | X
Your valuation: £ 35

Menu X | 🍚 | X
Your valuation: £ 10

Menu X | X | 🍗
Your valuation: £ 10

B. Summary screen

Please take a moment to review your answers (the NEXT button will only appear in a few seconds):

Menu	Value for Menu
🥦 | X | X	£35
🥦 | 🍚 | X	£35
🥦 | 🍚 | 🍗	£35
X | 🍚 | 🍗	£28
🥦 | X | 🍗	£28
X | X | 🍗	£10
X | 🍚 | X	£10

⚠ Normally, your **most preferred menu(s) should get the highest value** and your **least preferred menu(s) should get the lowest value**.

- If you want to change your choices, press < GO BACK to adjust your responses.
- If you are satisfied, click NEXT once the button appears.

Note: You cannot go back once you confirmed your choices.

Figure 3. Illustration of the interface for the monetary valuations method. Panel A shows the decision screen, where the participant entered £35 for GOR, GO and G, £28 for OR and GR, and £10 for O and R. Panel B shows the summary screen with the menus ordered in terms of their valuations.

[8] Indeed, while a standard decision-maker who strictly prefers G to GOR has a strict incentive to rank G better than GOR, someone who is indifferent between G and GOR would be equally happy with any probability distribution over these options.

[9] We chose a lower bound of £0 rather than £1 to distinguish participants who wish to opt out from the challenge altogether from participants who have a very low valuation. Notably, we do not allow for negative values of WTA i.e., we cannot identify respondents who would request to be paid to consume certain meals (Krajbich et al., 2012). However, our prior is that only a negligible number of respondents will indicate £0; furthermore, what we are mostly interested in is the ranking of the options.

(updated whenever choices are revised). By ordering the menus, we hope to reduce the cognitive burden and minimize erroneous answers.

The truthful reporting of all valuations is incentivized via a two-stage procedure. In stage 1, one menu M is randomly selected for consideration. In stage 2, the Becker-DeGroot-Marschak (BDM) mechanism is applied for this selected menu i.e., a random number X is drawn from $U[0,35]$ and compared to v_M: if $X \geq v_M$, the respondent receives £X without participating in the challenge; if $X < v_M$, she participates in the challenge with menu M. Note that given the two-stage procedure and randomness of the BDM mechanism, participants effectively face a compound lottery. This implementation was chosen for two reasons. First, we cannot require people to pay for the challenge participation, and so we have to measure WTA instead of WTP for logistical reasons. Second, we ask respondents to value each menu in absolute terms instead of relative to a reference menu e.g., GOR. While the latter approach would have been a more direct way of measuring preferences for commitment vs. flexibility, we wanted to keep the framing as similar as possible to the iterative ranking. Furthermore, we conjectured that our inferences would be sensitive to the choice of reference menu.

2.3.3 Binary Choices (B)

The last method elicits preferences for commitment or flexibility through a list of 12 binary choice comparisons, where each row compares a menu to a proper subset.[10] We randomize at the individual level whether the smaller menu is shown on the left or the right side, and between rounds, the order of the 12 binary comparisons. For each binary comparison, respondents indicate whether they prefer the larger menu, the smaller menu, or are indifferent between the two. To align the cost of expressing indifferences across elicitation methods, we present the option "I like both equally" as the default.

To minimize the risk that participants accidentally submit an indifference, a warning message is displayed whenever they expressed an indifference in at least one comparison. Figure 4 provides an illustration of the method, with the smaller menus presented on the left-hand side. In this example, the respondent chose GOR over all its subsets (first 6 rows), which we interpret straightforwardly as GOR $\succ_{B,r} M$ for all $M \in \mathscr{M} \setminus \{\text{GOR}\}$, selected "I like both equally" when comparing G to GO (i.e., G $\sim_{B,r}$ GO), and chose the smaller menu otherwise.

To elicit truthful reports, we implement the following lottery. We first randomly select one of the 12 binary comparisons and implement the expressed preferences. If a respondent indicated "I like both equally", one of the two menus is chosen at random. Hence, it is (weakly) dominant for participants to report their preferences truthfully.

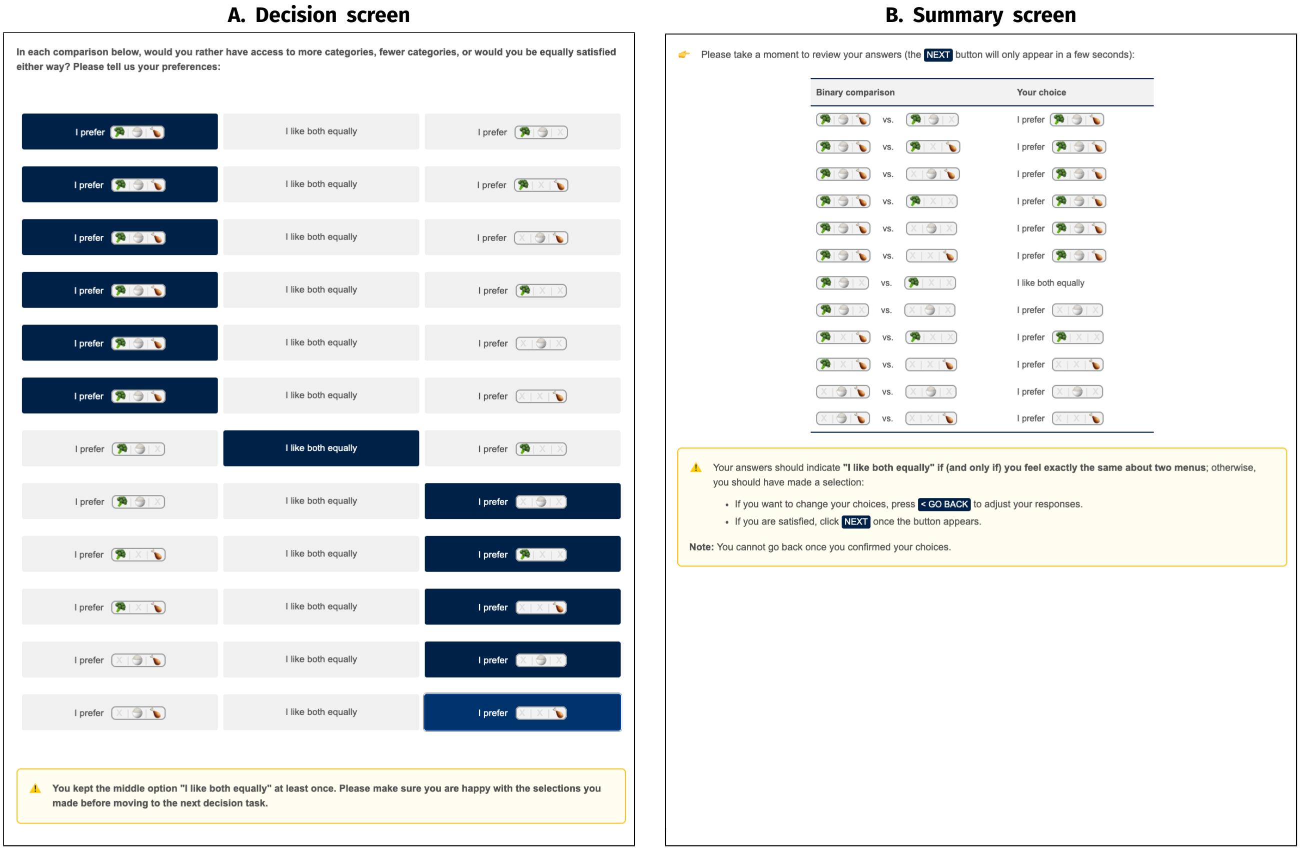
Unlike the other two methods, this binary choice procedure generates an incomplete and possibly intransitive order $\succsim_{B,r}$ on $\mathscr{M}$. First, it is incomplete because respondents are only asked to consider the 12 binary comparisons that involve two *nested* menus ($M, M' \in \mathscr{M}$ such that $M' \subset M$), instead of all $21 = (7 \times 6)\,/\,2$ binary comparisons between any two elements of $\mathscr{M}$. Second, this procedure can generate preference cycles of the form GOR $\succ_{B,r}$ GO, GO $\succ_{B,r}$ G, and G $\succ_{B,r}$ GOR. While asking only for a subset of comparisons without imposing transitivity precludes us from making one-to-one comparisons

[10] The set is
$\mathscr{P} := \{(\text{GOR},\text{GO}),(\text{GOR},\text{GR}),(\text{GOR},\text{OR}),(\text{GOR},\text{G}),(\text{GOR},\text{O}),(\text{GOR},\text{R}),(\text{GO},\text{G}),(\text{GO},\text{O}),(\text{GR},\text{G}),(\text{GR},\text{R}),(\text{OR},\text{O}),(\text{OR},\text{R})\}$.

with the other two methods, we considered this choice to strike a good balance between parsimony and informativeness of the elicitation. Indeed, examining all 21 binary comparisons twice would make the procedure much more tedious relative to the other two methods, thus undermining comparability. At the same time, the 12 comparisons we examine are sufficient to assess the robustness of commitment preferences over $\mathscr{M}$, which is the central focus of this paper. In addition, with this subset of 12 comparisons, we have sufficient information to directly test for violations of transitivity, and thus better understand the extent to which imposing transitivity (e.g., by requiring a ranking) might be restrictive.

2.4 Interpreting Expressed Preferences

Choosing the Preferred Elicitation Method. After reporting their preferences twice using the three elicitation methods, we introduce respondents to a final incentivized task based on the 12 binary comparisons between a larger menu M and a smaller menu M' such that $M' \subset M$. The purpose of this task is to investigate whether participants prefer one of the three elicitation methods. To mitigate any confounding effects, they were not informed about this procedure before expressing their preferences using the three methods.



A. Decision screen

In each comparison below, would you rather have access to more categories, fewer categories, or would you be equally satisfied either way? Please tell us your preferences:

[✓ I prefer 🥦 | 🍚 | 🍗]	I like both equally	I prefer 🥦 | 🍚 | X
[✓ I prefer 🥦 | 🍚 | 🍗]	I like both equally	I prefer 🥦 | X | 🍗
[✓ I prefer 🥦 | 🍚 | 🍗]	I like both equally	I prefer X | 🍚 | 🍗
[✓ I prefer 🥦 | 🍚 | 🍗]	I like both equally	I prefer 🥦 | X | X
[✓ I prefer 🥦 | 🍚 | 🍗]	I like both equally	I prefer X | 🍚 | X
[✓ I prefer 🥦 | 🍚 | 🍗]	I like both equally	I prefer X | X | 🍗
I prefer 🥦 | 🍚 | X	[✓ I like both equally]	I prefer 🥦 | X | X
I prefer 🥦 | 🍚 | X	I like both equally	[✓ I prefer X | 🍚 | X]
I prefer 🥦 | X | 🍗	I like both equally	[✓ I prefer 🥦 | X | X]
I prefer 🥦 | X | 🍗	I like both equally	[✓ I prefer X | X | 🍗]
I prefer X | 🍚 | 🍗	I like both equally	[✓ I prefer X | 🍚 | X]
I prefer X | 🍚 | 🍗	I like both equally	[✓ I prefer X | X | 🍗]

⚠ You kept the middle option "I like both equally" at least once. Please make sure you are happy with the selections you made before moving to the next decision task.

B. Summary screen

👉 Please take a moment to review your answers (the NEXT button will only appear in a few seconds):

Binary comparison			Your choice
🥦 | 🍚 | 🍗	vs.	🥦 | 🍚 | X	I prefer 🥦 | 🍚 | 🍗
🥦 | 🍚 | 🍗	vs.	🥦 | X | 🍗	I prefer 🥦 | 🍚 | 🍗
🥦 | 🍚 | 🍗	vs.	X | 🍚 | 🍗	I prefer 🥦 | 🍚 | 🍗
🥦 | 🍚 | 🍗	vs.	🥦 | X | X	I prefer 🥦 | 🍚 | 🍗
🥦 | 🍚 | 🍗	vs.	X | 🍚 | X	I prefer 🥦 | 🍚 | 🍗
🥦 | 🍚 | 🍗	vs.	X | X | 🍗	I prefer 🥦 | 🍚 | 🍗
🥦 | 🍚 | X	vs.	🥦 | X | X	I like both equally
🥦 | 🍚 | X	vs.	X | 🍚 | X	I prefer X | 🍚 | X
🥦 | X | 🍗	vs.	🥦 | X | X	I prefer 🥦 | X | X
🥦 | X | 🍗	vs.	X | X | 🍗	I prefer X | X | 🍗
X | 🍚 | 🍗	vs.	X | 🍚 | X	I prefer X | 🍚 | X
X | 🍚 | 🍗	vs.	X | X | 🍗	I prefer X | X | 🍗

⚠ Your answers should indicate **"I like both equally" if (and only if) you feel exactly the same about two menus**; otherwise, you should have made a selection:

- If you want to change your choices, press < GO BACK to adjust your responses.
- If you are satisfied, click NEXT once the button appears.

Note: You cannot go back once you confirmed your choices.

Figure 4. Illustration of the interface for the binary choice elicitation method. Panel A shows the decision screen, where the participant indicated preferring GOR over all its proper subsets, GO $\sim_B$ G (middle row), and the singleton menu was preferred in the 5 remaining cases. Panel B shows the summary screen with a summary of the participant's choices.

The procedure works as follows. First, participants indicate which method j and round r should determine their menu in case this final task is the one used for implementation.[11] Because we do not wish to enforce the choice of a method-round pair, we also offer participants a costless randomization device to select a method and/or round. If this task determines their menu, one of the 12 binary comparisons will be drawn at random and, based on the chosen method-round pair (j, r) and comparison (M, M'), the respondent will receive: menu M if $M \succ_{j,r} M'$, or an equal chance of M or M' if $M \sim_{j,r} M'$.[12]

To illustrate the procedure, suppose a participant indicated a preference for the ranking method and second round $(j = R, r = 2)$. Suppose further that the randomly drawn comparison is GO vs. R. If the respondent ranked GO better than G, expressing GO $\succ_{R,2}$ G, she receives GO as her platform version (and similarly for G ranked better than GO); if she assigned them equal ranks (i.e., GO $\sim_{R,2}$ G), each platform version is selected with equal probability.

Debriefing Questions. In the last section of the survey, we ask debriefing questions to facilitate the interpretation of the choice data. First, we ask participants which meals they would prefer to try, and how many meal categories they would prefer the platform to show – we use answers to these simple unincentivized questions to perform a prediction exercise (see Section 5.3). We also ask them to report any difficulty with making up their mind on the best menu. Next, we elicit participants' subjective perceptions of each elicitation method in terms of difficulty, tediousness of the process, and certainty about their answers. To learn more about intransitivities, we ask participants to rate the choice coherence of a transitivity violation in some hypothetical situation. The survey concludes with respondents' comments about the challenge concept and decision tasks. All variables collected in this survey are available in Table C.1, Appendix C.

3 Main Outcomes and Benchmark Data

In this section, we provide a roadmap of our planned analyses, with a definition of our main outcome measures and a justification of the analytical choices we propose. We lay out all components of our dataset, including the benchmark data (simulated data and expert survey) against which our main outcomes will be judged. Finally, we explain how we make use of prior pilot data to present our analyses in the next sections.

3.1 Main Outcome Measures

Our main outcomes pertain to the frequency of commitment choices, the consistency of expressed preferences across methods, and the stability of expressed preferences between two rounds of the same method. Below we provide definitions of the index measures we construct to capture each of these outcomes.

[11] Participants are informed that this final task has the same chance of determining their assigned menu as the other 6 decision tasks.

[12] One alternative to this procedure would have been to let participants alter the implementation probability of the 6 decision tasks e.g., by allowing them to assign a 50% chance of selection to the ranking method in round 1 (instead of a 1/6 chance). However, doing so would introduce multiple confounds because each method generates different probabilities of obtaining a given menu. To address this issue, we fix the binary comparison (M, M') to which the elicited preferences are applied.

3.1.1 Prevalence of Commitment

Definition: Preference Indicator. For each respondent i, elicitation method $j \in \{R, V, B\}$, round $r \in \{1, 2\}$ and menu pair $p = (M, M')$ such that $M' \subset M$, we introduce a *preference indicator* $P^i_{j,r}(p)$, which captures the respondent's preference

$$P^i_{j,r}(p) \; = \; \begin{cases} -1 & \text{if } M' \succ_{j,r} M \quad \text{(preference for commitment)} \\ 0 & \text{if } M \sim_{j,r} M' \quad \text{(indifference)} \\ 1 & \text{if } M \succ_{j,r} M' \quad \text{(preference for flexibility)} \end{cases} \tag{1}$$

For example, $P^i_{R,1}(\text{GOR}, \text{GO}) = 1$ if respondent i assigns a better rank to GOR than GO in round 1 and $P^i_{V,2}(\text{GOR}, \text{GO}) = -1$ if she assigns a lower valuation to GOR than GO in round 2.

Definition: Preference Index. In most analyses, we will be interested in evaluating the (relative) prevalence of commitment preferences at the individual level. To this end, we calculate a *commitment*, *indifference*, and *flexibility index* for each respondent i in the following way:

$$C^i_{j,r} \; := \sum_{p \in \mathscr{P}} \mathbb{1}\left[P^i_{j,r}(p) \; = \; -1\right] \qquad I^i_{j,r} \; := \sum_{p \in \mathscr{P}} \mathbb{1}\left[P^i_{j,r}(p) \; = \; 0\right] \qquad F^i_{j,r} \; := \sum_{p \in \mathscr{P}} \mathbb{1}\left[P^i_{j,r}(p) \; = \; 1\right]. \tag{2}$$

For instance, $C^i_{j,r}$ counts the number of times respondent i expressed a preference for commitment when using method j at round r across all 12 menu pairs $p \in \mathscr{P}$.[13] Hence, each index takes a value between 0 and 12, where for any given method j, round r, and individual i, $C^i_{j,r} + I^i_{j,r} + F^i_{j,r} = 12$. In the main text, we will concentrate our analyses on the distribution of the commitment index and report the corresponding numbers for the indifference and flexibility indices in Appendix F.

3.1.2 Consistency of Expressed Preferences

Two methods might generate similar prevalence rates of commitment choices without necessarily producing consistent answers at the individual level.[14] To study this issue, we construct summary indices that measure the frequency with which a given respondent expresses the same preferences when comparing two or all three methods.

[13] As a reminder, the set of 12 binary comparisons is:
$\mathscr{P} := \{(\text{GOR}, \text{GO}), (\text{GOR}, \text{GR}), (\text{GOR}, \text{OR}), (\text{GOR}, \text{G}), (\text{GOR}, \text{O}), (\text{GOR}, \text{R}), (\text{GO}, \text{G}), (\text{GO}, \text{O}), (\text{GR}, \text{G}), (\text{GR}, \text{R}), (\text{OR}, \text{O}), (\text{OR}, \text{R})\}$.

[14] For instance, suppose that the preferences expressed by respondent i in round 1 using methods $j \in \{R, V\}$ are as follows. Using the ranking, the respondent indicates $M' \succ_{R,1} M$ if $M' \in \{\text{GO}, \text{GR}, \text{OR}\}$ and $M = \text{GOR}$, and $M' \sim_{R,1} M$ otherwise. Using the monetary valuation task, the respondent indicates $M' \succ_{V,1} M$ if $M' \in \{\text{G}, \text{O}, \text{R}\}$ and $M = \text{GOR}$, and $M' \sim_{V,1} M$ otherwise. Then $C^i_{R,1} = C^i_{V,1} = 3$ but the ordinal ranking and monetary valuations entirely disagree on the comparisons that generate commitment.

Definition: Consistency Indices. For each respondent i and round r, we define *aggregate consistency indices*, which count the number of menu pairs (between 0 and 12) at which the expressed preferences agree across all three methods

$$\varGamma^i_{3,r} := \sum_{p\in\mathscr{P}} \mathbb{1}\left[P^i_{R,r}(p) = P^i_{B,r}(p) = P^i_{W,r}(p)\right], \tag{3}$$

between exactly two methods

$$\varGamma^i_{2,r} := \sum_{\substack{j,j',j''\in\{B,R,V\} \\ j\neq j'\neq j''}} \frac{1}{2} \sum_{p\in\mathscr{P}} \left[P^i_{j,r}(p) = P^i_{j',r}(p) \wedge P^i_{j',r}(p) \neq P^i_{j'',r}(p)\right], \tag{4}$$

or between none of the methods, $\varGamma^i_{0,r} := 12 - \varGamma^i_{3,r} - \varGamma^i_{2,r}$. Similarly, we define a *bilateral consistency index* for each pair of elicitation methods (j,j') with $j \neq j'$ as

$$\gamma^i_r(j,j') := \sum_{p\in\mathscr{P}} \mathbb{1}\left[P^i_{j,r}(p) = P^i_{j',r}(p)\right]. \tag{5}$$

3.1.3 Stability of Expressed Preferences

We construct a similar set of indices to measure the stability of a respondent's answers, this time fixing the method and comparing round 1 vs. round 2.

Definition: Stability Index. For each respondent i and method j, we define the *stability index* as

$$\rho^i_j := \sum_{p\,\in\,\mathscr{P}} \mathbb{1}\left[P^i_{j,1}(p) = P^i_{j,2}(p)\right]. \tag{6}$$

The index counts the number of menu pairs at which the preferences expressed with method j agree between the two rounds and takes values in $\{0, 1, \ldots, 12\}$. This index is thus a measure of within-method stability.

3.2 Benchmark Datasets

To evaluate the prevalence of our outcomes of interest, we need to provide benchmarks. Clearly, setting the null hypothesis to zero is uninformative, since some commitment choices, inconsistencies, and instabilities will necessarily arise in the data. We propose to examine two types of benchmarks, one based on random choice data and the other based on the beliefs of experts.

3.2.1 Simulated Random Benchmark

The first benchmark allows us to examine the extent to which our findings could be rationalized by idiosyncratic noise in decision-making, without assuming an underlying core preference. Below we explain the properties of this feedback and discuss one variation that we will present in the appendix.

Proposed Benchmark (Pure Noise). Our proposed benchmark consists of simulated data from purely random choice, abstracting away from any correlation structure between methods or rounds. For the iterative ranking and monetary valuations, we draw a $\text{rank}_{M,r} \sim U\{1, 2, ..., 7\}$ and $\text{valuation}_{M,r} \sim U\{0, 1, ..., 35\}$ for each $M \in \mathscr{M}$ and round r, respectively. For binary choices, we draw a $\text{choice}_{p,r} \sim U\{-1, 0, 1\}$, reflecting a preference for commitment, indifference, or flexibility, for each menu pair $p \in \mathscr{P}$ and round r. We translate the resulting choice data into our main outcome measures as described in Section 3.1. Such a benchmark allows us to assess how the choice of answer grid in each decision task influences the level of commitment, inconsistency, or instability that can be produced.

Alternative Benchmark. Assuming idiosyncratic noise that is drawn uniformly from each corresponding choice grid has implications for the mass of indifferences that can be generated under each method. In particular, the prediction of this model is that respondents should express a larger mass of indifferences for $j = B$ due to the coarseness of the grid $(-1, 0, 1)$, and a much lower mass of indifferences for $j = V$ due to respondents being able to pick any of 36 numbers, making the chances of any two equal draws unlikely. For a more realistic noise structure, we will consider a slightly different benchmark where we assume that respondents draw their valuations from a discrete uniform with round numbers as support points.[15] We will discuss its properties and relation to the actual choice data in the appendix.

3.2.2 Expert Prediction Benchmark

The first benchmark makes no use of prior evidence on the prevalence of commitment decisions, the consistency of preference measurements across elicitation methods, or the stability of these measurements within a method. We wish to assess the extent to which our findings simply confirm prior knowledge or instead lead to non-trivial updating on the outcomes of interest. For this reason, we will also collect incentivized expert forecasts. Below we describe our sampling strategy and elicitation of forecasts.

Recruitment and Sample Size. For the expert prediction benchmark, we will host an expert survey on the Social Science Prediction Platform (SSPP) and collect information on whether the forecasters have done prior work on commitment and self-control or preference consistency/stability. We invite all potential forecasters to participate in a 15-minute survey. We will incorporate the forecasting data in our analyses as long as we manage to collect at least 30 responses. We set this parameter based on a review of several recent papers, which document that using crowds of 5 to 10 forecasters already provides informative signals on the current state of knowledge by significantly reducing the influence of extreme individual forecasts (DellaVigna and Pope, 2018; Otis, 2022; Iacovone et al., 2023). More details about recruitment can be found in Section E.1, Appendix E.

Elicitation of Forecasts. Experts will be first introduced to the study setup and food ordering platform, and offered an opportunity to experience the various elicitation methods as presented to study

[15] Indeed, valuations in our pilot data were mostly integers that are multiples of 5 (see Figure D.2, Appendix D). Thus, one approach could be to assume $\text{valuation}_{M,r} \sim U\{0, 5, 10, ..., 35\}$ for each $M \in \mathscr{M}$ and round r; alternatively, valuations could be drawn from a theoretical distribution that mimics the empirical distribution of the pilot data (e.g., with more mass on middle values).

Table 3. Overview of elicited expert forecasts.

Prediction set	Number of forecasts	Quantities predicted
(1)	9	For each method $j \in \{R, V, B\}$, average value of the commitment, indifference, and flexibility indices: $\bar{C}_{j,1}$, $\bar{I}_{j,1}$, and $\bar{F}_{j,1}$
(2)	3	Average number of comparisons where the expressed preference is consistent between all three methods, only two methods, and none: $\bar{\Gamma}_{3,1}$, $\bar{\Gamma}_{2,1}$, and $\bar{\Gamma}_{0,1}$
(3)	3	For each method $j \in \{R, V, B\}$, average value of the stability index $\bar{\rho}_j$
(4)	2	Most frequently chosen method and round

participants. After that, they will provide 17 forecasts organized in 4 prediction sets. Table 3 presents a summary of all elicited forecasts. First, experts will guess the average value across all respondents of the commitment, indifference, and flexibility indices for each method in round 1, that is, $\bar{C}_{j,1}$, $\bar{I}_{j,1}$ and $\bar{F}_{j,1}$ for $j \in \{R, V, B\}$ (where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X^i$ refers to the sample average for quantity X). Second, they will predict the average number of comparisons at which the preferences expressed in round 1 are consistent across all three methods, two methods only, and none of the methods i.e., $\bar{\Gamma}_{3,1}$, $\bar{\Gamma}_{2,1}$ and $\bar{\Gamma}_{0,1}$. Third, they will predict the average number of comparisons at which expressed preferences are stable between the rounds for each method i.e., estimate $\bar{\rho}_j$ for $j \in \{R, V, B\}$. Finally, they will predict the method and round most frequently favored by respondents in order to implement their preferences.[16] We will index all expert forecasts with a superscript E. Forecasts will be incentivized by offering £100 (for themselves or as a charity donation) to the 5 forecasters who made the closest guess in one randomly selected forecast (with ties broken randomly). To assist forecasters, the survey will provide them with statistics about the proportion of inconsistencies and instabilities that would result if decisions were made completely at random (simulated random benchmark). Table E.1 in Section E.5 provides an overview of the collected variables (including additional demographics and variables about the researcher's expertise). The survey instrument is available in Appendix E.

3.3 Statistical Decisions

When comparing two categorical variables in a contingency table, we will report the results of a χ^2 test (or Fisher exact test if more than 20% of cells have frequencies < 5). When comparing two distributions of (e.g., stability) indices, we will perform Wilcoxon signed-rank rests of equality of medians if the two distributions are paired (within-subject comparisons) and Kolmogorov-Smirnov tests if the distributions are unpaired (between-subject comparisons). Our non-parametric analyses will be complemented by linear regression models with standard errors clustered at the individual level (respondents, experts, simulation unit) in all our regressions. We will conduct hypothesis tests against our simulated and expert benchmarks in one-sample t-tests of equality of means with the benchmark values treated as constants.

[16] Ideally, we would also elicit researchers' confidence in their answers and allow them to express uncertainty. However, this would lengthen the survey and most likely lead to high attrition rates.

3.4 Sample Size Restrictions and Statistical Power

Inclusion/Exclusion Criteria. We plan to collect at least 200 complete responses (up to 400). To participate, respondents must confirm that they are at least 18 years old, are member of a University of Oxford college, agree to complete the study on a laptop or desktop computer, have no dietary restrictions, and agree to provide their college email address. We will remove all participants from our quota count who fail to meet these inclusion criteria.

The analyses presented in the main text will be performed on the full sample of complete responses. To examine sensitivity to outliers, we will provide robustness analyses in the appendix after removing all respondents who meet at least one of the following exclusion criteria: (i) the respondent finished the study in less than 5 minutes, (ii) the respondent expressed indifferences at all menu pairs in at least three decision tasks, and (iii) the respondent was not interested in the challenge i.e., indicated a value of *challenge_motivation* less than 30 (out of 100).[17] This reflects our concern that respondents with low interest in the study are more prone to random choice.

Statistical Power. Our quota of at least 200 complete responses (up to 400) was set partially for budgetary reasons and partially for logistical reasons. Here we provide information on minimum detectable effect sizes given a sample size of 200. In one-sample (two-sided) t-tests comparing our preference indices (commitment, consistency and stability) against benchmark values, we estimate that we will have 80% (90%) power to detect at the 5% significance level a difference of 0.6 (0.7) commitment decisions against a null-hypothesized value of 4 commitment decisions (out of 12), assuming a standard deviation of 3 in our sample. We consider these assumptions to be conservative based on pilot data we collected (see Section 3.5) and the predicted random benchmark values (between 4 and 5.5 depending on the method).[18]

3.5 Prior Piloting and Current State of Knowledge

In Section 4, we will make use of pilot data to illustrate our statistical analyses and figures. For transparency, below we explain how prior piloting informed our present design, highlighting similarities and major differences, and describe the modifications made to the pilot data to fit our current setup.

Piloting History. Our study design was partially informed by a series of pilots conducted between 2020 and 2022. For the most part, the pilots were unincentivized and conducted with one of two online panels (Pureprofile and Prolific). We used these pilots to learn how to (i) structure the meal challenge, (ii) design the meal categories, and (iii) present the elicitation procedure. During this piloting period, we experimented with different ways of eliciting the ranking, binary choices, and monetary valuation of

[17] Based on the distribution in our pilot data (Figure D.1, Appendix D), a value of 30 corresponds roughly to the lowest 10% of engagement.

[18] Our pilot data yielded standard deviations for the commitment indices between 2.1 and 2.8 depending on the method, with mean commitment decisions between 2.0 and 3.3. The predicted values based on the random benchmark can be seen from Table 4. For instance, making random decisions would yield a commitment take-up rate of about one third in the binary choice method (i.e., 4 commitment decisions out of 12), with a higher mean number of commitment decisions for the other two methods.

commitment vs. flexibility. Having considered several alternatives, we opted for the simplest and most neutral approach that would allow us to separate strict preferences from indifferences. Section D.1.3 of Appendix D provides a summary of several alternative approaches we considered for each method and how we converged to our current design choices. For transparency, the survey files and pilot data generated in order to develop this research will be made available on OSF upon publication.

Data Used for Illustration (Registered Report Only). To illustrate our statistical analyses and figures in Section 4, we will use data from the pilot study that is closest to our current design in terms of the elicitation methods used. We conducted a 25-minute survey on 13/01/2022 with $N_{\text{pilot}} = 91$ individuals recruited via Prolific. Section D.1.1, Appendix D outlines how this pilot study differs from our planned study design as described in Section 2. To foster comparability, we modified the pilot data as follows: First, we sampled with replacement $200 - N_{\text{pilot}} = 109$ individuals from the pilot dataset to arrive at a sample size of $N = 200$. Second, we rescaled the elicited monetary valuations from the pilot study to match the admissible range of monetary values in the current study i.e., $v_M = v_M^{\text{pilot}} / 100 \times 35$. Third, we arbitrarily generated round-2 choices by drawing round-1 choices (rank, valuation, binary choice) with 85% chance and a different choice with 15% chance.[19]

4 Consistency and Stability of Commitment Decisions

In this opening results section, we examine the robustness of commitment decisions from a dual perspective. In Section 4.1, we adopt a static single-round perspective to study the *consistency* of expressed preferences *across methods*: fixing the decision round, how comparable are the distributions of preferences generated by the three methods? Do they provide consistent answers at the individual level? In Section 4.2, we investigate the *stability* of expressed preferences *across decision rounds*: fixing the elicitation method, does the distribution of preferences remain similar with repetition? How stable are respondents' answers at the individual level? At the end of this section, we bring these two dimensions of robustness together to ask how they relate to each other: are they complementary or do they paint the same picture? To what extent do they reflect a common phenomenon?

4.1 Consistency of Commitment Decisions Across Methods

If individuals have well-defined preferences and procedural invariance holds, we should expect expressed preferences to be similar across all three methods. Section 4.1.1 examines consistency at the aggregate level and Section 4.1.2 proceeds with an individual-level analysis. Both sections focus on round-1 choices; the corresponding analyses for round 2 can be found in Appendix I.

4.1.1 Heterogeneity in the Prevalence Rate of Commitment Across Methods

Starting from the most disaggregated level, Figure 5 shows the proportion of respondents who expressed commitment preferences at a given menu pair. Despite some differences across comparisons, our simulations based on pilot data reveal a tendency for the iterative ranking method to generate higher proportions of commitment decisions, with the monetary valuation method generating the lowest proportions (and with a lower variance across comparisons). To see this more closely, we next examine how the aggregate distribution of expressed preferences for commitment changes with the elicitation method (see Figure F.1 in Appendix F for an analogous analysis of indifferences and flexibility preferences). Table 4 shows the overall prevalence rate of commitment by method, benchmarked against the simulated random data and the expert data. The corresponding prevalence rates for indifferences and flexibility preferences are reported in Table F.1, Appendix F. The monetary valuation method provides a lower bound at 17% overall and the iterative ranking method an upper bound at 28%, with binary choices providing an estimate close to monetary valuations, at 21%. Using one-sample t-tests of equality of means, we will test the null hypothesis that the population prevalence of commitment for method j in round 1, $\mu_{C_{j,1}}$, is equal to the random and expert benchmark prevalence

$$H_0 : \mu_{C_{j,1}} = \bar{C}^S_{j,1} \quad \text{and} \quad H_0 : \mu_{C_{j,1}} = \bar{C}^E_{j,1}. \tag{7}$$

We will also report the proportion of experts who correctly order the methods in terms of the frequency of commitment decisions. To further look at heterogeneity, Figure 6 shows quantile plots of the empirical commitment indices as defined in Equation 2. The corresponding quantile plots for indifferences and flexibility preferences are reported in Figure F.2, Appendix F. The median respondent expresses a preference for the smaller menu in none of the 12 binary comparisons when using monetary valuations, relative to

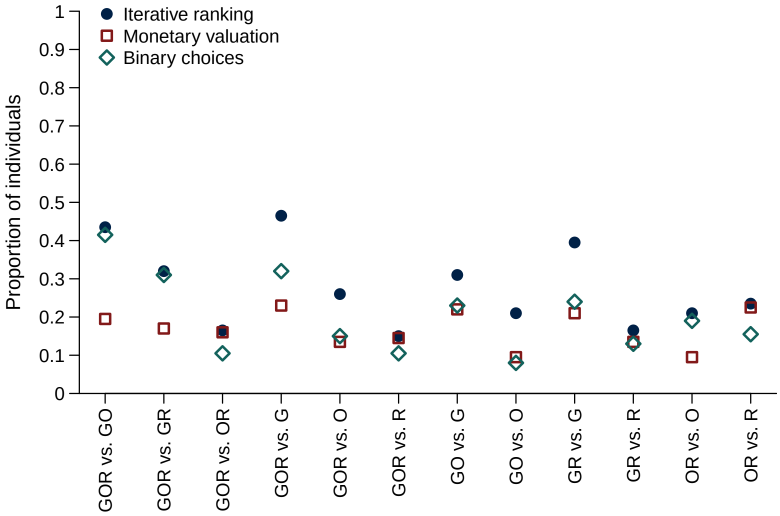


Figure 5. Proportion of individuals with commitment preferences (see Equation 1) at given menu pairs.

[19] For the iterative ranking and binary choice procedures, the choice selected with a 15% chance had an equal chance of being any choice not selected in round 1. For the monetary valuation procedure, the counterfactual valuation selected with 15% chance was drawn uniformly for each menu M from the interval $[v_{M,1} - 15, v_{M,1} + 15]$, where $v_{M,1}$ is the round 1 valuation for menu M.

Table 4. Prevalence rate of expressed commitment preferences from round-1 choices for each of the 12 binary comparisons across the three elicitation methods.

	Iterative ranking	Monetary valuations	Binary choices
Empirical	0.28	0.17	0.21
$N = 2,400$	(0.23)	(0.22)	(0.17)
Simulated	0.42	0.46	0.34
$N = 2,400$	(0.22)	(0.23)	(0.13)
Expert	0.xx	0.xx	0.xx
$N = N_E$	(0.xx)	(0.xx)	(0.xx)

Notes: Standard errors in parentheses. Since there are 12 binary comparisons per respondent, we have $200 \times 12 = 2,400$ total observations. Entries 0.xx will be populated once we collected the data.

3 and 4 comparisons for the binary choice and ranking methods, respectively. Across all methods, the top quartile of the distribution has a commitment rate of 41.6% (5 out of 12), while the bottom quartile is at 0%.

Quantitative Assessment of Overall Differences. As our previous analyses only shed light on the qualitative differences, we will perform regression analyses to learn about the quantitative magnitudes. To this end, we collect the individual round-1 commitment indices for all methods into a $(3N \times 1)$ vector $\boldsymbol{C}_1$ and regress them on method dummies, controlling for the order in which the methods were presented and the position (left or right) of the smaller menus in the binary choice task

$$\boldsymbol{C}_1 = \beta_0 + \beta_1 \mathbb{1}[j = R] + \beta_2 \mathbb{1}[j = V] + \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{\eta}$ is a vector of controls for order effects and $\boldsymbol{\epsilon}$ is a mean-zero error that is independent across, but possibly correlated within, respondents. The coefficients β_1 and β_2 represent the average change in the commitment index when the corresponding method is compared to the reference method i.e., binary choice. We will test the null hypothesis $\beta_1 = \beta_2 = 0$ using an F-test.

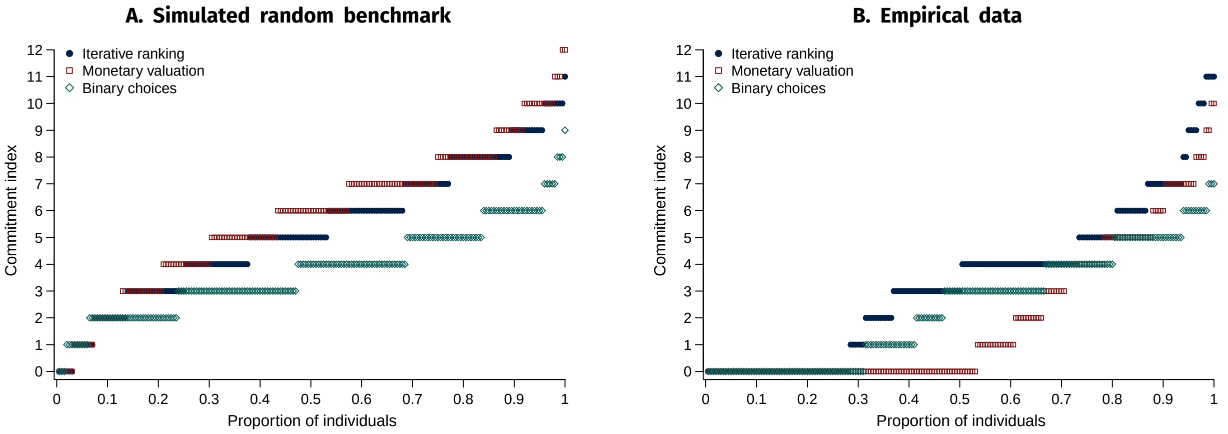


Figure 6. Quantile plots of commitment indices (see Equation 2) across the three elicitation methods.

Table 5. Consistency of expressed preferences across elicitation methods.

	$\bar{\Gamma}_{3,1}$	$\bar{\Gamma}_{2,1}$	$\bar{\Gamma}_{0,1}$
Empirical	0.48	0.47	0.05
$N = 2,400$	(0.05)	(0.06)	(0.03)
Simulated	0.13	0.67	0.20
$N = 2,400$	(0.02)	(0.04)	(0.03)
Expert	0.xx	0.xx	0.xx
$N = N_E$	(0.xx)	(0.xx)	(0.xx)

Notes: Standard errors in parentheses. The first, second and third column respectively indicates the proportion of binary comparisons at which expressed preferences agree across all three methods, agree for only two methods, and differ across all three methods. Since there are 12 binary comparisons per respondent, we have $200 \times 12 = 2,400$ total observations. Entries 0.xx will be populated once we collected the data.

4.1.2 Frequency and Size of Inconsistencies at the Individual Level

Even if the aggregate distribution of preferences looks fairly similar across methods, homogeneity in the aggregate could hide substantial variation at the individual level. Below we examine the *frequency* of inconsistent choices made by each respondent and the *size* of the observed inconsistencies.

How Frequently Do Expressed Preferences Differ Between Methods? Averaging across all respondents and menu pairs, Table 5 shows the fraction of comparisons in round 1 for which the same preference is expressed across all three methods, two methods only, or no methods. These consistency rates are contrasted with the simulated and expert benchmarks. Consistency across all three methods is achieved nearly 50% of the time, a proportion that is much higher than the random benchmark, and full inconsistency occurs only 5% of the time. Using one-sample t-tests of equality of means, we will test the null hypothesis that the population consistency level in round 1 across all three methods, $\mu_{\Gamma_{3,1}}$, is equal to the random and expert benchmark consistency levels

$$H_0 \; : \; \mu_{\Gamma_{3,1}} \; = \; \bar{\Gamma}^{S}_{3,1} \quad \text{and} \quad H_0 \; : \; \mu_{\Gamma_{3,1}} \; = \; \bar{\Gamma}^{E}_{3,1}. \tag{9}$$

To examine whether consistency is larger for a given pair of methods, Figure 7 presents the distribution of the bilateral consistency indices defined in Section 3.1 for round-1 choices. We will examine whether consistency varies between the three method pairs by testing the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ using an F-test in a regression of the bilateral consistency indices stacked into a vector $\boldsymbol{\gamma}$ $(3N \times 1)$ on method-pair dummies

$$\boldsymbol{\gamma} \; = \; \beta_0 + \beta_1 \mathbb{1}\left[(j,j') \; = \; (R,V)\right] + \beta_2 \mathbb{1}\left[(j,j') \; = \; (R,B)\right] + \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{10}$$

where $\mathbb{1}\left[(j,j') = (V,B)\right]$ is the reference category, $\boldsymbol{\eta}$ is a vector of controls for order effects, and $\boldsymbol{\epsilon}$ is a mean-zero error independent across individuals, but probably correlated within individuals.

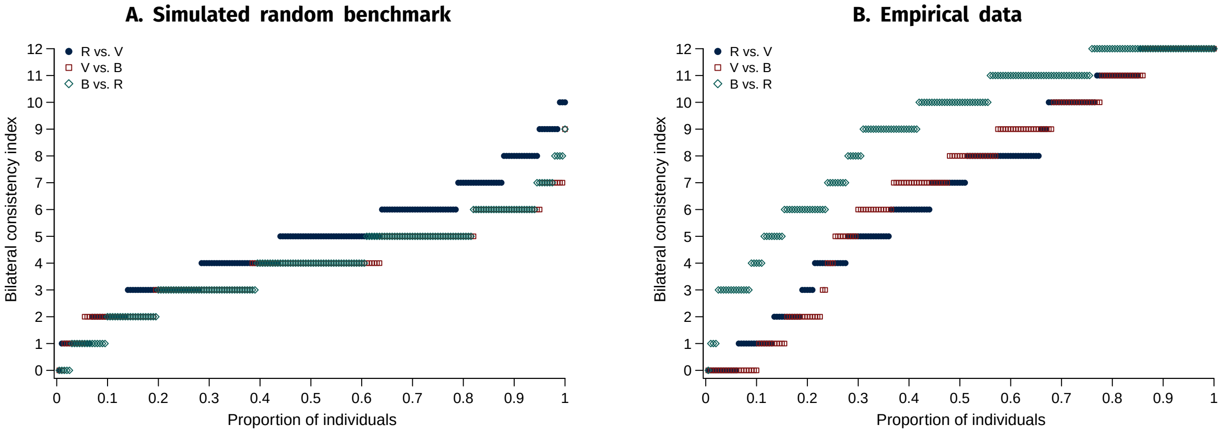


Figure 7. Quantile plots bilateral consistency indices (see Equation 5) calculated from round-1 choices across the three elicitation methods.

How Large are the Observed Inconsistencies? To measure the size of the observed inconsistencies, we will make use of the cardinal information contained in valuations. For all menu pairs (M,M') with $M' \subset M$, we will examine the distribution of the difference in round-1 valuations $v^i_{M,1} - v^i_{M',1}$ conditional on the preference expressed in method $j \in \{R,B\}$ i.e., $P^i_{j,1}(M,M') \in \{-1,0,1\}$. For instance, if $P^i_{R,1}(M,M') = -1$ (or, equivalently, $M' \succ^i_{R,1} M$), a positive difference $v^i_{M,1} - v^i_{M',1} > 0$ would indicate an inconsistency, as respondent i would assign a higher valuation to the larger menu despite having assigned a higher rank to the smaller menu. Panel A (B) of Figure 8 shows the distribution of $v^i_{M,1} - v^i_{M',1}$ conditional on commitment, indifference, and flexibility preferences being expressed in the ranking (binary choice) method. Conditional on expressing a preference for commitment, the median difference in valuations between the larger and the smaller menu is zero both for ranking and binary choice. The interquartile range conditional on commitment is small and (weakly) negative only in Panel B (binary choices). The median difference in valuations is zero conditional on expressing an indifference, but with a large interquartile range, and only slightly positive when expressing a preference for flexibility. We will test in a linear regression whether the observed inconsistencies differ in size depending on the preferences expressed.[20]

4.1.3 Nature of the Observed Inconsistencies

The previous section examined the overall frequency of inconsistencies between methods and the average size of these inconsistencies from a monetary point of view. Next, we wish to ask whether these inconsistencies are more likely to emerge conditional on the respondent having expressed a preference for commitment. In other words, should commitment decisions be trusted less than expressed indifferences or decisions to maintain flexibility? Or is noise symmetric across types of decisions?

[20] That is, we will look at the subset of all comparisons (M,M') with an inconsistency ($P^i_{V,1}(M,M') \neq P^i_{j,1}(M,M')$ for $j \in \{R,B\}$) and test whether the absolute difference in valuations $|v^i_{M,1} - v^i_{M',1}|$ differs depending on $P^i_{j,1}(M,M') \in \{-1,0,1\}$ using an F-test in a linear regression of $|v^i_{M,1} - v^i_{M',1}|$ on dummies for each value of the preference indicator $P^i_{j,1}(M,M')$.

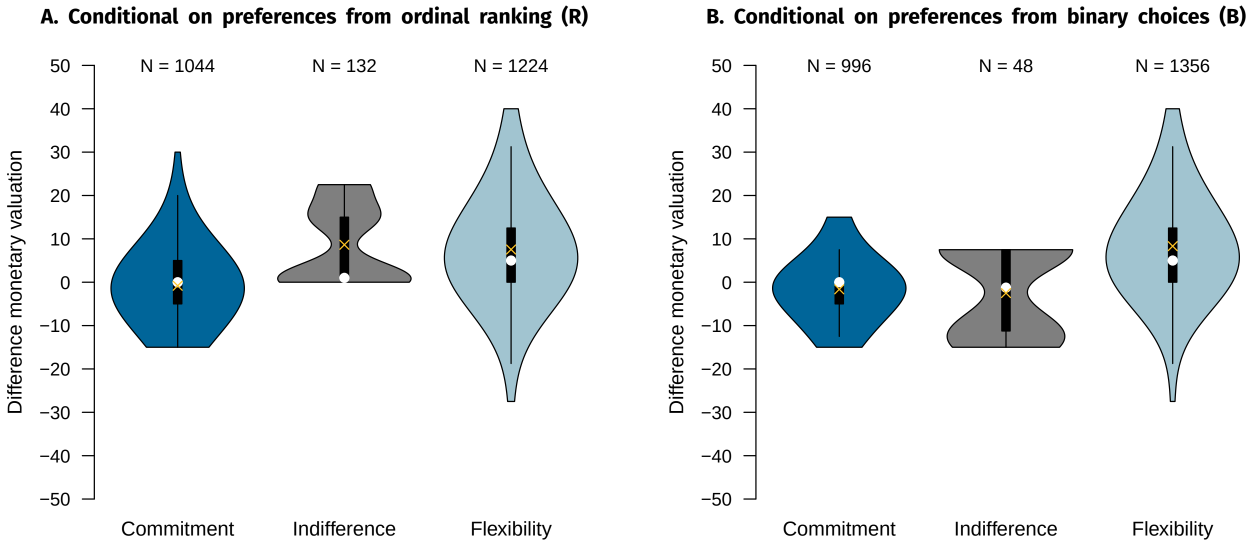


Figure 8. Violin plots of the difference in monetary valuations $v^i_{M,1} - v^i_{M',1}$ conditional on expressed commitment, indifference, and flexibility preferences ($P^i_{j,1}(M, M') \in \{-1, 0, 1\}$), pooling across all respondents i and all 12 menu pairs (M, M') such that $M' \subset M$. The white dot and orange cross indicate the median and mean, respectively.

Are Commitment Choices More Likely to Flip? Figure 9 shows the joint distribution of preference indicators (as defined in Equation 1) from round-1 choices across method pairs $(j,j') \in \{(R,B),(R,V),(V,B)\}$, pooling all binary comparisons (M,M') such that $M' \subset M$. Using this information, we can infer the type of inconsistencies and quantify the probabilities of minor inconsistencies (e.g., $M \succ_j M'$ and $M \sim_{j'} M'$) and major inconsistencies (e.g., $M \succ_j M'$ and $M \prec_{j'} M'$). For example, consider Panel A, where the marginal probability of observing a commitment preference in the ranking method is 0.28 ($= 0.10 + 0.01 + 0.17$). The probability that the expressed preference changes from a commitment to a flexibility type in the binary choice method is 0.36 ($= 0.10\,/\,0.28$).

To assess whether initial commitment choices are more likely to be undone, we will run regression analyses. For a given method pair (j,j'), let j_1 index the elicitation method that was shown first to the

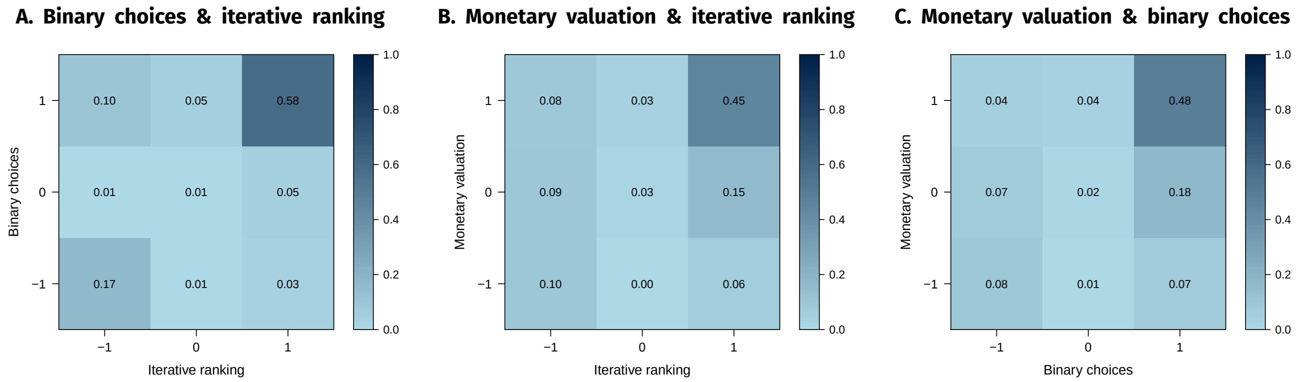


Figure 9. Joint probability distribution of preference indicators $P^i_{j,1}(M, M') \in \{-1, 0, 1\}$ from round-1 choices across the three method pairs $(j, j') \in \{(R, B), (R, V), (V, B)\}$.

respondent. We will regress the vector of individual consistency indices over all menu pairs on dummies for the preferences expressed in method j_1

$$\mathbb{1}\left[\boldsymbol{P}_{j_1,r} \neq \boldsymbol{P}_{j',r}\right] = \beta_0 + \beta_1 \mathbb{1}\left[\boldsymbol{P}_{j_1,r} = 0\right] + \beta_2 \mathbb{1}\left[\boldsymbol{P}_{j_1,r} = 1\right] + \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{11}$$

where the dependent variable is now a $(|\mathscr{P}|N \times 1)$ vector, $\boldsymbol{\eta}$ is a vector of controls for order effects, and $\boldsymbol{\epsilon}$ is a mean-zero error, independent across individuals but possibly correlated within individuals. Using an F-test, we will test the null hypothesis $\beta_1 = \beta_2 = 0$ against the alternative $\beta_1 < 0$ and/or $\beta_2 < 0$ i.e., that an initial commitment choice is more likely to be undone than an indifference or a choice of flexibility.

Are Committers More Inconsistent Than Non-Committers? In an exploratory analysis, we will study whether inconsistencies across methods are higher for those who make more commitment decisions, which would provide suggestive evidence that commitment preferences are more fragile. For each respondent, we will calculate the total proportion of commitment choices made in round 1 across all three methods

$$\bar{C}_1^i = \frac{1}{3 \times |\mathscr{P}|} \sum_{j \in \{R,V,B\}} C_{j,1}^i, \tag{12}$$

and take a median split of respondents into a low- and a high-commitment group, $\mathscr{I}_{\text{low}} := \left\{i : \bar{C}_1^i < \text{med}(\bar{C}_1^i)\right\}$ and $\mathscr{I}_{\text{high}} := \left\{i : \bar{C}_1^i \geq \text{med}(\bar{C}_1^i)\right\}$. Using t-tests of equality of means, we will investigate whether the aggregate consistency index in round 1, $\Gamma_{3,1}$, differs between $\mathscr{I}_{\text{low}}$ and $\mathscr{I}_{\text{high}}$ types.

4.2 Stability of Commitment Decisions Between Rounds

Two classes of mechanisms might explain discrepancies in expressed preferences between methods. First, respondents might have difficulty formulating a preference over the set of possible menus, thus leading to inconsistent answers regardless of the method. Second, even when respondents have a clear preference, different methods might extract information about their preferences with different degrees of reliability. If the observed inconsistencies are simply due to differences in methods e.g., triggering different heuristics or eliciting different preferences, there should be almost perfect stability within a method (absent learning and fatigue effects). In the following, we evaluate the level of within-method stability of expressed preferences and its relationship with the inconsistencies between methods discussed in the previous section.

4.2.1 Variation in the Prevalence Rate of Commitment Between Rounds

We start by assessing how the overall prevalence of commitment decisions changes between the two rounds when using the same method (see Section G.2, Appendix G for an analogous analysis of indifferences and flexibility preferences). Pooling across all respondents and binary choice comparisons, we will first contrast the prevalence rate of commitment in round 1 vs. round 2 for each method. To examine the correlation in commitment behavior between rounds, Figure 10 shows bubble plots of the empirical

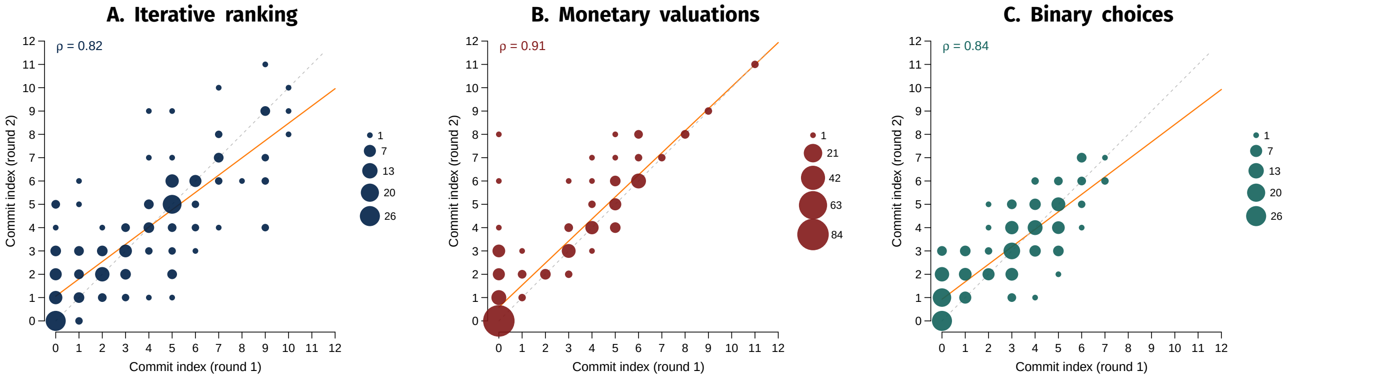


Figure 10. Scatter plots of commitment indices (see Equation 2) between round-1 and round 2-choices. Bubble sizes indicate the number of individuals with the same pair of commitment indices. See Figure G.1 for the corresponding scatter plots of the indifference and flexibility indices.

commitment indices for round 1 vs. round 2. We will test in a regression framework whether the prevalence of commitment decisions significantly differs between the two rounds for the same method and whether there are any differences between methods. To this end, we will stack the commitment indices to a $(6N \times 1)$ vector $\boldsymbol{C}$ and regress them on method dummies, a round dummy, and interactions between the two

$$\begin{aligned}\boldsymbol{C} = {} & \beta_0 + \beta_1 \mathbb{1}[j = R] + \beta_2 \mathbb{1}[j = V] + \beta_3 \mathbb{1}[r = 2] \\ & + \beta_4 \mathbb{1}[r = 2, j = R] + \beta_5 \mathbb{1}[r = 2, j = V] + \boldsymbol{\eta} + \boldsymbol{\epsilon}, \end{aligned} \tag{13}$$

where $\boldsymbol{\eta}$ is a vector of controls for order effects and $\boldsymbol{\epsilon}$ is a mean-zero error independent across, but possibly correlated within, respondents. We will test the null hypothesis $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ using an F-test.

4.2.2 Overall Frequency and Nature of the Instabilities

How Frequently Do Expressed Preferences Differ Between Rounds? Similar to the analysis of consistency between methods, we now examine the frequency with which each respondent expresses identical

Table 6. Stability of expressed preferences between rounds.

	Iterative ranking	Monetary valuation	Binary choices
Empirical	0.85	0.91	0.84
$N = 2,400$	(0.16)	(0.14)	(0.10)
Simulated	0.40	0.48	0.33
$N = 2,400$	(0.19)	(0.17)	(0.14)
Expert	0.xx	0.xx	0.xx
$N = N_E$	(0.xx)	(0.xx)	(0.xx)

Notes: Standard errors in parentheses. Since there are 12 binary comparisons per respondent, we have $200 \times 12 = 2,400$ total observations. Entries 0.xx will be populated once we collected the data.

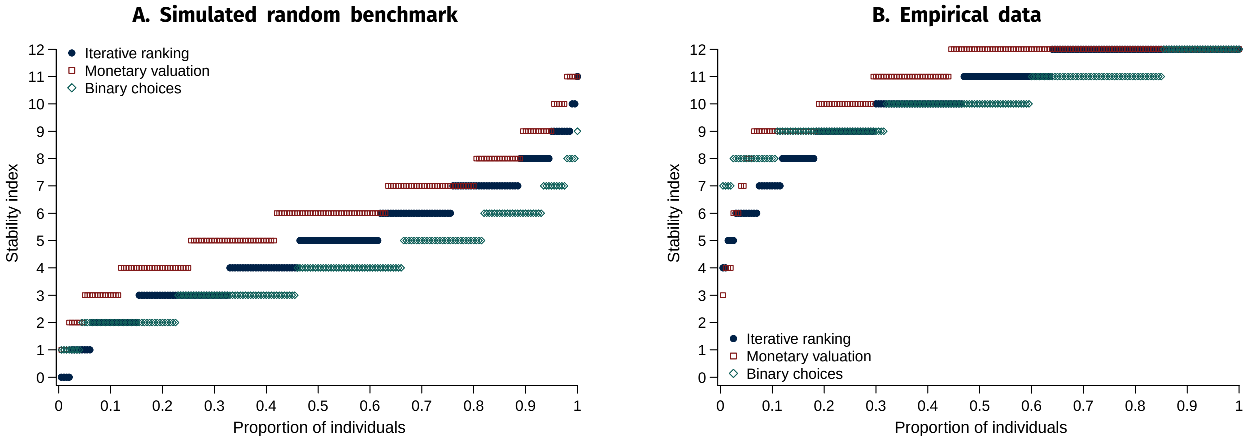


Figure 11. Quantile plots of stability indices (see Equation 6) across the three elicitation methods.

preferences between rounds. Table 6 displays the level of stability exhibited by respondents' answers in each of the methods relative to the two benchmarks. Using one-sample t-tests of equality of means, we will test the null hypotheses that mean stability for method j, μ_{ρ_j}, is equal to the random and expert benchmarks

$$H_0 : \mu_{\rho_j} = \bar{\rho}_j^S \quad \text{and} \quad H_0 : \mu_{\rho_j} = \bar{\rho}_j^E. \tag{14}$$

We will also report the proportion of experts who correctly order the methods in terms of the stability of the answers they produce. To test whether instabilities are correlated with the elicitation method, we will stack the stability indices to a $(3N \times 1)$ vector $\boldsymbol{\rho}$ and regress them on method dummies

$$\boldsymbol{\rho} = \beta_0 + \beta_1 \mathbb{1}[j = R] + \beta_2 \mathbb{1}[j = V] + \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{15}$$

where $\boldsymbol{\eta}$ is a vector of controls for order effects, and $\boldsymbol{\epsilon}$ is a mean-zero error that is independent across individuals, but probably correlated within individuals. The coefficients β_1 and β_2 quantify how the stability indices change relative to the preferences expressed via binary choice. We will test the null hypothesis $\beta_1 = \beta_2 = 0$ using an F-test. Finally, we will assess the level of heterogeneity in the stability of individual behavior by examining the quantile plots shown in Figure 11 of the stability indices for each method.

Are Initial Commitment Choices More Likely to be Undone? To examine the nature of the instabilities, we will test whether preferences for commitment as expressed in round 1 using a given method are more likely to be overturned than preferences for flexibility or indifferences. To this end, Figure 12 shows the joint distribution of preference type indicators in round 1 and round 2 by elicitation method. For example, we can quantify the probability of an expressed commitment preference in round 2 conditional on an expressed commitment preference in round 1. The probabilities are 0.85 (=0.23/0.27) for the iterative ranking, 0.85 (=0.17/0.20) for binary choices, and 1 (=0.16/0.16) for monetary valuations, respectively.

We will add quantitative evidence via regression analyses. For each method j, we will regress an indicator for preference instability at a given menu pair $p = (M, M')$ on dummies for the preferences

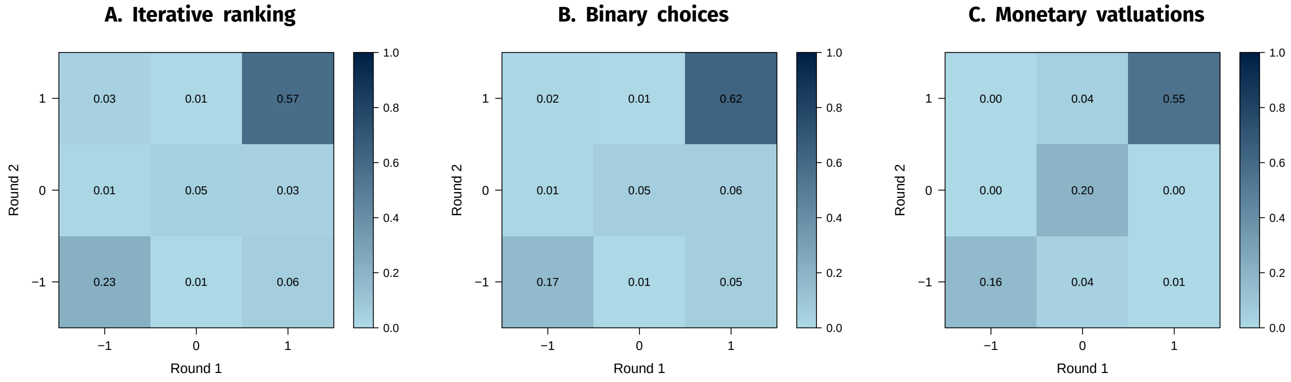


Figure 12. Joint distribution of preference indicators between round 1 and round 2, by elicitation method.

expressed in round 1. Stacking the menu-pair and individual regressions into a $(12N \times 1)$ vector, we obtain

$$\mathbb{1}\left[\boldsymbol{P}_{j,1} \neq \boldsymbol{P}_{j,2}\right] = \beta_0 + \beta_1 \mathbb{1}\left[\boldsymbol{P}_{j,1} = 0\right] + \beta_2 \mathbb{1}\left[\boldsymbol{P}_{j,1} = 1\right] + \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{16}$$

where $\boldsymbol{\eta}$ is a vector of controls for order effects and $\boldsymbol{\epsilon}$ is a mean-zero error, independent across individuals but possibly correlated within individuals. The interpretation of the coefficients is straightforward. For instance, $\beta_2 > 0$ means that those who expressed a preference for flexibility in round 1 are less likely to stick with the same answer than those who initially expressed a preference for commitment. We will test the null hypothesis $\beta_1 = \beta_2 = 0$ using an F-test. For the ranking and valuation methods, an examination of the size of deviations between rounds will be provided in Appendix G.

Do Committers Exhibit More Unstable Behavior? In an exploratory analysis, we will also analyze whether instabilities across methods are higher for those who commit in more menu pairs in round 1. We will take a median split into $\mathscr{I}_{\text{low}}$ and $\mathscr{I}_{\text{high}}$ commitment types based on Equation 12 and explore whether the stability indices for each of the three methods are different conditional on the respondent's type; we will do so by conducting t-tests of equality of means for the stability indices in the $\mathscr{I}_{\text{low}}$ and $\mathscr{I}_{\text{high}}$ groups.

4.2.3 Relationship Between Consistency and Stability

We conclude this section by examining the extent to which between-methods consistency and within-method stability are related phenomena. Figure 13 shows correlations between the bilateral consistency and stability indices. In addition, we will perform a regression analysis to better understand how much of the variation in the number of inconsistencies can be explained by the level of instability within each method. Formally, we will regress the aggregate consistency index across all three methods in round $r \in \{1, 2\}$ on the stability indices for each method

$$\Gamma_{3,r} = \beta_0 + \beta_1 \rho_R + \beta_2 \rho_V + \beta_3 \rho_B + \eta + \epsilon, \tag{17}$$

where η is a vector of controls for order effects. To assess the explanatory power of each stability index, we will examine the R^2 in simple regressions with each index ρ_j entered separately as well as in the full regression of Equation 17 (with and without controls).

5 Which Method(s) to Use?

If the three methods do not lead to the same expressed preferences, the next natural question is which method(s) should be used.[21] A priori, the answer to this question depends on the objective function. In this section, we evaluate each method according to three performance criteria. First, Section 5.1 evaluates the three methods in terms of the plausibility of the structural assumptions such as completeness or transitivity that each imposes on preferences. Next, Section 5.2 turns to respondents' subjective views of each method and preferences for which method to use. Finally, Section 5.3 evaluates the ability of the three methods (or combination thereof) to pin down a respondent's preferred platform version in a prediction exercise.

5.1 Plausibility of Structural Assumptions

The three methods put a different amount of structure on preferences and require different assumptions for the elicitation procedure to be incentive compatible. Table 7 summarizes the main characteristics of each elicitation method and structural assumptions imposed, including completeness and transitivity. In the following, we examine the plausibility of these assumptions in our environment and assess the extent to which a possible breakdown of any of them might explain the variability observed in the expression of preferences, including the problem of separating strict preferences from indifferences.

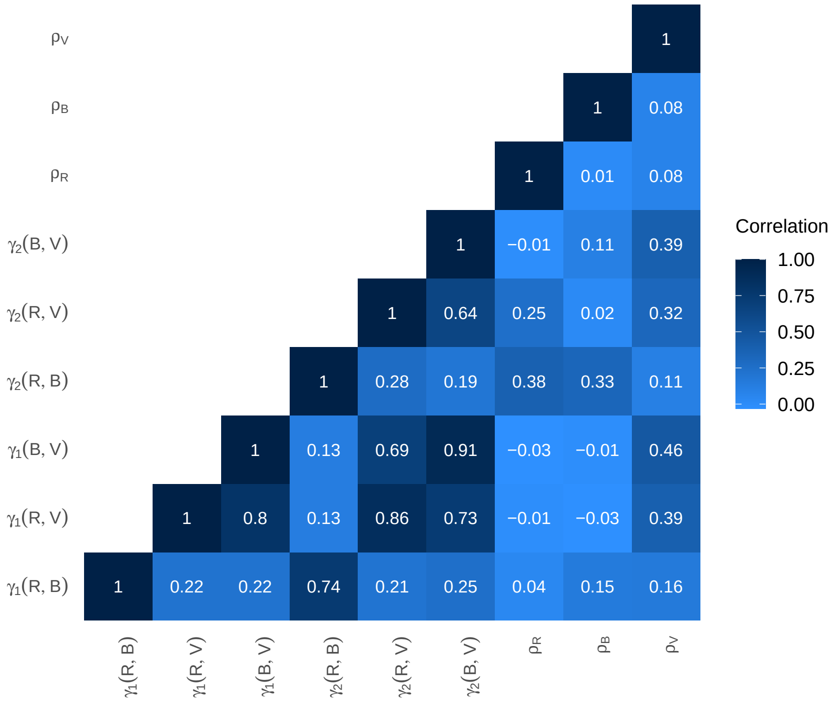


Figure 13. Correlation matrix of bilateral consistency (see Equation 5) and stability indices (see Equation 6).

[21] If we cannot establish any inconsistencies or instabilities, we will conclude that all methods can be used interchangeably. Since we use complex objects i.e., we elicit preferences over menus, such a finding would be interesting per se and suggest that procedural invariance may hold also in less complex settings. However, this would not be in line with the current evidence in the literature.

Table 7. Main characteristics of each elicitation method

	Iterative ranking	Monetary valuations	Binary choices
Decision-making Procedure	Step-by-step selection of the menus assigned rank #1, #2, etc. until all menus are ranked	Choices of amount (£0–£35) for giving up a spot in the challenge if offered a given menu	Sequence of choices between two menus (or coin flip) presented on the same page
Required choices	7	7	12
Complete ordering	Yes	Yes	No
Intransitivities	None by design	None by design	Possible and identifiable
Elicitation of indifferences	Same rank assigned Easier to report (fewer steps)	Same amount requested Similarly easy to report	Choice of "I like both equally" Easier to report (set as default)
Allocation rule	One-stage lottery	Compound lottery	Compound lottery*
Monotonicity in money required	No	Yes	No

*Compound lottery if indifferences are expressed and one-stage lottery otherwise.

5.1.1 Incompleteness of Preferences and Stochastic Choice

Plausibility of the Completeness Assumption. By design, the ranking and monetary valuation methods require respondents to express a complete preference ordering on $\mathscr{M}$, while the binary choice method allows to remain silent on 9 of the 21 binary comparisons.[22] Arguably, comparisons in this study are more complex than typical comparisons e.g., involving monetary lotteries or streams of monetary payoffs. For instance, consider the comparison between menus R and GO. To arrive at a preference, respondents must first determine whether they would prefer a meal bundle from category R vs. [G or O], a challenging task if meals from one category are not uniformly preferred to meals from the other two categories. Second, respondents must decide how much they value the possibility to keep their options open until the ordering stage vs. commit to a meal category right now. Given the multiple layers of decision-making, the completeness assumption might not hold, making forced decisions difficult to interpret. On the other hand, if respondents are able to determine their own preferences, the additional information extracted via the ranking and valuation exercises will offer a richer picture of the nature of commitment choices.[23]

Can Instability be Explained by Incompleteness? If respondents are forced to make decisions despite being unsure of their own preferences, then the instability of their answers could be a manifestation of this incompleteness. To test this conjecture, we ask respondents at the end of the survey to rate on a scale from 0 to 100 how difficult they found the task of comparing the various platforms and making up

[22] These 9 binary comparisons are (G,O), (G,R), (O,R), (GO,GR), (GO,OR), (GR,OR), (G,OR), (O,GR), and (R,GO).

[23] For instance, if G $\succ_j$ GO and O $\succ_j$ GO, then knowing the subject's preference for G vs. O would allow one to identify which type(s) of meals are more tempting. To achieve this with the binary choice method, one would need to collect information on additional comparisons.

their mind on which one(s) they prefer (irrespective of the method). We will report the distribution of responses in the appendix. We will test whether behavior instability might be the product of incomplete preferences by correlating a respondent's choice difficulty rating with the stability index ρ_j^i for each method j to see if correlations are negative.

5.1.2 Measurement Error on Indifferences

The Relevance of Indifferences in a Menu Choice Context. In most decision contexts, one could reasonably argue that indifferences represent a knife-edge case that can be ignored without loss. However, in the context of menu choice, indifferences are unlikely to be a measure zero phenomenon. Indeed, a standard decision maker whose utility only depends on the meals consumed and who has complete preferences over meal bundles will be indifferent between all menus that contain her preferred bundle(s). For such an individual, one should observe $C_{j,r}^i = 0$ and $I_{j,r}^i = 12$ for all $j \in \{R, V, B\}$ and $r \in \{1, 2\}$, absent any measurement error. In the presence of measurement error, researchers or policymakers who care about identifying tempted individuals will want to know the extent to which different methods might alter their judgment. Which method to use then depends on the amount of Type I and Type II errors that one is willing to tolerate i.e., falsely inferring a desire for commitment versus failing to identify a true desire for commitment. Below we go back to how the elicitation of indifferences in the three methods might affect the amount of Type I and Type II errors and examine empirical differences in the prevalence of indifferences reported.

Differences Between Methods in the Elicitation of Indifferences. Each method incentivizes the reporting of indifferences differently. In the binary choice and iterative ranking procedures, indifferences are simply made easier to report. For binary choice, the option "I like both equally" is selected by default. For the ranking, respondents must actively indicate an indifference, but doing so allows them to save time by skipping one round of iteration per indifference. In contrast, the effort cost of reporting an indifference vs. a strict preference is the same in the valuation exercise; instead, the identification of indifferences relies on the assumption of strict monotonicity in money. If small differences in monetary costs or effort costs are perceived as negligible, then each method may underestimate the true proportion of indifferences. On the other hand, all methods employ a random mechanism for implementation (in the form of a simple or compound lottery), which may fail to be incentive compatible and lead people to falsely report an indifference. For instance, a respondent with incomplete preferences might prefer a coin flip over committing to a given menu (Agranov and Ortoleva, 2022). Depending on the salience of randomization opportunities across the three methods, the true proportion of indifferences might be overestimated.

Prevalence of Expressed Indifferences Across Methods. Having clarified differences in the elicitation of indifferences across methods, we examine how the prevalence rate of expressed indifferences varies between methods and rounds (see Appendix F and Appendix I). The proportion of expressed indifferences appears largest for valuations and similar for the iterative ranking and binary choice procedures, suggesting a minor role of default effects and increased ease of reporting indifferences. In the valuation exercise, respondents tend to enter monetary amounts that are multiples of 5 (see Figure D.2, Appendix D), suggesting that subtle differences in preferences between menus might not be detected (also see Figure 8). We will test whether some of the reported indifferences might reflect a desire for randomization driven by preferences being incomplete. We will do so by correlating the indifference index $I^i_{j,r}$ of respondent i for each method j in round r with his choice difficulty rating (see Section 5.1.1); the correlation matrix will be reported in the appendix.

5.1.3 Lack of Transitivity

Unlike the other two methods, the binary choice procedure allows for transitivity violations. This is an undesirable feature if intransitivities are non-voluntary or mistakes. If, however, respondents wish to express intransitivities, then binary choices may have an edge in capturing the underlying preference structure. Below we study the bite of transitivity from a positive and a normative point of view by examining the prevalence of transitivity violations, and the support for transitivity as a normative property.

Transitivity as a Positive Property: Prevalence of Transitivity Violations. To test for violations of transitivity, one needs to identify all triplets (M, M', M''), with $M \neq M' \neq M''$, at which the condition $(M \succeq_{B,r} M') \wedge (M' \succeq_{B,r} M'') \Longrightarrow M \succeq_{B,r} M''$ fails. With $|\mathscr{M}| = 7$, there are $\binom{7}{3} =$ 35 triplets at which transitivity could be violated in principle. However, since the binary choice procedure only elicits preferences over the subset of 12 binary comparisons (M, M') such that $M' \subset M$, there are at most 6 triplets at which a preference cycle can be revealed: $\{(\mathrm{G}, \mathrm{GO}, \mathrm{GOR}), (\mathrm{G}, \mathrm{GR}, \mathrm{GOR}), (\mathrm{O}, \mathrm{GO}, \mathrm{GOR}), (\mathrm{O}, \mathrm{OR}, \mathrm{GOR}), (\mathrm{R}, \mathrm{GR}, \mathrm{GOR}), (\mathrm{R}, \mathrm{OR}, \mathrm{GOR})\}$. To capture the prevalence of transitivity violations, we define a *transitivity index* $\tau^i_r \in \{0, 1, \ldots, 6\}$, which counts the number of triplets at which transitivity is satisfied by individual i in round r. For example, if $\tau^i_r = 6$, then the respondent satisfied transitivity at every triplet in round r. The construction and distribution of this index will be presented in Appendix H, with our key findings reported in the main text.

Transitivity as a Normative Property: Support from Respondents. Eliciting preferences through binary choices is a useful approach if respondents who violate transitivity do this intentionally. If, however, transitivity violations are mistakes, this feature may hamper comparisons with other elicitation methods. To assess whether respondents regard transitivity as a desirable choice property, we ask them at the end of the survey to rate the coherence (from 0 to 100) of a hypothetical transitivity violation $\left(\mathrm{A} \succ^i_B \mathrm{B} \wedge \mathrm{B} \succ^i_B \mathrm{C} \wedge \mathrm{C} \succ^i_B \mathrm{A}\right)$ expressed in the binary choice task by a hypothetical respondent for some generic menus A, B, and C. We will compare the coherence ratings of respondents whose choices fully conform with transitivity in each round (i.e., the sets $\mathscr{I}_{T_1} := \{i : \tau^i_1 = 6\}$ and $\mathscr{I}_{T_2} := \{i : \tau^i_2 = 6\}$) to those whose choices do not (sets $\mathscr{I}_{NT_1} := \{i : \tau^i_1 < 6\}$ and/or $\mathscr{I}_{NT_2} := \{i : \tau^i_2 < 6\}$), using a two-sided t-test of equality of means.[24] If we do find a difference, then we may conjecture that violations of transitivity are at least partly deliberate.

[24] Since the hypothetical transitivity violation that we presented to respondents is strict, we will conduct a robustness exercise in which we compare the ratings of the subsample of respondents who violated transitivity in a strict preference cycle to the other respondents.

Are Transitivity Violations Predictive? A natural question is whether transitivity violations in binary choices are predictive of inconsistencies between the binary choice method and the other two methods: are respondents who want to express intransitivities more inconsistent simply because the ranking and monetary valuation procedures do not allow them to express those intransitivities while the binary choice method does? To test this conjecture, we will compare the distributions of the bilateral consistency indices $\gamma^i_{R,B}$ and $\gamma^i_{V,B}$ for respondents in $\mathscr{I}_{T_r}$ vs. $\mathscr{I}_{NT_r}$ using Kolmogorov-Smirnov tests. For completeness, we will also examine differences between $\mathscr{I}_{T_r}$ and $\mathscr{I}_{NT_r}$ for the third bilateral index, $\gamma^i_{R,V}$, although respondents could construct the same forced transitive order in each case. Figure H.2, Appendix H and Figure I.7, Appendix I illustrate the relationship between the transitivity and bilateral consistency indices. To keep the analyses tractable, we will collapse the round-specific sets $\mathscr{I}_{T_r}$ and $\mathscr{I}_{NT_r}$ to $\mathscr{I}_T := \{i : \tau^i_1 = 6 \wedge \tau^i_2 = 6\}$ and $\mathscr{I}_{NT} := \{i : \tau^i_1 < 6 \vee \tau^i_2 < 6\}$, respectively, if individual transitivity indices are not significantly different between rounds.

5.2 Differences in Subjective Experience and Preferred Method

Even if core preferences satisfy all structural assumptions, expressed preferences may differ between methods if respondents perceive them as differently tedious or complex to use. In this case, they may trust their responses with a certain method more than with some other(s). In this section, we investigate respondents' subjective evaluations of each method and their preferences for which method to use.

5.2.1 Differences in Complexity and Cognitive Load

Procedural Differences Between Methods. Setting issues of incentive compatibility aside, the three elicitation procedures differ in the presentation and number of required decisions as well as in the precision of the information requested. With binary choices, respondents can focus on a narrow frame as they evaluate the 12 menu pairs one by one. However, they have to make almost double the number of decisions relative to the other two procedures (7 decisions) and, despite this, only part of the preference relation is recovered. At the opposite end of the spectrum, expressing preferences through monetary valuations provides far more information with fewer steps, but might be less intuitive. Similar to the iterative ranking, respondents must consider the set of 7 menus all at once. In addition, they have to leave the goods domain and compute a monetary amount for each possible menu relative to the value of opting out of the challenge. In between, the iterative ranking procedure breaks down the number of decision steps but still requires respondents to evaluate the set of menus holistically.

Subjective Perceptions of Each Method. The procedural differences outlined in the previous paragraph likely moderate the perceived complexity and cognitive load of each task. In order to better understand potential differences in perceptions, we ask respondents to rate each method on the dimensions of difficulty, tediousness, and confidence in answers.[25] To control for individual differences in the perception of the scales, we will normalize each measure by subtracting the respondent's mean rating across the

[25] More precisely, we ask respondents to answer the following questions on a scale from 0 to 100: "How difficult was it for you to come up with answers in each decision task?", "How tedious did you find the process of completing each decision task?", and "How certain are you about the answers you gave in each decision task?"

Table 8. Joint distribution of respondents' preferred method-round pairs including marginal probabilities.

	Iterative ranking	Monetary valuations	Binary choices	Any method	Total
Round 1	0.xx	0.xx	0.xx	0.xx	0.xx
Round 2	0.xx	0.xx	0.xx	0.xx	0.xx
Any round	0.xx	0.xx	0.xx	0.xx	0.xx
Total	0.xx	0.xx	0.xx	0.xx	0.xx

Notes: The columns "Any method" and "Any round" refer to the decision to pick one randomly. Entries 0.xx will be populated once we collected the data.

three methods and dividing by the across-method standard deviation. We will present violin plots of the three normalized measures and, for each dimension, conduct paired t-tests of equality of means across methods. In the appendix, we will also present a correlation matrix showing the relationship between all three raw measures (difficulty, tediousness, confidence).

Relationship between Stability and Subjective Ratings. We will leverage these subjective ratings to investigate whether complexity and cognitive load are important drivers of any instabilities documented in Section 4.2. For each method j, we will regress the stability index on the normalized measures of perceived difficulty, tediousness, and confidence in answers

$$\rho_j \;=\; \beta_0 + \beta_1\,\text{difficulty}_j + \beta_2\,\text{tediousness}_j + \beta_3\,\text{confidence}_j + \eta + \epsilon, \tag{18}$$

where η is a vector of controls for order effects. For each elicitation method, we will report three linear regressions with each covariate entered separately, as well as the regression in Equation 18 including all covariates. For example, $\beta_1 < 0$ indicates that perceived difficulty in method j is predictive of instabilities.

5.2.2 Which Method do Respondents Prefer?

The subjective ratings are informative about three dimensions that we considered to be particularly relevant for method selection. To capture other unobserved factors, we introduce an incentivized task at the end of the survey. After completing all 6 decision tasks, respondents can indicate a method-round pair that might become relevant for the selection of their food ordering platform, or opt for the method and/or round to be randomly selected. Respondents' choices will be reported in Table 8. We will perform a χ^2 test to assess whether some option(s) were selected significantly more often than others and qualitatively compare actual selections to the predictions made by the experts.

To understand respondents' method selection, we will first test whether the chosen method is also considered to be the most reliable one (variable *reliability_{method}* in Table C.1, Appendix C) in determining preferences for the various menus. To this end, we will construct two categorical variables d_1 for the preferred method(s) and d_2 for the method(s) judged most reliable, with $d = 1$ for $j = R$, $d = 2$ for $j = V$, $d = 3$ for $j = B$, and $d = 4$ in case of ties between two or all three methods (i.e., random selection of method for d_1 and equal reliability rating for d_2). We will present bar graphs of these two categorical variables and conduct a χ^2 (or Fisher exact) test of independence in a 2×2 contingency table. Second, to assess the explanatory power of perceived difficulty, tediousness, and confidence in answers, we will regress an indicator for whether method j was selected on the three (normalized) ratings

$$\text{method\_}j\text{\_selected} = \beta_0 + \beta_1 \,\text{difficulty}_j + \beta_2 \,\text{tediousness}_j + \beta_3 \,\text{confidence}_j + \eta + \epsilon, \tag{19}$$

where η is a vector of controls for order effects. Third, we will explore whether the preferred method also tends to be the one that generates the most stable answers; this will be done by again performing a χ^2 test relating the categorical variable d_1 for the preferred method(s) to a corresponding categorical variable d_3 for the most stable method(s) based on their respective stability indices.

5.3 Predictive Performance

Having contrasted respondents' subjective views of each method and measured their preferences for which one to use, we close this section by comparing their performance in a prediction exercise. As noted in the previous sections, each method conveys a different amount of information, which in turn might be differently affected by measurement error. Below we examine which method(s) or combination thereof can best pin down the respondents' preferred platform version as measured with two simple questions.

Predicting Respondents' Self-Reports. In the debriefing section of the survey, we ask respondents to answer two simple unincentivized questions about the meal categories they would like to see on the platform. Although not incentivized, these questions are short and direct, thus limiting the scope for decision noise. In the first question, respondents indicate which type of meals they would prefer to try for the challenge among the three meal categories.[26] In the second question, they are asked to indicate their preferences for being assigned a platform showing only their chosen meal category, all meal categories, or either of the two options (equal preference). We combine information from these two questions to infer a preference $\succsim_{SR}$ (SR for "self-report") over restricted vs. unrestricted menus.[27] To illustrate, consider a respondent who reports a preference for trying meals from G and for ordering on a platform that shows all three meal categories; for this respondent, we infer GOR $\succ_{SR}$ G. We will assess the predictive performance of the three methods by counting the frequency with which each decision task (method and round combination) agrees with the relation $\succsim_{SR}$. For our example respondent, we would hence check whether GOR $\succ_{j,r}$ G holds for each j and r. The findings will be reported in Table 9, including the results from a χ^2 test of independence.

Additional Exploratory Analysis: LASSO Regressions. A plain comparison of proportions as in Table 9 ignores the fact that the ranking and valuation procedures extract more information about preferences

[26] Q1: "If you had to make a choice now, which type of meals would you like to try for the challenge next week?" Answers options include "G", "O", and "R".

[27] Q2: "Would you prefer to be assigned a platform that only shows you the meals from [selected category] or, to keep your options open, one that gives you access to all meal categories?" Answers include "Only show me the meals from [selected category]", "Show me the meals from all categories", and "I am equally happy with both options".

Table 9. Proportion of cases in which decision task (j, r) agrees with $\succsim_{SR}$.

	Ordinal ranking	Monetary valuations	Binary choices	All methods
Round 1	0.xx	0.xx	0.xx	0.xx
Round 2	0.xx	0.xx	0.xx	0.xx
All rounds	0.xx	0.xx	0.xx	0.xx

Notes: A value of 1 would indicate that the method-round pair predicts the same relation as $\succsim_{SR}$ for every individual. Entries 0.xx will be populated once we collected the data.

than binary choices. We thus propose to conduct a finer exploratory analysis that exploits all the information generated by each method, and additionally assess whether a combination of methods is superior in predicting $\succsim_{SR}$ over the set $\mathscr{M}^{\mathrm{L}} := \{\mathrm{G}, \mathrm{O}, \mathrm{R}, \mathrm{GOR}\}$. We will use a multinomial logistic regression framework with LASSO penalization (Tibshirani, 1996) to recover the most predictive covariates. We will classify individual behavior into one of nine mutually exclusive categories, $\mathrm{GOR} \succ_{\mathrm{SR}} \mathrm{G}, \mathrm{GOR} \succ_{\mathrm{SR}} \mathrm{O}, \mathrm{GOR} \succ_{\mathrm{SR}} \mathrm{R}, \mathrm{GOR} \sim_{\mathrm{SR}} \mathrm{G}, \mathrm{GOR} \sim_{\mathrm{SR}} \mathrm{O}, \mathrm{GOR} \sim_{\mathrm{SR}} \mathrm{R}, \mathrm{GOR} \prec_{\mathrm{SR}} \mathrm{G}, \mathrm{GOR} \prec_{\mathrm{SR}} \mathrm{O}, \mathrm{GOR} \prec_{\mathrm{SR}} \mathrm{R},$ based on the following covariates: ranks $\{\mathrm{rank}_{r,M}\}_{M \in \mathscr{M}^{\mathrm{L}}}$, monetary valuations $\{v_{r,M}\}_{M \in \mathscr{M}^{\mathrm{L}}}$, differences in ranks $\{\mathrm{rank}_{r,M} - \mathrm{rank}_{r,M'}\}_{M,M' \in \mathscr{M}^{\mathrm{L}}, M \neq M'}$, differences in monetary valuations $\{v_{r,M} - v_{r,M'}\}_{M,M' \in \mathscr{M}^{\mathrm{L}}, M \neq M'}$, method-specific commitment indices $\{C_{j,r}\}_{j \in \mathscr{J}}$, method-specific indifference indices $\{I_{j,r}\}_{j \in \mathscr{J}}$, and method-specific flexibility indices $\{F_{j,r}\}_{j \in \mathscr{J}}$. We will randomly select 80% of our sample as training set (using the remaining 20% as test set), standardize the covariates (removing the mean and scaling to unit variance), and select the regularization parameter via cross validation as the parameter that minimizes the mean squared error in the training set.[28] We will report results from round-1 choices in the main text and relegate results from round-2 choices to the appendix while discussing potential differences.

6 Discussion

Although the literature on commitment demand is extensive, little is known about the factors that might explain the large variation in observed take-up rates. In this paper, we propose to examine how the choice of the elicitation procedure might affect conclusions regarding the extent of commitment demand. We compare three main methods used in the literature to measure commitment take-up: ranking, monetary valuations, and binary choices. We assess the consistency of the preferences expressed via these three methods and its link to within-method stability. Doing so allows us to bound commitment demand and to analyze whether commitment decisions are noisier than choices expressing an indifference or a preference for flexibility. We then provide guidance on method selection. We do this by combining a theoretical discussion of the properties of each method with new data on respondents' perceptions of the methods and with an exercise evaluating their predictive performance. As part of this assessment, we develop a novel incentivized procedure to identify respondents' preferred elicitation method, which may

[28] Given our limited sample size, we may need to apply stratified sampling to ensure that our distribution of covariates will not be biased during the split.

be used in other settings. Throughout, we contrast our findings with expert forecasts to assess whether researchers correctly anticipate how certain methodological choices might affect scientific conclusions.

Examining these questions is important from a policy perspective because offering commitment devices may cause more harm than good, especially if chosen by mistake. For instance, customers might find themselves trapped in rigid contracts that aimed to do good but come with large cancellation fees, simply because their evaluations were swayed by the framing of the sign-up questions. Even when innocuous, firms or policymakers might be reluctant to make changes to a default choice architecture if they cannot be firmly convinced that the interest of customers outweighs the implementation costs. In such cases, a lack of confidence in the measured level of support for a proposed change might lead uninformed decision makers to favor the status quo. Even more, private information about the non-neutral impact of the chosen elicitation procedure could lead experts with biased interests (whether researchers, consultants, or technocrats) to select a method that generates the outcome favoring the position they defend. Defining the scope for agenda manipulation requires a solid understanding of whether such degrees of freedom do really exist and whether experts are aware of their existence. To our knowledge, the potential for "design hacking" has received little-to-no attention so far in the metascience literature, unlike p-hacking and other questionable research practices. Our study presents a step in this direction.

Our approach is systematic and combines various sources of data to offer a holistic perspective on decisions to restrict one's choice set in a meal selection context. However, our work also comes with limitations pertaining to the complexity of our chosen design and the generalizability of the insights produced. We end this section by discussing these limitations and highlighting areas for future research.

Design Complexity and Robustness. While the within-subject variation offered by our experimental design allows us to paint a rich picture of preferences, it does not come for free. If we observe a substantial fraction of inconsistencies and/or instabilities, one legitimate question is whether they are simply a consequence of the complexity of our research design. In particular, one conjecture might be that respondents get annoyed after being asked to repeat the three types of decision tasks. To assess this, we will test whether observed inconsistencies between methods are larger in round 2 than in round 1.[29] In addition, we will exploit the fact that the order of presentation of the three methods is randomized between subjects to test whether our conclusions regarding the prevalence of commitment decisions for each method are robust to only using answers from the very first decision task completed. Given the sophisticated nature of our dataset and the many layers of decisions, another concern could be that the incentives we offer are insufficient to ensure high decision quality. While we cannot eliminate this concern, we note that our chosen incentives compare favorably to recent online studies and our unincentivized pilot data may provide a reasonable upper bound on the level of noise to be expected.

Sensitivity to Implementation Details. We examine the performance of one specific implementation of each of the three different methods. As such, our design offers a joint test of both the choice of method and the chosen implementation procedure. One important question is therefore whether our

[29] We will do this by conducting paired t-tests of equality of means for the aggregate and bilateral consistency indices in round 1 vs. round 2.

conclusions would be robust to changes in implementation details, including the chosen framing and incentive mechanism. For instance, there is evidence that the BDM mechanism may produce particularly noisy answers (Berg et al., 2005; James, 2007; Cason and Plott, 2014). As a result, one possibility is that our study gives an upper bound on the frequency of instabilities using WTP/WTA methods (ceteris paribus). We also made non-innocuous choices in the elicitation of indifferences e.g., generating possible default effects in the binary choice setup. Would our conclusions be reversed if the default setup was removed? Given the myriad of possible design variations, we did not attempt to exogenously vary the implementation procedure within a method. However, a recent study shows that design heterogeneity might contribute significantly to the variation in observed outcomes across experiments (Huber et al., 2023). Future research should examine how sensitivity to implementation details might affect the robustness of our conclusions.

Measurement Error and Preference Stability. In our study, we propose to take two measurements using the same method in order to assess whether expressed preferences are stable. Although we do not make use of these techniques, recent work shows that repeated measurements may be particularly useful to address experimental measurement error by serving as instrumental variables for the original measures (Gillen et al., 2019). A downside of taking repeated measurements of the exact same decision is that they might appear unnatural, especially if taken within a narrow time window. Given the very short time span, our study has also very little to say about the temporal stability of commitment preferences. Future studies that aim to study the dynamics of commitment decisions might want to consider repeated measurements of the same commitment decision (constraining some future behavior at a given time t) in order to separate true changes in preferences from pure measurement error.

Relevance to Other Choice Domains. Bringing decision theory to a field setting implies that our choice options – menus of meal bundles – are more complex in nature than those of traditional experiments involving monetary lotteries, streams of monetary payoffs, or simple objects such as mugs. Evaluating these options under different elicitation methods might create unmodelled interaction effects. For example, evaluating complex objects using a more sophisticated elicitation method e.g., monetary valuations via the BDM mechanism, may be particularly error-prone. Furthermore, individuals with incomplete preferences might express their indecisiveness differently depending on the elicitation method. Future research should study how the distribution of inconsistencies and/or instabilities varies with the complexity of the objects under comparison. Similarly, if we observe that a specific elicitation method is preferred by a majority of study participants, an open question is whether preferences over the method depend on the choice domain. We hope that future work will provide new answers to these questions.

References

Agranov, Marina, and Pietro Ortoleva. 2022. “Revealed Preferences for Randomization: An Overview.” *AEA Papers and Proceedings* 112 (5): 426–30. DOI: 10.1257/pandp.20221093. [8, 33]

Bai, Liang, Benjamin Handel, Edward Miguel, and Gautam Rao. 2021. “Self-Control and Demand for Preventive Health: Evidence from Hypertension in India.” *Review of Economics and Statistics* 103 (5): 835–56. DOI: 10.1162/rest_a_00938. [4]

Berg, Joyce, John Dickhaut, and Kevin McCabe. 2005. “Risk Preference Instability Across Institutions: A Dilemma.” *Proceedings of the National Academy of Sciences* 102 (11): 4209–14. DOI: 10.1073/pnas.0500333102. [40]

Bryan, Christopher J., David S. Yeager, and Joseph M. O’Brien. 2019. “Replicator Degrees of Freedom Allow Publication of Misleading Failures to Replicate.” *Proceedings of the National Academy of Sciences* 116 (51): 25535–45. DOI: 10.1073/pnas.1910951116. [8]

Bryan, Gharad, Dean Karlan, and Scott Nelson. 2010. “Commitment Devices.” *Annual Review of Economics* 2 (1): 671–98. DOI: 10.1146/annurev.economics.102308.124324. [4]

Carrera, Mariana, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky. 2022. “Who Chooses Commitment? Evidence and Welfare Implications.” *Review of Economic Studies* 89 (3): 1205–44. DOI: 10.1093/restud/rdab056. [4]

Casaburi, Lorenzo, and Rocco Macchiavello. 2019. “Demand and Supply of Infrequent Payments as a Commitment Device: Evidence from Kenya.” *American Economic Review* 109 (2): 523–55. DOI: 10.1257/aer.20180281. [4]

Cason, Timothy N., and Charles R. Plott. 2014. “Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing.” *Journal of Political Economy* 122 (6): 1235–70. DOI: 10.1086/677254. [40]

Cheung, Stephen L., Agnieszka Tymula, and Xueting Wang. 2022. “Present Bias for Monetary and Dietary Rewards.” *Experimental Economics* 25 (4): 1202–33. DOI: 10.1007/s10683-022-09749-8. [7]

Chuang, Yating, and Laura Schechter. 2015. “Stability of Experimental and Survey Measures of Risk, Time, and Social Preferences: A Review and Some New Results.” *Journal of Development Economics* 117 (11): 151–70. DOI: 10.1016/j.jdeveco.2015.07.008. [8]

Costa-Gomes, Miguel A., Carlos Cueva, Georgios Gerasimou, and Matúš Tejiščák. 2022. “Choice, Deferral, and Consistency.” *Quantitative Economics* 13 (3): 1297–318. DOI: 10.3982/QE1806. [8]

DellaVigna, Stefano, and Devin Pope. 2018. “Predicting Experimental Results: Who Knows What?” *Journal of Political Economy* 126 (6): 2410–56. DOI: 10.1086/699976. [8, 18]

Enke, Benjamin, and Thomas Graeber. 2023. “Cognitive Uncertainty.” *Quarterly Journal of Economics*, 1–47. DOI: 10.1093/qje/qjad025. [8]

Frey, Renato, Andreas Pedroni, Rui Mata, Jörg Rieskamp, and Ralph Hertwig. 2017. “Risk Preference Shares the Psychometric Structure of Major Psychological Traits.” *Science Advances* 3 (10): e1701381. DOI: 10.1126/sciadv.1701381. [7]

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. “Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study.” *Journal of Political Economy* 127 (4): 1826–63. DOI: 10.1086/701681. [40]

Giné, Xavier, Dean Karlan, and Jonathan Zinman. 2010. “Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation.” *American Economic Journal: Applied Economics* 2 (4): 213–35. DOI: 10.1257/app.2.4.213. [4]

Grether, David M., and Charles R. Plott. 1979. “Economic Theory of Choice and the Preference Reversal Phenomenon.” *American Economic Review* 69 (4): 623–38. eprint: http://www.jstor.org/stable/1808708. [7]

Gul, Faruk, and Wolfgang Pesendorfer. 2001. “Temptation and Self-Control.” *Econometrica* 69 (6): 1403–35. DOI: 10.1111/1468-0262.00252. [4]

Hascher, Joshua, Nitisha Desai, and Ian Krajbich. 2021. “Incentivized and Non-Incentivized Liking Ratings Outperform Willingness-To-Pay in Predicting Choice.” *Judgment and Decision Making* 16 (6): 1464–84. DOI: 10.1017/S1930297500008500. [7]

Holzmeister, Felix, and Matthias Stefan. 2021. “The Risk Elicitation Puzzle Revisited: Across-Methods (In)Consistency?” *Experimental Economics* 24 (2): 593–616. DOI: 10.1007/s10683-020-09674-8. [7]

Huber, Christoph, Anna Dreber, Jürgen Huber, ..., and Felix Holzmeister. 2023. “Competition and Moral Behavior: A Meta-Analysis of Forty-Five Crowd-Sourced Experimental Designs.” *Proceedings of the National Academy of Sciences* 120 (23): e2215572120. DOI: 10.1073/pnas.2215572120. [8, 40]

Iacovone, Leonardo, David McKenzie, and Rachael Meager. 2023. “Bayesian Impact Evaluation with Informative Priors: An Application to a Colombian Management and Export Improvement Program.” Working paper 10274. The World Bank. DOI: 10.1596/1813-9450-10274. [18]

James, Duncan. 2007. “Stability of Risk Preference Parameter Estimates Within the Becker-DeGroot-Marschak Procedure.” *Experimental Economics* 10 (2): 123–41. DOI: 10.1007/s10683-006-9136-y. [40]

John, Anett. 2020. “When Commitment Fails: Evidence from a Field Experiment.” *Management Science* 66 (2): 503–29. DOI: 10.1287/mnsc.2018.3236. [4]

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan. 2015. “Self-Control at Work.” *Journal of Political Economy* 123 (6): 1227–77. DOI: 10.1086/683822. [7]

Krajbich, Ian, Dingchao Lu, Colin Camerer, and Antonio Rangel. 2012. “The Attentional Drift-Diffusion Model Extends to Simple Purchasing Decisions.” *Frontiers in Psychology* 193 (3): DOI: 10.3389/fpsyg.2012.00193. [12]

Laibson, David. 1997. “Golden Eggs and Hyperbolic Discounting.” *Quarterly Journal of Economics* 112 (2): 443–78. DOI: 10.1162/003355397555253. [4]

Lichtenstein, Sarah, and Paul Slovic. 1971. “Reversals of Preference Between Bids and Choices in Gambling Decisions.” *Journal of Experimental Psychology* 89 (1): 46–55. DOI: 10.1037/h0031207. [7]

Loomes, Graham, and Ganna Pogrebna. 2017. “Do Preference Reversals Disappear When We Allow for Probabilistic Choice?” *Management Science* 63 (1): 166–84. DOI: 10.1287/mnsc.2015.2333. [8]

O’Donoghue, Ted, and Matthew Rabin. 1999. “Doing It Now or Later.” *American Economic Review* 89 (1): 103–24. DOI: 10.1257/aer.89.1.103. [4]

Otis, Nicholas. 2022. “Policy Choice and the Wisdom of Crowds.” DOI: 10.2139/ssrn.4200841. [18]

Robinson, Carly D., Gonzalo A. Pons, Angela L. Duckworth, and Todd Rogers. 2018. “Some Middle School Students Want Behavior Commitment Devices (but Take-Up Does Not Affect Their Behavior).” *Frontiers in Psychology* 9 (2): 206. DOI: 10.3389/fpsyg.2018.00206. [4]

Sadoff, Sally, and Anya Samek. 2019. “Can Interventions Affect Commitment Demand? A Field Experiment on Food Choice.” *Journal of Economic Behavior & Organization* 158 (2): 90–109. DOI: 10.1016/j.jebo.2018.11.016. [6, 7]

Sadoff, Sally, Anya Samek, and Charles Sprenger. 2020. “Dynamic Inconsistency in Food Choice: Experimental Evidence from Two Food Deserts.” *Review of Economic Studies* 87 (4): 1954–88. DOI: 10.1093/restud/rdz030. [6, 7]

Schwartz, Janet, Daniel Mochon, Lauren Wyper, Josiase Maroba, Deepak Patel, and Dan Ariely. 2014. “Healthier by Precommitment.” *Psychological Science* 25 (2): 538–46. DOI: 10.1177/0956797613510950. [6]

Strotz, Robert H. 1955. “Myopia and Inconsistency in Dynamic Utility Maximization.” *Review of Economic Studies* 23 (3): 165–80. DOI: 10.2307/2295722. [4]

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection Via the Lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88. DOI: 10.1111/j.2517-6161.1996.tb02080.x. [38]

Toussaert, Séverine. 2018. “Eliciting Temptation and Self-Control Through Menu Choices: A Lab Experiment.” *Econometrica* 86 (3): 859–89. DOI: 10.3982/ECTA14172. [7]

Toussaert, Séverine. 2019. “Revealing Temptation Through Menu Choice: Field Evidence.” eprint: https://severinetoussaert.com/wp-content/uploads/2019/02/LS-2014-v7.pdf. [7]

Appendix A Commitment Take-Up in the Literature

A.1 Construction of Summary Figures

Selection Procedure of Papers. To identify a comprehensive list of economics papers that measure commitment take-up, we pursued the following strategy: (i) systematic database search, (ii) search of reference lists of reviews written on the topic. For (i), we used three databases: Web of Science, IDEAS/RePEc, and Google Scholar.[30] To be as broad as possible, we employed the search query outlined in Listing 1:

Listing 1. Stylized search query for literature search.

```
(    commitment device OR commitment contract
  OR commitment demand OR demand for commitment
  OR pre-commit*       OR precommit*           )
AND experiment
```

The search query was set up to identify well-known experimental studies on commitment devices, balancing specificity and inclusiveness. Using Google Scholar, we applied the same set of keywords to the papers referencing Ariely and Wertenbroch (2002), Ashraf et al. (2006), or Bryan et al. (2010), which are cited very frequently in the literature on commitment devices. To further refine our literature search, we searched the reference list of the review article of Bryan et al. (2010) as well as the references discussed in Schilbach (2019), John (2020), and Carrera et al. (2022), each of which presents a summary table of a selected set of papers. The database search was performed in October 2022, which resulted in 1,612 papers after excluding clear duplicates. An initial screening of these papers was performed based on the abstract to remove irrelevant papers (i.e., purely theoretical papers). This left 225 potentially relevant papers, which were examined in much more detail. Based on our inclusion criteria and after excluding ambiguous cases, this second screening stage yielded 107 empirical papers on commitment devices.[31] As a reference, Carrera et al. (2022) lists 33 empirical papers reporting the take-up rates for (weakly) dominated commitment contracts offered at no cost building on Schilbach (2019, Table 1) and John (2020, Table 1).

To increase comparability between studies, we further refined the selection by excluding papers that (i) were not written by economists or published in an economics journal/repository (19 excluded), (ii) do not have data available to calculate comparable take-up rates (6 excluded), (iii) focus on soft commitment devices that do not strictly restrict the choice set or involve no financial costs (14 excluded), (iv) use

[30] We used IDEAS/RePEc and Google Scholar to identify relevant unpublished papers.

[31] Our list of inclusion criteria were: (i) the paper is empirical (possibly including some theory), (ii) subjects have a choice to commit (commitment not imposed on them), (iii) information about commitment demand was produced (even if not directly reported in the paper), and (iv) the paper is not an older version of a published study (or of a most recent working paper). We also excluded papers that describe marriage or public transport usage as a commitment device, which we deemed to be ambiguous cases.

an impure commitment device[32] (11 excluded), and (v) only elicit purely hypothetical commitment decisions (3 excluded). This left us with 54 papers for Figure 1 and Table A.1.

Elicitation Methods. Below we explain how the three different types of methods have been used in the literature to measure commitment take-up, sometimes jointly, but most often separately. When papers leverage a combination of methods, we explain how we extracted different measures of commitment take-up to separate the methods.

- **Direct Choice:** In the vast majority of cases, participants face the binary choice to take up a commitment device or not.[33] One example is the study of Brune et al. (2016), in which participants choose whether or not to transfer money to special accounts that disallowed withdrawals until a set date chosen by the account owner. While some studies only involve a single binary choice of taking up the device, other studies involve multiple binary choices e.g., Multiple Price Lists (in this latter case, we consider take-up at a price of 0 as direct choice - see "Combination of Methods" below). Overall, there is some but limited variation in the implementation of this method.
- **Ranking:** In the 4 respective studies, participants are offered multiple options, some more restrictive than others, and they are asked to rank each of these according to their preferences. Demand for commitment corresponds to ranking one of the restricted options strictly above the unrestricted option. Few papers use this method. Among them, there is little variation in the implementation procedure. There are minor differences in the number of options that need to be ranked (3 out of 4 studies elicit rankings over 3 options only, the other one over 7 options). All papers allow for the expression of a weak preference ranking by enabling respondents to assign the same rank number to several options.
- **Valuation:** In the respective 13 papers, people are offered a commitment device at varying prices and asked to explicitly consider whether they would be willing to pay a certain price for commitment, so choosing the costly option reveals a positive valuation for commitment. This could be a direct monetary cost but also e.g., a time cost of receiving a payment as in Zhang and Greiner (2021), or an effort cost of performing additional tasks as in Breig et al. (2020). Common approaches are the use of the BDM mechanism and Multiple Price Lists (MPL). Studies on valuations differ in whether they ask for WTP for commitment by adding a cost to the commitment option as in Alan et al. (2021) or for WTA not to have the commitment by adding extra earnings to the non-commitment option as in Augenblick et al. (2015). Overall, the implementation varies a lot even within the same method.
- **Combination of Methods:** A small set of studies combine elicitation methods to measure a given commitment decision. The most typical combination is direct choice and valuation, in which people are presented with multiple decisions, with the cost of the commitment device varying each time

[32] A commitment device is said to be impure if it comes with other desirable features relative to the outside option (e.g., additional monetary reward) such that even a rational agent would be interested in the product; in this case, usage would not necessarily signal demand for commitment.

[33] One exception is Krawczyk and Wozny (2017), where participants are offered to choose among the three menus {A}, {B}, and {A, B} i.e., they have two commitment options.

(e.g., MPL or sequence of discrete choices). In these cases, we infer a direct choice (valuation) take-up rate by calculating the fraction of people who chose the commitment option when its price was equal to O (greater than 0). One example is Augenblick et al. (2015) who reports a 59% take-up when the commitment option costs nothing and a 9% take-up at even the smallest price of $0.25. In such cases, take-up rates across measures are not independent and differences in take-up rates across decisions can be mechanical if people satisfy monotonicity in money. Another combination used by Toussaert (2018) and Roy-Chowdhury (2023) is ranking and valuation, with WTP information being used to separate strict preferences from indifferences.

Construction of Take-Up Rates. We define take-up as the extensive margin of commitment device usage i.e., the share of people choosing a commitment option over a more flexible option, not the intensity of usage. When participants were not given a choice to receive a certain product with commitment features but were simply offered it, we code take-up as usage of the product. For direct choice, take-up is the share of individuals willing to take up the commitment device at all. For ranking, take-up is the share of individuals who rank any restrictive option strictly above the fully non-restrictive option. For valuation, take-up is the share of individuals willing to commit at a positive price. When studies report take-up rates for multiple different commitment treatments, we built an average take-up rate across all treatment groups (unweighted).

Repeated Measurement. In total, 10 papers out of 54 measure commitment decisions repeatedly and the temporal distance between decisions ranges from a few minutes or a few weeks up to 6 months or even 2 years. Unlike the present study, these papers take one measurement for each new decision, not repeated measurements for a single decision. For instance, in Kaur et al. (2015), workers choose daily whether they want to take up a contract that penalizes low output. In Chow (2011), video-game players can choose before and during every game whether they want to use a device that blocks the option to continue playing. Due to the lack of papers that take any repeated measurement and the large variation among those that do, we did not exploit this dimension.

Commitment Type. To allow for a better comparison of take-up rates across elicitation methods, we split commitment devices into two broad categories. We classify a commitment device as *hard* if it strictly removes certain options from a person's choice set. For instance, in Chow (2011), using the device blocks the player from continuing to play the video game. A commitment is said to be *financial* if it entails a costs whenever a person deviates from her commitment. One example is the study of Erev et al. (2022), in which participants can choose to deposit money if they want to be committed to performing a breathing exercise on the next day – if they perform the exercise, they earn a small amount and otherwise, they lose the deposit. A small number of papers include both types of commitment devices, in which case we report a separate take-up rate for each category. In total, 34 papers study hard commitment devices and 17 papers study financial commitment devices, 4 study both.

Limitations. Comparisons of take-up rates across elicitation methods are limited by the fact that commitment devices come with multiple characteristics that could be correlated with the elicitation method and affect take-up. One characteristic is the *intensity* of commitment i.e., whether the consequences of committing or breaking a commitment are large or minor. Other examples of potentially relevant dimensions are the type of population (highly educated students vs. farmers in a developing country), the type of experiment (lab or field), or the domain (savings, smoking, etc.), all of which could be correlated with the choice of elicitation method. Trying to robustly capture all dimensions of heterogeneity is beyond the scope of this analysis but we note that heterogeneity in commitment features, population, or setting could potentially account for a large part of the observed variation in take-up.

Another caveat is the classification of used methods into direct choice and not valuation when studies elicit demand for a costly commitment device at a fixed cost (one decision only). It is not trivial to know where to draw the line because the presence of a price can make people think in terms of how much they value the commitment device and we discard these observations for the valuation method. However, by monotonicity in the aggregate, we should mechanically observe lower take-up rates in (between-subject) treatments with costly commitments and we chose to be more conservative in our classification. Inclusion of these cases would only strengthen our conclusion that WTP generates lower take-ups.

A.2 Overview of Take-Up Rate Studies

For completeness, below we provide a table of take-up rates broken down by paper and method within a paper. In case of multiple treatments, we report the average take-up as well as a range of take-up rates. We also indicated the number of studies that measure commitment decisions repeatedly, the type of commitment (hard or financial), as well as the domain (alcohol, exercise, finance, food, health, risk, sleep, smoking, social media, weight loss, Work and effort, other) and the type of study (artefactual field, framed field, lab, lab in the field, natural field, online, online field, online framed field).

Table A.1. Take-up rates of commitment devices in the literature.

Paper	Method	Avg. take-up	Take-up range	Repeated	Commitment	Domain	Study type
Acland and Chow (2018)	Direct Choice	0.24		yes	Hard	Other	Online Field
Afzal et al. (2019)	Direct Choice	0.16	0.15 - 0.19	no	Financial	Finance	Framed Field
Alan and Ertac (2015) *	Direct Choice	0.69		no	Hard	Food	Framed Field
Alan et al. (2021)	Valuation	0.44		no	Hard	Food	Artefactual Field
Allcott et al. (2022)	Direct Choice	0.89		yes	Hard	Social media	Online Field
Allcott et al. (2022)	Valuation	0.58		yes	Hard	Social media	Online Field
Anderberg et al. (2018)	Direct Choice	0.43			Financial	Work and effort	Framed Field
Andreoni and Serra-Garcia (2021)	Direct Choice	0.20	0.13 - 0.26	no	Hard	Finance	Lab
Andreoni et al. (2020)	Direct Choice	0.35	0.29 - 0.41	no	Hard	Other	Lab
Ashraf et al. (2006)	Direct Choice	0.28		no	Hard	Finance	Framed Field
Augenblick et al. (2015)	Direct Choice	0.59		no	Hard	Work and effort	Lab
Augenblick et al. (2015)	Valuation	0.09		no	Hard	Work and effort	Lab
Avery et al. (2022)	Direct Choice	0.44	0.13 - 0.60	yes	Financial	Sleep	Framed Field
Bai et al. (2021)	Direct Choice	0.14	0.14 - 0.14	no	Financial	Health	Framed Field
Barton (2015)	Direct Choice	0.56		no	Hard	Work and effort	Lab
Beshears et al. (2020)	Direct Choice	0.73	0.68 - 0.79	no	Financial	Finance	Online Framed Field
Beshears et al. (2020)	Direct Choice	0.82		no	Hard	Finance	Online Framed Field
Bettega et al. (2023)	Direct Choice	0.35		no	Hard	Risk	Lab
Bhatia et al. (2021)	Direct Choice	0.76		no	Hard	Work and effort	Lab
Bhattacharya et al. (2015)	Direct Choice	0.23	0.22 - 0.23	no	Financial	Exercise	Framed Field
Bisin and Hyndman (2020)	Direct Choice	0.39	0.29 - 0.54	yes	Financial	Work and effort	Online
Bonein and Denant-Boèmont (2015)	Direct Choice	0.41	0.34 - 0.46	no	Financial	Work and effort	Lab
Breig et al. (2023)	Direct Choice	0.66		no	Hard	Work and effort	Lab and Online
Breig et al. (2023)	Valuation	0.36		no	Hard	Work and effort	Lab and Online
Brune et al. (2016) *	Direct Choice	0.21		no	Hard	Finance	Framed Field
Brune et al. (2021)	Direct Choice	0.42		no	Hard	Finance	Framed Field
Carrera et al. (2022)	Direct Choice	0.40	0.27 - 0.64	no	Financial	Exercise	Framed Field
Casaburi and Macchiavello (2019)	Direct Choice	0.90	0.86 - 0.93	no	Hard	Finance	Framed Field
Casari (2009)	Direct Choice	0.62		no	Hard	Finance	Lab
Chow (2011)	Direct Choice	0.79		yes	Hard	Other	Online
Chow (2011)	Valuation	0.10		no	Hard	Other	Online
Dupas and Robinson (2013) *	Direct Choice	0.76	0.65 - 0.97	yes	Hard	Finance	Framed Field
Dykstra (2020)	Direct Choice	0.34	0.31 - 0.42	no	Hard	Finance	Framed Field
Ek and Samahita (2023)	Valuation	0.27		no	Hard	Work and effort	Lab
Erev et al. (2022)	Direct Choice	0.56		no	Financial	Health	Framed Field
Exley and Naecker (2017)	Direct Choice	0.53	0.41 - 0.65	no	Financial	Work and effort	Framed Field
Francis (2018)	Direct Choice	0.81		no	Hard	Finance	Framed Field
Francis (2018)	Valuation	0.66		no	Hard	Finance	Framed field
Giné et al. (2010)	Direct Choice	0.08		no	Financial	Smoking	Framed Field
Goldhaber-Fiebert et al. (2010)	Direct Choice	0.58	0.56 - 0.61	no	Financial	Exercise	Framed Field
Houser et al. (2018)	Direct Choice	0.30	0.13 - 0.45	yes	Hard	Work and effort	Lab
John (2020)	Direct Choice	0.42		no	Hard	Finance	Framed Field
John (2020)	Direct Choice	0.27		no	Financial	Finance	Framed Field
Karlan and Linden (2022)	Direct Choice	0.44		no	Hard	Finance	Framed Field
Karlan and Zinman (2018)	Direct Choice	0.23		no	Hard	Finance	Framed Field
Kaur et al. (2015)	Direct Choice	0.36		yes	Financial	Work and effort	Framed Field
Krawczyk and Wozny (2017)	Direct Choice	0.21		no	Hard	Food	Framed Field
Krawczyk and Wozny (2017)	Ranking	0.31		no	Hard	Food	Framed Field
Krügel and Uhl (2023)	Direct Choice	0.60	0.57 - 0.63	no	Hard	Sleep	Framed Field
Le Cotty et al. (2019)	Direct Choice	0.05	0.02 - 0.08	yes	Hard	Finance	Natural Field
Lichand and Thibaud (2023)	Direct Choice	0.89	0.83 - 0.93	no	Hard	Other	Lab in the Field
Lichand and Thibaud (2023)	Valuation	0.92		no	Hard	Other	Lab in the Field
McIntosh et al. (2021)	Direct Choice	0.11		no	Financial	Finance	Framed Field
Palacios-Huerta and Volij (2021)	Valuation	0.07		no	Hard	Food	Lab
Roy-Chowdhury (2023)	Ranking	0.70		no	Hard	Work and effort	Online
Roy-Chowdhury (2023)	Valuation	0.30		no	Hard	Work and effort	Online
Royer et al. (2015)	Direct Choice	0.12		no	Financial	Exercise	Framed Field
Sadoff and Samek (2019)	Direct Choice	0.36	0.20 - 0.50	no	Hard	Food	Framed Field
Sadoff et al. (2020)	Direct Choice	0.56	0.34 - 0.78	yes	Hard	Food	Natural Field
Schilbach (2019)	Direct Choice	0.46	0.44 - 0.52	yes	Financial	Alcohol	Framed Field
Schilbach (2019)	Valuation	0.35	0.31 - 0.40	yes	Financial	Alcohol	Framed Field
Sjåstad and Ekström (2022)	Direct Choice	0.53	0.46 - 0.59	no	Hard	Work and effort	Online
Toussaert (2018)	Ranking	0.53		no	Hard	Work and effort	Lab
Toussaert (2018)	Valuation	0.40		no	Hard	Work and effort	Lab
Toussaert (2019)	Direct Choice	0.65		no	Financial	Weight loss	Framed Field
Toussaert (2019)	Ranking	0.61		no	Hard	Weight loss	Framed Field
Zhang and Greiner (2021)	Direct Choice	0.02	0.00 - 0.07	no	Financial	Finance	Lab
Zhang and Greiner (2021)	Direct Choice	0.27	0.16 - 0.31	no	Hard	Finance	Lab
Zhang and Greiner (2021)	Valuation	0.10	0.03 - 0.19	no	Hard	Finance	Lab

Notes. Papers marked with an asterisk * report sign-up to the commitment device and not necessarily its usage. Avg. take-up is the simple arithmetic average of all take-up rates if multiple are reported. Range is the respective range in these cases.

Appendix B Experimental Design

B.1 Timeline and Instructions

Timeline. Figure B.1 shows the timeline of the study. Participants complete a 5-minute register of interest survey approximately one week before the main study. We will recruit a minimum of 200 participants, who will then receive the main survey approximately one week before the field experiment. After completing the main survey, we will randomly select 50% of participants to continue with the field experiment. These participants will receive the link to the dedicated ordering platform on Monday at 8:00 am and will have until 11:00 am to order their 5 meals for the week. Once orders are submitted, participants will receive a confirmation email to redeem their lunches in the college halls from Monday to Friday. Research assistants stationed at the college halls' registers will cross-reference participants as they pick up their meals. This allows us to closely monitor whether participants have picked up their meals.

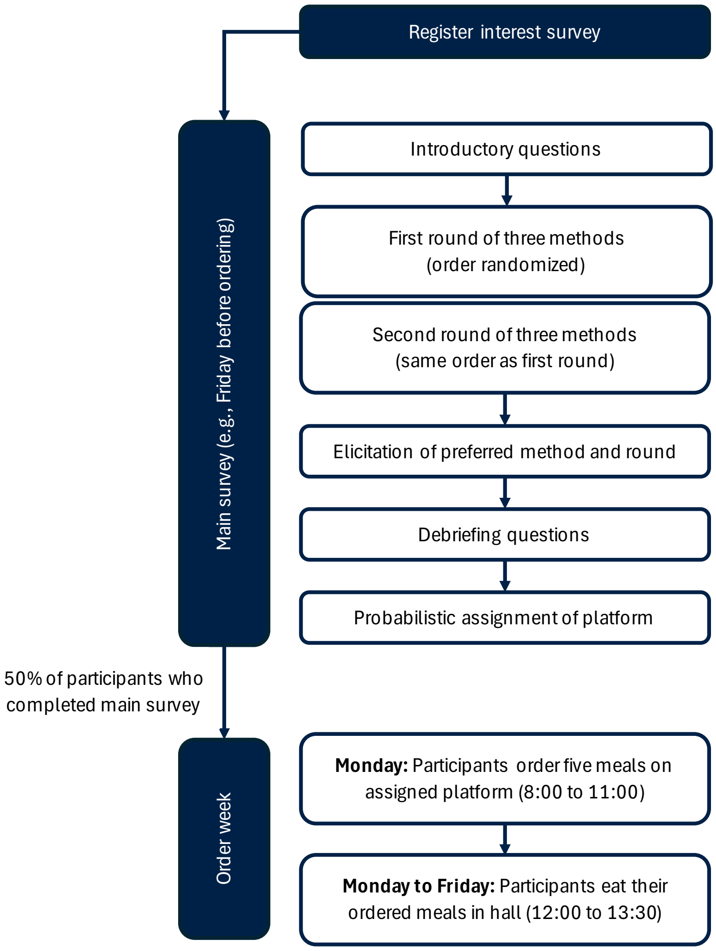
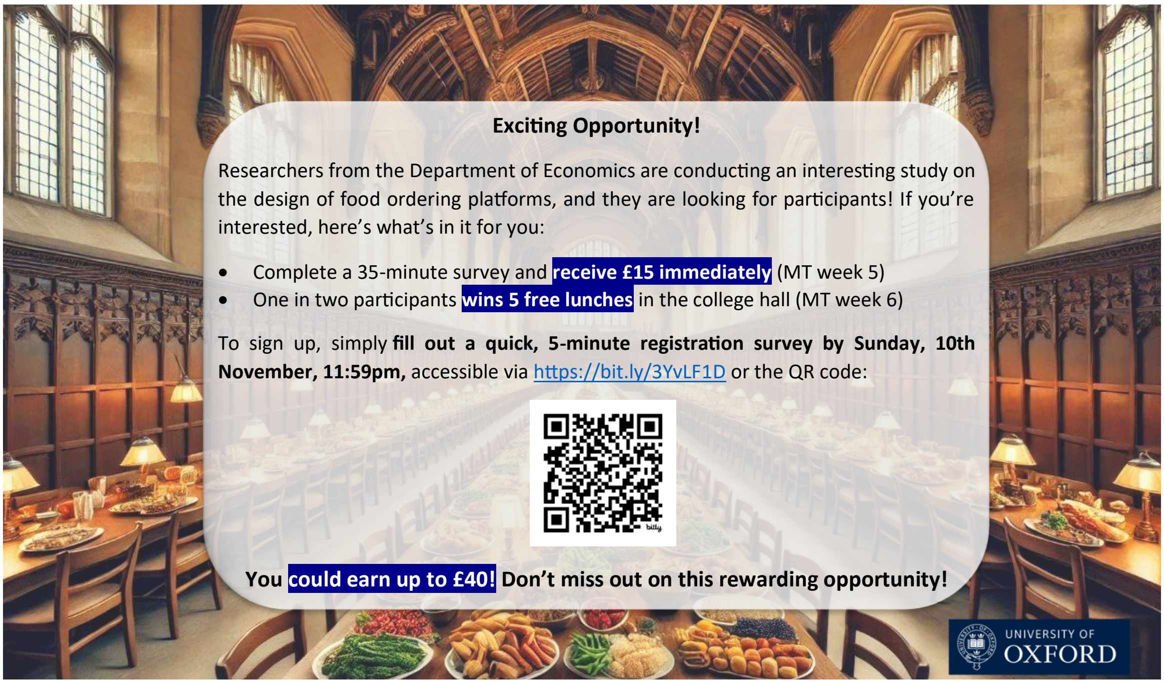

Register interest survey
Main survey (e.g., Friday before ordering)
Introductory questions
First round of three methods (order randomized)
Second round of three methods (same order as first round)
Elicitation of preferred method and round
Debriefing questions
Probabilistic assignment of platform
50% of participants who completed main survey
Order week
Monday: Participants order five meals on assigned platform (8:00 to 11:00)
Monday to Friday: Participants eat their ordered meals in hall (12:00 to 13:30)

Figure B.1. Illustrative timeline of the study.

B.2 Register-to-Interest Survey

Figure B.2 shows the recruitment flyer that leads to the registration survey.



Exciting Opportunity!

Researchers from the Department of Economics are conducting an interesting study on the design of food ordering platforms, and they are looking for participants! If you're interested, here's what's in it for you:

- Complete a 35-minute survey and **receive £15 immediately** (MT week 5)
- One in two participants **wins 5 free lunches** in the college hall (MT week 6)

To sign up, simply **fill out a quick, 5-minute registration survey by Sunday, 10th November, 11:59pm,** accessible via https://bit.ly/3YvLF1D or the QR code:

You could earn up to £40! Don't miss out on this rewarding opportunity!

Figure B.2. Recruitment flyer for the registration survey.

Table B.1. Illustration of possible meals in each category (the example represents a weekly menu at St John's).

Day	Daily Harvest (G) (Vegetarian)	Carb Powerhouse (O) (Carbohydrates)	Carnivore Corner (R) (Protein)
Mon	Lentil & Root Vegetable Pie	Roasted Tomato Arrabbiata	Jerk Chicken with Peas
Tue	Jackfruit & Sweet Potato Curry	Butternut Squash Sage Lasagna	Fish Pie with Leek
Wed	Spicy Bean Burger	Roasted Vegetable Pasta Bake	Chicken Burger
Thu	Sweet & Sour Tofu	Tomato Pasta Bake	Chicken Parmigiana
Fri	Moroccan Vegetable Filo Parcels	Cauliflower Penne Gratin	Battered Cod & Tartare Sauce

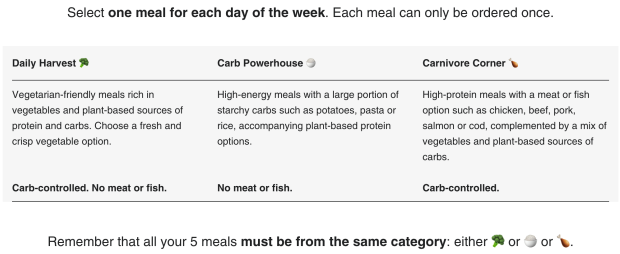
B.3 Food Ordering Platform

In the following, we describe the architecture of the food ordering platform.

Overview of Available Meals. Table B.1 presents an overview of the options available under a typical weekly menu at St John's College. The chosen composition of each category reflects the underlying principle of the challenge: meal category G is carb-controlled and contains no meat or fish, meal category O contains carbs but no meat or fish, and option R contains meat and fish but limited carbs. The meals in each row are only available on a given day. Participants decide for one category and order all meals from this category.

Landing and Selection Pages. The landing page of each platform version shows the available meals and the categories they belong to. Figure B.3 Panel A shows the landing page of a platform representing the menu GOR. Participants can select all their meals from a single category and then proceed to the checkout.

A. Overview meal categories



Select **one meal for each day of the week**. Each meal can only be ordered once.

Daily Harvest 🥦	Carb Powerhouse ⚪	Carnivore Corner 🍗
Vegetarian-friendly meals rich in vegetables and plant-based sources of protein and carbs. Choose a fresh and crisp vegetable option.	High-energy meals with a large portion of starchy carbs such as potatoes, pasta or rice, accompanying plant-based protein options.	High-protein meals with a meat or fish option such as chicken, beef, pork, salmon or cod, complemented by a mix of vegetables and plant-based sources of carbs.
Carb-controlled. No meat or fish.	**No meat or fish.**	**Carb-controlled.**

Remember that all your 5 meals **must be from the same category**: either 🥦 or ⚪ or 🍗.

B. Overview meal selection

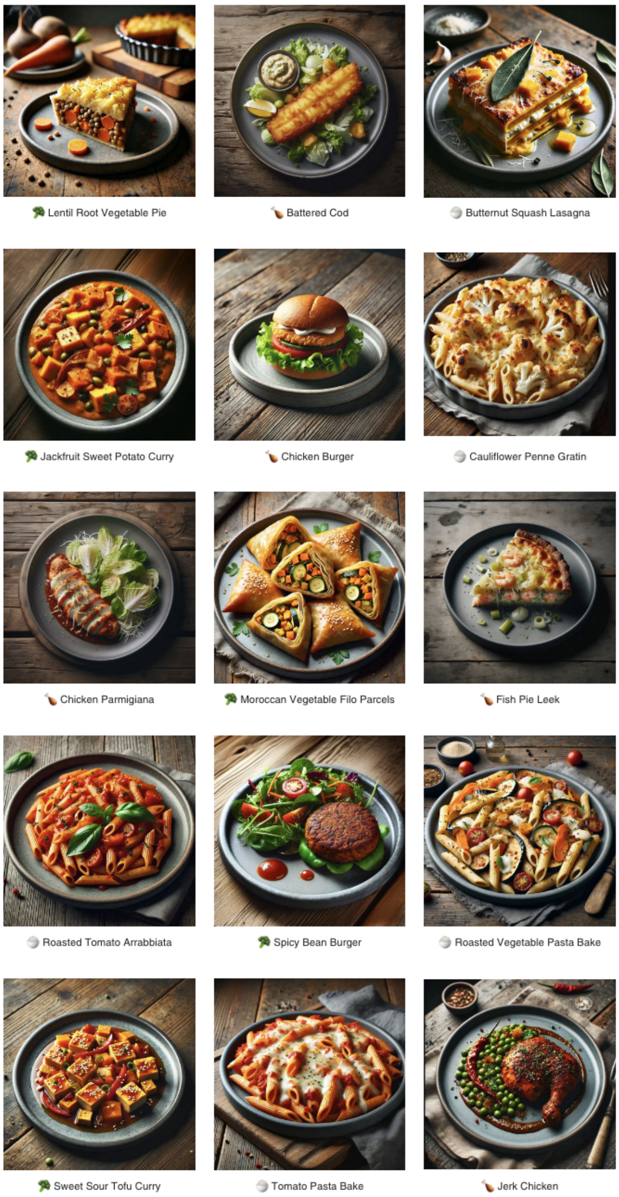

🥦 Lentil Root Vegetable Pie
🍗 Battered Cod
⚪ Butternut Squash Lasagna
🥦 Jackfruit Sweet Potato Curry
🍗 Chicken Burger
⚪ Cauliflower Penne Gratin
🍗 Chicken Parmigiana
🥦 Moroccan Vegetable Filo Parcels
🍗 Fish Pie Leek
⚪ Roasted Tomato Arrabbiata
🥦 Spicy Bean Burger
⚪ Roasted Vegetable Pasta Bake
🥦 Sweet Sour Tofu Curry
⚪ Tomato Pasta Bake
🍗 Jerk Chicken

Figure B.3. Landing and selection page of a platform representing an example menu GOR from St John’s college. The page shows all 15 available meals.

Appendix C Collected Variables in the Participant Survey

Table C.1 summarizes the variables collected in the participant survey. Variables 1–12 are demographic and background information. Variables 13–64 are choice variables, and variables 75–82 provide information to interpret participants' choices. Questions 13–26, 27–50, and 51–64 were block-randomized (fixed order between rounds).

Table C.1. Collected variables in the participant survey.

No.	Variable name	Description	Values
1	*gender*	Demographics	female (1), male (0), other
2	*college*	College affiliation of participant	Balliol, St Anne's, St John's
3	*jcr_mcr_member*	JCR or MCR member in college	JCR, MCR
4	*program_year*	Year in current study program	1, 2, 3, 4, 5
5	*num_visits_hall*	Number of times respondents go to college hall for lunch (in a typical week during term time)	1, 2, 3, 4, 5
6	*percentage_veggies*	Percentage of meals over the past 6 months that contained at least one portion of vegetables	0%, 10%, ..., 100%
7	*percentage_carbs*	Percentage of meals over the past 6 months that contained at least one portion of starchy foods as the main carbohydrates source	0%, 10%, ..., 100%
8	*percentage_meat*	Percentage of meals over the past 6 months that contained at least one portion of meat as the main protein source	0%, 10%, ..., 100%
9	*current_cons_veggies*	Current consumption of vegetables relative to desired	far too low, a bit too low, about right, a bit too high, far too high
10	*current_cons_carbs*	Current consumption of carbs relative to desired	far too low, a bit too low, about right, a bit too high, far too high
11	*current_cons_meat*	Current consumption of meat relative to desired	far too low, a bit too low, about right, a bit too high, far too high
12	*challenge_motivation*	Personal motivation to participate in the food challenge	0 (not motivated at all), 10, ..., 100 (very motivated)
13–26	*rank_{round}_{menu}*	Assigned rank in {round} to {menu}	1, 2, ..., 7
27–50	*bc_{round}_{menupair}*	Assigned preference in {round} for {menupair}	-1 (smaller), 0 (indifference), 1 (larger)
51–64	*wta_{round}_{menu}*	Assigned monetary valuation in {round} to {menu}	£1, £2, ..., £35
65	*preferred_method*	Preferred method in determining the menu	rank, bc, wta, random
66	*preferred_round*	Preferred round in determining the menu	round 1, round 2, random
67–69	*reliability_{method}*	Self-report about the reliability of {method} as a way to determine preferences for the various menus	0% (not at all), ..., 100% (extremely)
70	*hyp_meals_challenge*	Self-report of the meal category that a participant wants to try in the challenge	G, O, R
71	*categories_platform*	Self-report of the number of meal categories that a participant wants to see on the food platform	category chosen in *hyp_meals_challenge*, all meal categories, indifferent
72	*indecisiveness_rating*	Self-report on difficulty to compare the meal platforms	0, 10, ...10
73–75	*difficulty_{method}*	Self-report about difficulty to come up with an answer in {method}	0 (not at all), 10, ..., 100 (extremely)
76–78	*tediousness_{method}*	Self-report about tediousness of completing answers in {method}	0 (not at all), 10, ..., 100 (extremely)
79–81	*certainty_{method}*	Self-report on certainty about answers given in {method}	0 (not at all), 10, ..., 100 (extremely)
82	*coherence*	Agreement whether intransitive relation is coherent	0 (not at all), 10, ..., 100 (fully)
83	*instructions_clarity*	How difficult or easy were the instructions	{very easy, somewhat easy, neither easy or difficult, somewhat difficult, very difficult}
84	*study_feedback*	Personal opinion whether challenge concept was interesting and thought process of decision tasks	free text field
85	*incentive_feedback*	Suggestions to make future participation more attractive	free text field

Notes: The expressions in curly brackets can take the following values: {method} ∈ {rank, wta, bc}, {round} ∈ {round_1, round_2}, {menu} ∈ {gor, go, gr, or, g, o, r}, {menupair} ∈ {gor_go, gor_gr, gor_or, gor_g, gor_o, gor_r, go_g, go_o, gr_g, gr_r, or_o, or_r}

Appendix D Pilot Data and Simulated Benchmark

D.1 Information on Pilot Data

D.1.1 Differences Between the Pilot Study and Current Design

Below we provide more specific information on the pilot data used to build the figures and tables presented in this registered report. A link to this prior survey is posted on https://osf.io/robustness_commitment_demand. In particular, we highlight similarities and key differences with the current design to help the reader understand the extent to which findings from the pilot can be extrapolated to our current setting. Besides being hypothetical and using a different participant pool, this pilot study differed with respect to the structure of the food challenge, the procedure used to elicit preferences and the measurements taken in order to interpret the preferences expressed.

Food Challenge. The duration of the food challenge was also 5 consecutive days but included 10 meals (both lunch and dinner). The meals were provided by a meal-prep company. Two separate orders of 5 meals had to be placed instead of a single order of 5 meals. Including delivery fees, the total market value of the meals offered was £72 vs. £35 in the present design.

Another major difference is that participants could order mixed bundles containing meals from different meal categories if the options were available on their assigned platform. For example, if participants received the platform GOR, they could order 3 meals from G, 5 meals from O, and 2 meals from R. Consequently, GOR $\neq$ G $\cup$ O $\cup$ R in the space of meal bundles. This contrasts with our current design in which participants who receive GOR must order all 5 meals from a single category e.g,. 5 meals from G, 6 meals from O, or 5 meals from R. We made this design choice to obtain a tighter mapping between our theoretical framework and study design. One implication of this change is that we expect the prevalence of flexibility choices (i.e., favoring a larger menu over a subset) to decrease with this new design.

Elicitation Procedure. In the pilot study, participants did not repeat the decision tasks i.e., they indicated their preferences only once in each method. Overall, the design of the elicitation methods was less symmetric in the pilot study: (i) the order of presentation of the methods was fixed (ordinal ranking first, binary choices second, monetary valuations last), instead of being randomized, (ii) we did not show a summary page after participants completed the binary choice and monetary valuation tasks, and (iii) participants were only able to revise their choices in the ordinal ranking. We focused our attention on the measurement of an ordinal ranking, as this information is both necessary and sufficient to test the theoretical framework used in our main field experiment (which requires the elicitation of a weak order $\succeq$ on the set $\mathscr{M}$ of all menus).

The elicitation methods were also slightly different. The binary choice method used a drag-and-drop interface in which participants made only an active choice if they preferred a menu. In the monetary valuation method, the menus were not re-ordered according to the assigned values and the range of admissible values was £0 to £100.

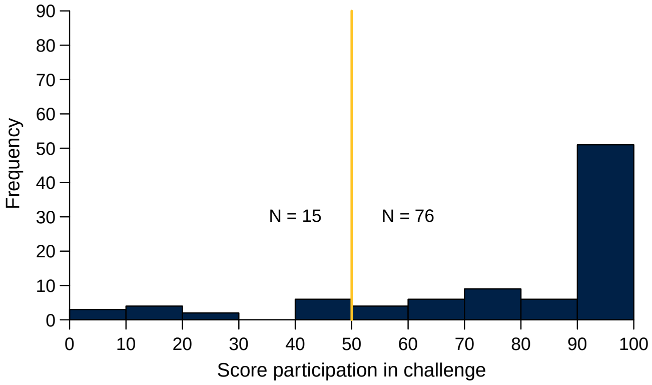


Figure D.1. Histogram distribution willingness to participate in challenge in the pilot data.

Interpretation of Expressed Preferences. The debriefing questions in the pilot study focused on the motivations for choosing a particular menu and participating in the challenge. Hence, we did not collect any of the variables described in Section 2.4 pertaining to subjective perceptions of each elicitation method or preferences for a particular method.

D.1.2 Basic Descriptive Statistics

Below we provide basic descriptive statistics from our pilot data.

Willingness to Participate in the Challenge. Figure D.1 plots the motivation to participate in the food challenge. We offer no financial incentives for succeeding in the challenge, and hence we regard the willingness to participate in the challenge as a proxy for how seriously participants regarded the challenge.

Distribution of Monetary Valuations. Figure D.2 shows the distribution of monetary valuations pooled across all menus. Although respondents could indicate any number between £0 and £100, most monetary valuations are multiples of 5.

Distribution of Ranking Steps. Figure D.3 shows the distribution of steps taken to complete the ranking. Most respondents (over 50%) completed the ranking in 7 steps i.e., indicated no indifferences.

D.1.3 Piloting of Alternative Elicitation Procedures

Our elicitation procedure for measuring preferences over food ordering platforms was refined after a series of pilots that took place between 2020 and 2022. For transparency, below we explain alternative designs we considered for each method and how we converged towards the present design.

Ordinal Ranking. We considered a simple drag-and-drop format in which menus are listed in a random order and respondents must reorder the 7 menus with their most (least) preferred option at the top (bottom) of the list. We tried this drag-and-drop format both as a single-step and as a step-by-step procedure, with the latter asking respondents to first rank the group of singleton menus $\mathrm{G}, \mathrm{O}, \mathrm{R}$ and then the group of doubleton menus $\mathrm{GO}, \mathrm{GR}, \mathrm{OR}$. A major downside of this procedure is that indifferences cannot be measured. Since we observed some inconsistencies in how respondents ranked singletons and doubletons alone vs. in the full-list format and were unable to tell apart noise from indifferences, we decided to opt for the iterative procedure. The iterative procedure has two main advantages: first, it is progressive, allowing respondents to reassess their decisions at any step of the ranking. Second, it allows them to express indifferences between several menus in a natural way i.e., by making multiple selections from the list for a given rank.

Binary Choices. We considered designs with direct binary choices that did not offer the option to express indifferences. These binary choices were presented in a simple matrix format with a bipolar scale (two columns of radio buttons with menu labels presented on the left- and right-hand side of the buttons). The list of decisions was presented in a compact but compressed manner, which we worried might lead to more mistaken selections. As a result, we opted instead for our current multiple-choice

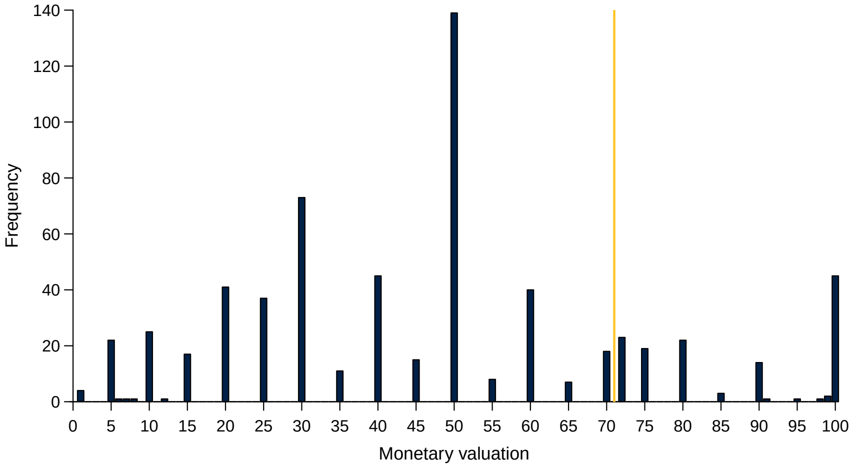


Figure D.2. Distribution of monetary valuations of $N = 91$ participants. The vertical line shows the market value of the meal package, £72. Note, that participants in the pilot study ordered more than 6 meals.

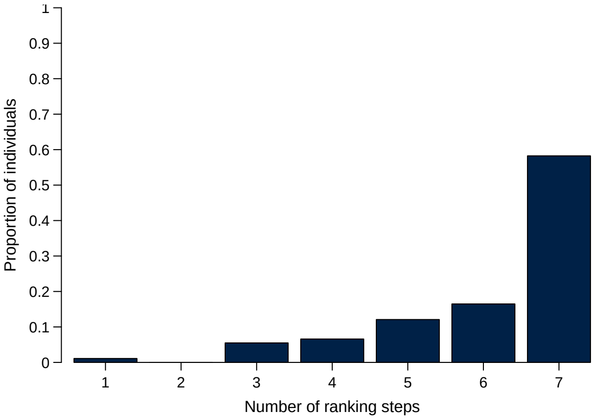


Figure D.3. Distribution of the number of steps that a respondent took to complete his ranking. A ranking in 7 steps indicates the expression of a strict preference $\succ_R$ on $\mathscr{M}$, while a ranking in one step indicates full indifference $\sim_R$.

format, which gives more breathing space and more clearly highlights selections. We also considered a drag-and-drop format in which respondents had to bring their favorite option inside a box in order to express a preference: not dragging an option was then interpreted as an indifference. With this format, we were concerned that respondents might more easily miss a comparison or skip one due to the tediousness of the task. For all these reasons, we decided to use the simplest format that would allow us to measure indifferences without imposing a large burden on respondents.

Monetary Valuations. The biggest challenge was figuring out how to obtain a money-metric measure of the commitment or flexibility value of a menu. At the outset, there are two main possible approaches:

(1) The first approach asks respondents to indicate how much more or less they value a certain menu relative to some reference menu, thus directly eliciting a difference in valuations Δv. Our very first design considered an option of this type by asking respondents to indicate using a slider the amount of money Δv_N they would request to give up their top-ranked menu (reference menu) and instead receive the menu they ranked $\# N$ for all $N > 1$. This procedure has several important limitations. First, differences in valuations are only directly elicited for 6 pairs (top ranked-menu vs. the other 6); thus, inferences for the remaining comparisons can be drawn only indirectly by taking differences in measurements, possibly leading to substantial measurement error. Second, the elicitation method is one-sided: respondents are constrained to report a (weakly) positive number, which does not allow one to test for possible preference reversals when switching between different methods of elicitation.[34] Third, there might be important reference effects depending on which menu is chosen as the benchmark.

(2) The second approach, which we ended up adopting, consists in asking respondents for their overall valuation v_M of each menu M and inferring the value of commitment vs. flexibility by calculating the difference $v_M - v_{M'}$ for each menu pair (M, M') such that $M' \subset M$. Doing so also allows us to obtain an estimate of respondents' valuation of the challenge (if offered menu M) and thus assess WTP for commitment in relative terms. The two main downsides are that if valuations are measured with error, then taking differences in valuations will have a compounding effect on the amount of measurement error. In addition, valuations might anchor more around the market value of the meals, possibly leading to an overestimation of indifferences.

In terms of incentivization procedure, we only considered the BDM mechanism to pin down a precise valuation while minimizing the number of elicitation steps. Other methods such as MPLs would have yielded more coarse measures and imposed a prohibitive time cost on respondents.

Current Information on Each Design Element. We conclude this section by summarizing what elements of the current design have been piloted and what remains unknown to us:

- We have never piloted our new challenge format with meal bundles constrained to be from one meal category only. As a result, we remain uncertain about the overall prevalence of commitment vs.

[34] We considered a two-sided elicitation procedure allowing respondents to request a positive amount of money to keep the option ranked $\# N$ instead of the top option, but we could not find an approach that would use a single question and remain intuitive to respondents.

flexibility choices. However, since we removed the diversification motive, we expect that respondents' propensity to favor larger platforms will go down.

- The iterative ranking procedure in its current form was piloted several times and refined to present summary tables of choices at each step. For this reason, we have some preliminary insights on respondents' tendency to make multiple selections (i.e., reveal indifferences).
- The binary choice procedure has never been piloted in its current format and we are therefore unable to assess what will be the impact of setting the default choice on the middle button "I like both equally".
- The present monetary valuation procedure was piloted several times with menus either presented in the order of the submitted rankings or in descending order of menu size. As a result, we are able to provide some insights on the potential clustering of valuations around certain values and tendency to express indifferences.
- None of our previous piloting efforts studied the stability of responses within a given method by repeating the same task twice. We have no insights on the level of stability that could be expected or how it might interact with the method.
- We have not piloted any of the questions presented in Section 2.4 to understand how respondents perceive each elicitation method and which one they prefer.

Appendix E Expert Survey

E.1 Recruitment and Incentives

Eligibility Criteria. Experts are eligible to participate in the survey if they are (i) a Ph.D. candidate, postdoctoral researcher, or (assistant/associate) professor, (ii) work in the field of economics and/or psychology. We will inform prospective respondents that the survey will stay open for two weeks, and send a reminder after one week. Responses that arrive after the two-week deadline as well as incomplete responses will be excluded from the analyses.

Participation and Incentives. Experts will be required to provide consent for their participation. We will incentivize predictions as follows: out of the total of 17 forecasts, we will select one at random and compare it to the empirical data we collect. Among those experts who submitted a correct prediction, five will be randomly selected to receive a prize of £100 which will be donated to a charity of their choice or distributed as gift card.

E.2 Survey Design Details

The expert survey can be found in the supplementary material or under https://osf.io/robustness_commitment_demand. To experience the expert survey directly, you can use the following link.

E.3 Expert Sample Characteristics

Here we will present descriptive statistics about our population of forecasters, including information on their academic position, main field, degree of expertise in the relevant topics, and their mode of recruitment.

E.4 Distribution of Expert Forecasts and Sensitivity Analyses

Descriptive Statistics. We will present violin plots of the distribution of expert forecasts for each prediction we elicit, allowing us to document heterogeneity in responses.

Alternative Null Hypotheses. In the main text, we plan to use the average of the expert forecasts as the null-hypothesized value. Here we will report whether our conclusions change if we use the median expert forecast instead.

E.5 Collected Variables in the Expert Survey

Table E.1 summarizes the variables collected in the expert survey. Variables 1–17 are the expert forecasts used in our analyses. Variables 18–21 are demographics and background information about the experts.

Table E.1. Collected variables in the expert forecast survey.

No.	Variable name	Description	Values
1–3	*exp_commit_{method}*	Average number of comparisons in which participants expressed a preference for commitment in round 1 using {method} i.e., $\bar{C}^E_{j,1}$	0, 1, …, 12
4–6	*exp_indiff_{method}*	Average number of comparisons in which participants expressed indifference in round 1 using {method} i.e., $\bar{I}^E_{j,1}$	0, 1, …, 12
7–9	*exp_flex_{method}*	Average number of comparisons in which participants expressed a preference for flexibility in round 1 using {method} i.e., $\bar{F}^E_{j,1}$	0, 1, …, 12
10	*exp_consist_three*	Average number of comparisons in which expressed preferences in round 1 will agree across all of the three methods i.e., $\bar{\Gamma}_{3,1}$	0, 1, …, 12
11	*exp_consist_two*	Average number of comparisons in which expressed preferences in round 1 will agree for two of the three methods i.e., $\bar{\Gamma}_{2,1}$	0, 1, …, 12
12	*exp_consist_zero*	Number of comparisons in which expressed preferences in round 1 will agree for none of the three methods i.e., $\bar{\Gamma}_{0,1}$	0, 1, …, 12
13–15	*exp_stability_{method}*	Average number of comparisons in which expressed preferences will be stable between round 1 and 2 using {method} i.e., $\bar{\rho}_j$	0, 1, …, 12
16	*exp_preferred_method*	Participants' most frequently chosen method to determine the food ordering platform	rank, wta, bc, random
17	*exp_preferred_round*	Participants' most frequently chosen round to determine the food ordering platform	round_1, round_2, random
18	*exp_current_position*	Current academic position	Ph.D. candidate, postdoctoral researcher, assistant professor / untenured, associate professor, full professor, other
19	*exp_main_field*	Main research field of the expert	applied microeconomics, behavioral / experimental economics, development economics, econometrics, microeconomic theory, macroeconomics, psychology, other
20	*exp_knowledge_pref*	Self-report on knowledgeability on the topic of preference elicitation (relative to other topics)	0 (not at all), 10, …, 100 (extremely)
21	*exp_knowledge_commit*	Self-report on knowledgeability on the topic of commitment devices (relative to other topics)	0 (not at all), 10, …, 100 (extremely)
22	*exp_comment_predictions*	Free-text comment on reasons why different methods lead to different prevalence rates, stability indices, and might be preferred by respondents	
22	*exp_donation*	Selected charity for the donation of the £50 incentive	Cancer Research UK, WaterAid UK, World Food Programme, other
23	*exp_comment_study*	Free-text comment about the forecasting survey	

Notes: Variable {method} ∈ {rank, wta, bc}.

Appendix F Prevalence of Indifferences and Flexibility Preferences

Figure F.1, Table F.1, and Figure F.2 about the prevalence of indifferences and flexibility preferences (as defined in Equation 1 and Equation 2) are the counterparts to those presented in Section 4.1 for commitment preferences.

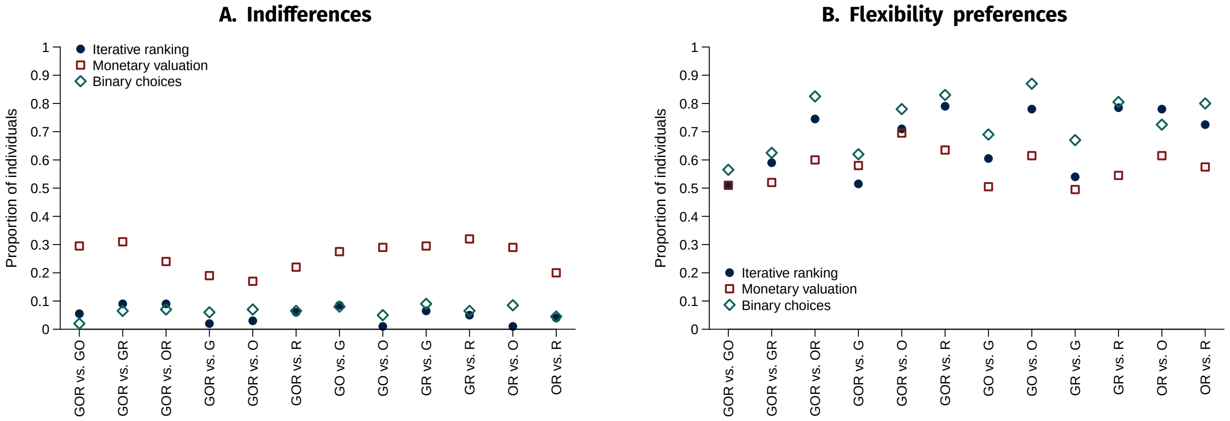


Figure F.1. Proportion of individuals with expressed indifferences and flexibility preferences at given menu pairs.

Table F.1. Prevalence rate of expressed indifferences and flexibility preferences from round-1 choices across the three elicitation methods.

	Indifferences			Flexibility preferences		
	Iterative ranking	Monetary valuation	Binary choices	Iterative ranking	Monetary valuation	Binary choices
Empirical $N = 2,400$	0.05 (0.12)	0.26 (0.32)	0.06 (0.19)	0.67 (0.25)	0.57 (0.34)	0.73 (0.24)
Simulated $N = 2,400$	0.13 (0.09)	0.02 (0.04)	0.33 (0.15)	0.43 (0.21)	0.50 (0.22)	0.35 (0.13)
Expert $N = N_E$	0.xx (0.xx)	0.xx (0.xx)	0.xx (0.xx)	0.xx (0.xx)	0.xx (0.xx)	0.xx (0.xx)

Notes: Standard errors in parentheses. Entries 0.xx will be populated once we collected the data.

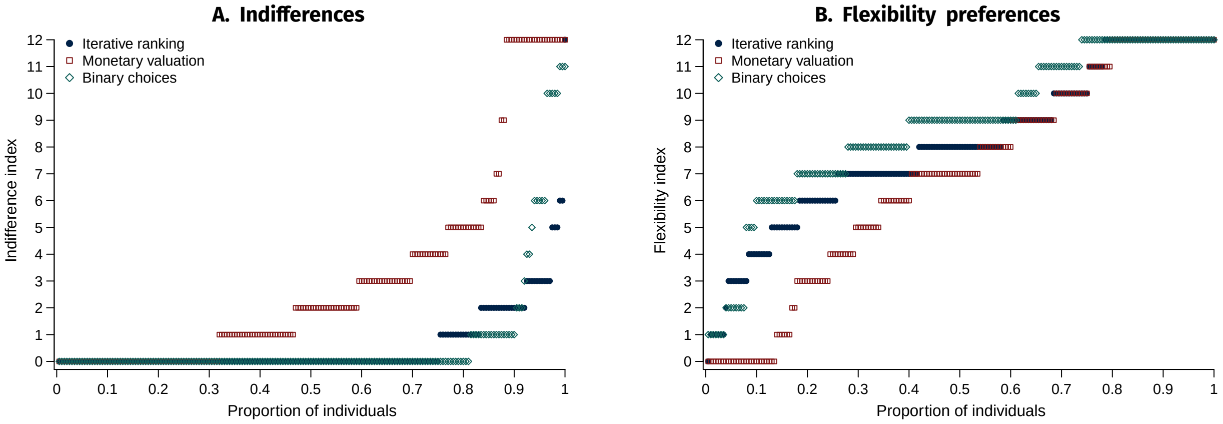


Figure F.2. Quantile plots of indifferences and flexibility preference indices across the three elicitation methods.

Appendix G Stability

G.1 Variation in Prevalence Rate of Preference Indices Between Rounds

In Section 4.2, we present scatter plots of the commitment indices (as defined in Equation 2) of each individual in round 1 vs. round 2. Below we report similar scatter plots for indifferences (Figure G.1) and flexibility preferences (Figure G.2).

G.2 Size of Instabilities

For the ranking and monetary valuation methods, we complement the qualitative evidence presented in the main text by quantifying the size of instabilities. For all menu pairs (M, M') with $M' \subset M$, we will examine the distribution of the difference in round-2 valuations $v^i_{M,2} - v^i_{M',2}$ conditional on the preference expressed via monetary valuations in round 1 i.e., $P^i_{V,1}(M, M') \in \{-1, 0, 1\}$. If the preferences expressed between the two rounds remain stable, then the distributions of this difference should be centered at zero. Larger deviations from zero would indicate larger instabilities. For completeness, we perform a similar exercise for the iterative ranking procedure, this time by computing the difference in ranks between menus M and M' in round 2 conditional on the preference expressed via ranking in round 1 i.e.,

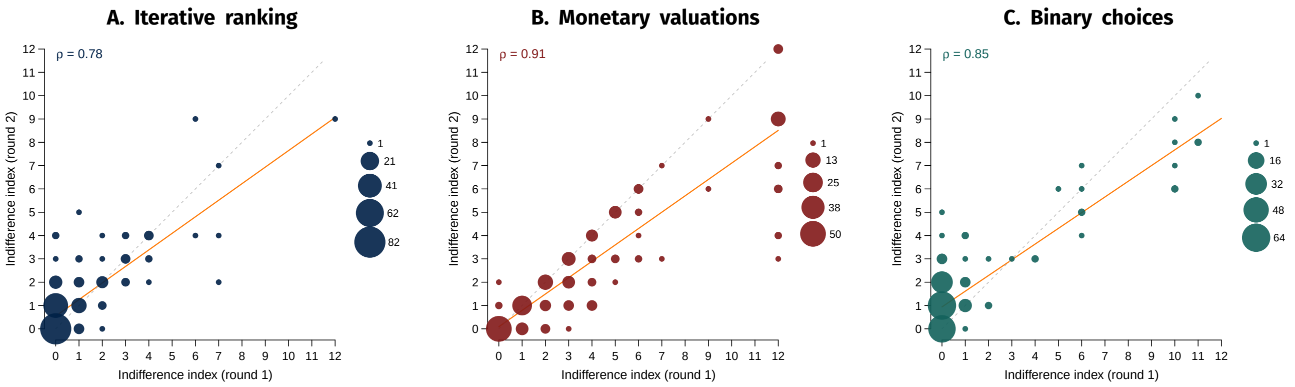


Figure G.1. Scatter plots of indifference indices between round 1 and round 2 choices. Bubble sizes indicate the number of individuals with the same pair of indifference indices.

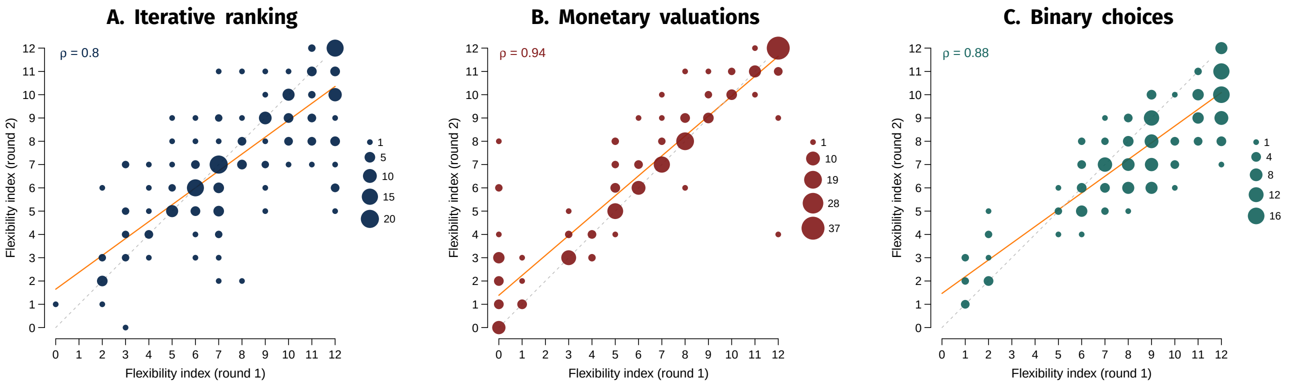


Figure G.2. Scatter plots of flexibility indices between round-1 and round-2 choices. Bubble sizes indicate the number of individuals with the same pair of flexibility indices.

$P^i_{R,1}(M,M') \in \{-1,0,1\}$. The information is presented in panel A (panel B) of Figure G.3 for the iterative ranking (monetary valuations).

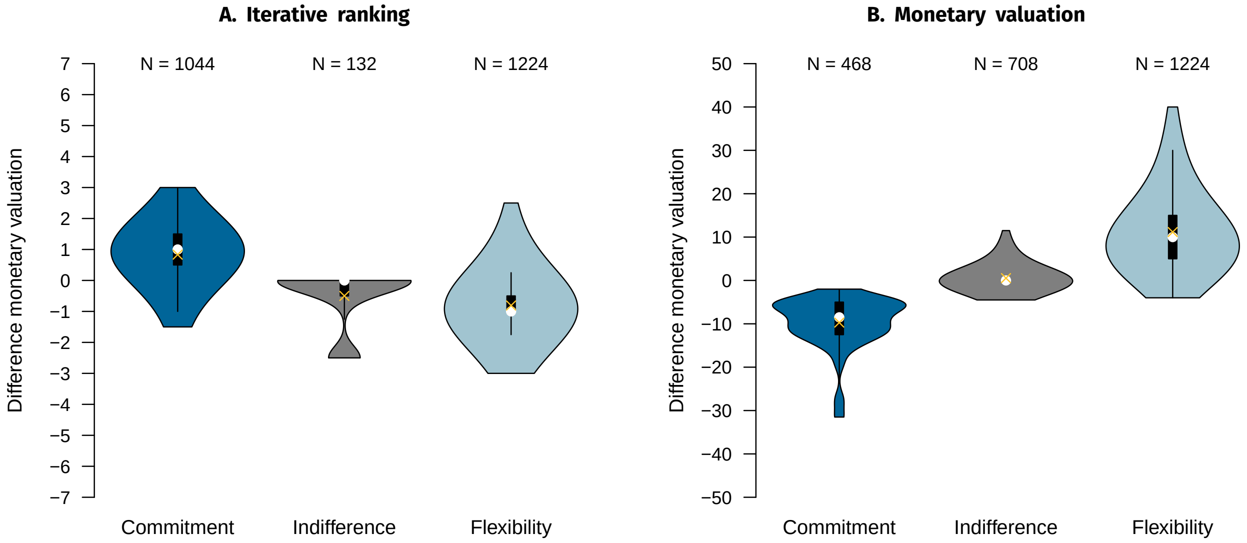


Figure G.3. Violin plots of difference of ranks and monetary valuations (V) conditional on expressed commitment, indifference, and flexibility preferences between round 1 and round 2. The white dot and orange cross indicate the median and mean, respectively.

Appendix H Analysis of Transitivity

To test for violations of transitivity in our setting, we check 6 possible preference cycles: $\{(G, GO, GOR), (G, GR, GOR), (O, GO, GOR), (O, OR, GOR), (R, GR, GOR), (R, OR, GOR)\}$. In Section 5.1.3, we define a *transitivity index* $\tau_r^i \in \{0, 1, \ldots, 6\}$ for the number of triplets at which transitivity is satisfied by individual i in round r. Figure H.1 shows the distribution of this index for the random benchmark (panel A) and empirical data (panel B). In our pilot, over 75% of people satisfied transitivity perfectly. Figure H.2 contrasts the distribution of the bilateral consistency indices (as defined in Equation 5) for perfectly transitive individuals (set $\mathscr{I}_{T_1}$) and those who fail to satisfy transitivity at least once (set $\mathscr{I}_{NT_1}$).

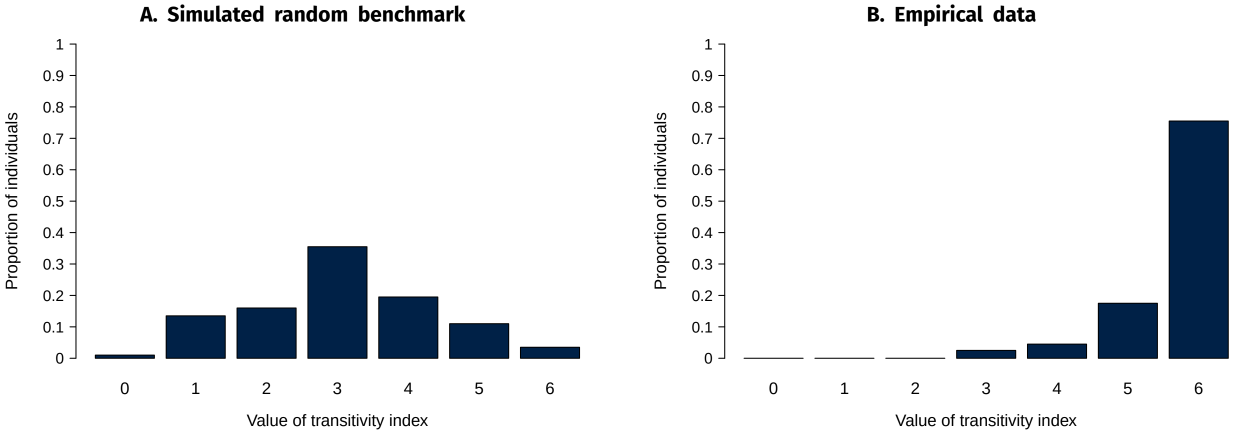


Figure H.1. Distribution of transitivity indices from round-1 binary choices.

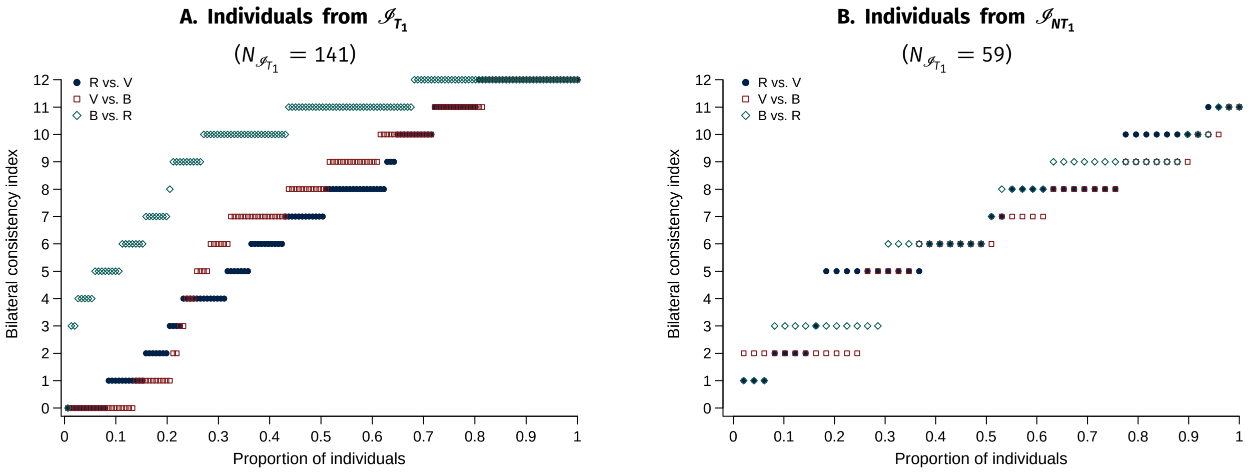


Figure H.2. Quantile plots of bilateral consistency indices calculated from round-1 choices across three elicitation methods, split by individuals from $\mathscr{I}_{T_1}$ and $\mathscr{I}_{NT_1}$.

Appendix I Round-2 Choices

In the following, we present the main figures and tables for round-2 choices.

I.1 Commitment, Indifferences and Flexibility Preferences in Round 2

Table I.1. Prevalence rate of expressed commitment preferences, indifferences and flexibility preferences across the three elicitation methods from round-2 choices.

	Commitment preferences			Indifferences			Flexibility preferences		
	R	V	B	R	V	B	R	V	B
Empirical	0.29 (0.21)	0.21 (0.23)	0.23 (0.15)	0.08 (0.11)	0.19 (0.25)	0.12 (0.15)	0.62 (0.23)	0.61 (0.31)	0.65 (0.19)
Simulated	0.44 (0.21)	0.48 (0.22)	0.31 (0.14)	0.15 (0.11)	0.02 (0.04)	0.34 (0.14)	0.44 (0.22)	0.51 (0.22)	0.32 (0.13)

Notes: Standard errors in parentheses. Columns R, V, and B denote iterative ranking, monetary valuations, and binary choices, respectively. All calculated numbers are based on $200 \times 12 = 2,400$ observations.

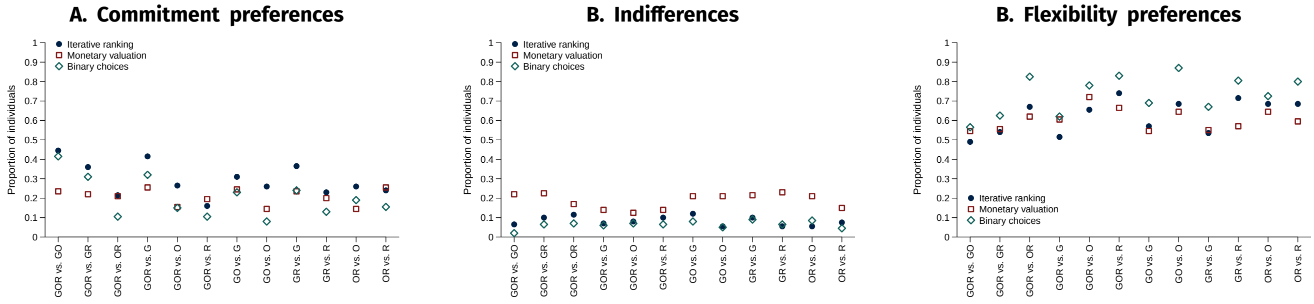


Figure I.1. Scatterplots of proportion of individuals with expressed commitment preferences, indifferences, and flexibility preferences (see Equation 1) by menu pairs from round-2 choices.

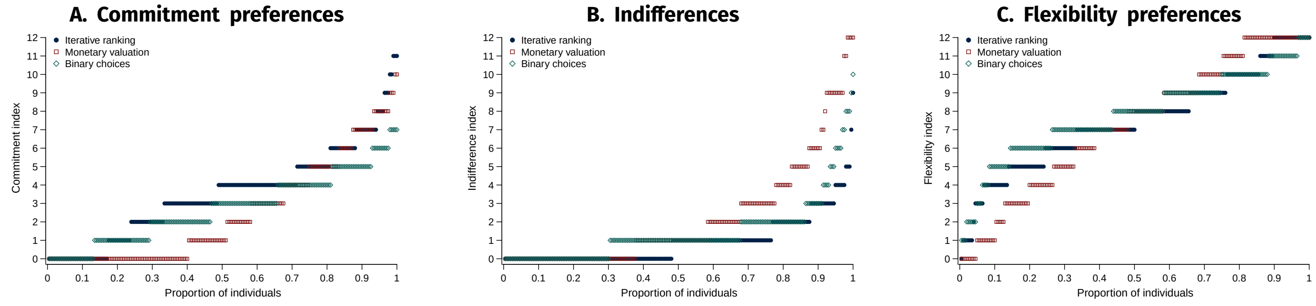


Figure I.2. Quantile plots of commitment, indifference, and flexibility preference indices (see Equation 2) across the three elicitation methods from round-2 choices.

I.2 Consistency Across the Three Methods in Round 2

Table I.2. Consistency of expressed preferences across elicitation methods from round-2 choices.

	$\bar{r}_{3,2}$	$\bar{r}_{2,2}$	$\bar{r}_{0,2}$
Empirical	0.38	0.53	0.09
$N = 2,400$	(0.05)	(0.06)	(0.03)
Simulated	0.14	0.65	0.21
$N = 2,400$	(0.06)	(0.04)	(0.03)

Notes: Standard errors in parentheses. The first, second and third column, respectively, indicates the proportion of binary comparisons at which expressed preferences agree across all three methods, agree for only two methods, and differ across all three methods.

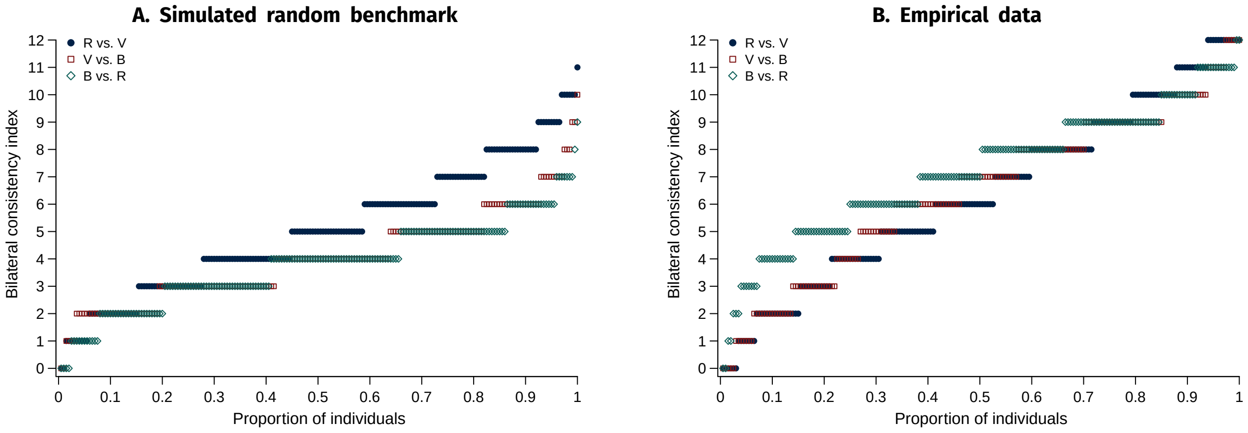


Figure I.3. Quantile plots of bilateral consistency indices across the three elicitation methods from round-2 choices.

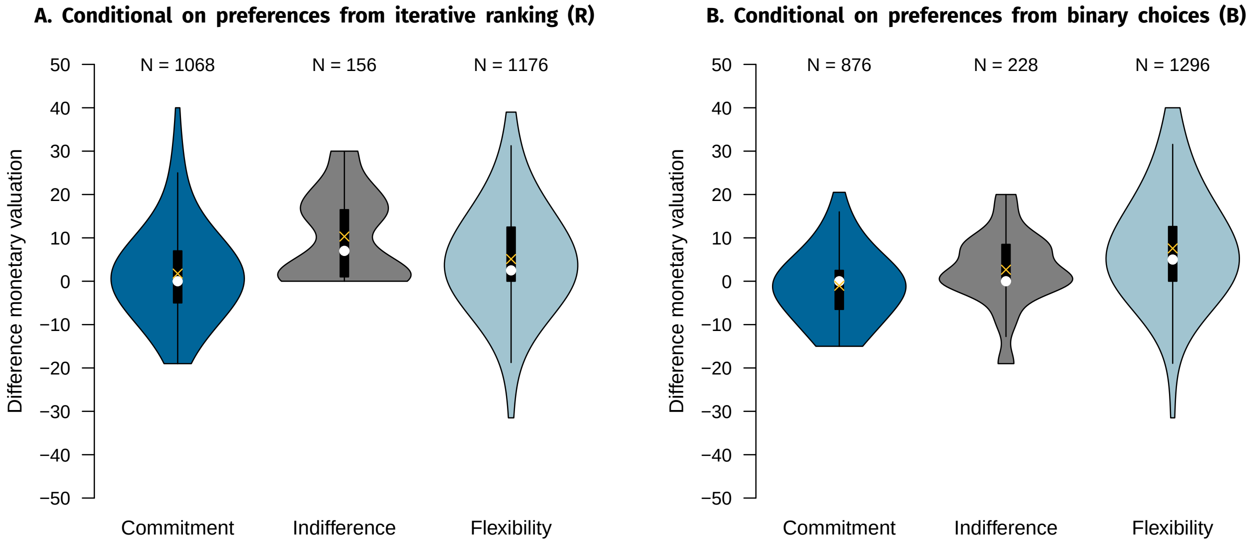


Figure I.4. Violin plots of the difference in monetary valuations $v^i_{M,2} - v^i_{M',2}$ conditional on expressed commitment, indifference, and flexibility preferences ($P^i_{j,2}(M, M') \in \{-1, 0, 1\}$), pooling across all respondents i and all 12 menu pairs (M, M') such that $M' \subset M$. The white dot and orange cross indicate the median and mean, respectively.

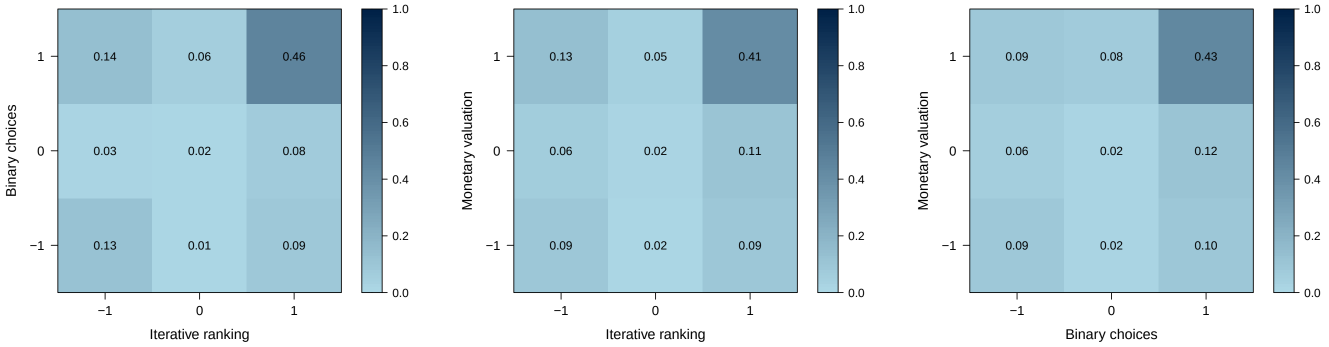


Figure I.5. Joint probability distribution of preference indicators $P^i_{j,2}(M, M') \in \{-1, 0, 1\}$ from round-2 choices across the 3 method pairs $(j, j') \in \{(R, B), (R, V), (V, B)\}$.

I.3 Transitivity Analyses for Round 2

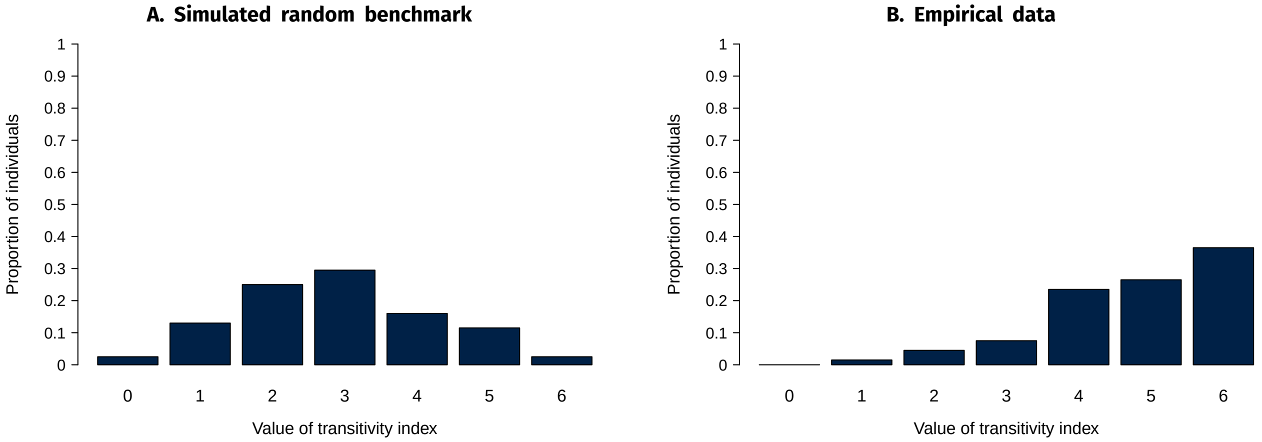


Figure I.6. Distribution of transitivity indices from round-2 binary choices.

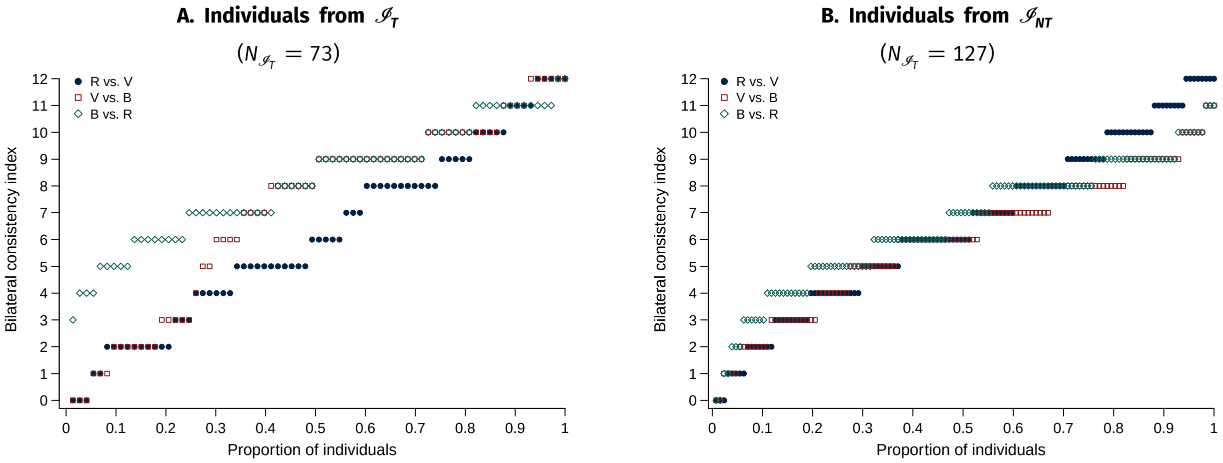


Figure I.7. Quantile plots of bilateral consistency indices calculated from round-2 choices across three elicitation methods, split by individuals from $\mathscr{I}_T$ and $\mathscr{I}_{NT}$.

References

Acland, Dan, and Vinci Chow. 2018. "Self-Control and Demand for Commitment in Online Game Playing: Evidence From a Field Experiment." *Journal of the Economic Science Association* 4 (1): 46–62. DOI: 10.1007/s40881-018-0048-3. [47]

Afzal, Uzma, Giovanna D'Adda, Marcel Fafchamps, Simon Quinn, and Farah Said. 2019. "Implicit and Explicit Commitment in Credit and Saving Contracts: A Field Experiment." Working paper w25802. National Bureau of Economic Research. DOI: 10.3386/w25802. [47]

Alan, Sule, and Seda Ertac. 2015. "Patience, Self-Control and the Demand for Commitment: Evidence From a Large-Scale Field Experiment." *Journal of Economic Behavior & Organization* 115 (7): 111–22. DOI: 10.1016/j.jebo.2014.10.008. [47]

Alan, Sule, Seda Ertac, and Inci Gumus. 2021. "Does a Forward-Looking Perspective Affect Self-Control and the Demand for Commitment? Results From an Educational Intervention." *Economic Inquiry* 59 (4): 1533–46. DOI: 10.1111/ecin.13001. [44, 47]

Allcott, Hunt, Matthew Gentzkow, and Lena Song. 2022. "Digital Addiction." *American Economic Review* 112 (7): 2424–63. DOI: 10.1257/aer.20210867. [47]

Anderberg, Dan, Claudia Cerrone, and Arnaud Chevalier. 2018. "Soft Commitment: A Study on Demand and Compliance." *Applied Economics Letters* 25 (16): 1140–46. DOI: 10.1080/13504851.2017.1400648. [47]

Andreoni, James, Deniz Aydin, Blake Barton, B. Douglas Bernheim, and Jeffrey Naecker. 2020. "When Fair Isn't Fair: Understanding Choice Reversals Involving Social Preferences." *Journal of Political Economy* 128 (5): 1673–711. DOI: 10.1086/705549. [47]

Andreoni, James, and Marta Serra-Garcia. 2021. "Time Inconsistent Charitable Giving." *Journal of Public Economics* 198 (6): 104391. DOI: 10.1016/j.jpubeco.2021.104391. [47]

Ariely, Dan, and Klaus Wertenbroch. 2002. "Procrastination, Deadlines, and Performance: Self-Control by Precommitment." *Psychological Science* 13 (3): 219–24. DOI: 10.1111/1467-9280.00441. [43]

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines." *Quarterly Journal of Economics* 121 (2): 635–72. DOI: 10.1162/qjec.2006.121.2.635. [43, 47]

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. "Working over Time: Dynamic Inconsistency in Real Effort Tasks*." *Quarterly Journal of Economics* 130 (3): 1067–115. DOI: 10.1093/qje/qjv020. [44, 45, 47]

Avery, Mallory L., Osea Giuntella, and Peiran Jiao. 2022. "Why Dont We Sleep Enough A Field Experiment Among College Students." NBER Working Paper w30375. National Bureau of Economic Research. DOI: 10.3386/w30375. [47]

Bai, Liang, Benjamin Handel, Edward Miguel, and Gautam Rao. 2021. "Self-Control and Demand for Preventive Health: Evidence from Hypertension in India." *Review of Economics and Statistics* 103 (5): 835–56. DOI: 10.1162/rest_a_00938. [47]

Barton, Blake. 2015. "Interpersonal Time Inconsistency and Commitment." Unpublished manuscript. [47]

Beshears, John, James J. Choi, Christopher Harris, David Laibson, Brigitte C. Madrian, and Jung Sakong. 2020. "Which Early Withdrawal Penalty Attracts the Most Deposits to a Commitment Savings Account?" *Journal of Public Economics* 183 (3): 104144. DOI: 10.1016/j.jpubeco.2020.104144. [47]

Bettega, Paul, Paolo Crosetto, Dimitri Dubois, and Rustam Romaniuc. 2023. "Hard vs. Soft Commitments: Experimental Evidence From a Sample of French Gamblers." Working Paper 05/2023. GAEL. eprint: https://gael.univ-grenoble-alpes.fr/sites/gael/files/doc-recherche/WP/A2023/gael2023-05.pdf. [47]

Bhatia, Sudeep, Megan M. Crawford, Rebecca Louise McDonald, Miguel A. Moreno, and Daniel Read. 2021. "Inconsistent Planning and the Allocation of Tasks Over Time." Preprint. Open Science Framework. DOI: 10.31219/osf.io/b4mg7. [47]

Bhattacharya, Jay, Alan M. Garber, and Jeremy D. Goldhaber-Fiebert. 2015. "Nudges in Exercise Commitment Contracts: A Randomized Trial." NBER Working Paper w21406. National Bureau of Economic Research. DOI: 10.3386/w21406. [47]

Bisin, Alberto, and Kyle Hyndman. 2020. "Present-Bias, Procrastination and Deadlines in a Field Experiment." *Games and Economic Behavior* 119 (1): 339–57. DOI: 10.1016/j.geb.2019.11.010. [47]

Bonein, Aurélie, and Laurent Denant-Boèmont. 2015. "Self-Control, Commitment and Peer Pressure: A Laboratory Experiment." *Experimental Economics* 18 (4): 543–68. DOI: 10.1007/s10683-014-9419-7. [47]

Breig, Zachary, Matthew Gibson, and Jeffrey G Shrader. 2020. "Why Do We Procrastinate? Present Bias and Optimism." IZA DP 13060. eprint: https://docs.iza.org/dp13060.pdf. [44]

Breig, Zachary, Matthew Gibson, and Jeffrey G. Shrader. 2023. "Why Do We Procrastinate? Present Bias and Optimism." Working Paper. eprint: https://jeffreyshrader.com/papers/present_bias_and_optimism.pdf. [47]

Brune, Lasse, Eric Chyn, and Jason Kerwin. 2021. "Pay Me Later: Savings Constraints and the Demand for Deferred Payments." *American Economic Review* 111 (7): 2179–212. DOI: 10.1257/aer.20191657. [47]

Brune, Lasse, Xavier Giné, Jessica Goldberg, and Dean Yang. 2016. "Facilitating Savings for Agriculture: Field Experimental Evidence from Malawi." *Economic Development and Cultural Change* 64 (2): 187–220. DOI: 10.1086/684014. [44, 47]

Bryan, Gharad, Dean Karlan, and Scott Nelson. 2010. "Commitment Devices." *Annual Review of Economics* 2 (1): 671–98. DOI: 10.1146/annurev.economics.102308.124324. [43]

Carrera, Mariana, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky. 2022. "Who Chooses Commitment? Evidence and Welfare Implications." *Review of Economic Studies* 89 (3): 1205–44. DOI: 10.1093/restud/rdab056. [43, 47]

Casaburi, Lorenzo, and Rocco Macchiavello. 2019. "Demand and Supply of Infrequent Payments as a Commitment Device: Evidence from Kenya." *American Economic Review* 109 (2): 523–55. DOI: 10.1257/aer.20180281. [47]

Casari, Marco. 2009. "Pre-Commitment and Flexibility in a Time Decision Experiment." *Journal of Risk and Uncertainty* 38 (2): 117–41. DOI: 10.1007/s11166-009-9061-5. [47]

Chow, Vinci Y. C. 2011. "Demand for a Commitment Device in Online Gaming." eprint: https://www.ticoneva.com/econ/jobmarket/wow.pdf. [45, 47]

Dupas, Pascaline, and Jonathan Robinson. 2013. "Why Don't the Poor Save More? Evidence from Health Savings Experiments." *American Economic Review* 103 (4): 1138–71. DOI: 10.1257/aer.103.4.1138. [47]

Dykstra, Holly. 2020. "Patience Across the Payday Cycle." eprint: https://scholar.harvard.edu/files/holly-dykstra/files/payday.pdf. [47]

Ek, Claes, and Margaret Samahita. 2023. "Too Much Commitment? An Online Experiment with Tempting YouTube Content." *Journal of Economic Behavior & Organization* 208 (4): 21–38. DOI: 10.1016/j.jebo.2023.01.019. [47]

Erev, Ido, Maximilian Hiller, Stefan Klößner, Gal Lifshitz, Vanessa Mertins, and Yefim Roth. 2022. "Promoting Healthy Behavior Through Repeated Deposit Contracts: An Intervention Study." *Journal of Economic Psychology* 92 (10): 102548. DOI: 10.1016/j.joep.2022.102548. [45, 47]

Exley, Christine L., and Jeffrey K. Naecker. 2017. "Observability Increases the Demand for Commitment Devices." *Management Science* 63 (10): 3262–67. DOI: 10.1287/mnsc.2016.2501. [47]

Francis, Eilin L. 2018. "Paying to Repay? Experimental Evidence on Repayment Commitment." eprint: http://barrett.dyson.cornell.edu/NEUDC/paper_504.pdf. [47]

Giné, Xavier, Dean Karlan, and Jonathan Zinman. 2010. "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation." *American Economic Journal: Applied Economics* 2 (4): 213–35. DOI: 10.1257/app.2.4.213. [47]

Goldhaber-Fiebert, Jeremy, Erik Blumenkranz, and Alan Garber. 2010. "Committing to Exercise: Contract Design for Virtuous Habit Formation." NBER Working Paper w16624. National Bureau of Economic Research. DOI: 10.3386/w16624. [47]

Houser, Daniel, Daniel Schunk, Joachim Winter, and Erte Xiao. 2018. “Temptation and Commitment in the Laboratory.” *Games and Economic Behavior* 107 (1): 329–44. DOI: 10.1016/j.geb.2017.10.025. [47]

John, Anett. 2020. “When Commitment Fails: Evidence from a Field Experiment.” *Management Science* 66 (2): 503–29. DOI: 10.1287/mnsc.2018.3236. [43, 47]

Karlan, Dean, and Leigh L. Linden. 2022. “Loose Knots Strong Versus Weak Commitments to Save for Education in Uganda.” Working Paper. eprint: https://www.povertyactionlab.org/sites/default/files/research-paper/WP1730_Loose-Knots-in-Uganda_Karlan-et-al-Dec22.pdf. [47]

Karlan, Dean, and Jonathan Zinman. 2018. “Price and Control Elasticities of Demand for Savings.” *Journal of Development Economics* 130 (1): 145–59. DOI: 10.1016/j.jdeveco.2017.10.004. [47]

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan. 2015. “Self-Control at Work.” *Journal of Political Economy* 123 (6): 1227–77. DOI: 10.1086/683822. [45, 47]

Krawczyk, Michal, and Lukasz Patryk Wozny. 2017. “An Experiment on Temptation and Attitude Towards Paternalism.” *SSRN Electronic Journal*, DOI: 10.2139/ssrn.2912427. [44, 47]

Krügel, Sebastian, and Matthias Uhl. 2023. “Is Only One of My Selves Authentic? An Empirical Approach.” *Journal of Behavioral and Experimental Economics* 102 (2): 101971. DOI: 10.1016/j.socec.2022.101971. [47]

Le Cotty, T, E Maître d’Hôtel, R Soubeyran, and J Subervie. 2019. “Inventory Credit as a Commitment Device to Save Grain Until the Hunger Season.” *American Journal of Agricultural Economics* 101 (4): 1115–39. DOI: 10.1093/ajae/aaz009. [47]

Lichand, Guilherme, and Juliette Thibaud. 2023. “Parent-Bias.” *SSRN Electronic Journal*, DOI: 10.2139/ssrn.3737685. [47]

McIntosh, Craig, Isaac Meza, Joyce Sadka, and Enrique Seira. 2021. “The Limits of Self-Commitment and Private Paternalism.” eprint: https://isaacmeza.github.io/personal/files/donde.pdf. [47]

Palacios-Huerta, Ignacio, and Oscar Volij. 2021. “Temptation and Stochastic Preferences.” *SSRN Electronic Journal*, DOI: 10.2139/ssrn.3868041. [47]

Roy-Chowdhury, Vivek. 2023. “Dynamic Consistency in Intrinsic Information Preferences.” DOI: 10.1257/rct.10865-2.0. [45, 47]

Royer, Heather, Mark Stehr, and Justin 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 (3): 51–84. DOI: 10.1257/app.20130327. [47]

Sadoff, Sally, and Anya Samek. 2019. “Can Interventions Affect Commitment Demand? A Field Experiment on Food Choice.” *Journal of Economic Behavior & Organization* 158 (2): 90–109. DOI: 10.1016/j.jebo.2018.11.016. [47]

Sadoff, Sally, Anya Samek, and Charles Sprenger. 2020. “Dynamic Inconsistency in Food Choice: Experimental Evidence from Two Food Deserts.” *Review of Economic Studies* 87 (4): 1954–88. DOI: 10.1093/restud/rdz030. [47]

Schilbach, Frank. 2019. “Alcohol and Self-Control: A Field Experiment in India.” *American Economic Review* 109 (4): 1290–322. DOI: 10.1257/aer.20170458. [43, 47]

Sjåstad, Hallgeir, and Mathias Ekström. 2022. “Ulyssean Self-Control: Pre-Commitment Is Effective, but Choosing It Freely Requires Good Self-Control.” Preprint. PsyArXiv. DOI: 10.31234/osf.io/w24eb. [47]

Toussaert, Séverine. 2018. “Eliciting Temptation and Self-Control Through Menu Choices: A Lab Experiment.” *Econometrica* 86 (3): 859–89. DOI: 10.3982/ECTA14172. [45, 47]

Toussaert, Séverine. 2019. “Revealing Temptation Through Menu Choice: Field Evidence.” eprint: https://severinetoussaert.com/wp-content/uploads/2019/02/LS-2014-v7.pdf. [47]

Zhang, Qing, and Ben Greiner. 2021. “Time Inconsistency, Sophistication, and Commitment: An Experimental Study.” *Economics Letters* 206 (9): 109982. DOI: 10.1016/j.econlet.2021.109982. [44, 47]