Probabilistic Beliefs, Choice Reversals, and Commitment*

Alexander Coutts† Séverine Toussaert‡

July 2026

Abstract

Experimental evidence of dynamic inconsistency is often interpreted as indicating widespread naivety, based on point predictions of future choices. In the canonical effort paradigm, we revisit this interpretation by eliciting full probabilistic beliefs to study sophistication and menu demand. While point forecasts replicate apparent naivety, probabilistic beliefs reveal substantial uncertainty; accounting for it bridges over half of the belief–behavior gap implied by point predictions. At the same time, this approach uncovers that naivety is less widespread but more sharply concentrated: 40% of reversals arise from misspecified respondents, who account for 21% of the sample and assign zero probability to the effort they ultimately choose. This measurement shift also reshapes our understanding of commitment demand. While flexibility choices track anticipated behavior, commitment demand does not: almost all removed options are assigned zero probability, a pattern inconsistent with purely consequentialist models such as the standard present-bias framework, even in stochastic form.

Keywords: prediction, uncertainty, naivety, reversal, commitment, flexibility.

*This document builds on a complete set of pre-registered analyses: AEARCTR-0015407. Ethics approval was secured from the University of Oxford (ECONCIA20-21-22) and York University (2025-035). We are extremely grateful to Yves Le Yaouanq and Dmitry Taubinsky for their comments on the manuscript. We also thank Ted Bergstrom, Kim Sau Chung, Ignacio Esponda, Erik Eyster, Lorenz Goette, Benson Tsz Kin Leung, Xiaoyue Shan, Sevgi Yuksel, and Songfa Zhong, as well as seminar and conference participants at the Barcelona Summer Forum ("Preferences, Bounded Rationality, and Strategic Interactions", 2026), Canadian Economic Association meetings (Vancouver, 2026), ESA Asia/Pacific Meeting (Osaka, 2025), Hong Kong Baptist University, Loyola Marymount University, Nanyang Technological University, National University of Singapore, New York University, Southwest Economic Theory Conference (Los Angeles, 2026), Southwestern University of Finance and Economics, the University of California, Santa Barbara, the University of California, San Diego, and the Virtual Asian Meeting of Behavioral and Experimental Research, for helpful comments and suggestions. We gratefully acknowledge expert research assistance from Abdessamad Ait Mbarek, Axel La Pira, Stefania Merone, Hamid Tahir, Matthew Tan, and Le Wu. Toussaert gratefully acknowledges funding from the Economic and Social Research Council (ES/V003461/1); Coutts gratefully acknowledges funding from the Schulich School of Business, York University.

†Schulich School of Business, York University; E-mail: acoutts@schulich.yorku.ca

‡University of Oxford; E-mail: severine.toussaert@economics.ox.ac.uk

1 Introduction

Across a wide range of domains, including gym attendance, effort provision, savings, and dietary choice (DellaVigna and Malmendier, 2006; Augenblick and Rabin, 2019; John, 2020; Sadoff, Samek and Sprenger, 2020), individuals often behave differently from what they initially predict or plan, e.g., attending the gym far less than intended, and make limited use of commitment devices that might help them follow through. These patterns are commonly taken as evidence of substantial naivety and are consistent with a broader tendency toward overconfidence and misprediction documented across many economically consequential settings (Grubb, 2009; Oster, Shoulson and Dorsey, 2013; Huffman, Raymond and Shvets, 2022; Orhun, Cohn and Raymond, 2024).

A common thread in this literature is how naivety is measured. In experimental studies of dynamic inconsistency, the standard approach compares a point prediction of future behavior with the subsequent choice. Consider a gym-goer who plans to attend the gym seven times over a week, predicts that they will do so, but ultimately attends only twice. The discrepancy between the prediction and realization is classified as evidence of naivety. By this metric, naivety appears pervasive: structural estimates of perceived present bias often suggest that individuals view themselves as much closer to being time-consistent than they truly are (Augenblick and Rabin, 2019; Carrera et al., 2022; Fedyk, 2025).[1]

Yet a point prediction fully characterizes subjective beliefs only under certainty. The same gym-goer might perceive a 40% chance of attending seven times, more than any other single outcome, and a 60% chance of attending less, including some chance of attending twice. If asked for one prediction, they may naturally report their current plan and the most likely individual outcome, seven, even though they perceive a downward reversal to be more likely overall. If they end up going only twice, this single event by itself does not mean that their probabilistic beliefs were wrong, as long as attending twice was assigned positive probability. To assess calibration, the analyst must instead compare reported probabilities with realized frequencies, or mean predictions with average behavior. By contrast, average modal predictions need not equal average behavior even when beliefs are perfectly calibrated, creating *apparent naivety*, analogous to the apparent

[1] In this paper, we use the terms "reversal" and "time-consistent choice" descriptively, not as statements about underlying preferences; in particular, a time-consistent *choice* need not reflect time-consistent *preferences*. When linking our results to established theory, we retain the conventional term *present bias* (O'Donoghue and Rabin, 1999) for continuity with that literature, without implying a normative judgment; Bernheim and Taubinsky (2018) instead use the more neutral, descriptive term *present focus*.

overconfidence identified by Benoît and Dubra (2011).

A deterministic interpretation of point predictions implies that any subsequent deviation lies outside the support of the agent's beliefs. This assumption of *misspecified* beliefs, widely used in models of time inconsistency (O'Donoghue and Rabin, 1999; DellaVigna and Malmendier, 2006), is not innocuous. A growing literature shows how misspecified subjective models shape behavior and equilibrium outcomes (Esponda and Pouzo, 2016; Bohren and Hauser, 2023), and how such beliefs may fail to converge to the truth even asymptotically (Bohren and Hauser, 2021; Frick, Iijima and Ishii, 2023). In models of self-control, Heidhues and Kőszegi (2009) show that whether beliefs assign positive probability to true preferences can have first-order welfare consequences. Relatedly, Spiegler (2011) demonstrates that, when firms can price discriminate, whether greater sophistication benefits consumers depends on the precise modeling of naivety. Yet whether individuals' beliefs are in fact misspecified, and for whom, remains largely untested.

We revisit the canonical framework for studying dynamic inconsistency in effort provision, the most studied non-monetary domain (Imai, Rutter and Camerer, 2021), by complementing point predictions with full probability distributions over future choices. We develop a framework linking these beliefs to preferences over future choice sets and realized behavior. We find that this shift in measurement has first-order consequences for what we conclude about naivety, for how we interpret demand for commitment devices, and for the class of models needed to organize the data.

What we do We conduct a pre-registered online experiment ($N = 1,025$) that follows standard protocols for studying time inconsistency. At t_1, participants choose how much work to complete at a future date t_2, seven days later. At t_2, they make the same decision for immediate completion; comparing the two choices classifies each participant as consistent or as reversing downward (choosing less work at t_2) or upward (choosing more). After their initial choice, we elicit both a point prediction and a full probability distribution over future work amounts. We elicit the distribution in two steps, with built-in consistency checks that enable us to assess the reliability of these reports. A treatment without belief elicitation allows us to test whether the procedure itself affects subsequent behavior. We use the elicited beliefs to study (i) the extent of subjective uncertainty about future choices; (ii) how beliefs relate to realized choice reversals; and (iii) how beliefs relate to demand for commitment and flexibility.

To study the third question, we allow participants to *customize* the menu available at t_2 by paying either to remove options, creating commitment, or to add options, creat-

ing flexibility. This allows us to observe both margins of menu demand within a unified framework. Our main outcome is choice from the full menu, while customization decisions reveal whether participants prefer to restrict or expand their future choice set in ways that accord with their beliefs. At t_2, they choose from the full, commitment, and flexibility menus, one of which is randomly selected for payment.

To interpret these customization choices, we develop a menu-choice framework that builds on Gul and Pesendorfer (2001).[2] Unlike standard present-bias models, the framework allows unchosen temptations to affect utility through self-control costs, while nesting consequentialist present-bias models as the limiting case in which temptation becomes overwhelming. We introduce state-dependent shocks, which provide an additional source of choice reversals and allow the same individual to value both commitment and flexibility. Observed beliefs then yield empirically tractable tests of whether menu demand is driven solely by its consequences for eventual choice.

What we find Using point predictions alone, we replicate the canonical inference: 78% of participants predict time consistency, yet only 53% follow through, with downward reversals more common than upward ones (34% vs. 13%). Probabilistic beliefs qualify this conclusion. Only 22% express certainty about future effort; the rest assign positive probability to deviations from their initial plan. Subjective uncertainty strongly predicts realized reversals, and accounting for it closes 61% of the belief-behavior gap implied by point predictions. Point predictions are therefore not a sufficient statistic for expectations.

Yet full distributions reveal a sharper form of naivety. About one in five participants assigns zero probability to the effort they ultimately choose, including 40% of those who reverse. This misspecification is highly concentrated among reversals, especially downward ones: only 3% of consistent participants are misspecified, versus 28% and 45% among upward and downward reversals, respectively. Excluding these participants, predicted and realized consistency rates nearly coincide. The pattern is systematic: misspecified participants report narrower belief distributions, appear to treat point predictions as targets rather than forecasts, and exhibit larger reversals when constraints bind.

Turning to menu preferences, we find a clear asymmetry between flexibility and commitment. Flexibility demand tracks beliefs: participants add options they expect to use, and uncertainty predicts demand for larger choice sets. Commitment demand, by contrast, is unrelated to subjective uncertainty or anticipated reversals. Indeed, 90% of participants

[2] As Fudenberg and Levine (2006) show, their dual-self model with linear self-control costs is isomorphic to the Gul–Pesendorfer framework.

who remove at least one option assign zero probability to *all* removed options. This pattern is inconsistent with purely consequentialist models, including stochastic present-bias models, in which options are removed only to affect eventual choices. Explaining these choices therefore requires some form of menu dependence, whether due to self-control costs (Gul and Pesendorfer, 2001) or cognitive costs from large choice sets (Maćkowiak, Matějka and Wiederholt, 2023; Dean, Ravindran and Stoye, 2025), the two main motives supported by our data. Consistent with this interpretation, removed options are almost never chosen when available, whereas added options are frequently chosen ex post. In our setting, demand for flexibility therefore provides a more reliable revealed-preference proxy for sophistication than commitment take-up.

To assess whether these patterns were anticipated, we collected forecasts from 58 academic experts. Experts correctly predicted more downward than upward reversals, but overestimated reversal rates and expected misspecification to be similar across reversal types, whereas it is sharply diagnostic in the data. They also anticipated largely consequentialist commitment demand, predicting that 33% of removed options would receive positive probability, compared with only 4% in the data. Finally, they expected belief elicitation to shift behavior, whereas we find no effect.

Related literature Our paper contributes to the literature on dynamic inconsistency in intertemporal choice, where a central question is whether individuals correctly anticipate their future behavior. Table 1 situates our design within this literature. The standard approach compares initial choices, point predictions—typically elicited as a mode, modal interval, or "best guess"—and realized behavior to classify agents as naive or sophisticated (Augenblick and Rabin, 2019; Cerrone et al., 2025; Fedyk, 2025). A complementary revealed-preference approach infers beliefs from menu choices (Kopylov, 2012; Ahn et al., 2019; Ahn, Iijima and Sarver, 2020). Existing designs typically observe only some of these elements. When probabilistic beliefs are elicited at all, the generally concern coarse events, such as meeting a target or crossing a threshold, and therefore reveal neither whether the realized choice lay outside subjective support nor whether customized options were expected to be chosen. We instead jointly observe option-level beliefs, commitment and flexibility preferences,[3] and realized choices. Within our framework, this permits direct tests of whether menu demand can be explained solely by its anticipated consequences for future choice.

[3] Toussaert (2019) jointly studies commitment and flexibility through menu choice, but does not elicit beliefs or observe choices from those menus.

Table 1: Comparison across Experiments: Beliefs and Menu Preferences

		Beliefs		Menu preferences		
Paper	Domain	Point	Full	Commit.		Flex.
Augenblick, Niederle and Sprenger (2015)	Effort / Lab + online	✗	✗		✓[a]	
Augenblick and Rabin (2019)	Effort / Lab + online	✓	✗	✗		✗
Carrera et al. (2022)	Gym / Field	✓	≈[b]	✓		✗
Cerrone et al. (2025)	Effort / Online	✓	✗	✗		✗
Dean and McNeill (2020)	Effort / Online	✗	✗		✓[c]	
Fedyk (2025)	Effort / Online	✓	✗	✗		✗
Le Yaouanq and Schwardmann (2022)	Effort / Lab + online	✗	≈[d]	✗		✗
Sadoff, Samek and Sprenger (2020)	Food / Field	✗	✗	✓		✗
Toussaert (2018)	Effort / Lab	≈[e]	✗		✓[a]	
This paper	Effort / Online	✓	✓	✓		✓

✓ = explicitly measured; ✗ = not measured; ≈ = partial or indirect measurement.
[a] Design reveals either commitment or flexibility preference, but not both for the same individual.
[b] Subjective probability of meeting specific attendance thresholds (wave 3 only); not a full distribution.
[c] As in note (a). Consequentialism for flexibility tested using realized choice frequencies.
[d] Probability for a binary outcome (high vs. low effort); not a distribution over multiple effort levels.
[e] Incentivized belief about others' choices; qualitative self-assessment for own choices.

A related challenge concerns identification. Static designs identify nonstationarity rather than dynamic inconsistency, since implicit risk or liquidity constraints can generate nonstationary yet time-consistent choices (Halevy, 2015). Revision designs, including ours, elicit future-dated choices and later permit revision, but unobserved state changes can still rationalize revisions under time consistency (Strack and Taubinsky, 2026), just as unobserved task-completion payoffs prevent identification of present bias from completion timing (Heidhues and Strack, 2021).[4] Monetary valuations provide a cardinal yardstick for identifying time preference parameters under state-independent utility in money (Augenblick and Rabin, 2019; Strack and Taubinsky, 2026). Our approach instead takes anticipated future choices as directly elicited objects. This reveals whether revisions were anticipated and, combined with commitment choices, yields sign restrictions on perceived temptation, though not its cardinal magnitude. We are agnostic about whether revisions or misspecification reflect present bias, exogenous shocks, optimism, or other forces.[5]

[4] Freeman (2021) identifies naivety and sophistication by varying available completion opportunities.
[5] Breig, Gibson and Shrader (2023) use information about past behavior to distinguish present bias from optimism about future demands.

Our design also contributes to the literature on subjective expectations and provides direct evidence on misspecified beliefs. Probabilistic belief elicitation has a long tradition in applied microeconomics (Manski, 2004; Delavande, 2014; Fuster and Zafar, 2023) and macroeconomics (Engelberg, Manski and Williams, 2009; D'Acunto and Weber, 2024), but individual-level evidence that realized outcomes fall outside the support of stated beliefs remains scarce. Zero-probability realizations have been documented in professional density forecasts of macroeconomic outcomes (Kenny, Kostka and Masera, 2014; Diebold, Shin and Zhang, 2023) and in expectations of job loss and longevity (Stephens, 2004; Bago d'Uva, O'Donnell and Van Doorslaer, 2020), while related laboratory work examines how beliefs adjust when unforeseen outcomes materialize (Becker et al., 2026). In our setting, individuals have direct control over the outcome: we examine whether they assign positive probability to their *own* future action. We find that misspecification is both prevalent and systematic, lending empirical support to models in which the truth may lie outside subjective support.

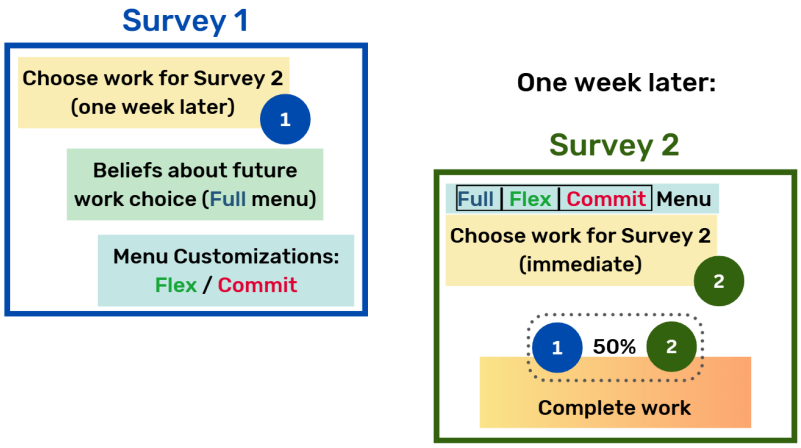
Finally, we contribute to the literature on commitment versus flexibility (Laibson, 2015; Toussaert, 2018, 2019; Le Lec and Tarroux, 2020; Sadoff, Samek and Sprenger, 2020; Dean and McNeill, 2020; Carrera et al., 2022). Existing designs typically observe only one of these two margins. By eliciting both margins for the same individual and linking them to anticipated and realized reversals, we show that commitment and flexibility demand are correlated but have different behavioral content. Demand for flexibility strongly predicts anticipated reversals, whereas commitment demand is unrelated to them. Commitment demand observed in isolation may therefore be mistaken for evidence of anticipated present bias. Our fine-grained customization design sharpens this distinction by allowing participants to add or remove any option at a positive cost and matching each adjustment to option-level beliefs. Consequentialism largely holds for flexibility but fails for commitment: participants remove options they assign zero probability to choosing and almost never choose when available, rejecting even stochastic present-bias models and making flexibility the more informative measure of sophistication. These distinctions cannot be detected in coarser menu designs (e.g., Toussaert, 2018).

The rest of the paper proceeds as follows. Section 2 describes the experimental design. Section 3 introduces our conceptual framework. Section 4 presents the empirical results. Section 5 discusses mechanisms and broader implications before concluding.

2 Experimental Design

We conducted an online experiment following a standard protocol for studying time inconsistency. At time t_1, individuals chose from 10 discrete options how much work to complete at time t_2, exactly seven days later, as in Augenblick, Niederle and Sprenger (2015); Fedyk (2025). At t_2, they chose how much work to complete immediately. We augmented this design with two key components. First, we elicited both point predictions and full probabilistic beliefs over future work choices. Second, we studied preferences for commitment and flexibility by allowing individuals to *customize* the menu of options faced at t_2. Figure 1 provides a simplified overview of the design. We describe in turn the work task, the belief elicitation procedure, and the elicitation of preferences for commitment and flexibility; full instructions are provided on OSF at `https://osf.io/79x6v/`.[6]

Figure 1: Experimental design



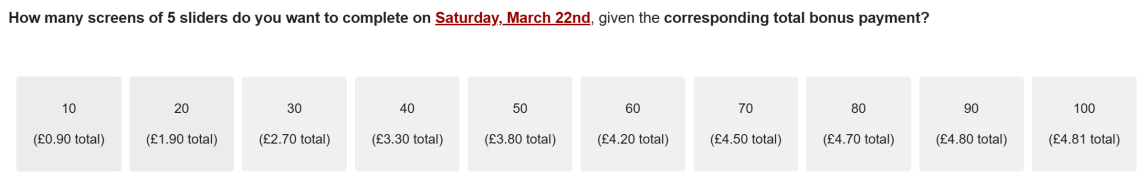
Notes: Overview of core design elements. In Survey 1, participants chose a work amount for t_2, entered beliefs and made two customizations to measure flexibility and commitment preferences. In Survey 2, they made their work decisions for all three menus, with one menu implemented at random (80% = full,10%,10%).

2.1 The effort task

The effort task was a "slider task", which involves manually dragging sliders to prespecified integers between 0 and 500 (Gill and Prowse, 2019). Participants chose how many screens of 5 sliders to complete for a fixed payment. Options ranged from 10 to 100 screens in multiples of 10, equating 50 to 500 individual sliders. Figure 2 shows the available options. We denote a generic effort choice by $e \in \mathcal{E} := \{10, 20, ..., 100\}$, with associated monetary payment m_e. For instance, completing $e = 40$ screens earned £3.30.

[6]Survey 1 can be accessed via the live Qualtrics link here: Survey 1.

Figure 2: Effort choices



Notes: Decision screen at t_1 showing effort options (number of screens and payment) for t_2.

We chose a discrete choice set to reduce the complexity of the instructions and belief elicitation. Payments were chosen to exhibit decreasing marginal returns, shifting choices toward interior solutions (Le Yaouanq and Schwardmann, 2022; Cerrone et al., 2025). Before making their choice, participants completed two practice screens to familiarize themselves with the task and were subsequently informed of the average time (in seconds) it took them to complete one screen. They were then shown a calculator allowing them to input their expected time per screen and obtain a (linear) projection of total time and hourly wage for each effort level, with inputs unrestricted, ensuring that the consequences of their choices were transparent.

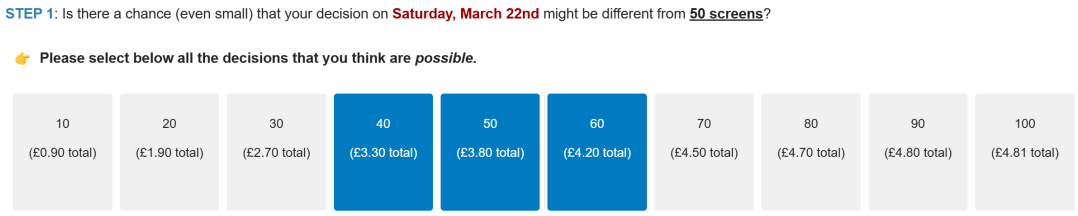
2.2 Beliefs about future effort

After making their effort choices, participants were asked to report a point prediction of the number of screens they would complete at t_2, by selecting one of the 10 options; we denote this prediction by $\hat{e}_2 \in \mathcal{E}$. This mirrors standard elicitation practices in the literature. Because we next elicit full belief distributions, we are able to infer which summary statistic participants have in mind (e.g., mean, median, or mode) when providing their point predictions, allowing us to measure the *apparent naivety* referenced in the introduction.[7]

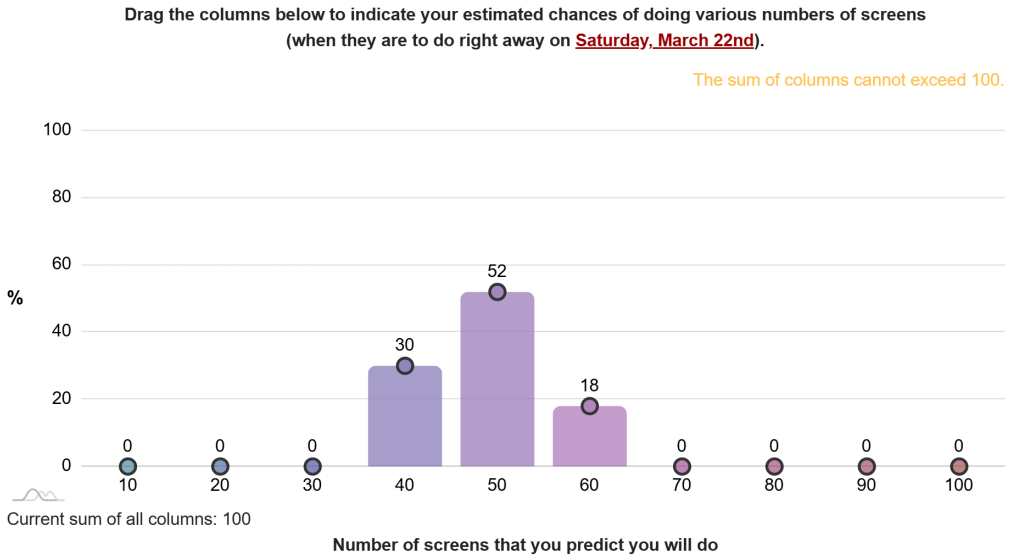
We then elicited a full belief distribution over future effort in two steps. In Step 1 (Figure 3a), participants selected all effort levels they considered *possible*. In Step 2 (Figure 3b), they allocated 100 probability points across effort levels using an interactive histogram, starting from the most likely outcomes. The two steps were not mechanically linked: participants could assign zero probability to options selected in Step 1 or positive probability to options not selected. Such inconsistencies triggered a warning message but were not constrained. We exploit these discrepancies as a diagnostic of measurement error.

[7]This necessitates that we do not incentivize point predictions, noting that prior work finds little effect of incentives in this context (Augenblick and Rabin, 2019; Fedyk, 2025). The elicitation language thus does not impose a specific summary statistic: "When the date of $[t = 2]$ arrives, what do you predict is the

Figure 3: Elicitation of belief distribution



(a) Step 1: Choice of support



(b) Step 2: Full distribution

The belief elicitation was incentivized using a Binarized Scoring Rule (BSR; Hossain and Okui, 2013). Unlike point predictions, we provided a small incentive to encourage engagement with the more complex task.[8] To reduce measurement error, participants were shown a summary table of their reported probabilities and could revise their answers before submission. We denote by $\hat{f}(e) \in [0, 1]$ the probability assigned to effort level e.

2.3 Preferences for commitment and flexibility

After eliciting beliefs, participants were offered to customize their menu of effort choices at t_2. Specifically, they were told that with 80% chance their choice would be drawn from

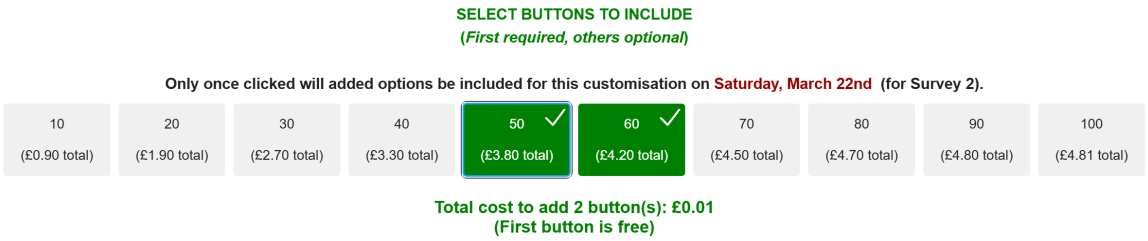
number of screens you will choose to do right away?".

[8]Following Danz, Vesterlund and Wilson (2022), who show that salient information about the elicitation mechanism can induce distortions, full details were not displayed by default but were available on demand. The BSR was applied to one randomly selected column, with a payment of £0.20 based on accuracy. As participants had control over the outcome, stakes were kept low to limit potential distortions.

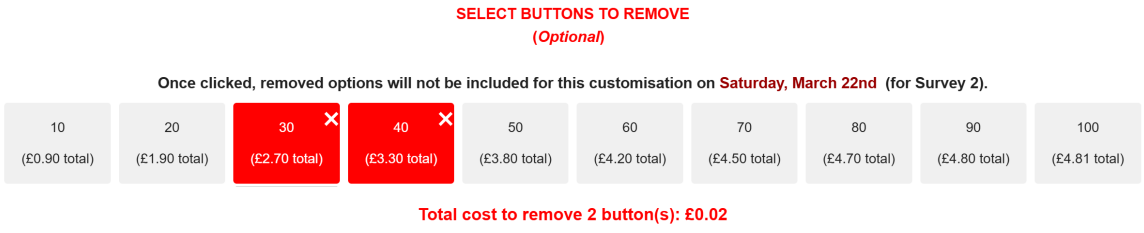
the full menu $\mathcal{E} = \{10, 20, ..., 100\}$, and with 20% chance from one of two menus they could customize (10% chance each). The first menu, $\mathcal{E}^+$, elicited preferences for flexibility. Participants started from an empty menu and could add options at a cost of £0.01 per option. They had to include at least one option and the first option was free. An example is shown in Figure 4a, where $\mathcal{E}^+ = \{50, 60\}$. The second menu, $\mathcal{E}^-$, elicited preferences for commitment. Participants started from the full menu $\mathcal{E}$ and could remove options at a cost of £0.01 per option, but had to leave at least one option available. An example is shown in Figure 4b, where $\mathcal{E}^- = \mathcal{E} \setminus \{30, 40\}$. Participants were prompted to confirm their decision if they removed or failed to include their previously chosen effort at t_1.

We elicit two menu customizations with a per-option cost rather than a single preferred menu at a fixed fee because preferred menus are often not unique. Under standard consequentialist preferences, once a menu contains all options a participant expects to choose, additional options are irrelevant. Whether the redundant options are included or excluded then depends on the tie-breaking rule: some participants might keep only the options they expect to use, while others might also keep options that are harmless. Charging for each addition or removal allows to break these ties: participants pay to change the menu only when the change matters to them. This provides a conservative measure of

Figure 4: Preferences for commitment and flexibility



(a) Menu $\mathcal{E}^+$: Preferences for flexibility



(b) Menu $\mathcal{E}^-$: Preferences for commitment

Notes: Screenshot showing menu customization decisions at t_1 (adding or removing options). In this example, $\mathcal{E}^+ = \{50, 60\}$ and $\mathcal{E}^- = \mathcal{E} \setminus \{30, 40\}$. Order of the two customizations was randomized.

demand for flexibility and commitment and allows us to examine whether participants both add and remove the same option, a useful coherence diagnostic.[9]

At the end of Survey 1, we probed the motives underlying flexibility and commitment decisions. Participants who reported certainty but added multiple options, or who removed options assigned zero probability, were asked to explain these decisions. We further flagged potential inconsistencies (e.g., both adding and removing the same option) and asked participants to account for them. Participants who expressed uncertainty were asked to report its sources. Finally, we collected information on expected availability for Survey 2, self-reported effort in customization decisions, and perceived task difficulty.

2.4 Decisions in Survey 2

Survey 2 took place one week after Survey 1 and could be started between 7:00 am and midnight GMT; participants were informed in advance and received up to three reminders. The survey began with a brief refresher: participants were reminded of their practice time and shown the same calculator as in Survey 1, but did not complete additional practice. They then made effort choices from each of the three menus, with the full menu $\mathcal{E}$ presented first and the customized menus $\mathcal{E}^+$ and $\mathcal{E}^-$ in randomized order, implementing the strategy method.[10] They were then asked whether earlier predictions influenced their decisions and to recall their point prediction.

Uncertainty was then resolved, with Survey 1 and Survey 2 decisions being implemented with equal probability. If Survey 2 was selected, one menu was drawn at random (80% $\mathcal{E}$, 10% $\mathcal{E}^+$, 10% $\mathcal{E}^-$), and the corresponding number of screens was completed.[11] Those who exhibited reversals reported their reasons and those who started at a different time than planned explained why.

[9] Adding and removing the same option is not necessarily inconsistent because the two tasks begin from different baseline menus (the full set for $\mathcal{E}^-$ and a singleton set for $\mathcal{E}^+$). For instance, an option could be valuable as a substitute or compromise in a sparse menu but become a costly distraction or temptation in a larger menu. Nevertheless, such cases require a sign reversal in the option's marginal value across menus and are therefore informative about either strong context dependence or decision noise.

[10] After choosing from $\mathcal{E}$, respondents reported decision difficulty. Our analysis focuses on choices from $\mathcal{E}$, which are comparable across subjects and allow us to study reversals between t_1 and t_2; we report some exploratory analyses leveraging choices from $\mathcal{E}^+$ and $\mathcal{E}^-$ in the discussion section and Appendix E.4.

[11] Participants had to complete all assigned screens for payment; the survey did not allow partial completion.

2.5 Treatment condition without beliefs

Our main treatment relates beliefs about future effort to (i) preferences for flexibility and commitment at t_1 and (ii) actual effort at t_2. Because predictions concern an event under participants' control, eliciting beliefs may itself affect subsequent behavior, e.g., by creating a reference point or by prompting individuals to reflect on their uncertainty. To assess this, we included a condition without belief elicitation. In this condition, participants in Survey 1 made only their effort choice for Survey 2 and their menu customization decisions. This allows us to test whether eliciting beliefs influences subsequent choices.

2.6 Sample size and recruitment

The experiment was conducted in April 2025 with 1,025 UK respondents, 722 in the main treatment and 303 in the no-belief condition, using the online platform Prolific.[12] Survey 1 included five mandatory comprehension questions; failure to answer any correctly within two attempts led to termination. Payment required completion of both surveys and was issued one week after Survey 2, including a show-up fee of £4. Consistent with our pre-registration, all analyses are restricted to participants who completed both surveys (see Appendix B.3 for sample size justification and attrition). The structure of the study is summarized in Table 2.

2.7 Expert forecasts

To benchmark our results, we collected forecasts from $N = 58$ academic experts on the Social Science Prediction Platform (`https://socialscienceprediction.org/`). Forecasters were Ph.D. candidates, postdoctoral researchers, or faculty members in economics or psychology who had not previously heard about the study. Data collection ran from February 27 to April 30, 2025—before and during our main data collection—and access to the forecast data was restricted throughout. The sample exceeds our pre-registered target of 30, set based on prior evidence on forecast aggregation (DellaVigna and Pope, 2018; Iacovone, McKenzie and Meager, 2025).

Forecasters were introduced to the experimental design and could experience the ac-

[12] Survey 1 ran from April 15 to 22, 2025, followed one week later by Survey 2 (April 22–29). We required a minimum Prolific approval rate of 99%, between 10 and 1,000 prior submissions, and enabled gender balancing. The no-belief condition was implemented via in-survey randomization. With advances in AI at the time of writing, we tested in February 2026 whether consumer-facing ChatGPT (GPT-5 series, agent-enabled) could complete the survey; it was unable to complete even a single screen of sliders, suggesting such systems could not meaningfully perform the task at the time.

Table 2: Structure of the study

Survey 1	**Introduction to effort task** Instructions, comprehension questions, and practice.
	Effort decision Choose number of screens for next week from $\{10, 20, \ldots, 100\}$.
	Predictions about Survey 2 effort Belief treatment ($N = 722$): point prediction and full distribution. No-Belief treatment ($N = 303$): no predictions.
	Menu customizations (order randomized) Flexibility: add options at £0.01 each (first free; ≥ 1 required). Commitment: remove options at £0.01 each (≥ 1 must remain).
	Follow-up questions
Survey 2 (one week later)	**Effort decisions** Choice from the full menu $\mathcal{E}$, flexibility menu $\mathcal{E}^+$, and commitment menu $\mathcal{E}^-$. Full menu first; order of flexibility and commitment menus randomized.
	Outcome and payment Survey 1 or 2 selected (50/50); if Survey 2, one menu drawn (80% full, 10% flex, 10% commit). Screens completed for payment.

tual decisions faced by participants before submitting their predictions. We collected 21 forecasts spanning all main quantities of interest: reversal rates, anticipated reversals (from both point predictions and belief distributions), belief uncertainty, misspecification rates, demand for commitment and flexibility, and the impact of belief elicitation on effort. The instructions are available on OSF at `https://osf.io/79x6v/`.[13] Accuracy was incentivized: one forecast was randomly selected, and the five closest forecasters each received £100 donated to a charity of their choice. The full list of forecasted quantities are compared with sample estimates in Table F.1 and Figure F.1; we highlight key comparisons in the discussion (Section 5.3).

3 Conceptual framework

3.1 Setup

Following models of menu choice (Dekel, Lipman and Rustichini, 2001; Gul and Pesendorfer, 2001), we consider a decision maker (DM henceforth) who must decide at t_1 on the

[13] The forecast survey can be accessed via the live Qualtrics link here: Forecasting survey.

effort option(s) $E \subseteq \mathcal{E}$ available to them at some future period t_2. The DM faces two considerations. First, uncertainty about future shocks $s \in S$ that affect the cost $c_s(e)$ of producing effort e. At t_1, the DM forms beliefs $\mathbf{p} \in \Delta(S)$ about these shocks. Subjective uncertainty creates a preference for flexibility (Kreps, 1979): the DM would prefer a larger set E at t_2 to tailor effort to the realized state. Second, the DM may face temptation at t_2 to complete a different number of tasks than planned at t_1. Temptation creates a preference for commitment, i.e., a smaller set E. We allow the DM to be tempted either by lower effort (at the expense of a lower payment) or by larger payments (requiring higher effort).

Utility function We assume that the DM's preferences at t_1 over the menu E admit the following representation:

$$W(E) = \sum_{s \in S} p_s \left[\max_{e \in E} [U_s(e) + V_s(e)] - \max_{e' \in E} V_s(e') \right]$$

The first term $U_s(e) = m_e - c_s(e)$ is the normative (free-of-temptation) utility of effort e under state s: the difference between the monetary payment and effort cost. Following models of costly self-control (Gul and Pesendorfer, 2001; Dekel, Lipman and Rustichini, 2009), the second term $V_s(e) = m_e - \gamma c_s(e)$ captures temptation, where $\gamma > 0$ may magnify the effort cost ($\gamma > 1$, temptation to do less) or weaken it ($\gamma < 1$, temptation to do more).[14] The difference $V_s(e) - \max_{e' \in E} V_s(e') < 0$ is the self-control cost of resisting e' when choosing e. For each $s \in S$, we assume that $c_s(e)$ is strictly increasing in e, that U_s, V_s, and $U_s + V_s$ each have a unique maximizer, and that there is no statewise-dominant option (i.e., choices are responsive to the state).

Effort at t_1 versus t_2 At t_1, the DM commits to an effort level $e_1 \in \mathcal{E}$ for t_2, before uncertainty is resolved, equivalently choosing the singleton menu $\{e_1\}$ that maximizes $W(\{e\}) = \sum_{s \in S} p_s U_s(e)$ (as self-control costs vanish for singleton menus). At t_2, the state s is realized and the DM chooses e_2 from the full set $\mathcal{E}$ to maximize $U_s(e) + V_s(e)$. We say that choice is *time consistent* if $e_2 = e_1$ and that a *choice reversal* occurs if $e_2 \neq e_1$; we distinguish *downward reversals* ($e_2 < e_1$) from *upward reversals* ($e_2 > e_1$). Reversals can arise from two sources: uncertainty ($|S| > 1$), which may render e_1 suboptimal ex post, and temptation, which distorts t_2 choices toward lower effort ($\gamma > 1$, downward reversals) or higher effort ($\gamma < 1$, upward reversals).

[14] As an alternative framing, one could assume $V_s(e) = \beta_m m_e - \beta_e c_s(e)$ where $\beta_m, \beta_e \in [0, 1]$ are present bias parameters; $\beta_m < 1, \beta_e = 1$ corresponds to $\gamma > 1$, while $\beta_m = 1, \beta_e < 1$ corresponds to $\gamma < 1$. For parsimony, we take γ as deterministic.

Menu preferences The DM's preferred menus $\mathcal{E}^+$ (flexibility) and $\mathcal{E}^-$ (commitment) depend on the interaction between uncertainty and temptation. Without temptation ($\gamma = 1$), the DM values only flexibility: $\mathcal{E}^+$ grows with the number of subjective states (Dekel, Lipman and Rustichini, 2001), and $\mathcal{E}^- = \mathcal{E}$. Without uncertainty ($|S| = 1$), the DM values only commitment: $\mathcal{E}^+ = \{e_1\}$ and options are removed from $\mathcal{E}^-$ to reduce self-control costs (Gul and Pesendorfer, 2001), with lower effort levels removed when $\gamma > 1$ and higher effort levels when $\gamma < 1$. When both forces are present ($|S| > 1$, $\gamma \neq 1$), the DM may simultaneously exhibit preference for flexibility ($|\mathcal{E}^+| > 1$) and commitment ($\mathcal{E}^- \subsetneq \mathcal{E}$). We provide a parametrized example illustrating these comparative statics in Appendix A.2.

3.2 Link between menu preferences and beliefs

We now articulate testable implications relating preferences over menus $\mathcal{E}^+$ and $\mathcal{E}^-$ to anticipated choice from the full menu, denoted $\hat{f}(e) = \hat{f}^{\mathcal{E}}(e)$.

Preference for flexibility and uncertainty In this model, the value of flexibility is purely instrumental: it allows the DM to condition effort on the realized state. There is no intrinsic value to choosing from a larger set, such as a preference for autonomy or decision rights (Sen, 1988; Bartling, Fehr and Herz, 2014; Ferreira, Hanaki and Tarroux, 2020). This creates a natural link between demand for flexibility and subjective uncertainty. If the DM expects different effort levels to be optimal in different states, access to multiple options becomes valuable (Ahn and Sarver, 2013; Dean and McNeill, 2020). We therefore expect a preference for flexibility ($|\mathcal{E}^+| > 1$) to be positively associated with non-degenerate beliefs ($|\operatorname{supp}(\hat{f})| > 1$) and anticipated reversals ($\hat{f}(e_1) \neq 1$). The converse is only partial: uncertainty need not generate demand for flexibility if customization costs are high enough or if larger menus create sufficiently strong temptation costs.

H1: *Preference for flexibility is positively associated with subjective uncertainty and anticipated choice reversals.*

Preference for commitment and uncertainty The relationship between uncertainty and commitment demand ($\mathcal{E}^- \subsetneq \mathcal{E}$) is more subtle. In purely consequentialist models such as models of present bias, commitment is only valuable if it can alter future behavior, and hence is naturally linked to anticipated reversals (Laibson, 1997; O'Donoghue and Rabin, 1999). In costly self-control models, however, the DM may also value commitment because removing tempting options lowers self-control costs, even when those options are not expected to be chosen and no reversals are expected (Gul and Pesendorfer, 2001;

Toussaert, 2018). The link between anticipated reversals and commitment demand should therefore be weaker than for flexibility. Moreover, uncertainty can increase or decrease commitment demand depending on how it interacts with temptation (see Appendix A.2).

H2: *Preference for commitment is positively associated with subjective uncertainty and anticipated choice reversals, albeit less strongly than preference for flexibility.*

We test H1 and H2 at the extensive margin, by comparing menu demand across participants with degenerate and non-degenerate beliefs, and at the intensive margin, by relating menu size to $|\operatorname{supp}(\hat{f})|$.

Option-level consequentialism and global tests We next ask whether the specific options added or removed are consequentialist relative to the DM's beliefs about their future choice. To this end, we define the following two properties:

$$\hat{c}\text{-FLEX:} \quad e \in \mathcal{E}^+ \Rightarrow \hat{f}^{\mathcal{E}}(e) > 0, \qquad \hat{c}\text{-COMMIT:} \quad e \notin \mathcal{E}^- \Rightarrow \hat{f}^{\mathcal{E}}(e) > 0.$$

Both conditions capture the same basic idea: options that the DM pays to add or remove should be relevant under their own beliefs, where relevance is defined as being assigned a positive probability of choice from the full menu $\mathcal{E}$.

Individual violations of these full-menu properties are not, however, sufficient to reject consequentialism. Indeed, an option that is irrelevant for choice from the full menu may become relevant once other options are absent, for instance as a compromise or as a temptation that is chosen only after a stronger temptation has been removed.[15] For this reason, the primitive consequentialist restriction formulated in the literature (Ahn and Sarver, 2013; Dean and McNeill, 2020) is local: for any menu E with $e \notin E$, a marginal addition or removal of e should require e to matter in the enlarged menu, i.e., if $E \cup \{e\} \succ E$ or $E \succ E \cup \{e\}$, then $\hat{f}^{E\cup\{e\}}(e) > 0$. Testing this restriction directly would be onerous, as it requires beliefs for every relevant counterfactual menu. We instead use the stronger full-menu diagnostic, replacing $\hat{f}^{E\cup\{e\}}$ with $\hat{f}^{\mathcal{E}}$, which provides an upper bound on the prevalence of consequentialism violations and a useful first-pass test.

Consequentialist commitment is rejected, however, if violations of $\hat{c}$-COMMIT are *systematic*, that is, if $\hat{f}^{\mathcal{E}}(e) = 0$ for *all* $e \notin \mathcal{E}^-$. In such a case, one is ensured that $\hat{f}^{\mathcal{E}^-\cup\{e\}}(e) =$

[15] For example, a DM who believes they would choose 30 and 50 with equal probability from the full menu may choose $\mathcal{E}^+ = \{40\}$ as a compromise. Similarly, a DM who expects to choose 100 from the full menu, so that $\hat{f}^{\mathcal{E}}(90) = 0$, may nevertheless expect 90 to be chosen once 100 is removed. See Appendix A.2 for an illustration and discussion.

0 also holds. To see this, notice that if $\hat{f}^{\mathcal{E}}(e) = 0$ for all $e \notin \mathcal{E}^-$, the restricted menu $\mathcal{E}^-$ still contains all the options that the DM expected to choose from the full menu under some state, i.e., $\text{supp}(\hat{f}^{\mathcal{E}}) \subseteq \mathcal{E}^-$. Since the date-2 selection rule does not depend on the menu, the same $U_s + V_s$-maximizers remain available and continue to be chosen in each anticipated state. Hence, adding back any removed option cannot change anticipated behavior, i.e., $\hat{f}^{\mathcal{E}^- \cup \{e\}}(e) = 0$. A violation of $\hat{c}$-COMMIT is therefore also a violation of the primitive menu-specific condition. Equivalently, paid restrictions of options assigned zero chance of being chosen have no value under purely consequentialist commitment; any benefit from the restriction must come from the menu itself, such as reduced self-control costs.

A sufficient condition for consequentialist commitment to be violated is thus that the restricted menu $\mathcal{E}^-$ preserves the support, $\text{supp}(\hat{f}^{\mathcal{E}}) \subseteq \mathcal{E}^-$.[16] For flexibility, a similar intuition holds, but with an additional requirement: if $\text{supp}(\hat{f}^{\mathcal{E}}) \subseteq \mathcal{E}^+$ *and* $\mathcal{E}^+ \setminus \text{supp}(\hat{f}^{\mathcal{E}}) \neq \emptyset$, then any included option outside the support cannot improve anticipated choice and can only add temptation or customization costs. Hence, it must be included for reasons outside the model, such as a preference for autonomy or noise.

Importantly, while we present our consequentialism tests within an extension of the Gul-Pesendorfer model, they do not lean on the specific functional form. They apply to any model in which the t_2 selection rule is a fixed, menu-independent function of the state, which implies that the Weak Axiom of Revealed Preference (WARP) is satisfied state by state. This class spans expected and non-expected utility, random utility, and present-bias models, even in stochastic form – provided in each case that realizations are independent of the menu presented. What our tests do not cover is models in which the selection rule directly depends on the composition of the menu or its size, e.g., self-control cost that scales with the salience or number of tempting alternatives present. We discuss evidence bearing on this class of violations in Section 5.

Empirically, we report both option-level violations and subject-level patterns across all customized options. We also use the fact that the full-menu commitment diagnostic coincides with the primitive local condition when only one option is removed, since then $\mathcal{E}^- \cup \{e\} = \mathcal{E}$. If apparent violations were mainly driven by options becoming relevant

[16] More generally, consequentialism violations can be identified whenever the removal of an option is *choice-preserving*; support preservation is a sufficient but not necessary condition. For example, suppose $\text{supp}(\hat{f}^{\mathcal{E}}) = \{30, 50\}$ and $\mathcal{E}^- = \mathcal{E} \setminus \{10, 50\}$. Removing 50 may change anticipated choice because it is a support option. However, under single-peaked preferences, the retained support option (30) prevents the off-support option (10) from being chosen. Hence, 10 remains irrelevant, and its costly removal violates consequentialism. This "shield" argument identifies additional violations and can also provide sign restrictions on the direction of temptation. See Appendix A.1.

only after other options are removed, they should be rare among one-option removals and increase with the number of removed options.

3.3 Link between beliefs and actual behavior

We now turn to the consistency between choice expectations $\hat{f}(e)$ and actual behavior $f(e)$, necessary to identify naivety.

Misspecification and calibration We allow for inaccurate beliefs, arising from misperceptions of either the state distribution ($\hat{\mathbf{p}} \neq \mathbf{p}$) or the temptation parameter γ, so that $\hat{f}(e)$ may differ from $f(e)$. We propose two classes of tests. First, we say beliefs are *misspecified* if $\hat{f}(e_2) = 0$, i.e., the DM assigns zero probability to the effort chosen. We assess misspecification at the individual level and its association with reversals. Second, we say beliefs are *miscalibrated* if $\hat{f}(e) \neq f(e)$: predicted and actual choice probabilities diverge. Even imperfectly calibrated beliefs may predict reversals. We test this by comparing reversal rates when participants anticipated possible reversal ($\hat{f}(e_1) \neq 1$) versus certainty ($\hat{f}(e_1) = 1$), and by comparing the distribution of $\hat{f}(e_1)$ across reversal types.

Netting out uncertainty The DM's beliefs could be a noisy signal of their future behavior for two reasons. First, uncertainty itself creates an information loss: even perfectly calibrated beliefs will not predict the realized choice with certainty. Second, beliefs might be misspecified or biased in a particular direction. To disentangle these forces, we introduce a "behavior drawn from beliefs" (BDB) benchmark in which each participant's t_2 effort is generated by sampling from their stated belief distribution $\hat{f}(\cdot)$, with reversals classified by comparing the sampled effort to e_1. Because $\hat{f}$ is discrete, the BDB quantities we report are computed exactly, in closed form.[17] The BDB benchmark thus provides a ceiling on how closely behavior can match beliefs: any gap between BDB predictions and actual behavior reflects miscalibration rather than uncertainty.

Menu preferences and actual behavior We also test realized counterparts of the two full-menu diagnostics:

$$c\text{-FLEX:} \quad e \in \mathcal{E}^+ \Rightarrow f^{\mathcal{E}}(e) > 0, \qquad c\text{-COMMIT:} \quad e \notin \mathcal{E}^- \Rightarrow f^{\mathcal{E}}(e) > 0.$$

[17] We depart from this only when reporting a range alongside a benchmark, which we obtain by simulation. The range indicates where the realized statistic would typically fall if behavior were drawn from beliefs. We draw one $e_2^r \sim \hat{f}(\cdot)$ per participant in $r = 1{,}000$ replications and report the 2.5^{th} and 97.5^{th} percentiles. These ranges reflect sampling variation under the BDB exercise and are not confidence intervals.

These tests ask whether options included in the flexibility menu, or removed from the commitment menu, are in fact chosen from the full menu. Because we observe only one realized choice per individual, these are aggregate relevance checks rather than individual-level revealed-preference tests. We implement them by calculating, among all individuals who included or excluded a given effort level e, the proportion who ultimately chose e from $\mathcal{E}$. These realized diagnostics either reinforce or qualify their belief-based counterparts: if options assigned zero probability are also rarely chosen, this confirms the case against consequentialism; if options assigned zero probability are often chosen ex post, this would point instead to misspecified beliefs.

4 Results

Below we present the main results for the 722 respondents in the belief treatment who completed both surveys.[18] Most analyses follow a detailed pre-analysis plan (PAP) developed using data from a pilot study conducted in November 2023. The pilot data were used to write the analysis code. Details on the piloting strategy and any deviations from the PAP are provided in Appendix B. The PAP replication code is available on OSF at `https://osf.io/79x6v/`.

We document four sets of findings. Section 4.1 examines the relationship between point predictions and choice reversals. Section 4.2 studies belief uncertainty and reversals. Section 4.3 analyzes residual discrepancies between beliefs and behavior. Finally, Section 4.4 links beliefs about future effort to preferences for commitment versus flexibility.

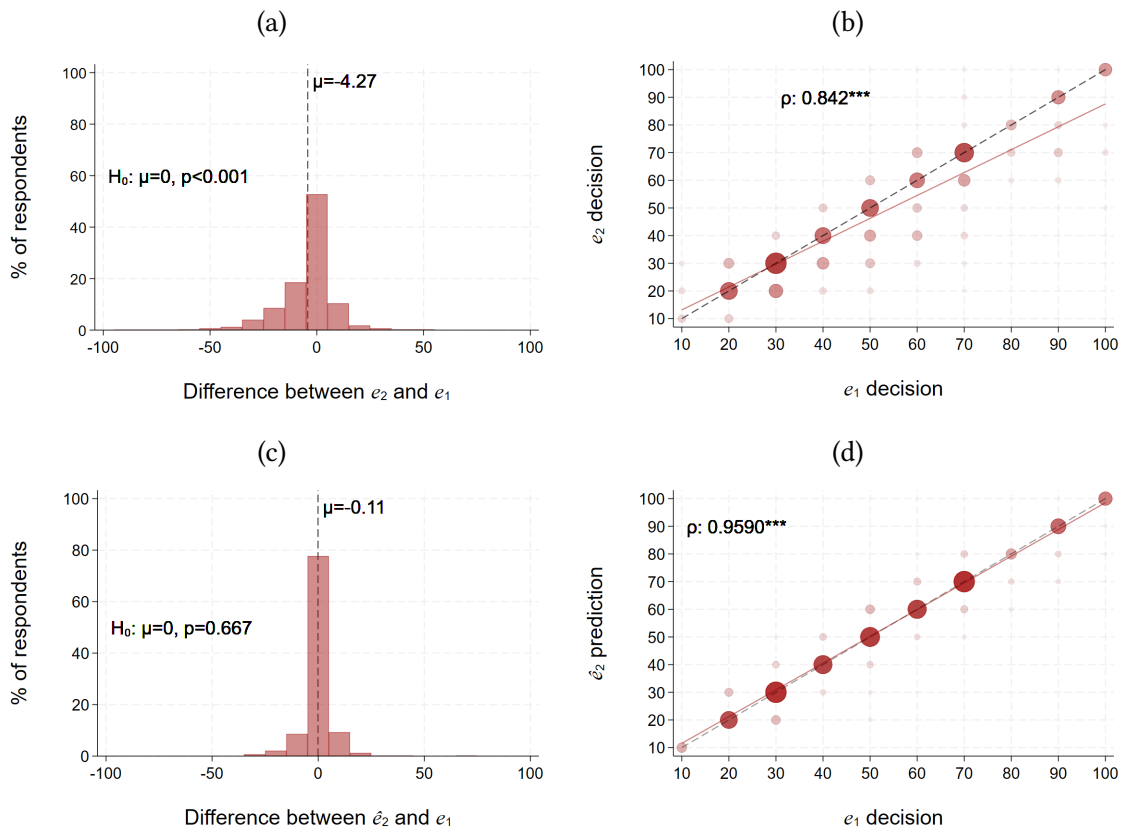
4.1 Inferences of Naivety under Point Predictions

We first examine the extent to which participants exhibit choice reversals between t_1 and t_2, and mispredict their future behavior when considering point predictions only.

Result 1. *Using point predictions, we replicate the canonical inference of near-complete naivety. Effort revisions are predominantly downward (34% versus 13% upward), yet 78% of participants forecast time-consistent behavior, even though only 53% ultimately choose e_1 at t_2. Point beliefs therefore imply a substantial upward bias in anticipated effort.*

[18] A total of 803 participants completed Survey 1, with an attrition rate of 10%. Table B.2 compares effort and beliefs between participants who did not return for Survey 2 and those who completed both surveys. Surveys 1 and 2 lasted on average 22 and 38 minutes, respectively. Participants earned £7.72 on average, including the £4 show-up fee; median hourly wage was £8.43.

Figure 5: Realized and Predicted Effort Relative to Initial Choices



Notes: Panels (a) and (c) show histograms of $e_2 - e_1$ and $\hat{e}_2 - e_1$, respectively, with a vertical bar for the mean and p-value from a paired t-test of a zero mean difference. Panels (b) and (d) plot e_2 and $\hat{e}_2$, respectively, against e_1, with Pearson correlations. Bubble sizes range from $N = 1$ to $N = 87$ respondents. $N = 722$.

As shown in Figure 5a, the mean number of tasks declines between the two weeks, from 53.0 at t_1 to 48.7 at t_2, a reduction of 4.3 (95% CI: $[3.3, 5.2]$). While most deviations are small in magnitude, they are more pronounced at the extremes of the effort distribution (Figure C.1). Figure 5b further shows that effort decisions display a high level of stability between t_1 and t_2 ($\rho = 0.842$). However, 47% of participants revise their initial choice and 72% of these deviations are below the 45° line, indicating a downward reversal.

Unlike realized behavior, predicted deviations from e_1 are centered around and statistically indistinguishable from zero (Figure 5c), indicating that predictions remain strongly anchored to initial plans. Accordingly, the near-perfect correlation between $\hat{e}_2$ and e_1 ($\rho = 0.959$; Figure 5d) exceeds both the correlation between e_2 and e_1 (0.842) and that between $\hat{e}_2$ and e_2 (0.827). Overall, 78% of participants predict time-consistent behavior. Predicted reversals are nearly symmetric (11% downward and 11% upward), whereas realized rever-

sals are predominantly downward (34% vs.13%). Point predictions therefore exceed actual effort e_2 by 4.2 on average (95% CI: [3.1, 5.2]), indicating a clear upward bias.

A natural question is how informative predicted reversals are about realized reversals. Figure C.2 cross-tabulates point predictions against outcomes (Downward, Consistent, Upward), and Table C.2 translates these frequencies into prior-to-posterior updates (see Panel A). Among participants who predicted time consistency, about 57% were indeed consistent; errors skew toward downward reversals (roughly 3:1 relative to upward). Accordingly, learning that a participant predicts consistency increases the posterior probability of actual consistency by only 4.2 percentage points relative to the 53% base rate (LR = 1.19). By contrast, when a reversal is anticipated, the prediction is substantially more informative: the posterior probability of the corresponding realized reversal increases by more than 20 percentage points (likelihood ratios of 2.6 for downward and 3.5 for upward reversals). Thus, most reversals are unanticipated, but the rare anticipated reversals are meaningfully diagnostic of behavior.

Taken together, these patterns replicate the standard inference in the literature. Effort declines on average between t_1 and t_2, and most participants forecast time-consistent behavior using point predictions. Viewed through this lens, reversals appear largely unanticipated, suggesting near-complete naivety as in Augenblick and Rabin (2019) and Fedyk (2025). In the next section, we ask whether this inference persists once participants' subjective uncertainty about future effort is taken into account.

4.2 Probabilistic Beliefs and Subjective Uncertainty

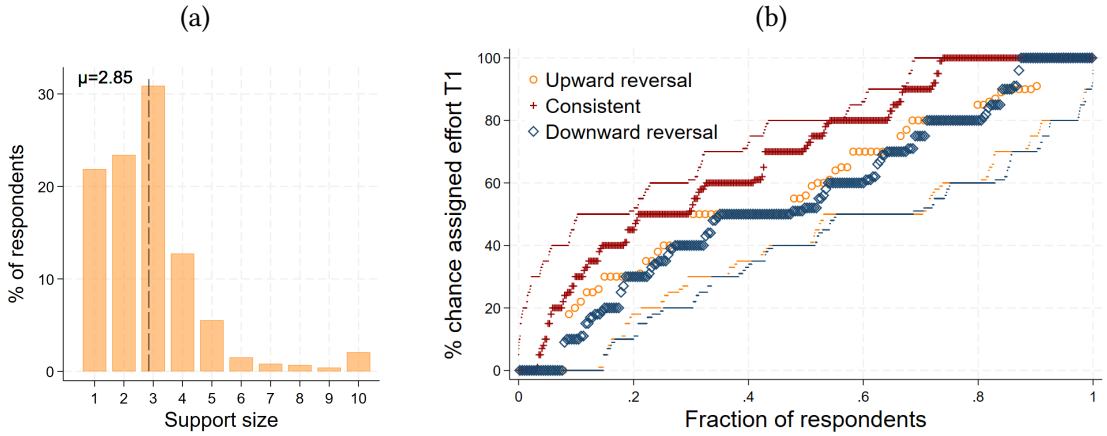
We next incorporate participants' probabilistic beliefs about their future effort and examine the role of subjective uncertainty in explaining choice reversals.[19]

Result 2. *Belief distributions reveal substantial uncertainty about future effort: when asked to form probabilistic beliefs, only 22% express certainty. This uncertainty predicts realized reversals and explains about 60% of the belief–behavior gap implied by point predictions. Point predictions are therefore not a sufficient statistic for expectations.*

Figure 6a reports how many effort levels participants deemed possible (i.e., $\hat{f}(e) > 0$). On average, they assign positive probability to 2.85 effort levels (median 3.0). Conditional on support, uncertainty is moderate: mean entropy, normalized by support size, is 0.30

[19]Unless otherwise noted, all belief-related analyses use the elicited belief distributions (Step 2); measurement considerations are discussed in Appendix D.1.

Figure 6: Belief uncertainty and predicted effort



Notes: (a) Histogram of belief support size $|\text{supp}(\hat{f})|$; the vertical line marks the mean. $N = 722$. (b) Quantile plots of the mass assigned to e_1, $\hat{f}(e_1)$, by reversal type. Thick markers show empirical distributions; thin markers the "behavior drawn from beliefs" (BDB) benchmark, computed in closed form from each participant's $\hat{f}(\cdot)$ (Section 3.3). $N = 244$ (downward), 381 (consistent), and 97 (upward) for empirical distributions.

(Table C.1). Among participants with degenerate beliefs ($|\text{supp}(\hat{f})| = 1$), 89% place all probability on their t_1 choice ($\hat{f}(e_1) = 1$), as predicted when temptation is perceived to be limited (i.e., for $\hat{\gamma}$ sufficiently close to 1; see Table A.1). On average, respondents place 64% probability on their point prediction $\hat{e}_2$, and 25% assign less than 50% probability to it. Importantly, non-degenerate beliefs imply that most participants anticipate the possibility of a reversal, even if they may miscalibrate its likelihood or magnitude.

Which feature of participants' subjective belief distributions do their point predictions capture? The point prediction $\hat{e}_2$ coincides with a mode for 85% of participants, with the median for 78%, and with the (rounded) mean for 69%. As noted in the introduction, this distinction matters because the average point prediction identifies the average subjective expectation only when participants report the means of their belief distributions. If some instead report a mode or median, a discrepancy between average point predictions and average realized effort may arise even when beliefs are correctly calibrated on average. In Section 5.1, we assess how much this *apparent naivety* affects classification and find that the reported statistic accounts for part, but not all, of the miscalibration we document.

We next test whether subjective uncertainty predicts actual reversals. Table 3 compares reversal rates for participants who expressed uncertainty about their future effort ($|\text{supp}(\hat{f})| > 1$) with those who did not ($|\text{supp}(\hat{f})| = 1$). Fewer than half of uncertain participants were time consistent in their choices, compared with about two thirds of

Table 3: Conditional probabilities of reversal

	Consistent	Upward reversal	Downward reversal
(1) No belief uncertainty	65.8%	8.9%	25.3%
BDB Benchmark	88.6%	5.7%	5.7%
(2) Belief uncertainty	49.1%	14.7%	36.2%
BDB Benchmark	55.2%	19.4%	25.5%
	[51.8, 58.9]	[16.7, 22.0]	[22.7, 28.5]
N	381	97	244
p**-value (1) = (2)**	< 0.001	0.064	0.010

Notes: Columns are $e_2 = e_1$, $e_2 > e_1$ and $e_2 < e_1$. Row (1) conditions on $|\text{supp}(\hat{f})| = 1$ and row (2) on $|\text{supp}(\hat{f})| > 1$. Exact p-values from proportions tests. The BDB benchmark is computed in closed form from $\hat{f}$ (see Section 3.3); for (1), beliefs are degenerate and the BDB benchmark is a point prediction. Brackets give the range in which the realized share would fall 95% of the time if behavior were drawn from beliefs, from 1,000 Monte Carlo replications; they are not confidence intervals.

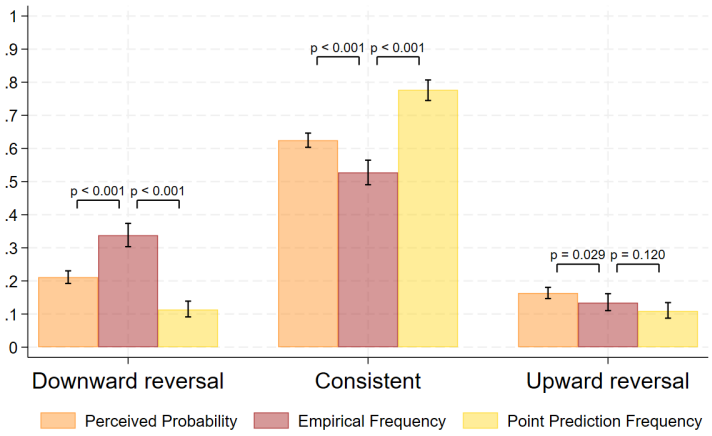
those without uncertainty (−16.7 pp; $p < 0.001$). Both upward and downward reversals are more frequent among uncertain participants, with no clear asymmetry.[20] Consistent with this pattern, Figure 6b compares $\hat{f}(e_1)$—the stated probability of choosing e_1—across reversal types and shows clear first-order stochastic dominance: time-consistent participants place more probability on e_1 ex ante (Kolmogorov–Smirnov test; $p < 0.001$). The point-biserial correlation between $\hat{f}(e_1)$ and time consistency at t_2 (indicator for $e_2 = e_1$) is 0.219 ($p < 0.001$), indicating that probabilistic beliefs predict reversal status.

Beliefs about reversals become substantially more accurate once subjective uncertainty is incorporated. Figure 7 compares actual reversal rates (Empirical Frequency, center bars) with predictions based on elicited belief distributions (Perceived Probability, left bars) and point predictions (Point Prediction Frequency, right bars). While point predictions substantially overstate time consistency, incorporating subjective uncertainty brings predicted and realized reversal rates much closer: the gap between point-predicted and actual consistency is 25 percentage points, but narrows to 10 percentage points when uncertainty is taken into account, a 61% reduction.

While uncertainty is a strong predictor of reversals, it remains an imperfect one. To assess the remaining belief-behavior gap, we compare the data with the "behavior drawn from beliefs" (BDB) benchmark in Table 3, the counterfactual of perfectly calibrated beliefs. Under BDB, the gap in time consistency between certain and uncertain participants nearly doubles (33.4 pp vs. 16.7 pp in the data), and the separation between reversal types

[20] While the extensive margin of uncertainty is predictive, its magnitude does not predict reversals.

Figure 7: Predicted versus actual frequency of reversals



Notes: Predicted reversals based on belief distributions (orange, left) and point predictions (yellow, right) versus observed reversals (maroon, center). Error bars are 95% CIs: conventional for mean probabilities and exact binomial otherwise. Two-sided p-values use paired t-tests for probabilities vs. behavior and exact McNemar tests for point predictions vs. behavior. $N = 244$ (downward), 381 (consistent), and 97 (upward).

in Figure 6b (thin vs. thick markers) is correspondingly wider. This comparison suggests that much of the predictive power of subjective uncertainty is muted by biases in belief formation rather than uncertainty per se. Consistent with this interpretation, the correlation between $\hat{f}(e_1)$ and actual time consistency ($e_2 = e_1$) would rise from 0.219 in the data to 0.606 [0.570, 0.642] under the BDB benchmark (also see Table C.2, Panels B vs. C). In the next section, we show that this residual gap is largely driven by misspecified beliefs.

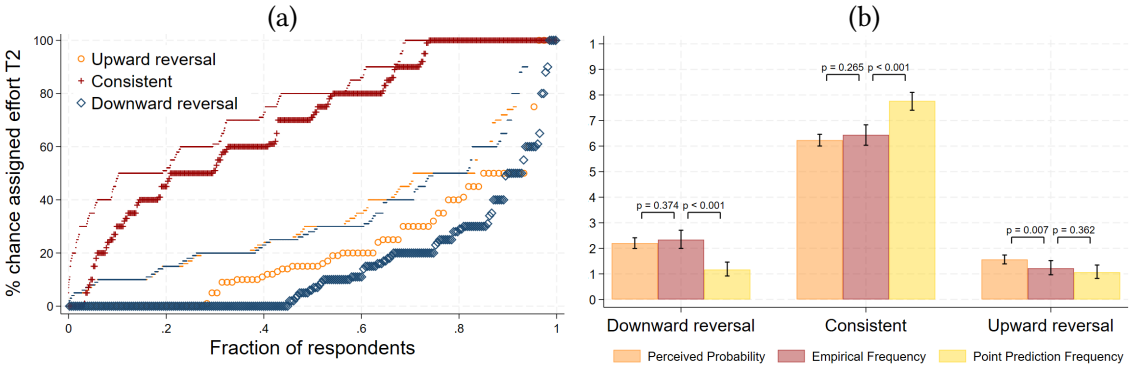
4.3 The Structure of Residual Naivety

We now turn to why uncertainty does not fully close the gap between beliefs and behavior.

Result 3. *About one in five participants assign zero probability to their chosen effort. Misspecification is rare among time-consistent individuals (3%) but much more frequent among downward and upward reversals (45% and 28%, respectively). Once beliefs assign positive probability to the realized effort, predicted and actual consistency almost perfectly align.*

Figure 8a shows the distribution of probability mass assigned to the chosen effort, $\hat{f}(e_2)$, by reversal type. Among time-consistent participants the median is 71%, against 80% under BDB. About 26.2% report certainty ($\hat{f}(e_2) = 1$, vs. 31.0% under BDB), and only 3.1% assign zero probability. By contrast, $\hat{f}(e_2)$ is typically low among participants who reverse, with a sizeable fraction assigning $\hat{f}(e_2) = 0$, especially among downward reversals.

Figure 8: (a) Misspecification, and (b) prediction after excluding misspecified beliefs

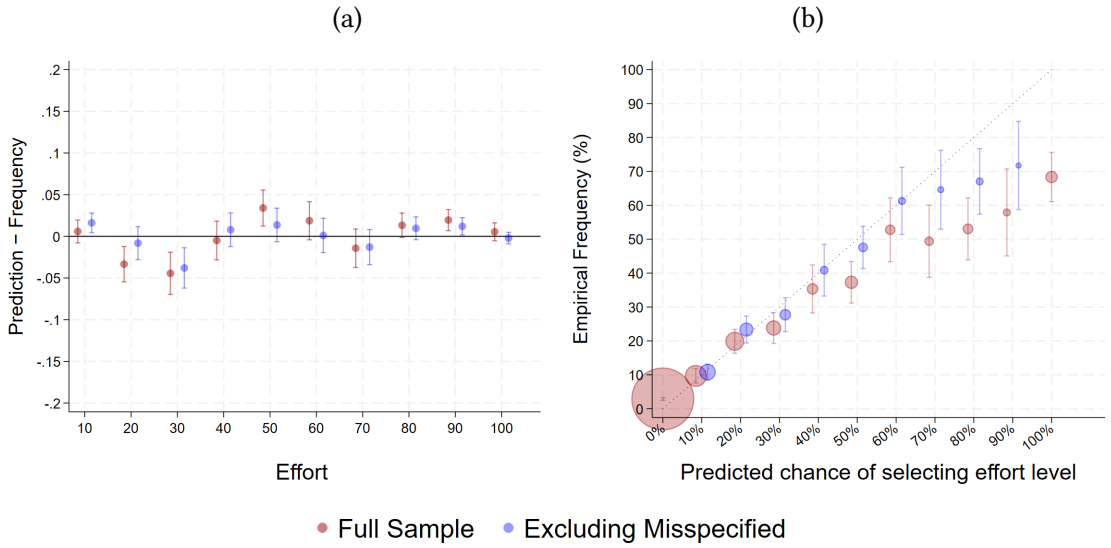


Notes: (a) Quantile plots of the probability mass $\hat{f}(e_2)$ assigned to effort chosen at t_2, by reversal type. Thick markers show empirical distributions; thin markers show closed-form BDB benchmarks implied by stated beliefs (Section 3.3). Empirical $N = 244$, 381, and 97 for downward, consistent, and upward types. (b) Predicted reversals based on belief distributions (orange, left) and point predictions (yellow, right) versus observed reversals (maroon, center), excluding participants who assigned zero probability to their realized effort. Error bars are 95% CIs: conventional for mean probabilities and exact binomial otherwise. Two-sided p-values use paired t-tests for probabilities versus behavior and exact McNemar tests for point predictions versus behavior. $N = 134$ (downward), 369 (consistent), and 70 (upward).

Misspecification is significantly more prevalent among downward reversals than upward reversals ($p = 0.003$), and far more common in both groups than among time-consistent participants ($p < 0.001$ for both). These patterns suggest that many participants fail to anticipate their eventual effort choice. We examine them further in Section 5.1.

We next reassess calibration after excluding the 21% of misspecified participants. While the direction of the improvement is unsurprising, its magnitude is not: Figure 8b shows that predicted and observed consistency rates now almost perfectly align. The same holds at the effort level. Figure 9a plots, for each effort level, the gap between its average predicted probability and realized frequency, $\bar{\hat{f}}(e) - f(e)$. Miscalibration is already modest in the full sample: the average absolute deviation is 0.02, and perfect calibration cannot be rejected for 6 of the 10 effort levels, including 10 and 100. Excluding misspecified participants reduces the average deviation to 0.01, with calibration no longer rejected in 7 cases. In cumulative terms, the gap between the predicted and realized effort distributions peaks at 7.7 percentage points at $e \leq 40$ and shrinks by about 61% once misspecified participants are excluded. Figure 9b shows that, while realized frequencies fall short of participants' most confident predictions in the full sample, they move much closer to the 45° line after misspecified participants are excluded. Significant deviations remain only at predicted probabilities of 80% and 90%.

Figure 9: Calibration of beliefs against effort



Notes: (a) Difference between the average probability assigned to each effort level and its empirical frequency; 95% CIs use robust standard errors. (b) For each assigned-probability bin, the fraction of cases in which the corresponding effort level was chosen, estimated by binomial GLM with cluster-robust 95% CIs. Probabilities are binned within ±10 percentage points, except exact 0% and 100%; predictions in $(0, 5)$ and $(95, 100)$ are excluded (1% of all predictions). Bubbles scale with sample size. Endpoints are omitted when excluding misspecified beliefs because degenerate predictions mechanically match realizations. Full sample $N = 722$; excluding misspecified $N = 573$.

Overall, these findings suggest that most of the residual naivety arises from misspecification (i.e., failing to anticipate one's eventual choice) rather than from large distortions in probability assessments conditional on those possibilities.

4.4 The Belief Foundations of Commitment and Flexibility

We now examine how beliefs about future effort relate to participants' demand for commitment and flexibility, and how these customization decisions relate to realized choices.

Result 4. *Demand for flexibility is positively associated with subjective uncertainty and anticipated reversals. Demand for commitment shows little relationship with uncertainty and participants frequently remove options they assign zero probability of choosing, a pattern inconsistent with purely consequentialist models of temptation satisfying WARP, including models of (even stochastic) present bias.*

Customization decisions Table 4 reports the prevalence of customization decisions. About 40% of participants remove at least one option from their choice set ($\mathcal{E}^- \subsetneq \mathcal{E}$), while 50% add options ($|\mathcal{E}^+| > 1$); 26% do both. Commitment demand is somewhat less prevalent

Table 4: Demand for commitment and flexibility

	Did not add	Added	Total
Did not remove	36.3%	23.4%	59.7%
Removed	13.9%	26.5%	40.4%
Total	50.2%	49.9%	100%

Notes: Percentage of participants who added at least one option ($|\mathcal{E}^+| > 1$) and/or removed at least one option ($\mathcal{E}^- \subsetneq \mathcal{E}$). $N = 722$.

than demand for flexibility (diff = 0.10, $p < 0.001$), but the intensive margin of commitment is stronger: those restricting their choice set remove an average of 4.3 options, while those expanding it add 2.2 options (diff = 2.1, $p < 0.001$). We find a slight positive correlation of 0.158 ($p < 0.001$) between the number of options added and removed.

Figure E.6 shows which effort levels are most often added or removed. Commitment decisions concentrate on extreme effort levels, with 10 and 100 excluded by 55% and 73% of participants who restrict their menus. These patterns are consistent with some participants expecting to be tempted by lower effort ($\hat{\gamma} > 1$) and others by larger payments ($\hat{\gamma} < 1$), an interpretation we return to in the discussion (see Section 5.2).[21] By contrast, flexibility decisions tend to retain intermediate effort levels such as 50 and 70 (48% and 39% of participants, respectively).

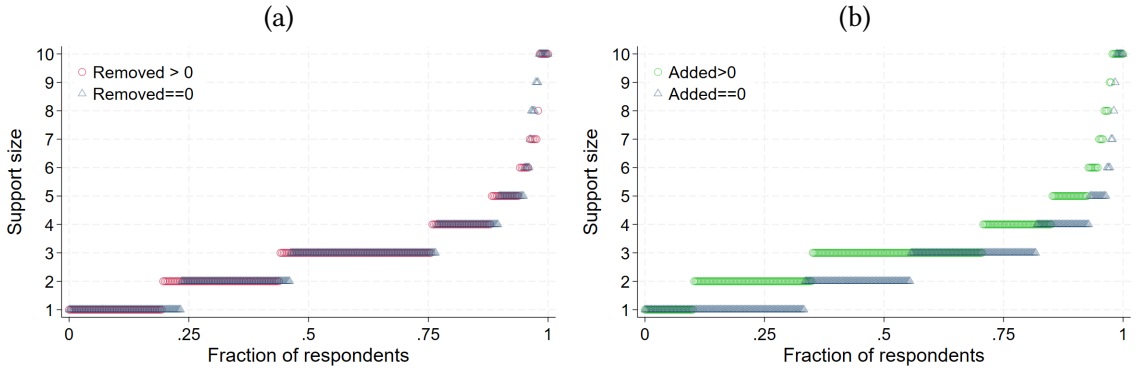
Overall, these patterns are internally coherent and prevalence rates are broadly consistent with those documented in related studies such as Toussaert (2018). Importantly, our procedures and consistency checks suggest that the observed patterns are unlikely to be driven by noise (Carrera et al., 2022). First, to ensure that participants did not select commitment by mistake, the interface required a correct answer to a comprehension question clarifying that customization was optional. Second, virtually no participant chose to both remove and add a given effort option ($N = 11$; 2% of participants). Third, 96% of the participants who included only one option added their effort chosen at t_1 (i.e., $\mathcal{E}^+ = \{e_1\}$).[22]

Customization and beliefs Figure 10 relates customization decisions to subjective uncertainty, measured by belief support size $|\text{supp}(\hat{f})|$. Commitment demand is unrelated to

[21] See Appendices A.1 and E.1 for further details on the nature of customization decisions and the implications for the sign of the perceived temptation parameter $\hat{\gamma}$.

[22] Additionally, very few participants cited confusion as a reason for their customization decisions (Figure E.9). We observe slight differences in the level of demand for commitment and flexibility depending on the (randomized) order in which the customization tasks appeared, but the qualitative results remain unchanged. See Appendix E.5 for further discussion of measurement error.

Figure 10: Support size by demand for commitment and flexibility



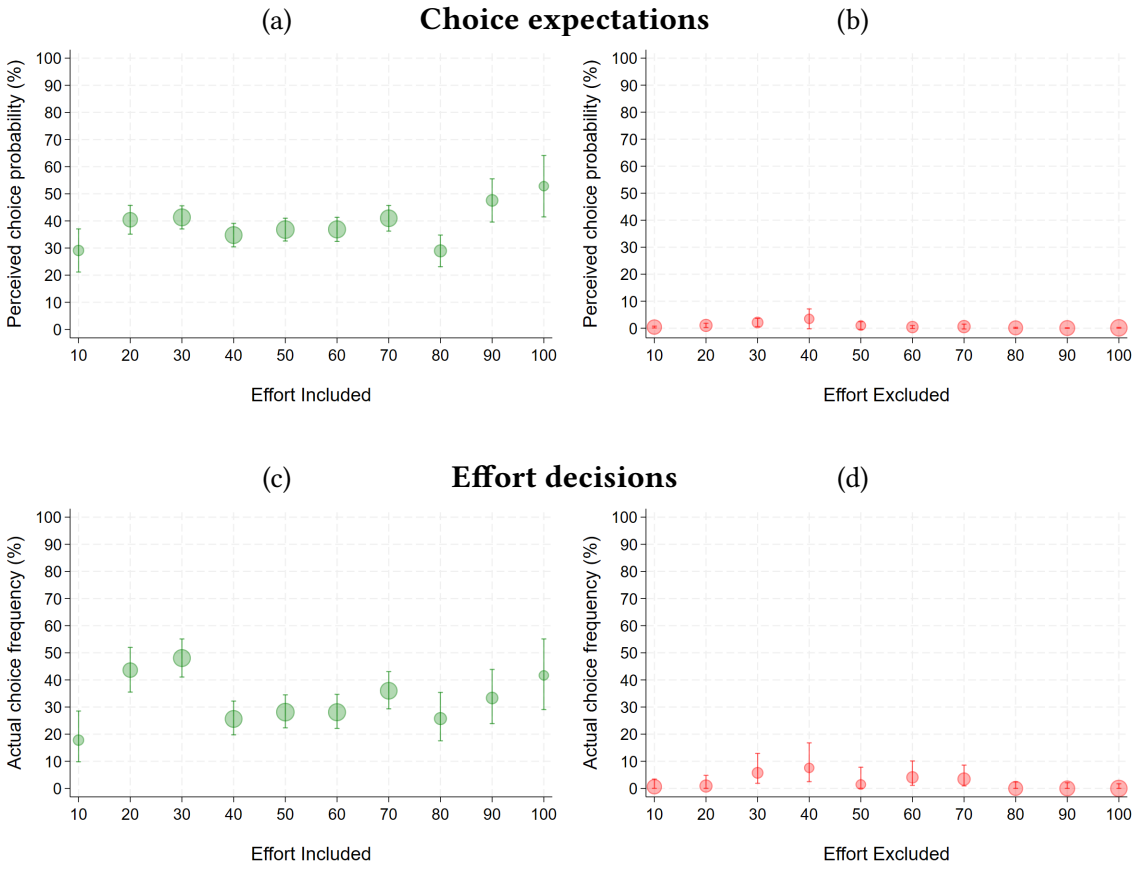
Notes: Quantile plots of belief-support size, $|\text{supp}(\hat{f})|$, by demand for (a) commitment (options excluded vs. none) and (b) flexibility (options added vs. none). $N = 722$.

uncertainty at the extensive margin: degenerate beliefs are equally common among participants who exclude options and those who do not (20% vs. 23%, $p = 0.234$). By contrast, preference for flexibility is strongly associated with uncertainty. The share reporting degenerate beliefs falls from 33% among those who add no options to 10% among those who do ($p < 0.001$), while mean support size rises from 2.5 to 3.2 ($p < 0.001$). Among flexibility demanders, the number of added options is positively correlated with support size (ρ = 0.199, $p < 0.001$); among commitment demanders, the number removed is negatively correlated with it (ρ = -0.152, $p = 0.009$). These findings are robust to alternative measures of uncertainty (entropy of $\hat{f}$ or anticipated reversals $1 - \hat{f}(e_1)$) and to using customization intensity instead of the extensive margin (Table E.8, Panels A and B).

We next ask whether customization targets options participants expect to choose. To this end, we turn to our full-menu diagnostics, ĉ-FLEX and ĉ-COMMIT, which require included and excluded options, respectively, to receive positive choice probability. As explained in Section 3.2, an isolated violation of either condition need not reject consequentialism, because an option assigned zero probability in the full menu may become relevant once other options are removed. A rejection is warranted, however, when customization is support-preserving: if $\text{supp}(\hat{f}) \subseteq \mathcal{E}^-$, excluded options cannot affect anticipated choice; similarly, if $\text{supp}(\hat{f}) \subsetneq \mathcal{E}^+$, the additional options have no material value.

A particularly stark case concerns the 140 participants (19.4% of the sample) who expect to choose e_1 with certainty. Of these, 47 (33.6%) nevertheless demand commitment, and 46 retain e_1 in the restricted menu. For these participants, WARP implies that shrinking the menu cannot change anticipated choice, pointing directly to a non-

Figure 11: Choice expectations and effort decisions across flexibility and commitment



Notes: Panels (a)–(b): Choice expectations. Average of $\hat{f}(e)$ for each effort level e, conditional on inclusion in the flexibility menu $\mathcal{E}^+$ (a) or exclusion from the commitment menu $\mathcal{E}^-$ (b). OLS estimates from linear models with robust 95% confidence intervals, clustered at the individual level. Panels (c)–(d): Effort decisions. Empirical frequency of choosing effort level e at t_2, conditional on inclusion in the flexibility menu (c) or removal from the commitment menu (d). In all panels (a)–(d), bubble size is proportional to sample size. $N = 722$.

consequentialist motive for commitment.

This pattern extends beyond this extreme case. Figures 11a and 11b relate customization decisions to average beliefs. Included options receive an average probability of 39%, ranging from 29% to 53% across effort levels. Excluded options receive only 1% on average, ranging from 0% to 3%. Consistent with this asymmetry, the likelihood of inclusion rises monotonically with $\hat{f}(e)$, whereas exclusion does not (Figures E.7a and E.7b).

At the option level, ĉ-FLEX is violated in 18% of cases where it has bite (270/1519), compared with 96% for ĉ-COMMIT (1193/1248). Among flexibility demanders, 63% satisfy ĉ-FLEX for every included option ($N = 360$). By contrast, 96% of commitment demanders

violate ĉ-COMMIT at least once ($N = 291$), and 90% assign zero probability to *every* option they exclude. For this group, $\mathcal{E}^-$ preserves the full-menu support, which is sufficient to reject consequentialist commitment within the framework.[23] The analogous condition for flexibility is met by 30% of flexibility demanders, whose menus strictly contain their belief support and thus violate consequentialism at least partially. By contrast, 38% choose menus equal to their support, a pattern fully consistent with consequentialist flexibility.

Taken together, flexibility choices generally track participants' expectations, whereas commitment demand overwhelmingly targets options outside their subjective support.

Customization and realized reversals We conclude this section by examining how preferences over choice sets relate to realized choices, and whether the full-menu properties documented above for beliefs extend to actual behavior.

Table 5 shows how the frequency of actual choice reversals varies with participants' customization decisions. Time-consistent participants are generally less inclined to customize their choice sets, whether by adding or removing options. While the unconditional probability of reversal is 0.47, the posterior probability conditional on demanding flexibility, $P(e_2 \neq e_1 \mid |\mathcal{E}^+| > 1)$, is 0.55. The corresponding probability conditional on commitment demand, $P(e_2 \neq e_1 \mid \mathcal{E}^- \subsetneq \mathcal{E})$, is 0.54. Regression results confirm that both are significantly associated with realized reversals, with demand for flexibility emerging as the stronger and more robust predictor across specifications both on the extensive and intensive margins (Table E.8, Panels C and D).

We next examine the realized counterparts of the two full-menu diagnostics, c-FLEX and c-COMMIT, at the aggregate level. Figures 11c and 11d show, for each added or removed effort level, the empirical probability that it is chosen at t_2. The pattern mirrors the belief-

Table 5: Choice reversals by demand for commitment and flexibility

Actual reversals	**Did not add**	**Added**	**Total**
Did not remove	37.4%	50.9%	42.7%
Removed	45.0%	58.6%	54.0%
Total	39.5%	55.0%	47.2%

Notes: Percentage of participants who exhibited a choice reversal, conditional on adding at least one effort level ($|\mathcal{E}^+| > 1$) and/or removing at least one effort level ($\mathcal{E}^- \subsetneq \mathcal{E}$). $N = 722$.

[23] Table E.6 shows that this conclusion is stable across menu sizes. It also holds for 88.9% of participants who remove only one option, for whom the full-menu diagnostic coincides with the primitive local condition because $\mathcal{E}^- \cup \{e\} = \mathcal{E}$.

based diagnostics: added options are chosen at substantial rates, whereas removed options are almost never chosen. Among participants who reversed after demanding flexibility, approximately 65% chose an added option, compared with only 7% who chose a removed option. Among those who reversed after demanding commitment, only 12% reversed to a removed option, while 48% reversed to an added option. Thus, commitment demand appears only weakly related to reversals toward the options participants chose to exclude. We return to the link between commitment and actual reversals in the discussion section.

5 Discussion

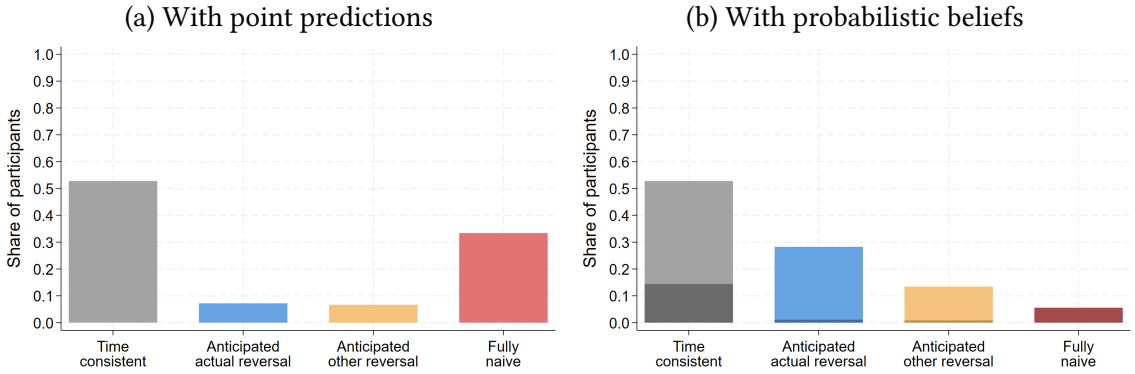
This paper revisits a classical framework for studying the consistency of effort decisions in intertemporal choice. By eliciting full belief distributions and combining them with preferences over opportunity sets, we provide new evidence on how beliefs, menu preferences, and realized behavior interact. Below we synthesize our main findings and supplement them with exploratory analyses around two questions: what does naivety look like once subjective uncertainty is measured? And what drives demand for restricting versus expanding choice sets? We end with a discussion of expert predictions and what we learn.

5.1 Reassessing naivety

From point predictions to belief distributions The standard classification of decision makers in intertemporal choice experiments relies on point predictions. A participant who chooses e_1 at t_1, predicts $\hat{e}_2 = e_1$, and then chooses $e_2 \neq e_1$ at t_2 is classified as fully naive. Under this approach, our data yield the canonical inference: as shown in Figure 12a, most of the 47% who reverse would be classified as fully naive in a standard model of present bias with deterministic beliefs. The picture substantially changes once participants can express subjective uncertainty about future effort. Reclassifying the same participants using probabilistic beliefs, Figure 12b shows that the "fully naive" category—those assigning probability one to their initial choice but reversing—shrinks from 33% to 6%. Most instead anticipate the possibility of deviating to the effort level they ultimately choose and, as discussed in Section 4.3, are broadly well calibrated. Point predictions are therefore not a sufficient statistic for expectations about future behavior.

Apparent naivety? One wrinkle noted earlier is that point predictions need not equal the mean of participants' belief distributions. If participants report a mode or median, point-based classifications may overstate naivety even when beliefs are well calibrated,

Figure 12: Classifying naivety



Notes: (a) Classifies participants using only point predictions: *Time consistent* ($e_2 = e_1$); *Anticipated actual reversal* ($\hat{e}_2 = e_2 \neq e_1$); *Anticipated other reversal* ($\hat{e}_2 \neq e_1$, $\hat{e}_2 \neq e_2$, $e_2 \neq e_1$); *Fully naive* ($\hat{e}_2 = e_1 \neq e_2$). (b) Reclassifies participants using belief distributions: *No reversal* ($e_2 = e_1$); *Anticipated actual reversal* ($e_2 \neq e_1$, $\hat{f}(e_2) > 0$); *Anticipated other reversal* ($e_2 \neq e_1$, $\hat{f}(e_2) = 0$, $\hat{f}(e_1) < 1$); *Fully naive* ($e_2 \neq e_1$, $\hat{f}(e_1) = 1$), i.e., anticipated no reversal. Within each bar, certain individuals ($\hat{f}(e_1) = 1$) are shown with darker shading. $N = 722$.

creating "apparent" naivety in the spirit of Benoît and Dubra (2011). To gauge its importance, we reclassify participants in Figure 12a using the rounded mean of the distribution instead of the point prediction. The fully naive share falls from 33% to 28% ($p < 0.001$), confirming that some of the naivety implied by point predictions is an artifact of the elicited summary statistic. Consistent with this, among participants with non-degenerate beliefs, those whose point prediction equals the rounded mean are better calibrated: realized effort matches their prediction in 54% versus 37% of cases ($p < 0.001$). Yet this substitution closes only about a fifth of the gap between the point-based classification and that based on full belief distributions (Figure 12b); the remainder reflects information beyond the mean, especially the positive probability many participants assign to their eventual choice. Moreover, even among those whose point prediction equals the mean, predictions exceed realized effort by 3.8 screens on average ($p < 0.001$), indicating genuine miscalibration beyond the summary-statistic confound.

Naivety or measurement error? A substantial share of participants who reverse assign zero probability to their eventual choice ($\hat{f}(e_2) = 0$). We interpret this as misspecification: despite being prompted to consider uncertainty, they fail to anticipate a relevant contingency. An alternative is measurement error in eliciting full belief distributions. Several design features mitigate this concern. Beliefs were elicited in two steps, first selecting the support and then assigning probabilities, with an opportunity to revise responses.

Consistency across steps was high (87%), and only 3% omitted their point prediction from the support.[24] Our results are robust to alternative definitions of belief support and sample restrictions (Table D.3). Finally, among misspecified participants, only 2% attribute their responses to confusion or mistakes in the belief elicitation task (based on follow-up explanations for their reported uncertainty in Survey 1). Further, misspecification is unrelated to perceived task difficulty or interface ratings (Appendix D.1).

Characterizing the misspecified Further evidence against pure measurement error is that misspecification is highly structured. Among those who reverse their choice, the misspecified place more mass on e_1 and report narrower belief supports (Figure D.3). They are less likely to report trying to predict accurately (45% vs. 62%, $p < 0.001$), and more likely to report setting higher targets to motivate themselves (30% vs. 18%, $p < 0.001$) or lower targets to ensure achievement (23% vs. 16%, $p = 0.034$; Figure D.4). That is, the misspecified appear to treat predictions less as forecasts and more as motivational targets.

A similar pattern emerges for constraints. Ex ante, both groups cite uncertainty about future mood or energy at the same rate (76%), but the misspecified report less schedule uncertainty (21% vs. 30%, $p = 0.089$), fewer unavailable hours (5.1 vs. 5.8, $p = 0.018$), and place less weight on "giving in to temptation" (15% vs. 25%, $p = 0.097$) or fluctuating motivation (38% vs. 54%, $p = 0.015$) as customization motives. Ex post, however, they reverse downward more often (80% vs. 66%, $p = 0.003$) and more sharply so (22.0 vs. 13.5 screens, $p < 0.001$), attributing reversals more to time constraints (57% vs. 45%, $p = 0.026$) and less to energy (50% vs. 68%, $p < 0.001$; Figure D.5). They also recall prior point predictions less accurately and, on average, as lower than originally stated, while reporting less influence of those predictions on their choices.[25] Taken together, the misspecified appear less like classically naive agents and more like individuals with narrow, motivated beliefs that downplay constraints ex ante and reverse sharply when they bind.

Does eliciting beliefs change behavior? If some participants treat predictions as motivational targets, a natural question is whether the act of eliciting beliefs itself shifts behavior. Comparing our main treatment to an otherwise identical condition without belief

[24] Among those with inconsistent supports, 1% included an option only in Step 1, 10% only in Step 2, and 2% had discrepancies in both directions.

[25] In Survey 2, participants recalled their Survey 1 point prediction (without accuracy incentives) and rated its influence on their decision from 1 ("Not at all") to 4 ("Strongly"). Misspecified individuals were less likely to recall their prediction correctly ($p < 0.001$) and exhibited a more negative recall bias (recalled minus original: -2.28 vs. 0.12, $p = 0.069$). They also reported lower perceived influence of their point prediction (2.2 vs. 2.5, $p = 0.002$). Recall-bias estimates are similar after excluding time-consistent individuals.

elicitation, we find no evidence of an effect on effort choices or reversal rates (Table D.4). The one exception is that elicitation increases demand for flexibility on the extensive margin, possibly because forming expectations makes future contingencies more salient. There are no differences in demand for commitment. Whether and when structured self-reflection supports self-discipline and forward-looking behavior remains an open question (Sial, Sydnor and Taubinsky, 2024).

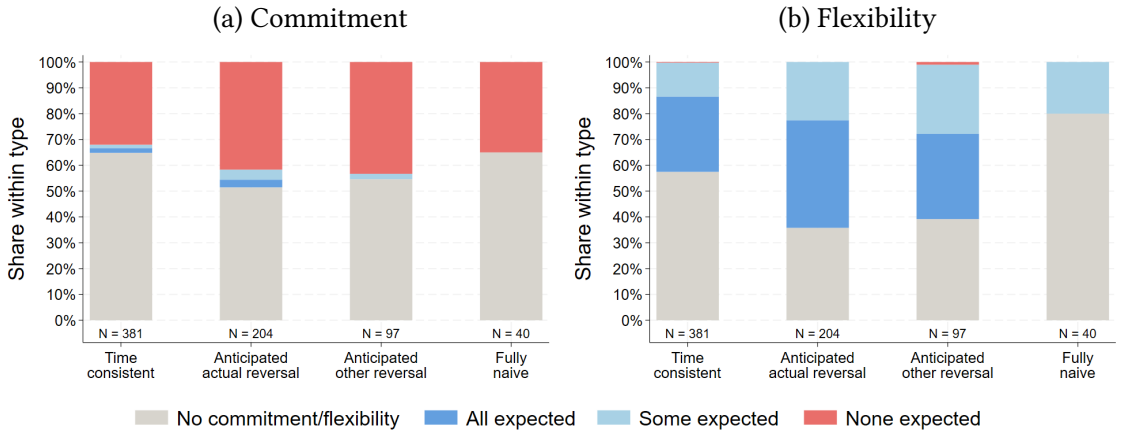
5.2 The nature of demand for commitment and flexibility

Menu demand as a measure of sophistication The literature on dynamic inconsistency commonly interprets demand for commitment as a behavioral marker of sophistication (Laibson, 1997; O'Donoghue and Rabin, 1999). Figure 13 allows us to test this directly by looking at the share of each belief–behavior type from Figure 12b whose customizations are (i) fully consequentialist, with every selected option assigned positive probability, $\hat{f}(e) > 0$ (*all expected*); (ii) fully non-consequentialist, with $\hat{f}(e) = 0$ for every selected option (*none expected*); or (iii) mixed (*some expected*).[26] If commitment reflected sophistication, consequentialist commitment (dark blue) should be concentrated among participants who anticipated their actual reversal and absent among the fully naive. Instead, non-consequentialist commitment (red), accounting for 90% of commitment demanders, dominates across all types with little variation (Figure 13a). Figure 13b shows the mirror image for flexibility: consequentialist demand (dark blue) closely tracks the sophistication gradient, highest among those who anticipated their actual reversal and nearly absent among the fully naive. Demand for flexibility is therefore a far better proxy for beliefs about future behavior than demand for commitment.

What drives commitment demand? We find that 90% of commitment demanders assign zero probability to *every* option they exclude. Measurement error in beliefs or commitment choices seems unlikely to explain this pattern. Although zeros could represent small probabilities rounded down, three findings argue against this. First, when asked about zero-probability removals, 73% said they knew they would not choose the option but still wanted it removed; only 23% answered "I might still choose it" (Figure E.10). Second, removed options were ultimately chosen by only 7% of commitment demanders, or 1.7% of the time per removed option, suggesting little concealed probability mass (Appendix E.3).

[26] We use these labels as shorthand for the belief-based categories formalized in Table E.7. The mapping onto consequentialism is exact for commitment, where *none expected* removals reject consequentialism; for flexibility, rejection requires the finer diagnostics in the table.

Figure 13: Commitment and flexibility demand by belief-behavior type



Notes: Each panel classifies participants into four mutually exclusive belief-behavior types: *Time consistent* ($e_2 = e_1$); *Anticipated actual reversal* ($e_2 \neq e_1$, $\hat{f}(e_2) > 0$); *Anticipated other reversal* ($e_2 \neq e_1$, $\hat{f}(e_2) = 0$, $\hat{f}(e_1) < 1$); *Fully naive* ($e_2 \neq e_1$, $\hat{f}(e_1) = 1$). Within each type, stacked bars show the share of participants with *No commitment/flexibility* demand (grey), *All expected*: full consequentialist demand (dark blue; all removed/added options had $\hat{f}(e) > 0$), *Some expected*: mixed demand (light blue; some options had $\hat{f}(e) > 0$ and some had $\hat{f}(e) = 0$), and *None expected*: full non-consequentialist demand (red; all removed/added options had $\hat{f}(e) = 0$). These labels are shorthand; Table E.7 gives the exact interpretation of each category, which differs between commitment and flexibility. Panel (a) shows commitment (paid to remove options) and panel (b) shows flexibility (paid to add options). $N = 722$.

Third, zero-probability removals are pervasive across menu sizes and locations relative to the belief support, and adjacent removals are no more likely than distant ones to elicit "I might still choose it" (Tables E.6 and E.5; Appendix E.3). Error in commitment choices also appears limited: only 1.5% of participants both added and removed the same option, 8% of removers cite confusion about the task, and 3% cite believing removal was required. The £0.01 fee further strengthens the test: under consequentialism, removing a true-zero option is a pure loss at any positive price, and most participants removed nothing.

The evidence instead points to non-consequentialist motives, under which utility depends on the menu itself. Two mechanisms emerge from the data. First, costly self-control (Gul and Pesendorfer, 2001) may lead individuals to remove alternatives they would otherwise need to resist. Indeed, 37% of commitment demanders cite temptation as a primary motive (Figure E.9a), more often among fully non-consequentialist than mixed cases (40% vs. 7%). These motives concern both low- and high-effort options: 41% of commitment demanders remove only levels above e_1, citing bonus pressure or concern about overdoing it (Figure E.11), and 33% of them cite temptation. Self-control concerns are thus not confined to avoiding low effort, challenging the view that temptation operates only toward

lower effort; our framework allows both directions ($\gamma \gtrless 1$; Section 3). Second, many respondents cite less screen clutter and easier decisions, pointing to cognitive or attentional costs (Maćkowiak, Matějka and Wiederholt, 2023; Dean, Ravindran and Stoye, 2025). This may explain why commitment predicts actual reversals despite targeting options that participants never expect to choose: volatile circumstances may both increase reversals and raise the value of simpler menus. The correlation would then reflect a common factor rather than targeted anticipated deviations.

Scope of the rejection Our falsification tests reject any model in which (i) utility depends only on the effort ultimately undertaken and (ii) choice from each menu satisfies WARP state by state. This is a broad class, spanning expected and non-expected utility, random utility, and stochastic present bias models. But some models relax WARP, e.g., under convex self-control costs (Fudenberg and Levine, 2006, 2011; Noor and Takeoka, 2010, 2015), the presence of a never-chosen option can erode resistance to others, so removing it changes choice among those that remain, making removal consequentialist in a broader sense. Three observations weigh against this channel. First, it predicts that commitment demand tracks anticipated reversals, which we do not observe (Figure 13). Second, it operates through WARP violations, and these are no more frequent when the menu is genuinely restricted than when participants face the identical menu twice (8.9% vs. 12.8%; Appendix E.4). Third, independently of realized violations, paying to remove an option on this theory requires anticipating that its absence will shift one's future choice among the remaining options: a second-order forecast, more demanding than anticipating temptation itself, about which our participants are demonstrably imperfect.

Flexibility as insurance Unlike commitment, flexibility demand largely follows a consequentialist logic: 63% of flexibility demanders assign positive probability to *every* option they add, and a substantial share of reversals are toward options added ex ante (65%). Reported motives are consistent with insurance against uncertainty about mood, energy, or schedule at t_2 (Figure E.9b). Flexibility demand also does not appear to reflect contemporaneous indecision. Although undecided agents might randomize across menus, generating choice inconsistencies (Cerreia-Vioglio et al., 2019), WARP violations between $\mathcal{E}$ and $\mathcal{E}^+$ are rare (8%; Table E.9), and self-reported uncertainty overwhelmingly concerns future states rather than difficulty forming preferences (Figure E.12).

Yet roughly one-third of flexibility demanders add both expected and unexpected options (light blue in Figure 13). Small subjective probabilities recorded as zero seem to be

a more plausible explanation for zero-probability additions than removals. Most such additions lie at the boundary of the reported support; their prevalence rises with menu size, consistent with participants "running out of probability points"; and participants almost never add a zero-probability option while omitting one assigned positive probability (Table E.6; Appendix E.3).[27] While some unexpected additions may reflect censored small probabilities, others may capture aspirational or signalling motives (Kopylov, 2012). Ex post, however, many reversals are toward options not added ex ante, especially among the misspecified (Figures E.8a and E.8b), suggesting some suboptimality (Ahn et al., 2019).

5.3 Taking stock

Did we know it all already? To benchmark the novelty of our findings, we compare our results to expert forecasts collected prior to observing the data (Section 2.7; Appendix F). Experts correctly anticipated the qualitative baseline, with more downward than upward reversals and point predictions overstating time consistency. However, they highly overestimated reversal rates (67% vs. 47%, $p < 0.001$), while underestimating belief uncertainty (67% vs. 78%, $p < 0.001$). The largest errors concern mechanisms. Experts predicted similar misspecification rates for downward reversals and time-consistent choices (32% vs. 24%), whereas in the data misspecification is highly diagnostic (45% vs. 3%). Experts also appear to have had largely consequentialist models in mind: they predicted that 33% of removed options would be assigned a positive probability of being chosen, compared to only 4% in the data. Finally, most experts expected belief elicitation to significantly affect effort choices, whereas the observed effect is insignificant. Thus, while experts anticipated the qualitative patterns established in the literature, they failed to foresee the mechanisms revealed in the data, especially the diagnostic role of misspecification and the prevalence of consequentialism violations.

Open questions Our findings raise two related sets of questions. On the belief side, our analyses point to structured misspecification: participants who assign zero probability to their eventual choice appear to form narrower, more motivated beliefs and reverse more sharply when constraints bind. Yet we cannot distinguish whether this reflects misperceptions of future shocks, underestimation of susceptibility to temptation, or other factors. Randomly varying information about external constraints and self-control demands could help separate these channels. Whether misspecification strengthens with stakes (e.g., a

[27] Free-text explanations from participants who reported certainty yet added options point the same way, e.g., one states: "I should have put 98% and 1% on the others."

higher cost of ignoring an option) or with ego relevance also remains an open question.

On the menu side, the largely non-consequentialist demand for commitment, for which Toussaert (2018) provides partial evidence in the lab, raises questions about whether it persists beyond low-stakes settings and whether it reflects self-control or cognitive costs. These channels might be separated by varying how alternatives are presented, for example whether removed options are hidden or merely unselectable, or how much information about them is available. The relationship between commitment and reversals is also unsettled. We find a positive correlation, perhaps because volatile environments generate both more reversals and greater costs from large choice sets, whereas Sadoff, Samek and Sprenger (2020) report a negative correlation in the field. Measurement may partly explain the difference. In that study, commitment is elicited as a binary choice over a fixed bundle, so those who decline may be rejecting rigidity rather than commitment per se. Extending our design to the field could help disentangle these mechanisms.

More generally, our framework could be extended to identify richer aspects of how menus shape behavior. Eliciting beliefs under multiple menus, rather than only the full menu, could reveal how individuals expect their choices to depend on the available set and, combined with choice data, help distinguish models that violate IIA or WARP. Building on Strack and Taubinsky (2026), supplementing these data with cardinal measures of the value of individual alternatives and the costs of adding or removing them could potentially point-identify temptation parameters jointly with the distribution of external shocks, whereas our approach provides only set identification. This could, in turn, support richer policy counterfactuals.

References

Ahn, David S., and Todd Sarver. 2013. "Preference for flexibility and random choice." *Econometrica*, 81(1): 341–361.

Ahn, David S., Ryota Iijima, and Todd Sarver. 2020. "Naivete about temptation and self-control: Foundations for recursive naive quasi-hyperbolic discounting." *Journal of Economic Theory*, 189: 105087.

Ahn, David S, Ryota Iijima, Yves Le Yaouanq, and Todd Sarver. 2019. "Behavioural Characterizations of Naivete for Time-Inconsistent Preferences." *The Review of Economic Studies*, 86(6): 2319–2355.

Augenblick, Ned, and Matthew Rabin. 2019. "An Experiment on Time Preference and Misprediction in Unpleasant Tasks." *The Review of Economic Studies*, 86(3): 941–975.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. "Working over Time: Dynamic Inconsistency in Real Effort Tasks." *The Quarterly Journal of Economics*, 130(3): 1067–1115.

Bago d'Uva, Teresa, Owen O'Donnell, and Eddy Van Doorslaer. 2020. "Who can predict their own demise? Heterogeneity in the accuracy and value of longevity expectations." *The Journal of the Economics of Ageing*, 17: 100135.

Bartling, Björn, Ernst Fehr, and Holger Herz. 2014. "The Intrinsic Value of Decision Rights." *Econometrica*, 82(6): 2005–2039.

Becker, Christoph K, Tigran Melkonyan, Eugenio Proto, Andis Sofianos, and Stefan T Trautmann. 2026. "Revising beliefs in light of unforeseen events." *Journal of the European Economic Association*, 24(2): 661–700.

Benoît, Jean-Pierre, and Juan Dubra. 2011. "Apparent Overconfidence." *Econometrica*, 79(5): 1591–1625.

Bernheim, B Douglas, and Dmitry Taubinsky. 2018. "Behavioral public economics." *Handbook of behavioral economics: Applications and Foundations 1*, 1: 381–516.

Bohren, J. Aislinn, and Daniel N. Hauser. 2021. "Learning with Heterogeneous Misspecified Models: Characterization and Robustness." *Econometrica*, 89(6): 3025–3077.

Bohren, J. Aislinn, and Daniel N. Hauser. 2023. "The Behavioral Foundations of Model Misspecification." CEPR Discussion Paper DP17884.

Breig, Zachary, Matthew Gibson, and Jeffrey G. Shrader. 2023. "Why Do We Procrastinate? Present Bias and Optimism." Working Paper.

Carrera, Mariana, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky. 2022. "Who Chooses Commitment? Evidence and Welfare Implications." *The Review of Economic Studies*, 89(3): 1205–1244.

Cerreia-Vioglio, Simone, David Dillenberger, Pietro Ortoleva, and Gil Riella. 2019. "Deliberately Stochastic." *American Economic Review*, 109(7): 2425–45.

Cerrone, Claudia, Anujit Chakraborty, Hyok Jung Kim, and Leonhard K. Lades. 2025. "Estimating Present Bias and Sophistication over Effort and Money." *Experimental Economics*, 28(6): 1191 – 1213.

D'Acunto, Francesco, and Michael Weber. 2024. "Why Survey-Based Subjective Expectations Are Meaningful and Important." *Annual Review of Economics*, 16: 329–357.

Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2022. "Belief Elicitation and Behavioral Incentive Compatibility." *American Economic Review*, 112(9): 2851–2883.

Dean, Mark, and John McNeill. 2020. "Preference for Flexibility and Random Choice: an Experimental Analysis." Working Paper.

Dean, Mark, Dilip Ravindran, and Jörg Stoye. 2025. "A Better Test of Choice Overload." Working Paper.

Dekel, Eddie, Barton L. Lipman, and Aldo Rustichini. 2001. "Representing Preferences with a Unique Subjective State Space." *Econometrica*, 69(4): 891–934.

Dekel, Eddie, Barton L. Lipman, and Aldo Rustichini. 2009. "Temptation-Driven Preferences." *Review of Economic Studies*, 76(3): 937–971.

Delavande, Adeline. 2014. "Probabilistic Expectations in Developing Countries." *Annual Review of Economics*, 6: 1–20.

DellaVigna, Stefano, and Devin Pope. 2018. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy*, 126(6): 2410–2456.

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *American Economic Review*, 96(3): 694–719.

Diebold, Francis X, Minchul Shin, and Boyuan Zhang. 2023. "On the aggregation of probability assessments: Regularized mixtures of predictive densities for Eurozone inflation and real interest rates." *Journal of Econometrics*, 237(2): 105321.

Engelberg, Joseph, Charles F. Manski, and Jared Williams. 2009. "Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters." *Journal of Business and Economic Statistics*, 27: 30–41.

Esponda, Ignacio, and Demian Pouzo. 2016. "Berk–Nash equilibrium: A framework for modeling agents with misspecified models." *Econometrica*, 84(3): 1093–1130.

Fedyk, Anastassia. 2025. "Asymmetric Naïveté: Beliefs About Self-Control." *Management Science*, 71(7): 6047–6068.

Ferreira, João V., Nobuyuki Hanaki, and Benoît Tarroux. 2020. "On the roots of the intrinsic value of decision rights: Experimental evidence." *Games and Economic Behavior*, 119: 110–122.

Freeman, David J. 2021. "Revealing Naïveté and Sophistication from Procrastination and Preproperation." *American Economic Journal: Microeconomics*, 13(2): 402–438.

Frick, Mira, Ryota Iijima, and Yuhta Ishii. 2023. "Belief Convergence under Misspecified Learning: A Martingale Approach." *The Review of Economic Studies*, 90(2): 781–814.

Fudenberg, Drew, and David K. Levine. 2006. "A Dual-Self Model of Impulse Control." *American Economic Review*, 96(5): 1449–1476.

Fudenberg, Drew, and David K. Levine. 2011. "Risk, Delay, and Convex Self-Control Costs." *American Economic Journal: Microeconomics*, 3(3): 34–68.

Fuster, Andreas, and Basit Zafar. 2023. "Chapter 4 - Survey experiments on economic expectations." In *Handbook of Economic Expectations.* , ed. Rüdiger Bachmann, Giorgio Topa and Wilbert van der Klaauw, 107–130. Academic Press.

Gill, David, and Victoria Prowse. 2019. "Measuring costly effort using the slider task." *Journal of Behavioral and Experimental Finance*, 21: 1–9.

Grubb, Michael D. 2009. "Selling to overconfident consumers." *American Economic Review*, 99(5): 1770–1807.

Gul, Faruk, and Wolfgang Pesendorfer. 2001. "Temptation and Self-Control." *Econometrica*, 69(6): 1403–1435.

Halevy, Yoram. 2015. "Time Consistency: Stationarity and Time Invariance." *Econometrica*, 83(1): 335–352.

Heidhues, Paul, and Botond Kőszegi. 2009. "Futile Attempts at Self-Control." *Journal of the European Economic Association*, 7(2-3): 423–434.

Heidhues, Paul, and Philipp Strack. 2021. "Identifying present bias from the timing of choices." *American Economic Review*, 111(8): 2594–2622.

Hossain, Tanjim, and Ryo Okui. 2013. "The Binarized Scoring Rule." *The Review of Economic Studies*, 80(3): 984–1001.

Huffman, David, Collin Raymond, and Julia Shvets. 2022. "Persistent Overconfidence and Biased Memory: Evidence from Managers." *American Economic Review*, 112(10): 3141–75.

Iacovone, Leonardo, David McKenzie, and Rachael Meager. 2025. "Bayesian Impact Evaluation With Informative Priors: An Application to a Colombian Management and Export Improvement Program." *Econometrica*, 93(5): 1915–1935.

Imai, Taisuke, Tom A. Rutter, and Colin F. Camerer. 2021. "Meta-Analysis of Present-Bias Estimation using Convex Time Budgets." *The Economic Journal*, 131(636): 1788–1814.

John, Anett. 2020. "When Commitment Fails: Evidence from a Field Experiment." *Management Science*, 66(2): 503–529.

Kenny, Geoff, Thomas Kostka, and Federico Masera. 2014. "How informative are the subjective density forecasts of macroeconomists?" *Journal of Forecasting*, 33(3): 163–185.

Kopylov, Igor. 2012. "Perfectionism and Choice." *Econometrica*, 80(5): 1819–1843.

Kreps, David M. 1979. "A Representation Theorem for "Preference for Flexibility"." *Econometrica*, 47(3): 565–577.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics*, 112(2): 443–478.

Laibson, David. 2015. "Why don't present-biased agents make commitments?" *American Economic Review*, 105(5): 267–272.

Le Lec, Fabrice, and Benoît Tarroux. 2020. "On Attitudes to Choice: Some Experimental Evidence on Choice Aversion." *Journal of the European Economic Association*, 18(5): 2108–2134.

Le Yaouanq, Yves, and Peter Schwardmann. 2022. "Learning About One's Self." *Journal of the European Economic Association*, 20(5): 1791–1828.

Manski, Charles F. 2004. "Measuring Expectations." *Econometrica*, 72(5): 1329–1376.

Maćkowiak, Bartosz, Filip Matějka, and Mirko Wiederholt. 2023. "Rational Inattention: A Review." *Journal of Economic Literature*, 61(1): 226–273.

Noor, Jawwad, and Norio Takeoka. 2010. "Uphill self-control." *Theoretical Economics*, 5(2): 127–158.

Noor, Jawwad, and Norio Takeoka. 2015. "Menu-dependent self-control." *Journal of Mathematical Economics*, 61: 1–20.

O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review*, 89(1): 103–124.

Orhun, Yeşim, Alain Cohn, and Collin B Raymond. 2024. "Motivated optimism and workplace risk." *The Economic Journal*, 134(663): 2951–2981.

Oster, Emily, Ira Shoulson, and E Ray Dorsey. 2013. "Optimal expectations and limited medical testing: Evidence from Huntington disease." *American Economic Review*, 103(2): 804–830.

Sadoff, Sally, Anya Samek, and Charles Sprenger. 2020. "Dynamic Inconsistency in Food Choice: Experimental Evidence from Two Food Deserts." *The Review of Economic Studies*, 87(4): 1954–1988.

Sen, Amartya K. 1988. "Freedom of choice: Concept and content." *European Economic Review*, 32: 269–94.

Sial, Afras, Justin Sydnor, and Dmitry Taubinsky. 2024. "Biased Memory and Perceptions of Self-Control." Working Paper.

Spiegler, Ran. 2011. "Dynamically Inconsistent Preferences III: Partial Naivete." In *Bounded Rationality and Industrial Organization*. Oxford University Press.

Stephens, Jr., Melvin. 2004. "Job loss expectations, realizations, and household consumption behavior." *Review of Economics and Statistics*, 86(1): 253–269.

Strack, Philipp, and Dmitry Taubinsky. 2026. "Dynamic Preference "Reversals" and Time Inconsistency." *American Economic Journal: Microeconomics*, 18(2): 426–68.

Toussaert, Séverine. 2018. "Eliciting Temptation and Self-Control Through Menu Choices: A Lab Experiment." *Econometrica*, 86(3): 859–889.

Toussaert, Séverine. 2019. "Revealing Temptation Through Menu Choice: Field Evidence." Working Paper.

Supplemental Appendix

A Theory: Menu Diagnostics and Illustration

A.1 Consequentialism diagnostics and identification

Section 3.2 outlines that removing an option assigned zero choice probability in the full menu does not necessarily violate consequentialism—removals must be *choice-preserving* in a sense made precise below (Proposition 1). Removals also inform the sign of the perceived temptation parameter $\hat{\gamma}$ under the maintained model (Proposition 2).[1]

Setup and Maintained Assumptions Let $T_s := U_s + V_s$ be the utility functional maximized by date-2 choice in state s, where $U_s(e) = m_e - c_s(e)$ and $V_s(e) = m_e - \hat{\gamma} c_s(e)$ are the normative and temptation-weighted payoffs, and $c_s(e)$ is strictly increasing in e and convex.[2] Let $e_T(s) := \arg\max_{e \in \mathcal{E}} T_s(e)$ be the resulting choice and $e_U(s) := \arg\max_{e \in \mathcal{E}} U_s(e)$ the normative optimum; throughout, we assume that $e_T(s)$ is unique for all s. Finally, assume that any removal $r \notin \mathcal{E}^-$ is *non-redundant* in the sense of being genuinely worth removing, i.e., $W(\mathcal{E}^-) - W(\mathcal{E}^- \cup \{r\}) > \varepsilon$ where $\varepsilon = 0.01$ is the removal cost.

Proposition 1 (When consequentialism is violated)**.** *Say $r \notin \mathcal{E}^-$ is* choice-preserving *if $\arg\max_{\mathcal{E}^-} T_s = \arg\max_{\mathcal{E}^- \cup \{r\}} T_s$ for all s, i.e., restoring it changes no state's date-2 choice. If non-redundant, this is inconsistent with consequentialism. Either of the following two conditions is sufficient for choice preservation:*

(i) $\mathcal{E}^-$ is support-preserving*, i.e.,* $\operatorname{supp}(\hat{f}) \subseteq \mathcal{E}^-$

(ii) r is shielded*: $r < \min \operatorname{supp}(\hat{f})$ and $\exists\, q \in \mathcal{E}^-$ such that $r < q \leq \min \operatorname{supp}(\hat{f})$; or, symmetrically, $r > \max \operatorname{supp}(\hat{f})$ and $\exists\, q \in \mathcal{E}^-$ such that $\max \operatorname{supp}(\hat{f}) \leq q < r$.*

Proof. (i) If $e_T(s) \in \operatorname{supp}(\hat{f}) \subseteq \mathcal{E}^-$ for every state s, a global maximizer of T_s already in $\mathcal{E}^-$ stays the maximizer of $\mathcal{E}^- \cup \{r\}$.
(ii) Suppose $e_T(s) \geq \min \operatorname{supp}(\hat{f}) \geq q > r$. Since T_s is single-peaked, $T_s(q) \geq T_s(r)$, as q sits between r and the peak $e_T(s)$; this holds whether or not $e_T(s) \in \mathcal{E}^-$, so q dominates r regardless, thus ensuring that r would not be chosen. The case $e_T(s) \leq \max \operatorname{supp}(\hat{f}) \leq q < r$ mirrors this. □

[1] We write $\hat{\gamma}$ rather than γ throughout: menu choices and beliefs are elicited ex ante, so they reveal the subject's *perceived* temptation, which need not equal the true γ if the subject is naive.

[2] Together with $\{m_e\}_{e \in \mathcal{E}}$ being a strictly concave schedule, this implies that U_s, V_s, and hence T_s, are all single-peaked, an assumption we leverage in the proof of Propositions 1(ii) and 2(b).

Proposition 2 (Sign of $\hat{\gamma}$)**.** *Let $r \notin \mathcal{E}^-$ be non-redundant.*

(a) If r is choice-preserving (Proposition 1), then $\hat{\gamma} > 1$ whenever $r < \min \operatorname{supp}(\hat{f})$, and $\hat{\gamma} < 1$ whenever $r > \max \operatorname{supp}(\hat{f})$. If both directions occur among choice-preserving, non-redundant removals, no single $\hat{\gamma}$ rationalizes the data.

(b) The same two signs extend to expected endpoints: $r = 100 \in \operatorname{supp}(\hat{f})$ gives $\hat{\gamma} < 1$, and $r = 10 \in \operatorname{supp}(\hat{f})$ gives $\hat{\gamma} > 1$, provided every other expected option e' has some $q \in \mathcal{E}^-$ with $e' \le q < r$ (or $r < q \le e'$ for $r = 10$).

Proof. (a) Take $r < \min \operatorname{supp}(\hat{f})$ and suppose by contradiction that $\hat{\gamma} \le 1$. Let $q = e_T(s)$ if r is support-preserving, or let q be the shield if r is shielded; either way $q \in \mathcal{E}^-$, $q > r$, and $T_s(q) \ge T_s(r)$.[3] Since

$$T_s(q) \ge T_s(r) \iff m_q - m_r \ge \tfrac{1+\hat{\gamma}}{2}\big[c_s(q) - c_s(r)\big],$$

and $\hat{\gamma} \le \frac{1+\hat{\gamma}}{2}$ with $c_s(q) > c_s(r)$, this gives $m_q - m_r \ge \hat{\gamma}[c_s(q) - c_s(r)]$, i.e., $V_s(q) \ge V_s(r)$. So restoring r raises neither T_s nor V_s in any state: $W(\mathcal{E}^-) = W(\mathcal{E}^- \cup \{r\})$, contradicting non-redundancy. Hence $\hat{\gamma} > 1$. The case $r > \max \operatorname{supp}(\hat{f})$ mirrors this, giving $\hat{\gamma} < 1$; both cases satisfied together is then a direct contradiction of a model with a single $\hat{\gamma}$.

(b) Take $r = 100$ with $100 \in \operatorname{supp}(\hat{f})$, so $e_T(s^*) = 100$ for at least one state s^* and suppose by contradiction that $\hat{\gamma} \ge 1$. Write $W_s(A) := \max_A T_s - \max_A V_s$ (so $W = \sum_s p_s W_s$); we show that there is no state s under which restoring 100 raises W_s.

In state s^*, $\hat{\gamma} \ge 1$ gives $e_U(s^*) \ge e_T(s^*) = 100$, so $e_U(s^*) = 100$. With $t' = \arg\max_{\mathcal{E}^-} T_{s^*}$,

$$\underbrace{T_{s^*}(100) - T_{s^*}(t')}_{\ge 0,\ \text{since } e_T(s^*)=100} \ \ge\ \underbrace{V_{s^*}(100) - V_{s^*}(t')}_{\text{since } U_{s^*}(100)\ge U_{s^*}(t')} \ \ge\ \underbrace{V_{s^*}(100) - \max_{\mathcal{E}^-} V_{s^*}}_{\text{since } V_{s^*}(t')\le \max_{\mathcal{E}^-} V_{s^*}}.$$

So $T_{s^*}(100) - T_{s^*}(t') \ge \max\left\{V_{s^*}(100) - \max_{\mathcal{E}^-} V_{s^*}, 0\right\} = \max_{\mathcal{E}^- \cup \{100\}} V_{s^*} - \max_{\mathcal{E}^-} V_{s^*}$. Rearranging, $T_{s^*}(100) - \max_{\mathcal{E}^- \cup \{100\}} V_{s^*} \ge T_{s^*}(t') - \max_{\mathcal{E}^-} V_{s^*}$, i.e, $W_{s^*}(\mathcal{E}^- \cup \{100\}) \ge W_{s^*}(\mathcal{E}^-)$. For every other state s', with $e_T(s') = e' \ne 100$, some retained t'' satisfies $T_{s'}(t'') \ge T_{s'}(100)$.[4] As in (a), since $\hat{\gamma} \ge \frac{1+\hat{\gamma}}{2}$, this gives $V_{s'}(t'') \ge V_{s'}(100)$ too, so t'' beats 100 on both counts and $W_{s'}(\mathcal{E}^-) = W_{s'}(\mathcal{E}^- \cup \{100\})$. Summing over states, $W(\mathcal{E}^-) \le W(\mathcal{E}^- \cup \{100\})$, contradicting non-redundancy. Hence $\hat{\gamma} < 1$. The case $r = 10$ mirrors this. □

[3] If $q = e_T(s)$, this is simply by global optimality of T_s; if q is the shield, this is by single-peakedness, since $r < q \le e_T(s)$.

[4] If e' itself is retained, $t'' = e'$ works trivially. If instead e' is shielded by some retained q with $e' \le q < 100$, single-peakedness (peak at e') gives $T_{s'}(q) \ge T_{s'}(100)$, so $t'' = q$ works.

Flexibility Counterpart We can similarly consider choice-preserving additions. The support-covering case, $\text{supp}(\hat{f}) \subseteq \mathcal{E}^+$, mirrors Prop 1(i): if $e_T(s) \in \mathcal{E}^+$ for every relevant state, no off-support addition $a \in \mathcal{E}^+ \setminus \text{supp}(\hat{f})$ is ever chosen. The shielding case mirrors Prop 1(ii): an off-support addition a is choice-preserving if some other added q satisfies $\max \text{supp}(\hat{f}) \leq q < a$ (or the mirror, $a < q \leq \min \text{supp}(\hat{f})$). For example, with $\text{supp}(\hat{f}) = \{30, 50\}$ and $60, 80 \in \mathcal{E}^+$, 60 shields 80: for any state s with $e_T(s) \in \text{supp}(\hat{f})$, $T_s(80) \leq T_s(60)$, so 80 would never be chosen. Either way, consequentialist flexibility is rejected provided a is non-redundant ($W(\mathcal{E}^+) - W(\mathcal{E}^+ \setminus \{a\}) > \varepsilon$). Unlike commitment, however, it never signs $\hat{\gamma}$: a choice-preserving addition can only weakly lower value, for any $\hat{\gamma}$.

A.2 Parametrized Example

Suppose $S = \{s_1, s_2\}$, $p = \text{P}\{s = s_1\}$, and $c_s(e) = \frac{1}{1000}e^s$, with s governing cost convexity. We set $s_1 = 2$, $s_2 = 1.6$. Table A.1 shows how e_1, $\mathcal{E}^+$, $\mathcal{E}^-$, and e_2 vary with p and the temptation parameter γ, assuming full sophistication ($\hat{\gamma} = \gamma$).

In the absence of uncertainty ($p \in \{0, 1\}$) and temptation ($\gamma = 1$), the DM simply chooses the effort level $e^* = e_1 = e_2$ that maximizes U_s under the unique state. If $s = s_1$, the effort cost is $c_{s_1}(e) = \frac{1}{1000}e^2$, yielding $e^*(s_1) = 30$; if $s = s_2$, the cost is $c_{s_2}(e) = \frac{1}{1000}e^{1.6}$ and $e^*(s_2) = 70$. Because $p \in \{0, 1\}$, the DM has no preference for flexibility ($\mathcal{E}^+ = \{e_1\}$), and since $\gamma = 1$, no preference for commitment ($\mathcal{E}^- = \mathcal{E}$). When $p = 0.5$, the DM has flexibility motives, $\mathcal{E}^+ = \{30, 70\}$; in addition, when forced to select at t_1, the DM chooses $e_1 = 40$, a compromise between 30 and 70.[5]

Now consider temptation. When $\gamma = 3$, the DM overweights the effort cost, choosing $e_2(s_1) = 20$ and $e_2(s_2) = 60$, i.e., a downward deviation by 10 tasks in both states relative to $\gamma = 1$. To mitigate this, the DM removes lower effort levels from $\mathcal{E}^-$. When $\gamma = 0.05$, the DM underweights the effort cost, choosing $e_2(s_1) = 50$ and $e_2(s_2) = 80$, i.e., upward deviations of 20 and 10 tasks, respectively. The DM now removes higher effort levels. For values $\gamma \in \{0.9, 1.7\}$ (i.e., close enough to 1), commitment demand is weaker (fewer options removed) and t_2 choices exhibit no temptation-driven deviation in either state. When uncertainty is low ($p = 0.9$), the DM favors at t_1 the effort level most likely to be chosen at t_2, and choices are time consistent with high probability.

[5] More generally, e_1 need not be optimal under any state at t_2: it may act as a compromise that minimizes expected losses from the inability to tailor choices to the state. For instance, with $S = \{s_1, s_2\}$, $e = 10$ is optimal under s_1 and $e = 90$ is optimal under s_2, the DM chooses $e_1 = 50$ provided $\frac{U_{s_1}(10) - U_{s_1}(50)}{U_{s_2}(50) - U_{s_2}(10)} < \frac{1-p}{p} < \frac{U_{s_1}(50) - U_{s_1}(90)}{U_{s_2}(90) - U_{s_2}(50)}$, and then deviates at t_2 with probability one.

Table A.1: Chosen menus and effort for different parameter values

Parameter values		Menu and effort choices			
p	γ	e_1	$\mathcal{E}^+$	$\mathcal{E}^-$	e_2 (from $\mathcal{E}$)
1	1	30	$\{30\}$	$\mathcal{E}$	30
0	1	70	$\{70\}$	$\mathcal{E}$	70
0.5	1	40	$\{30, 70\}$	$\mathcal{E}$	30 w.p. 0.5, else 70
1	3	30	$\{30\}$	$\mathcal{E} \setminus \{10, 20\}$	20
0	3	70	$\{70\}$	$\mathcal{E} \setminus \{30, 40, 50, 60\}$	60
0.5	3	40	$\{30, 60\}$	$\mathcal{E} \setminus \{10, 20, 40, 50\}$	20 w.p. 0.5, else 60
1	0.05	30	$\{30\}$	$\{10, 20, 30\}$	50
0	0.05	70	$\{70\}$	$\mathcal{E} \setminus \{80, 90, 100\}$	80
0.5	0.05	40	$\{40\}$	$\{10, 20, 30, 40\}$	50 w.p. 0.5, else 80
1	1.7	30	$\{30\}$	$\mathcal{E} \setminus \{20\}$	30
0	1.7	70	$\{70\}$	$\mathcal{E} \setminus \{60\}$	70
0.9	1.7	30	$\{30, 70\}$	$\mathcal{E} \setminus \{20, 60\}$	30 w.p. 0.9, else 70
1	0.9	30	$\{30\}$	$\mathcal{E}$	30
0	0.9	70	$\{70\}$	$\mathcal{E} \setminus \{80\}$	70
0.9	0.9	30	$\{30, 70\}$	$\mathcal{E} \setminus \{80\}$	30 w.p. 0.9, else 70

Notes: Because some options do not affect the DM, the menu maximizing W need not be unique. Among maximizers, $\mathcal{E}^+$ is the smallest menu and $\mathcal{E}^-$ the largest, since adding or removing options is costly. These menus are optimal when the 1-cent per-option cost is sufficiently small relative to the benefit; otherwise, $\mathcal{E}^+ = \{e_1\}$ and $\mathcal{E}^- = \mathcal{E}$.

Importantly, ĉ-COMMIT ($e \notin \mathcal{E}^-$ implies $\hat{f}(e) > 0$) is violated for some $e \notin \mathcal{E}^-$ whenever $\gamma \neq 1$, and $\mathcal{E}^-$ is support-preserving in several cases, e.g., $\gamma = 1.7$, $p = 0.9$, where both removed options (20, 60) have zero probability. Similarly, ĉ-FLEX ($e \in \mathcal{E}^+$ implies $\hat{f}(e) > 0$) is violated when temptation is strong enough, e.g., at $p = 0.5$, $\gamma = 3$, the DM includes $e = 30$ in $\mathcal{E}^+$ yet expects $e_2 = 20$ under s_1 (due to temptation), so $\hat{f}(30) = 0$ despite $30 \in \mathcal{E}^+$.

Uncertainty interacts with commitment demand. When the DM expects different choices in different states, the tempting options change with the state; as a result, the number of excluded options may increase with uncertainty, i.e., $|\mathcal{E}^-|$ may decrease. This can be seen by comparing $p \in \{0, 1\}$ vs. $p = 0.9$ when $\gamma = 1.7$ in Table A.1. However, this is not a general feature: when $\gamma = 0.05$, $|\mathcal{E}^-|$ is *higher* when $p = 0.5$ than when $p = 1$, because temptation and flexibility concerns conflict.[6]

[6] In this case, $\mathcal{E}^+ = \{40\}$ when $p = 0.5$, illustrating the tension.

B Piloting, Pre-Analysis Plan, and Sampling

B.1 Piloting

The final design emerged from a series of iterations between 2020 and 2023, which we evaluated through small-scale pilots. Our initial design, developed in 2020–21, was intended to support structural estimation and closely followed existing designs in the literature, with three effort dates, variation in piece rates, and labor supply chosen on a fine grid. Piloting revealed that this design was overly complex and generated both occasional non-monotonic responses to piece rates and substantial bunching at corner choices.

We changed direction in 2022–23, adopting a simpler design with two decision dates and a non-parametric approach. At the same time, we enriched the data through the elicitation of probabilistic beliefs and preferences over customized menus. We discretized the effort choice set to permit elicitation of a full belief distribution, replaced linear piece rates with a single concave wage schedule to reduce corner choices, and introduced commitment and flexibility decisions. We collected pilot data to refine the belief and menu-customization measures, minimize measurement error, and ensure sufficient variation in effort choices. Our earlier survey designs are available upon request.

The pre-analysis plan was developed using a pilot conducted in November 2023 on Prolific ($N = 81$ Survey 1; $N = 71$ Survey 2; 12% attrition). The pilot closely matched the final design, except that Survey 2 elicited choices directly from a single menu rather than using the strategy method across all three menus. We used it to estimate attrition, calibrate payments based on survey duration, and identify misunderstandings in the customization decisions. The resulting data were then used to prepare the populated PAP and analysis code posted on OSF. The data from this pilot can also be found on our OSF page.

B.2 Implementation of Pre-Analysis Plan and Deviations

We implemented our PAP with a limited number of omissions, modifications, and additions. The fully populated PAP is available here: https://osf.io/79x6v/overview.

Omitted analyses Some pre-registered analyses were not implemented or are not reported. First, we pre-registered sensitivity bounds for attrition assuming all (or no) attrited participants reversed; this was not triggered, as attrition remained low (11%). Second, we pre-registered robustness checks excluding participants who reported confusion about the prediction task; this question was inadvertently omitted from Survey 2. Third, replacing

support size with normalized entropy ($\rho = 0.873$) yields nearly identical results. Finally, we omit figures comparing predicted and actual reversal probabilities, which relied on small bins and revealed no systematic patterns.

Modifications to pre-registered analyses First, we extend the "behavior drawn from beliefs" (BDB) benchmark, which the PAP used once as a summary statistic (a point-biserial correlation between $\hat{f}(e_1)$ and time consistency). We define it explicitly in Section 3.3 and use it throughout Table 3 and Figures 6b and 8a. Second, in Figures 7 and 8b, the PAP mapped each belief distribution to a single draw, so predicted and realized outcomes could be compared with two-sample proportions tests. We instead take the perceived probability of each reversal type to be the mass $\hat{f}$ assigns to it and test it against realized behavior with a paired t-test; empirical frequencies are compared with point predictions by McNemar's exact test, as is commitment versus flexibility take-up in Section 4.4. The added power matters: participants' overstatement of upward reversals is significant in both figures ($p = 0.029$ and $p = 0.007$) but not under the pre-registered comparison ($p = 0.294$ and $p = 0.225$). The same substitution applies to the calibration tests in Figure 9a: the PAP compared realized frequencies with frequencies from a single belief draw, whereas we test, at each effort level, the mean respondent-level difference between assigned probability and realized choice with robust standard errors. No other conclusion changes. Finally, the forecasting survey contained a wording error: forecasters compared belief draws to realized effort e_2 rather than planned effort e_1, so this forecast is not used.

Unregistered analyses First, we place greater emphasis on misspecification, which emerged as an important feature of the data. We characterize participants who assign zero probability to their realized choice (Section D.2), report key results with and without them (Figures 8b and 9), and show how probabilistic beliefs alter inferences about naivety (Figure 12). Second, we revisit the interpretation of commitment demand. We develop theoretical results that sharpen when customization choices violate consequentialism and what removals imply for $\hat{\gamma}$ (Appendix A.1), and relate these diagnostics to anticipated behavior (Figure 13; Tables E.5, E.6, and E.7). We also exploit choices from all three menus, $\mathcal{E}$, $\mathcal{E}^+$, and $\mathcal{E}^-$, to test for WARP violations, which can arise under convex self-control costs and allow removals to be consequentialist even when the removed option has zero choice probability (Appendix E.4). Third, we introduce a new application of the BDB framework to measure the information content of beliefs (Table C.2).

B.3 Sample Size and Attrition

We targeted a final sample of 1,000 participants (700 belief treatment, 300 control), providing 80% power to detect a 9 percentage point difference in downward reversals from a baseline of 30%. Based on pilot data (Section B.1), we anticipated 10–15% attrition and set a recruitment target of 1,150 completed Survey 1 responses.

A total of 1,442 participants began Survey 1. We excluded 3% who attempted the survey multiple times and 247 (17.7%) who did not complete it (of whom 81 failed a comprehension check).[7] This yielded 1,149 completers (803 belief, 346 control), all invited to Survey 2. Of these, 1,025 completed Survey 2 (722 belief, 303 control), after excluding 2% with duplicate attempts and 3% non-completers.[8] The overall attrition rate was 11% (10% belief, 12% control), at the lower bound of our prediction interval.

As pre-registered, we examined whether attrition was associated with belief uncertainty, anticipated reversals, and demand for flexibility or commitment. As reported in Table B.2, we find virtually no systematic differences between attriters and non-attriters.

Table B.2: Attrition: effort and beliefs

	S1 only	Both	Diff	p-value
A. Effort				
e_1	58.4	53.0	5.4	0.066
B. Beliefs				
$\hat{e}_2$	58.4	52.9	5.5	0.063
$\lvert\text{supp}(\hat{f})\rvert$	3.02	2.85	0.17	0.505
$\text{entropy}(\hat{f})$	0.29	0.30	-0.01	0.814
$1-\hat{f}(e_1)$	0.38	0.38	0.00	0.833
C. Commitment and flexibility				
Added any	0.43	0.50	-0.07	0.258
Number added (if any)	2.5	2.2	0.3	0.245
Removed any	0.37	0.40	-0.03	0.568
Number removed (if any)	3.8	4.3	-0.5	0.319
N	81	722		

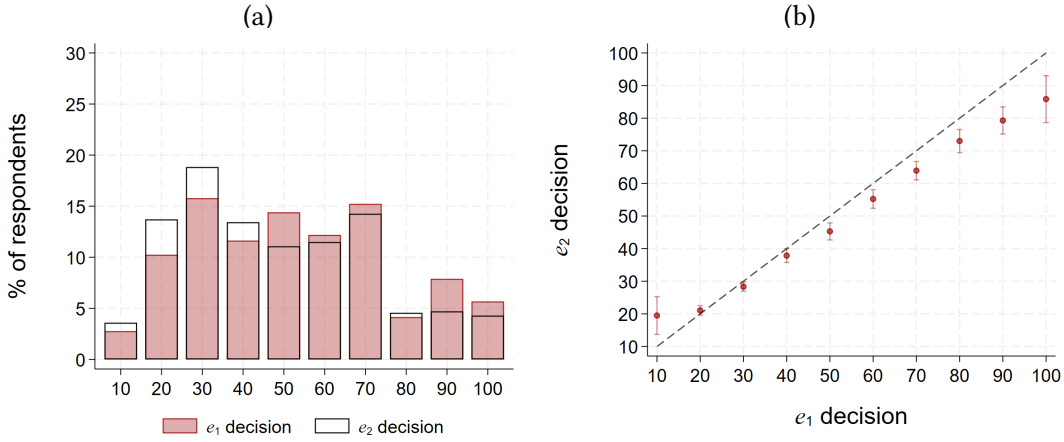
Notes: "S1 only": completed Survey 1 but did not return. "Both": completed both surveys. p-values from unpaired, unequal-variance t-tests. $\text{entropy}(\hat{f}) = -\frac{1}{\ln 10}\sum_{j=1}^{10}\hat{f}(e_j)\ln\hat{f}(e_j)$.

[7] Multiple attempts likely reflect device switching or technical issues rather than strategic behavior. We dropped all such participants because randomization occurred at the survey level.

[8] Most non-completers of Survey 2 ($N = 29$) appear to have dropped out during the slider task.

C Reversals, Point Predictions and Belief Uncertainty

Figure C.1: Effort decisions and choice reversals



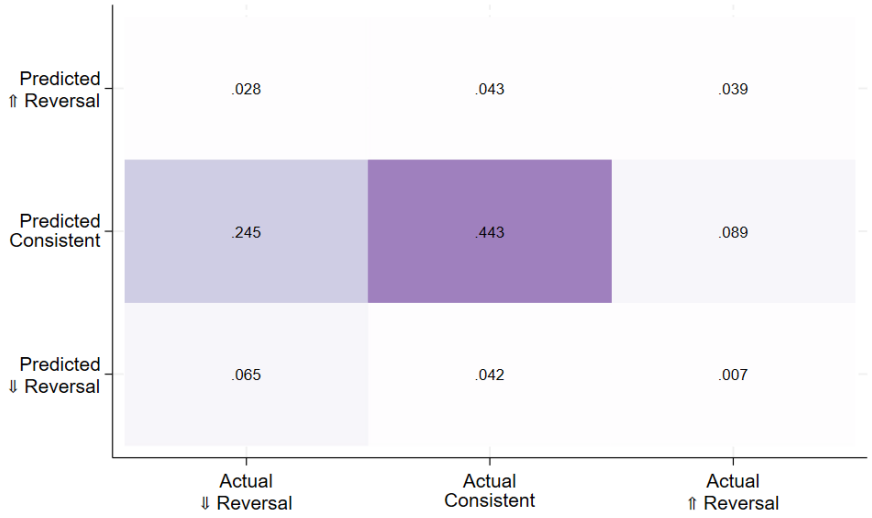
Notes: (a) Hist. of t_1 and t_2 effort, e_1 and e_2. (b) Mean of e_2 by e_1, with 95% CIs and dashed 45° line. $N = 722$.

Table C.1: Effort, Point Predictions and Belief Distributions

	Mean	SD	Min	Q25	Q50	Q75	Max	Diff vs. e_2	p-value
A. Effort									
e_1	53.0	24.1	10.0	30.0	50.0	70.0	100.0	-4.3	0.000
e_2	48.7	23.7	10.0	30.0	50.0	70.0	100.0	–	–
B. Predictors									
$\hat{e}_2$	52.9	23.9	10.0	30.0	50.0	70.0	100.0	-4.2	0.000
$E_{\hat{f}}(e)$	51.8	22.5	10.0	33.0	50.0	69.1	100.0	-3.1	0.000
$\text{mode}_{\hat{f}}(e)$	52.5	23.8	10.0	30.0	50.0	70.0	100.0	-3.8	0.000
$\hat{f}^{-1}(0.5)$	52.8	23.4	10.0	30.0	50.0	70.0	100.0	-4.1	0.000
C. Uncertainty									
$\text{sd}(\hat{f})$	6.19	5.56	0.00	3.00	5.36	8.05	36.07	–	–
$\text{entropy}(\hat{f})$	0.30	0.22	0.00	0.14	0.30	0.45	1.00	–	–
$\lvert\text{supp}(\hat{f})\rvert$	2.85	1.75	1.00	2.00	3.00	3.00	10.00	–	–
$\text{FP-skew}(\hat{f})$	-0.01	1.16	-9.85	-0.36	0.00	0.17	9.85	–	–
D. Probability mass at reference effort levels									
$\hat{f}(e_1)$	62.5	29.4	0.0	43.0	61.0	89.0	100.0	–	–
$\hat{f}(\hat{e}_2)$	62.5	29.4	0.0	43.0	61.0	89.0	100.0	–	–
$\hat{f}(e_2)$	44.4	36.2	0.0	10.0	40.0	80.0	100.0	–	–

Notes: Panel B compares predictors from point and probabilistic beliefs with realized effort, e_2; p-values are from Wilcoxon matched-pairs tests. Panels C–D summarize $\hat{f}$. Entropy is $-\frac{1}{\ln 10}\sum_j \hat{f}(e_j)\ln\hat{f}(e_j)$; FP-Skew denotes Fisher–Pearson skewness. $N = 722$.

Figure C.2: Heatmap of predicted vs. actual reversals



Notes: Heatmap of the estimated joint distribution $p(\widehat{R} = j, R = k)$, $j, k \in \{-1, 0, 1\}$, where -1, 0, and 1 denote downward reversal, consistency, and upward reversal, respectively. Predicted categories compare $\hat{e}_2$ with e_1; actual categories compare e_2 with e_1. Cell values are frequencies. $N = 722$.

Table C.2: Informativeness of respondents' claims about reversals

	Base rate	Posterior	Δ (pp)	Likelihood ratio
A. Point prediction vs. actual behavior				
Predicted Consistent	52.8%	57.0%	+ 4.2	1.19
Predicted Upward	13.4%	35.4%	+22.0	3.54
Predicted Downward	33.8%	57.3%	+23.5	2.63
B. Probabilistic beliefs vs. actual behavior				
Predicted Consistent	52.8%	57.9%	+ 5.1	1.23
Predicted Upward	13.4%	25.6%	+12.2	2.21
Predicted Downward	33.8%	50.9%	+17.1	2.03
C. Probabilistic beliefs vs. the BDB benchmark				
Predicted Consistent	62.5%	76.3%	+13.8	1.93
Predicted Upward	16.4%	49.0%	+32.6	4.91
Predicted Downward	21.1%	53.1%	+32.0	4.22

Notes: "Downward" means $x < e_1$, "Consistent" $x = e_1$, and "Upward" $x > e_1$, where e_1 is the effort chosen at t_1 and x the relevant effort variable. The prediction is the point forecast $\hat{e}_2$ (Panel A) or the belief distribution $\hat{f}(e)$ (Panels B and C); actual behavior is the realized choice e_2 (Panels A and B) or behavior drawn from beliefs, $\tilde{e}_2 \sim \hat{f}(\cdot)$ (Panel C, the BDB benchmark, computed in closed form). Base rate is the unconditional probability of each type (e.g., $P(e_2 < e_1)$ in Panels A and B; the mean belief mass on each type in Panel C). Posterior is the conditional probability that the actual type matches the predicted type (e.g., $P(e_2 < e_1 \mid \hat{e}_2 < e_1)$). Δ (pp) is Posterior minus Base rate, in percentage points. Likelihood ratio is $\text{LR} = \frac{P(\hat{e}_2 < e_1 | e_2 < e_1)}{P(\hat{e}_2 < e_1 | e_2 \geq e_1)}$; in Panels B and C, $P(\hat{e}_2 < e_1 \mid \cdot)$ is replaced by the average belief mass on the predicted type within each conditioning group. Values near 1 indicate little information, 2–4 moderate, and > 4 strong evidence.

D Belief Elicitation: Validity and Behavioral Effects

D.1 Role of Measurement Error and Robustness

As described in Section 2.2, probabilistic beliefs were elicited in two steps: selecting possible effort levels (Step 1) and assigning probabilities (Step 2). Consistency between the two steps is high (87%; see Section 5.1). Table D.3 examines robustness to alternative definitions of belief support and sample restrictions. Results are robust to (i) using Step 1 rather than Step 2 support, (ii) restricting to respondents with matching supports, and (iii) restricting to those who report accuracy as their primary prediction motive (Section D.2).[9]

Difficulty of the elicitation exercise Participants reported moderate difficulty in forming beliefs. On a scale from 1 (easy) to 10 (difficult), mean difficulty was 4.10 for point predictions and 4.22 for probabilistic beliefs. Although statistically significant (Wilcoxon signed-rank test, $p = 0.007$), the difference is only 5% of a standard deviation. This suggests that forming predictions, rather than assigning probabilities, is the main source of difficulty. Participants rated the interface and instructions highly: on a 1–5 scale, 85.6% rated the interface 4 or 5 for simplicity, and 71.5% rated the instructions 4 or 5 for clarity.

Misspecification, measured using Step 2 beliefs, is largely unrelated to inconsistencies between Steps 1 and 2, perceived task difficulty, and interface ratings. Correlations are small and insignificant for any inconsistency ($\rho = 0.023$, $p = 0.539$), the number of inconsistencies ($\rho = -0.057$, $p = 0.125$), difficulty of the point prediction ($\rho = 0.051$, $p = 0.171$), and ratings of the interface ($\rho = -0.006$, $p = 0.864$) and instructions ($\rho = 0.011$, $p = 0.769$). The only significant relationship is a modest correlation with difficulty expressing a full belief distribution ($\rho = 0.079$, $p = 0.033$).

D.2 Characteristics and Motives of the Misspecified

Narrower, more confident beliefs Among participants who reverse, the misspecified report more concentrated beliefs (Figure D.3): a median probability on e_1 of 80% versus 50% ($p < 0.001$), and a median support size of 2 rather than 3 ($p < 0.001$). Misspecification therefore appears to reflect excessive confidence rather than broad uncertainty that happens to exclude e_2.

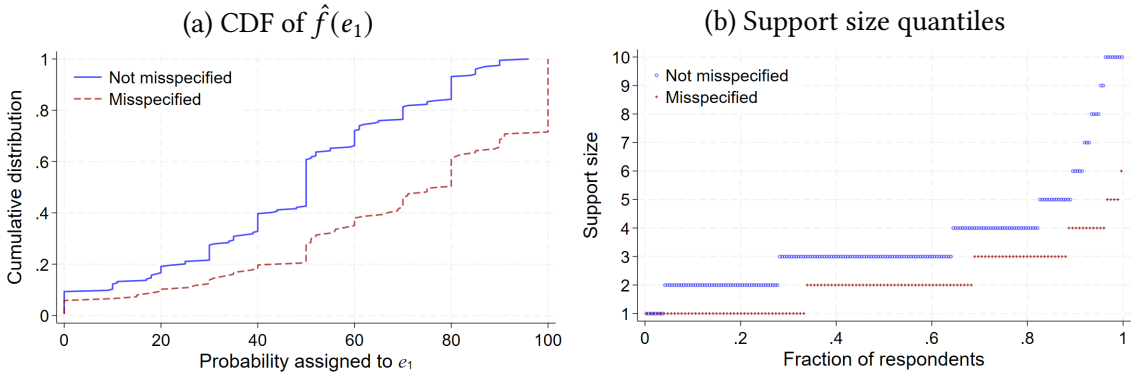
[9] Participants could also review a numerical summary of their entered beliefs and make adjustments. Only 2.4% changed at least one entry, typically through minor adjustments or rounding corrections.

Table D.3: Robustness of belief uncertainty, reversals, and customization choices

	Primary	Step 1	Consistent	Accuracy
A. Probabilistic beliefs and subjective uncertainty				
% with degenerate distribution	21.9%	30.1%	24.2%	19.9%
Mean # options in support	2.85	2.57	2.70	2.86
Median # options in support	3.0	2.0	3.0	3.0
% inclusion of $\hat{e}_2$	96.7%	91.6%	96.8%	97.2%
% inclusion of e_1	94.6%	89.9%	94.8%	95.0%
Consistent choice, Not uncertain (N)	65.8%	58.1%	66.0%	75.0%
Consistent choice, Uncertain (U)	49.1%	50.5%	50.8%	55.8%
p-*value* ($N = U$)	(< 0.001)	(0.074)	(0.001)	(0.001)
Upward reversal, Not uncertain (N)	8.9%	9.7%	9.2%	8.3%
Upward reversal, Uncertain (U)	14.7%	15.0%	15.5%	14.5%
p-*value* ($N = U$)	(0.064)	(0.057)	(0.060)	(0.154)
Downward reversal, Not uncertain (N)	25.3%	32.3%	24.8%	16.7%
Downward reversal, Uncertain (U)	36.2%	34.5%	33.7%	29.8%
p-*value* ($N = U$)	(0.010)	(0.607)	(0.046)	(0.019)
B. The structure of residual naivety				
Overall % misspecified	20.6%	24.8%	20.3%	15.8%
% misspecified among consistent choices	3.1%	6.3%	2.9%	2.8%
% misspecified among upward reversals	27.8%	28.9%	25.0%	21.4%
% misspecified among downward reversals	45.1%	52.0%	48.2%	41.7%
C. The belief foundations of commitment and flexibility				
Degenerate beliefs, Did not remove (NR)	23.4%	30.4%	25.8%	20.9%
Degenerate beliefs, Removed (R)	19.6%	29.6%	21.8%	18.3%
p-*value* ($NR = R$)	(0.234)	(0.869)	(0.255)	(0.537)
Degenerate beliefs, Did not add (NA)	33.4%	40.9%	37.3%	31.0%
Degenerate beliefs, Added (A)	10.3%	19.2%	10.9%	8.6%
p-*value* ($NA = A$)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
ĉ-FLEX violations	17.8%	22.5%	17.2%	16.3%
ĉ-COMMIT violations	95.6%	96.8%	96.8%	96.8%
c-FLEX violations	66.8%	–	65.4%	63.8%
c-COMMIT violations	98.3%	–	98.7%	98.6%

Notes: Columns report alternative definitions of beliefs or sample restrictions: Primary uses Step 2 beliefs ($N = 722$); Step 1 uses the coarser belief elicitation from Step 1 ($N = 722$); Consistent restricts the sample to respondents whose Step 1 and Step 2 belief supports coincide ($N = 631$); and Accuracy restricts the sample to respondents who report accuracy as their primary prediction motive ($N = 423$). All p-values are based on two-sided exact tests for equality of proportions (Fisher's exact tests).

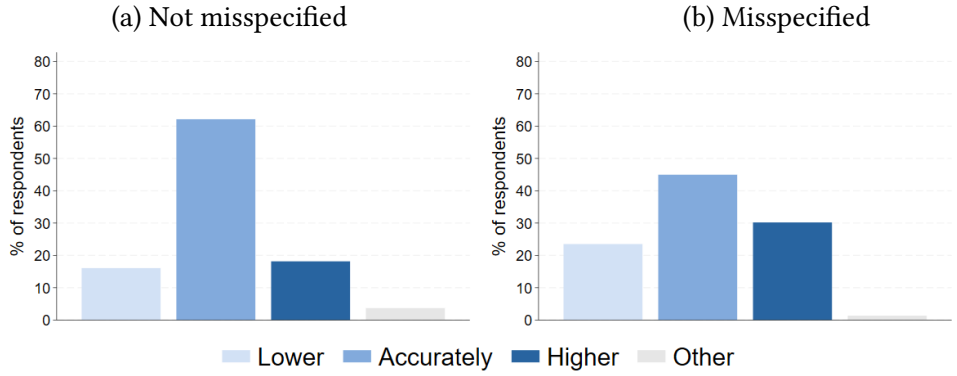
Figure D.3: Misspecification and belief confidence



Notes: Beliefs among those who reversed ($e_2 \neq e_1$), by misspecification status. Not misspecified reversals: $\hat{f}(e_2) > 0$ ($N = 204$); misspecified reversals: $\hat{f}(e_2) = 0$ ($N = 137$).

Prediction as a motivational target In Survey 2, participants were asked whether their Survey 1 predictions were set to be conservative (i.e., lower), accurate, or aspirational (i.e., higher). While a majority overall reported predicting as accurately as possible, Figure D.4 shows that the prediction strategy differs sharply by misspecification status. The misspecified are less likely to report predicting accurately and more likely to report using predictions as aspirational or conservative targets.

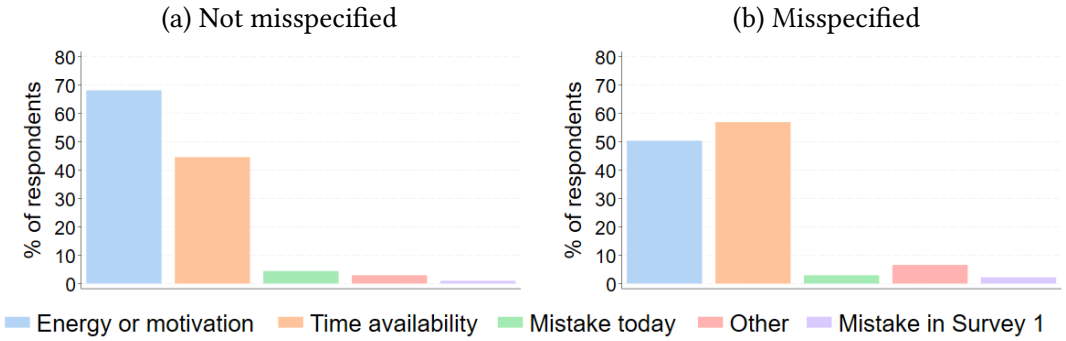
Figure D.4: Prediction strategy by misspecification status



Notes: Distribution of Survey 2 reports of motives for Survey 1 predictions: lower to ensure achievement, accurate, higher to set a goal, or other. (a) Not misspecified ($N = 573$). (b) Misspecified ($N = 149$).

Stated reasons for reversals Figure D.5 reports stated reasons for reversals, by misspecification status. Both groups overwhelmingly cite changes in energy, motivation, or time availability, and rarely report a mistake; among downward reversals, the misspecified cite time constraints more and energy less, consistent with binding external constraints

Figure D.5: Reported reasons for reversals at t_2



Notes: Stated reasons, multiple permitted, for revising effort choices between Surveys 1 and 2, among those who reversed choice. (a) not misspecified individuals ($N = 204$). (b) misspecified individuals ($N = 137$).

rather than shifts in internal motivation.

D.3 Does Belief Elicitation Affect Behavior?

Table D.4: Impact of belief elicitation

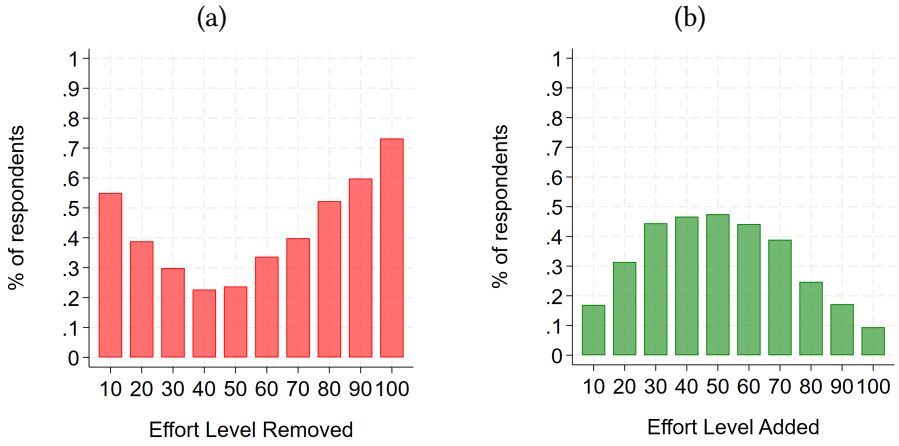
	No predictions	Predictions	Diff.	p-value
A. Effort and Reversals				
e_1	55.0	53.0	2.0	0.225
e_2	49.2	48.7	0.5	0.772
$\lvert e_2 - e_1 \rvert$	9.0	8.0	1.0	0.204
Consistent	0.51	0.53	-0.02	0.637
Upward reversal	0.12	0.13	-0.01	0.596
Downward reversal	0.37	0.34	0.03	0.383
B. Commitment/Flexibility				
Added any	0.36	0.50	-0.14	0.000
Number added (if any)	2.3	2.2	0.1	0.636
Removed any	0.44	0.40	0.04	0.245
Number removed (if any)	4.0	4.3	-0.3	0.299
N	303	722		

Notes: Columns compare participants who did not make any predictions to those who did. Entries are means; p-values are from unpaired unequal-variance t-tests. Reversal rates and effort distributions are similar across treatments (proportions test: $p = 0.637$; Kolmogorov–Smirnov tests for e_1 and e_2: $p = 0.576$ and 0.997). The e_1 comparison is a placebo test because beliefs were elicited after the initial choice.

E Nature and Interpretation of Customization Decisions

E.1 Nature and Direction of Customization Choices

Figure E.6: Proportion who chose to remove vs. add a given effort level



Notes: Proportion of participants who removed (a) or added (b) a given effort level, among those paying to remove ($N = 291$) or add ($N = 360$) at least one option.

Table E.5: Commitment removals and sign information

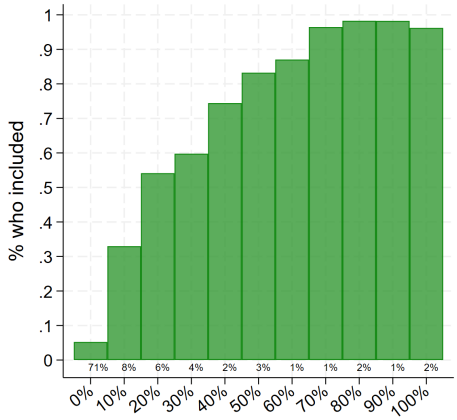
Classification	Condition on removals ($e \notin \mathcal{E}^-$)	N	%
$\hat{\gamma} < 1$: pulls toward more effort	None below $\text{supp}(\hat{f})$; at least one above[a] *e.g.*, $\text{supp}(\hat{f}) = \{30\}$, *removed* $\{100\}$	115	39.5
$\hat{\gamma} > 1$: pulls toward less effort	None above $\text{supp}(\hat{f})$; at least one below[a] *e.g.*, $\text{supp}(\hat{f}) = \{70, 100\}$, *removed* $\{60, 80\}$	65	22.3
Conflicting signs: no common $\hat{\gamma}$	Removals on both sides of $\text{supp}(\hat{f})$[a] *e.g.*, $\text{supp}(\hat{f}) = \{50, 60\}$, *removed* $\{10, 100\}$	91	31.3
Unsigned: no directional evidence	A removed option is itself expected (18); or every unexpected removal falls strictly between support points (2) *e.g.*, $\text{supp}(\hat{f}) = \{40\}$, *removed* $\{40\}$	20	6.9
All committers ($\mathcal{E}^- \subsetneq \mathcal{E}$)		291	100.0

Notes. $\text{supp}(\hat{f})$ is the full-menu belief support. Subjects removing an expected option are *Unsigned*, unless the exceptions in note a apply; otherwise location relative to $\min \text{supp}(\hat{f})$ and $\max \text{supp}(\hat{f})$ determines the row. [a] Includes cases signed via Appendix A.1, Prop. 2(b): (i) a removed endpoint (10 or 100) that is expected still signs if every *other* expected option is retained or shielded from it (e.g., $\text{supp}(\hat{f}) = \mathcal{E}$, removed only $\{100\}$); and (ii) an unexpected removal is still shielded in a menu that also removes an expected option (e.g., $\text{supp}(\hat{f}) = \{20, 30\}$, removed $\{10, 30\}$: 10 is shielded by retained 20). These move 10 subjects out of Unsigned (4 to $\hat{\gamma} < 1$, 5 to $\hat{\gamma} > 1$, and 1 signing both directions, hence Conflicting).

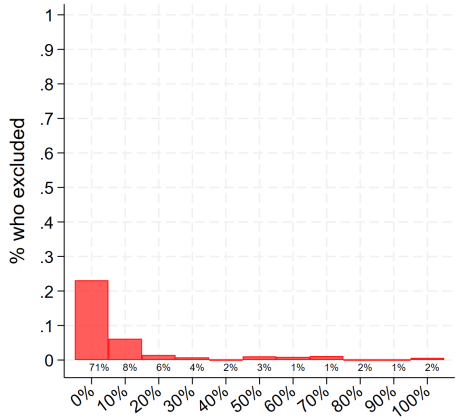
E.2 Belief-Based Tests of Consequentialism

Figure E.7: Predicted Choice Probabilities and Menu Construction

(a) Inclusion by predicted probability



(b) Removal by predicted probability



Notes: Share of effort levels included in (a) or removed from (b) the menu by predicted choice probability. Probabilities are binned in 10-percentage-point intervals (e.g., "20%" = [15,25]%), with 0% and 100% shown separately. Near-boundary predictions ([1,4]% and [96,99]%) are excluded (2% of all predictions). Numbers below bars give bin shares. $N = 722$ for added and $N = 291$ for removed.

Table E.6: Zero-probability rate among excluded and added options, by menu size

	Number of excluded options, $\lvert\mathcal{E} \setminus \mathcal{E}^-\rvert$									
	1	2	3	4	5	6	7	8	9	10
% $e \in \mathcal{E} \setminus \mathcal{E}^-$: $\hat{f}(e) = 0$	88.9	90.4	86.7	98.5	99.2	100.0	93.3	92.3	96.6	–
Subjects	54	26	35	33	48	39	30	13	13	–
Options	54	52	105	132	240	234	210	104	117	–
	Number of included options, $\lvert\mathcal{E}^+\rvert$									
	1	2	3	4	5	6	7	8	9	10
% $e \in \mathcal{E}^+$: $\hat{f}(e) = 0$	5.2	11.4	12.9	21.3	33.0	33.3	47.6	87.5	–	47.1
Subjects	362	136	124	54	20	8	9	2	–	7
Options	362	272	372	216	100	48	63	16	–	70

Notes. Top panel: excluded options ($e \notin \mathcal{E}^-$), among the 291 commitment demanders. Bottom panel: all options in the flexibility menu ($e \in \mathcal{E}^+$), among all 722 subjects. "Options" pools every excluded (included) option across all subjects with that exact menu size. "–" indicates no subject has that exact menu size.

Table E.7: Consequentialism diagnostics: commitment and flexibility demand ($N = 722$)

Category	Condition	N	%	Interpretation
A. Commitment: excluded options ($e \notin \mathcal{E}^-$)				
No exclusions	$\mathcal{E}^- = \mathcal{E}$	431	59.7	$\hat{\gamma} = 1$ not rejected
All expected	$\forall e \notin \mathcal{E}^- : \hat{f}(e) > 0$	13	1.8	$\hat{\gamma} \neq 1$; consistent with consequentialism
None expected	$\forall e \notin \mathcal{E}^- : \hat{f}(e) = 0$	263	36.4	Rejects consequentialism
Some expected	$\exists e, e' \notin \mathcal{E}^- : \hat{f}(e) > 0,\ \hat{f}(e') = 0$	15	2.1	Needs local checks
≥ 1 *shielded* [a]		6	0.8	Rejects consequentialism
None shielded		9	1.2	Ambiguous; compromise possible
B. Flexibility: included options ($e \in \mathcal{E}^+$)				
No extras	$\lvert\mathcal{E}^+\rvert = 1$ [b]	362	50.1	$\hat{\gamma} = 1$ not rejected
Certain	$\lvert \operatorname{supp}(\hat{f})\rvert = 1$	121	16.8	No uncertainty to hedge
Uncertain	$\lvert \operatorname{supp}(\hat{f})\rvert > 1$	241	33.4	Value below cost
All expected	$\forall e \in \mathcal{E}^+ : \hat{f}(e) > 0$	228	31.6	Consistent with consequentialism
None expected	$\forall e \in \mathcal{E}^+ : \hat{f}(e) = 0$	2	0.3	Ambiguous; compromise possible
Some expected	$\exists e, e' \in \mathcal{E}^+ : \hat{f}(e) > 0,\ \hat{f}(e') = 0$	130	18.0	Needs local checks
Support-covering	$\operatorname{supp}(\hat{f}) \subseteq \mathcal{E}^+$	108	15.0	Rejects consequentialism
Not covering	$\operatorname{supp}(\hat{f}) \not\subseteq \mathcal{E}^+$	22	3.0	Needs local checks
≥ 1 *shielded* [a]		13	1.8	Rejects consequentialism
None shielded		9	1.2	Ambiguous; compromise possible

[a] An unexpected removed (added) option r is *shielded* if some other retained (added) option q lies strictly between r and the belief support, e.g. $r < q \leq \min \operatorname{supp}(\hat{f})$ or $\max \operatorname{supp}(\hat{f}) \leq q < r$. See Appendix A.1, Prop. 1(ii) and its flexibility counterpart.
[b] Not necessarily $\{e_1\}$: 347 of the 362 have $\mathcal{E}^+ = \{e_1\}$; the remaining 15 kept a different single option (12 uncertain, 3 certain).

Notes: Each panel partitions the full sample; indented rows partition the row directly above. Flexibility rows include the free option; excluding it and counting only paid additions gives instead: 362 no extras, 239 all expected, 35 none expected, and 86 some expected.

E.3 Zero-Probability Customizations: Rounding or True Zeros?

A concern is that zero-probability options may reflect small probabilities rounded down when participants allocated 100 points. Since most reported distributions are single-peaked (86% of those with multiple support points), any mass that is rounded down should logically appear adjacent to the support. Adjacency sharply distinguishes additions from removals: 89% of zero-probability additions lie within one step of the support, compared with only 13% of removals.[10] Follow-up responses corroborate this reading (Section 5.2): the share saying they would not choose the zero-probability options they removed *increases* in the number removed ($p = 0.035$), contrary to what "running out of probability

[10] This difference is not driven by the extent of customization: among participants making a single zero-probability change, 82% of additions are adjacent, compared with 18% of removals ($p < 0.001$).

points" would predict, and is unrelated to location, with adjacent removals no more likely to draw "I might still choose it" than distant ones ($p = 0.59$). Zeros far from the support could instead reflect a small second mode associated with temptation, but disconnected supports are rare (6.6%). Moreover, realized behavior suggests that any concealed mass is small: removed options were ultimately chosen by 7% of commitment demanders, or 1.7% of the time on average per removed option (Figure 11).

E.4 Customization against Predicted and Realized Choices

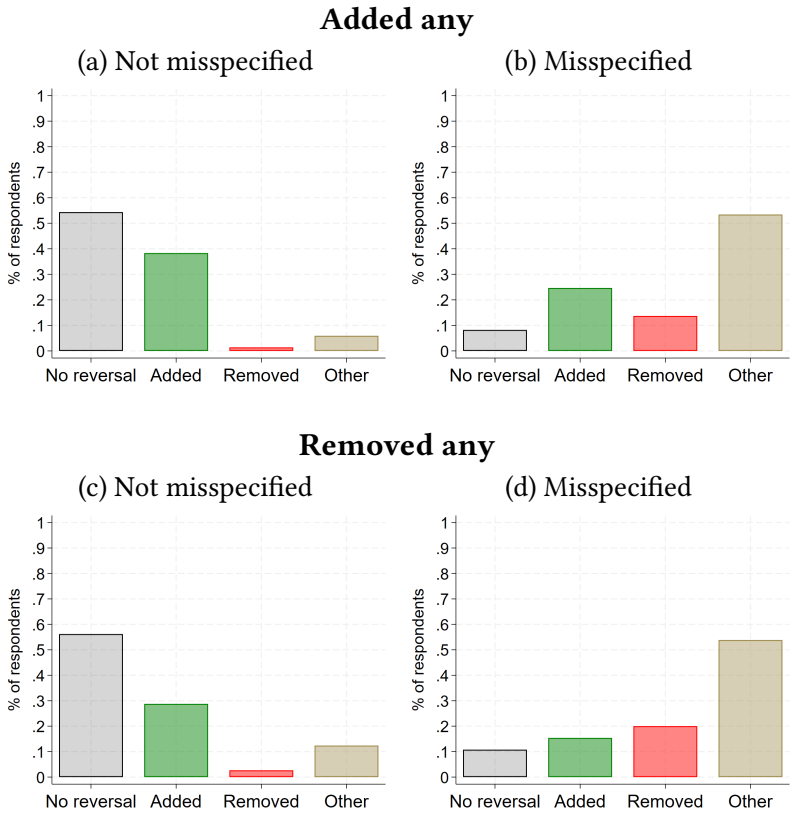
Table E.8: Reversals and preference for commitment/flexibility

	A. Predicted (point): $1\{\hat{e}_2 \neq e_1\}$						**B. Predicted (mass):** $\sum_{e \neq e_1} \hat{f}(e)$					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Added Any	0.16***		0.17***				0.14***		0.14***			
	(0.03)		(0.03)				(0.02)		(0.02)			
Removed Any		0.00	-0.04					0.04*	-0.00			
		(0.03)	(0.03)					(0.02)	(0.02)			
Number Added				0.04***		0.04***				0.04***		0.04***
				(0.01)		(0.01)				(0.01)		(0.01)
Number Removed					-0.01	-0.01*					-0.00	-0.00
					(0.01)	(0.01)					(0.00)	(0.00)
Constant	0.14***	0.22***	0.16***	0.18***	0.23***	0.19***	0.30***	0.36***	0.30***	0.33***	0.38***	0.34***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)

	C. Actual: $1\{e_2 \neq e_1\}$						**D. Size:** $\lvert e_2 - e_1 \rvert$					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Added Any	0.15***		0.14***				2.04**		1.54*			
	(0.04)		(0.04)				(0.86)		(0.88)			
Removed Any		0.11***	0.08**					2.35***	1.94**			
		(0.04)	(0.04)					(0.90)	(0.92)			
Number Added				0.05***		0.04***				0.79**		0.77**
				(0.01)		(0.01)				(0.32)		(0.32)
Number Removed					0.02***	0.02**					0.13	0.05
					(0.01)	(0.01)					(0.14)	(0.14)
Constant	0.40***	0.43***	0.37***	0.42***	0.44***	0.40***	6.96***	7.03***	6.42***	7.11***	7.76***	7.03***
	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.62)	(0.53)	(0.66)	(0.55)	(0.52)	(0.60)

Notes: Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Columns (1)–(6) and (7)–(12) are the same six specifications. All specifications have $N = 722$ observations.

Figure E.8: Does e_2 correspond to a removed or added effort level?



Notes: Bars show shares with no reversal or reversals to an added, removed, or other option. Panels (a)–(b) restrict to flexibility demanders: (a) not misspecified ($N = 287$), (b) misspecified ($N = 73$). Panels (c)–(d) restrict to commitment demanders: (c) not misspecified ($N = 226$), (d) misspecified ($N = 65$).

Choice changes and WARP violations Recall that participants chose their effort first from the full menu $\mathcal{E}$ and then from the commitment and flexibility menus, $\mathcal{E}^-$ and $\mathcal{E}^+$, presented in randomized order. A departure from e_2, their full-menu choice, was forced when e_2 was unavailable in the customized menu; when it remained available, choosing differently constituted an individual-level WARP violation. Table E.9 decomposes these cases. Among the 291 participants who removed options, 7.2% could not repeat e_2 and 8.9% of those who could violated WARP. Among the 360 who added options, 19.7% could not repeat e_2 and 8.0% of those who could violated WARP.

Most violations appear to reflect baseline choice instability rather than an effect of menu restriction, such as changes in the salience or temptation value of particular options. Among the 431 non-removers, who faced the same menu twice ($\mathcal{E}^- = \mathcal{E}$), 12.8% changed their choice. Moreover, violation rates are far below matched random-choice benchmarks

Table E.9: Choice consistency across menus

	N	Constrained	Opportunity	(i) Violated WARP	(ii) Benchmark	(i)/(ii)
Commitment						
Did not remove	431	0.0%	431	12.8% (55)	90.0%	0.142
Removed	291	7.2% (21)	270	8.9% (24)	77.8%	0.114
Flexibility						
Did not add	362	40.3% (146)[a]	216	0.0%[b]	0.0%[b]	n/a
Added	360	19.7% (71)	289	8.0% (23)	65.1%	0.122
All subjects	722	2.5% (18)[c]	704	12.8%[d]	89.3%	0.143

Notes. "Constrained" = e_2 unavailable in the restricted menu. (i) = violation rate among those with opportunity. (ii) = matched random-choice benchmark, using each subject's actual menu size. [a] Reflects $e_2 \neq e_1$, not customization. [b] Structurally zero: singleton menu. [c] Unavailable in both menus. [d] Violated WARP in at least one menu where e_2 was available.

and are similar across groups after accounting for menu size. Violators instead show signs of local indecision: they spread probability more evenly across two adjacent effort levels, whereas non-violators concentrate it on one,[11] and report greater difficulty choosing at t_2 (3.63 versus 3.85 on a scale from 1, very difficult, to 5, very easy; $p = 0.060$). Overall, WARP violations are rare, offering little evidence of convex—or broader menu-dependent—self-control (Fudenberg and Levine, 2011; Noor and Takeoka, 2010, 2015). Though consistent with some indecision, they cannot explain observed reversals, suggesting that uncertainty concerns future states rather than unstable rankings at the time of choice (Section E.6).

E.5 Reliability of Customization Decisions

Customization decisions were implemented with 20% probability, potentially weakening incentives. A minority (22%) reported exerting little effort.[12] Greater engagement is associated with stronger demand: among low-effort participants, 16% removed and 30% added at least one option, compared with 46% and 56% among the rest ($p < 0.001$ in both cases). If anything, the low implementation probability therefore appears to understate demand for commitment and flexibility. The presentation order was also randomized. Table E.10 shows that commitment demand is higher on both margins when $\mathcal{E}^-$ appears first; flexibility demand is also higher, significantly so on the intensive margin.

[11] Violators assign less probability to each realized effort individually (31% versus 47%, p<0.001), but more total mass to the pair (62% versus 47%, p<0.001), even when the commitment menu coincides with the full menu. Entropy and support size are similar across groups, suggesting that violators are not more uncertain overall but divide their beliefs more evenly between two adjacent options.

[12] Among those reporting low effort, 86% considered a 2-in-10 chance small, while 14% did not. Among the 78% reporting some effort, the corresponding shares were 79% and 21%.

Table E.10: Order effects on flexibility and commitment decisions

	$\mathcal{E}^-$ First	$\mathcal{E}^+$ First	Diff.	p-value
A. Extensive margin				
Added any	0.52	0.48	0.04	0.265
Removed any	0.44	0.36	0.08	0.030
B. Intensive margin				
Number added	2.47	1.95	0.52	0.001
Number removed	4.52	4.02	0.50	0.076

Notes: Order effects by the randomized sequence of the two customization screens. Panel A reports the share who customized; Panel B conditions on having done so, and "Number added" excludes the free first option. p-values from two-sided unequal-variance t-tests. $N = 354$ when the commitment menu ($\mathcal{E}^-$) was presented first, and $N = 368$ when the flexibility menu ($\mathcal{E}^+$) was presented first.

E.6 Self-Reported Motives and Sources of Uncertainty

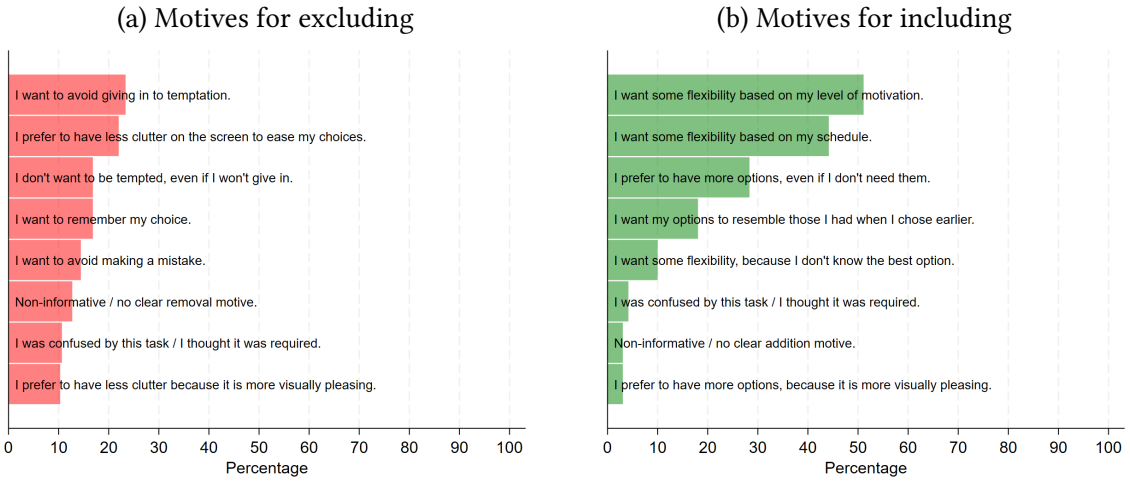
Motives for choice restrictions At the end of Survey 1, participants who removed at least one option were shown their removal choice(s) and asked to select all applicable motives from a structured list.[13]

Figure E.9a presents the results. The most common motive is "I want to avoid giving in to temptation," selected by 23% of removers. A distinct but related motive, "I don't want to be tempted, even if I won't give in", intended to capture pure self-control costs, was selected by 17%. Overall, 37% cited at least one temptation-related motive and only 3% selected both. Temptation motives are concentrated among participants whose removals are entirely non-consequentialist (40%) and nearly absent among those with at least one consequentialist removal (7%); confusion-related motives are less common (11%).

A natural question is why many participants selected a motive implying they might succumb to temptation when their belief data assigns 0% to the removed option. Part of the answer may be inattention to the subtle distinction between the two temptation options (the more popular one appeared first). But some participants do appear to have recognized residual risk: among the 278 who removed an option assigned 0% probability, 23% reported they might still choose it, while 73% stated they would not choose it

[13]The list included nine options: six randomized, followed by two temptation motives ("I want to avoid giving in to temptation"; "I don't want to be tempted, even if I won't give in"), "I prefer to have less clutter on the screen to ease my choices," and an open-ended "other." Placing the temptation motives near the bottom, unrandomized and with "giving in" first, works against finding temptation motives in general and costly self-control motives in particular. The clutter option was inadvertently not randomized.

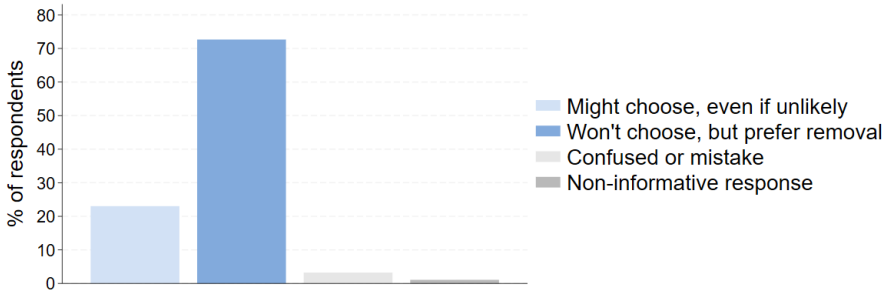
Figure E.9: Motives for menu customization



Notes: (a) motives for excluding options, among those who paid to remove at least one ($N = 291$); (b) motives for including options, among those who paid to add at least one ($N = 360$). In (a), the $N = 55$ "other" responses were classified using ChatGPT 5.2 Pro into two added categories: non-informative, and "reduce the chance of ending up with an undesired option," merged with "I want to avoid making a mistake" ($N = 3$). We reviewed all classifications and re-classified five. In (b), the $N = 12$ "other" responses were re-classified manually (11 non-informative, one confused). Details in the replication materials.

but preferred its removal (Figure E.10). Responses also differ by the specific temptation motive cited: among those who stated they would not "give in," 82% reported they would not choose a removed option assigned 0%, compared to 62% among those who said they "wanted to avoid giving in" (p-value: 0.029). Although based on smaller samples, this suggests the two temptation motives capture some distinct reasoning.

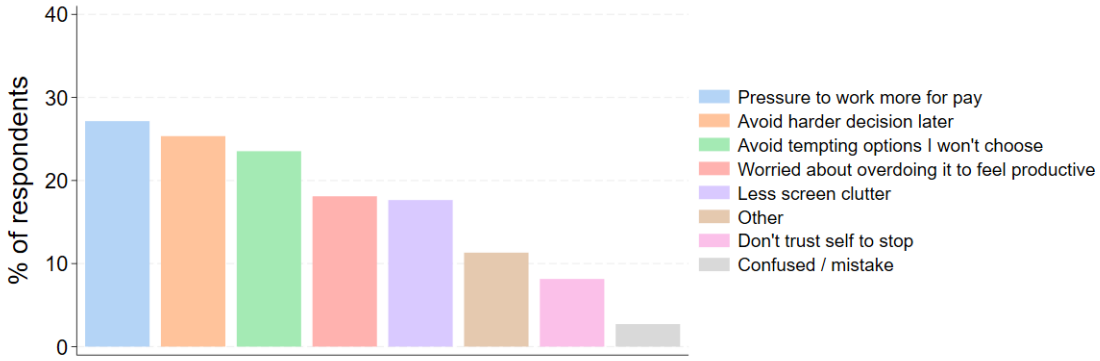
Figure E.10: Reason for removing an option assigned 0% probability



Notes: Responses to Survey 1 question on why participants chose to remove an option they had assigned 0% probability of choosing. The exact answer items were: "I might still choose it," "I know I won't choose it, but prefer to remove it anyway," or an open-ended "other". The $N = 22$ "other" responses were classified using GPT-5.4-mini into the two categories plus "confused or mistaken" and "non-informative"; we reviewed all of them and re-classified three (classifications are in the replication materials). Misunderstandings account for 3.2% of zero-probability removers. $N = 278$.

Heterogeneity by effort level and bidirectional temptation Commitment decisions frequently target high-effort options, not just low-effort ones. Among those who removed at least one option, temptation motives are common whether participants exclusively removed effort levels below e_1 (46%) or above e_1 (33%), though somewhat more prevalent in the former group ($p = 0.089$). To understand the latter group, we asked participants who removed an option above e_1 to explain their reasons (Figure E.11). Many cite temptation-related concerns: 24% wanted to avoid tempting options they would not ultimately choose, 27% worried that the bonus might push them to exert more effort than intended, and 18% felt they would be compelled to appear productive.

Figure E.11: Reasons for Removing Option to Do More Screens



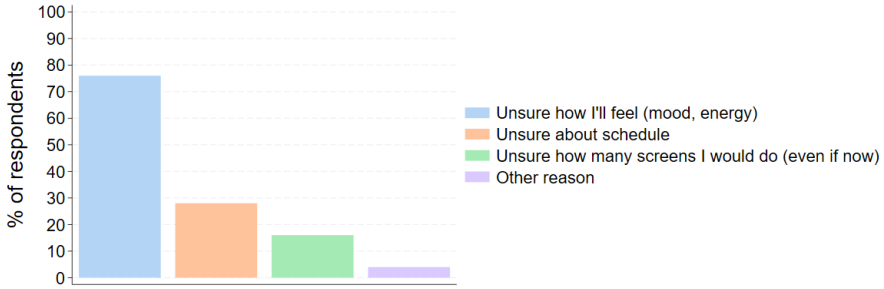
Notes: Reasons selected in Survey 2 by participants who removed higher effort options than their preferred e_1 in Survey 1; respondents could select multiple reasons. $N = 221$.

Motives for choice expansions As shown in Figure E.9b, participants who expand their choice set primarily offer consequentialist reasons, e.g., the desire to adjust decisions based on motivation or schedule. Non-consequentialist motives such as "I prefer to have more options, even if I don't need them" are less common but still chosen by 28% of respondents. Among those who added options, 51% selected only consequentialist motives, 18% selected only non-consequentialist motives, and 24% selected a mix. Reports of task confusion are rare, and we find little heterogeneity by effort level added.

Sources of uncertainty Figure E.12 reports stated sources of uncertainty from Survey 1, among participants who expressed uncertainty about their future effort choice (i.e., those who did not assign 100% probability to a single effort level). The vast majority report that uncertainty stems from how they will feel on the Survey 2 date (e.g., mood or energy; 76%). A smaller but non-trivial share cite uncertainty about their schedule (28%).

By contrast, relatively few attribute their uncertainty to contemporaneous preference uncertainty, namely being unsure how many screens they would choose even if they had to decide immediately (16%).[14] When asked in Survey 2 what best explained their earlier uncertainty about the effort level they ultimately chose, participants gave similar answers: the vast majority cited future unknowns such as mood, energy, or schedule, while reports of confusion, mistakes, or difficulty evaluating effort levels were rare (each below 6%).

Figure E.12: Sources of uncertainty about future effort (Survey 1)



Notes: Sources of uncertainty selected in Survey 1 by uncertain participants (those not assigning 100% probability to a single effort level); multiple selections allowed. $N = 564$.

[14] On a scale from 1 (very difficult) to 5 (very easy), the effort decision was perceived as moderately easy: the modal response was "somewhat easy" in both surveys, selected by roughly 44% (Survey 1) and 48% (Survey 2) of participants. Few participants rated the decision as very difficult (under 1% in both surveys).

F Expert Forecasts

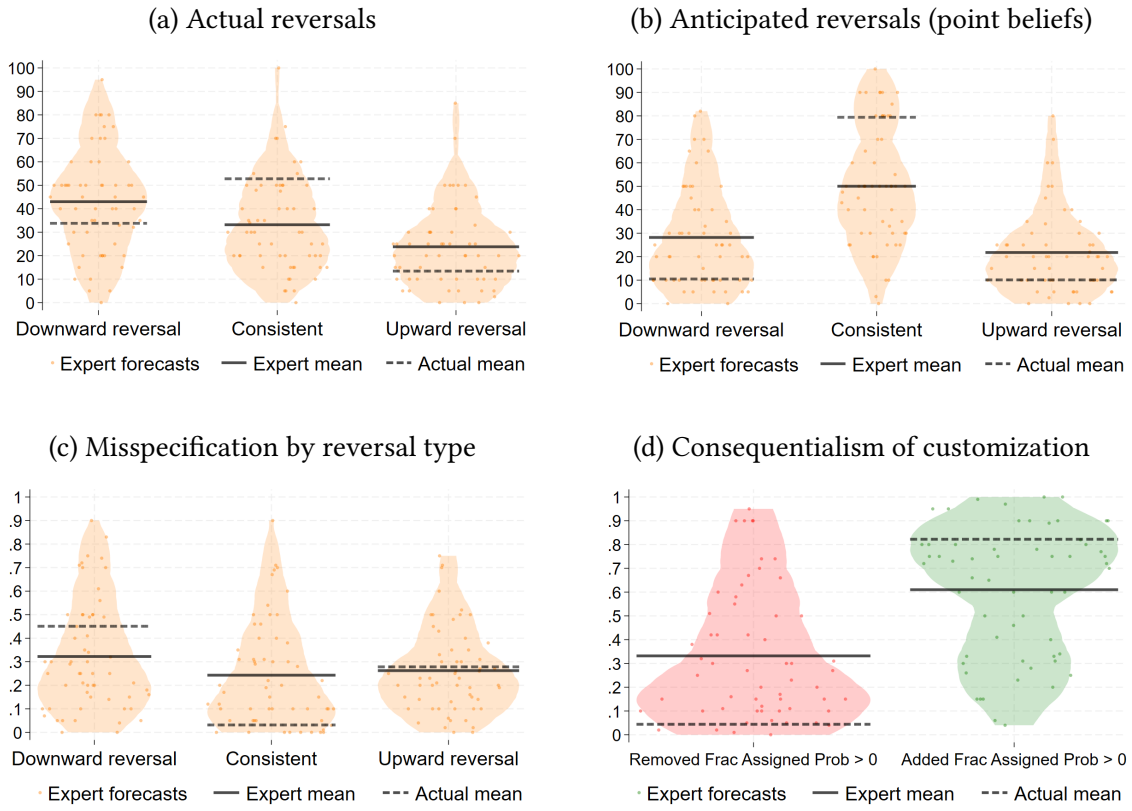
Table F.1: Actual vs. expert forecasts

	Data	Expert	Gap	Holm p
A. Choice reversals				
Downward reversal	0.338	0.430	0.092	< 0.001
No reversal	0.528	0.332	−0.196	< 0.001
Upward reversal	0.134	0.238	0.104	< 0.001
B. Anticipated reversals (point beliefs)				
Anticipated downward	0.105	0.282	0.177	< 0.001
Anticipated none	0.794	0.500	−0.294	< 0.001
Anticipated upward	0.101	0.218	0.117	< 0.001
C. Belief uncertainty				
Any uncertainty ($\lvert\text{supp}(\hat{f})\rvert > 1$)	0.781	0.671	−0.110	< 0.001
Mean support size	2.850	4.586	1.736	< 0.001
D. Reversal rate by uncertainty status				
Uncertain ($\lvert\text{supp}(\hat{f})\rvert > 1$)	0.509	0.588	0.079	< 0.001
Certain ($\lvert\text{supp}(\hat{f})\rvert = 1$)	0.342	0.446	0.104	0.032
E. Misspecification ($\hat{f}(e_2) = 0$) by reversal type				
Among downward reversals	0.451	0.322	−0.129	< 0.001
Among consistent choices	0.031	0.243	0.211	< 0.001
Among upward reversals	0.278	0.262	−0.016	0.718
F. Actual effort versus belief-distribution draw				
Effort < draw	0.342	0.389	0.047	0.032
Effort = draw	0.461	0.338	−0.123	< 0.001
Effort > draw	0.197	0.273	0.076	< 0.001
G. Menu customization				
Removed any option	0.403	0.333	−0.070	< 0.001
Added any option	0.499	0.459	−0.039	0.070
H. Consequentialism of customization				
Removed option with $\hat{f}(e) > 0$	0.044	0.332	0.288	< 0.001
Added option with $\hat{f}(e) > 0$	0.822	0.610	−0.212	< 0.001

Notes: $N = 722$. Gap = Expert − Data, so positive values indicate expert overestimation; Expert reports the mean forecast across $N = 58$ forecasters, treated as the population of interest (wisdom-of-crowds average). Each row tests H_0: the sample estimate equals this population mean, using a one-sample proportion test (dichotomous outcomes) or t-test (continuous outcomes); no sampling error is attributed to the forecast mean. Holm p-values adjust for multiple testing across all rows of the table.

Belief elicitation effect: Experts also forecast the significance level of a two-sided t-test for equality of mean effort e_2 between the belief and no-belief conditions. The actual p-value is 0.773 ($N_{\text{belief}} = 722$, $N_{\text{control}} = 303$). Expert forecasts: $p < 0.01$: 10.3% ; $p \in [0.01, 0.05)$: 37.9% ; $p \in [0.05, 0.10)$: 39.7% ; $p \geq 0.10$: 12.1% . Only 12.1% of experts correctly anticipated a null result.

Figure F.1: Distribution of expert forecasts



Notes: Each panel shows the distribution of individual expert forecasts ($N = 58$) as violin plots, with dots indicating jittered individual forecasts. The solid line denotes the expert mean. The dashed line indicates the realized value in the data ($N = 722$). Panels (a)–(b): experts substantially overestimated actual reversal rates (67% vs. 47%) and underestimated the anchoring of point predictions to e_1 (experts predicted 50% anticipating no reversal vs. 78% in the data). Panel (c): experts predicted similar misspecification rates across reversal types (≈25%), whereas in the data misspecification is concentrated among downward reversals (45%) and nearly absent among consistent choices (3%). Panel (d): experts predicted roughly one-third of removed options would carry positive choice probability, compared to only 4% in the data—the largest mechanism error in the forecast exercise.