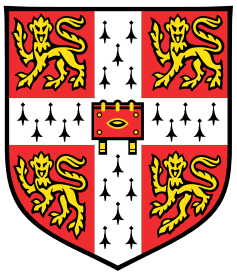
Empirical Tests of Non-standard Preferences for Beliefs and Information



Vivek Roy-Chowdhury

Faculty of Economics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Christ's College June 2024

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or, is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Abstract

This thesis consists of three chapters that conduct empirical tests of behavioural microeconomic theory. The first is a study of observational data, using reduced form and structural models, and the other two are based on laboratory experiments. The connecting thread between all three chapters is their focus on preferences for beliefs and information.

Chapter 1, *Biased Recall and The Dynamics of Beliefs: Evidence From Schools*, uses reduced form and structural methods to test theories of motivated beliefs in the field. These theories predict biased recall of ego-relevant information and thus provide a compelling explanation for persistent overconfidence in a variety of contexts. However, they have remained largely untested in the field until very recently. This chapter does so in a critical context for the formation of beliefs about ability: schooling.

Using within-individual variation, the analysis finds evidence of typical patterns of recall consistent with motivated beliefs theory: recall of grades is biased and becomes more biased as time passes, and biased recall is associated with overconfidence about future grades. A structural model is then estimated. The model allows for unobserved heterogeneity in tastes for biased recall, revealing large differences across types of students. The model is used to simulate the effects of an informational intervention correcting biases in recall, finding that it would reduce student welfare and widen existing attainment gaps.

Chapter 2, *Recall Distortion and Past Choices*, inherits Chapter 1's focus on recall and memory and marks the beginning of the part of the thesis that focuses on experiments. It is joint work with Rebecca Heath.

The basic question addressed by this chapter is whether beliefs are distorted to enable optimism about the consequences of past choices. The experiment finds support for this hypothesis and obtains suggestive evidence that belief distortions

increase after a few days due to memory loss. This has implications for the measurement of beliefs in many empirical settings, since it shows that beliefs can be endogenous to prior choices.

The analysis also shows that individuals distort beliefs in precisely the same way even when they know they will find out the truth at the end of the experiment, implying that the ability to sustain belief distortions indefinitely does not influence the present motive to distort beliefs.

Chapter 3, *Temptation to Consume Information*, was my job market paper. It makes several contributions to the growing literature on preferences and markets for information. In contrast to existing work, which focuses on when and why individuals avoid information, it instead explores whether individuals are *tempted* to consume information they would like to avoid. In doing so, it entails the first attempt to empirically examine whether preferences for information can be dynamically inconsistent. The analysis relies on an experiment, guided heavily by theory, to establish the prevalence of temptation for information. Almost half of all individuals, and 75% of those who want to avoid information, are classified as finding information tempting: undesirable but more difficult to resist at the time of access.

This chapter introduces a methodological innovation to the empirical study of temptation by showing theoretically that, when facing a tempting but resistible choice, *response time* increases in self-control costs. Using this method, the analysis shows that preferences for commitment are highly predictive of experienced self-control costs. The findings in this chapter suggest that the provision of commitment devices to avoid information could be welfare-improving. They also raise new questions on the feasibility of strategic information avoidance when such devices are unavailable. The chapter closes with a discussion of implications for important contemporary digital markets, including social media and news.

Acknowledgements

The three papers underpinning this thesis were immensely improved by my interactions with other researchers. My supervisor, Chris Rauh, gave me his time, his pragmatic advice on navigating the idiosyncrasies of academia, his flexibility and encouragement as I burrowed deeper into unfamiliar topics, and his patience with my repeated failure to add appropriate footnotes to figures and tables. My advisor, Julia Shvets, consistently devoted great effort to providing thoughtful feedback on my work. I learned a great deal from her methodical, rigorous approach. I'm also very grateful to have had the chance to travel and share my work with so many sharp, inspiring researchers across the world — in the US, Spain, Italy, and Germany.

I have been lucky to draw from a deep well of support from a wide circle of family and friends in the last four years. My parents made major sacrifices to give me the set of opportunities that have led me to this point. Many others graciously offered their services as experiment testers, sounding boards, and a welcome source of distraction. I especially want to mention my niece, Anya, her parents, Dinesh and Meera, and my friends, particularly Abheek, Constance, Jack, Josh, Lindsay, Sarah, and the two Toms, among many others. Finally, nobody else gave more of their time and effort to my thesis than my partner, Caroline. She proof-read countless drafts and listened carefully to hours of incoherent rambling on embryonic research proposals. Her generosity in shouldering some of the burden of my work has been so important.

My overall feeling is that writing this thesis has been enormously enriching. The liberty I've been given to exercise my intellectual curiosity has taken me in some unexpected directions, but it's fitting that the end result is a fairly narrow study of topics in behavioural economics. Over a decade ago, I was deeply taken with Kahneman's *Thinking, Fast and Slow*, and it was a pivotal factor in my decision to study economics. After exploring a few fields within economics, this doctorate has been my first opportunity to make proper contributions to the one that first grabbed my attention. It has not always been straightforward, but, as is true of most things, that is a big part of why it has been so enriching.

Table of contents

Chapter 1

Self-Confidence and Motivated Memory Loss: Evidence From Schools

1.1 Introduction

Beliefs about one's own characteristics are fundamental to a broad range of choices. However, a notable failure of the canonical model of belief updating in this domain is that it cannot accommodate the common observation that individuals are persistently overconfident. Overconfidence has economically important implications; it has been demonstrated to affect decisions in financial markets (De Bondt and Thaler, 1995), CEOs' investment allocations (Malmendier and Tate, 2005), personal health (Oster et al., 2013), and political behaviour (Ortoleva and Snowberg, 2015). A leading explanation of how overconfidence can functionally be sustained against persistent contradictory information comes from biased recall. In the context of self-image, biased recall captures the notion that favourable past information about oneself tends to come to mind more easily than unfavourable information.

This chapter exploits longitudinal data on teenagers' recall of their recent mathematics grades at school to examine the dynamic relationship between biased recall and beliefs. It therefore entails the first attempt to test whether motivated beliefs are important during schooling, a particularly sensitive time for the formation of an individual's beliefs about their ability. We begin by replicating recent empirical tests of theories of motivated beliefs with longitudinal field data: we test whether individuals recall favourable information about themselves more readily than unfavourable information; and whether a given individual is more likely to have

overconfident beliefs about future performance when they incorrectly recall high recent performance.[1]

Dynamic measures of both recall and beliefs then allow us to make several contributions to the existing literature on motivated beliefs. We provide the first field evidence of biased memory loss by demonstrating that recall of grades is substantially more positively biased at a longer time horizon. Then, in a dynamic structural model, we explore how biased recall shapes the impact of shocks to attainment and self-esteem, and how these dynamics vary across unobserved types of student. It also provides a natural environment in which to indirectly test the key conjecture of 'fragile self-esteem' theory: that positive or negative beliefs can be self-sustaining through associative recall (Koszegi et al., 2022).

The dataset we use is the Beginning School Study (BSS). The BSS was a longitudinal study of Baltimore City public schools in the 1980s, combining administrative data on attainment from schools with subjective survey data. The BSS is distinct from other comparable datasets in that it contains repeated measures of recalled grades alongside their factual counterparts.

Our results show that students in the BSS make positive (flattering) errors in recall much more often than negative (unflattering) ones. Using within-individual variation, we find that they are also much more likely to make recall errors when receiving low grades. Additionally, we find that the same student is much more likely to make an overconfident prediction of their next grade in periods where they recall a higher grade than they received. These analyses corroborate existing experimental and field evidence on the asymmetric recall of ego-relevant information and its link to overconfidence (Eil and Rao, 2011; Huffman et al., 2022). We then exploit an idiosyncrasy of the BSS to test for biased memory loss in the field, replicating Zimmermann (2020)'s study in the laboratory: we compare recall in Spring survey periods, which were about 2 months after the relevant report cards, to Fall periods which were about 4–5 months after report cards. Recall of grades is substantially more positive in Fall periods, indicating that memory loss is tilted towards ego enhancement.

We then estimate a discrete choice model which represents biased recall as a tool to optimise psychological welfare. Our parameter estimates support models of ego utility with reality-constrained belief distortion. In order to test Koszegi et al.

[1]See Bénabou and Tirole (2002) for a notable early theory. Related theories of belief distortion include Brunnermeier and Parker (2005) and Köszegi (2006).

(2022)'s distinctive assumption that beliefs influence recall, we allow preferences for positively biased recall to depend on prior academic self-esteem. We find evidence consistent with this mechanism, and that shocks to self-esteem can be very persistent: a poor report card in mathematics negatively affects the distribution of self-esteem for around 2–3 years.

In an extended version of the model, we permit unobserved heterogeneity in both preferences and state variables. We allow *mindset*, the student's belief in the malleability of their ability, to be a predictor of unobserved type. Mindset has been posited as a key factor underpinning the formation of self-confidence in developmental psychology (Dweck, 2002) and related concepts have received attention in the economics of education (Alan et al., 2019). The model reveals substantial heterogeneity across students and finds that having a growth mindset is a strong predictor of both unobserved type and attaiment. High achieving types with a growth mindset are estimated to have disproportionately strong preferences for recalling positive grades, meaning their self-esteem is substantially less responsive to actual attainment. Additionally, we empirically quantify the fragility of belief equilibria, and show that an equilibrium with unrealistically high self-esteem is more fragile for low achievers. Further analysis suggests that this may be because high achievers find it more threatening to admit to themselves that they have low ability; even though lower recalled grades are less likely to affect their self-esteem, their enjoyment of school is more sensitive to self-esteem.

Aside from testing prominent theories of ego utility and beliefs, our analysis also contributes to research on the economics of education. A key focus in recent years has been to understand gaps in educational attainment and why they widen over schooling years (Heckman et al., 2006). One strand of this literature explores the role of self-confidence as a non-cognitive skill (Cunha and Heckman, 2007; Cunha et al., 2010; Alan et al., 2019). Our results indicate that self-deception is inextricable from self-confidence at school.

One concern with unincentivised measures of recall is that reporting may not be truthful. However, two pieces of evidence provide reassurance that dishonesty is unlikely to play a major role in the results. The first is our finding that recall biases starkly increase when beliefs are elicited after a longer delay, which cannot be rationalised by dishonesty. The second is that recalled grades have a material impact on self-esteem, grades, and wellbeing at school, all of which are highly persistent through time.

Our results have several implications for research and policy focusing on the provision of ego-relevant information, both inside and outside of education. In some circumstances, implementing realistic beliefs about ability is desirable. Under this prerogative, our results on biased memory loss demonstrate that unfavourable grades may need to be reiterated over time. However, a simulation of our dynamic model suggests that an intervention to correct recall errors has economically significant, negative effects on both academic self-esteem and attainment, at least within the relatively short horizon measured in our sample. The effects are also heterogeneous across the population: our model suggests low achievers' future attainment is more sensitive to their awareness of recent attainment. This suggests a more complicated set of consequences from policies aiming to reduce informational frictions, for example between schools, students, and parents (Dizon-Ross, 2019). Our results therefore throw into sharp relief the importance of expanding the classical perspective on information — that more is better — to account for non-standard preferences for beliefs and information.

Outline. This chapter proceeds as follows. In Section 1.2, we describe the dataset. In Section 1.3, we outline results from reduced-form analysis testing the basic predictions of motivated beliefs theory. In Section 1.4, we outline and estimate our dynamic structural model. In Section 1.5, we conclude.

1.2 Data: The Beginning School Study

This study uses survey data on students' perceptions of their academic ability and attainment, as well as administrative data on their actual attainment, from the Beginning School Study (BSS).[2] The BSS began in 1982, tracking a representative cohort of 790 students attending 20 Baltimore City public schools in the USA.

The backdrop of the BSS was one of striking educational disadvantage. From the beginning of the study, Baltimore was in the midst of a protracted economic decline, with profound effects on sample students' households (Alexander et al., 2014). 61% of responding parents in the sample did not finish high school, 28% were sole earners without employment, and 25% were the sole resident parent (Table 1.1). In spite of these dispiriting circumstances, parents were relatively optimistic about their child's educational prospects: in the first run of the survey, 98% expected their

[2] Alexander and Entwisle (2003).

child to finish high school; only 70% eventually did so without initially dropping out.

Table 1.1 Sample demographics, Fall 1982 (first grade)

	Value
Child's race: Black	55%
Child's sex: Female	50%
Mother as parental respondent	86.3%
Parent school dropout	60.8%
Parent employed	60.4%
Parent not employed & sole earner	28.1%
Non-resident second parent	25.1%
Expect child not to finish high school	2%

The BSS conducted face-to-face interviews with students in almost every year of their education, sometimes twice, until leaving.[3] It then made attempts to track respondents after they left school and began adult life, once in 1998 (with most respondents aged around 20) and again in 2006.

BSS surveys asked students what grade they remembered getting in their last quarterly report card. Survey waves were implemented twice a year, in Fall and Spring. In Fall (around November), students were asked to try to remember the grade they received at the end of the last academic year in June. In Spring (mostly in April or May), students were asked to remember the grade they got a few weeks earlier, in the third quarter. Recall accuracy is around 70% in Spring, but only around 50% in Fall (Table 1.2). Notwithstanding this substantial seasonal variation, students' recall accuracy was largely stable over time, other than in fourth grade (which we omit from all ensuing analysis). Imperfect recall observed in Spring suggests that there may have been ample opportunities for adolescents to avoid absorbing information from report cards.[4] However, recall accuracy is significantly diminished in Fall. This

[3]The sample is substantially smaller in Fall '88 and Spring '89, when the BSS was unable to track many students who had moved from the initial set of elementary schools to a range of middle schools. The study made a concerted effort to locate most missing students by the end of the academic year, but not in time to collect viable measures of recall for these two survey sweeps.

[4]Accuracy of recall may still be somewhat better than in Dizon-Ross (2019), where 60% of parents in the Malawian sample say they did not know their child's last report card grade.

difference presents a valuable opportunity to explore the effects of the passage of time on memory, to be revisited in the analysis to follow.

Table 1.2 Summary statistics

Sweep	N	School grade (modal)	Mean mathematics mark (recalled)	Mean mathematics mark (actual)	Correct recall %
Fall '85	531	4	3	2.4	40.9
Fall '87	496	6	2.9	2.4	49.6
Spring '88	465	6	2.5	2.2	74
Fall '88	184	6	2.8	2.2	47.8
Spring '89	172	7	2.2	1.9	67.4
Fall '89	381	8	2.6	2.1	51.2
Spring '90	409	8	2.5	2.2	72.1
Fall '90	444	9	2.7	2.2	48
Spring '94	143	12	2.3	2	60.1

Note: Mean mathematics marks computed by assigning 4 to *Excellent*, 3 to *Good*, 2 to *Satisfactory*, and 1 to *Unsatisfactory*. Fall '85 is excluded from all the ensuing analysis.

The BSS contains a large set of variables that we do not use in our analysis. Almost invariably, our reason for ignoring them is infrequency or inconsistency in how the variable was recorded. For example, the BSS also surveyed parents and teachers on their beliefs about students' ability and, in the former case, their recall of students' grades. While students themselves were surveyed twice a year, parents and teachers were only sampled once a year or less. Combined with an unbalanced panel of students, this means longitudinal analyses using data from parents and teachers would be very poorly powered. Thus, our focus is on students' own recall and beliefs, other than where we allow cross-sectional variation in parents' 'mindset' to predict students' unobserved type in our structural model. Similarly, the BSS contains administrative data on report card grades, as well as measures of recall, for a range of other subjects: Reading, Writing, English, and Science. None of these subjects appears consistently through the sample, largely due to changes in the curriculum as students progress through school. Thus, all of our analysis centres on mathematics, for which both grades and recall were collected in all key periods and retained a consistent definition throughout.

Table 1.3 Grades and scores

g_{it}	Score
Excellent	90–100%
Good	80–89%
Satisfactory	70–79%
Unsatisfactory	<70%

1.3 Reduced form analysis

We now present some reduced form evidence consistent with motivated beliefs theory. In what follows, t denotes the survey sweep, so the subscript it denotes the last available observation of the given variable for individual i when survey t was collected. Throughout the chapter, g_{it} denotes report card grades in mathematics. r_{it} denotes recalled report card grades in mathematics. Grade recall was elicited using the following survey question: *"Remember the last report card you got when school ended for the summer? You could have gotten marks like E (Excellent), G (Good), S (Satisfactory), or U (Unsatisfactory). What mark did you get in Mathematics?"* As such, the domain for both g_{it} and r_{it} is $\{E, G, S, U\}$; see Table 1.3 for a translation into numerical marks.[5] The other notable variable we use later in this section is each student's expected grade in their next report card, also collected on the $\{E, G, S, U\}$ scale.

1.3.1 Recall errors

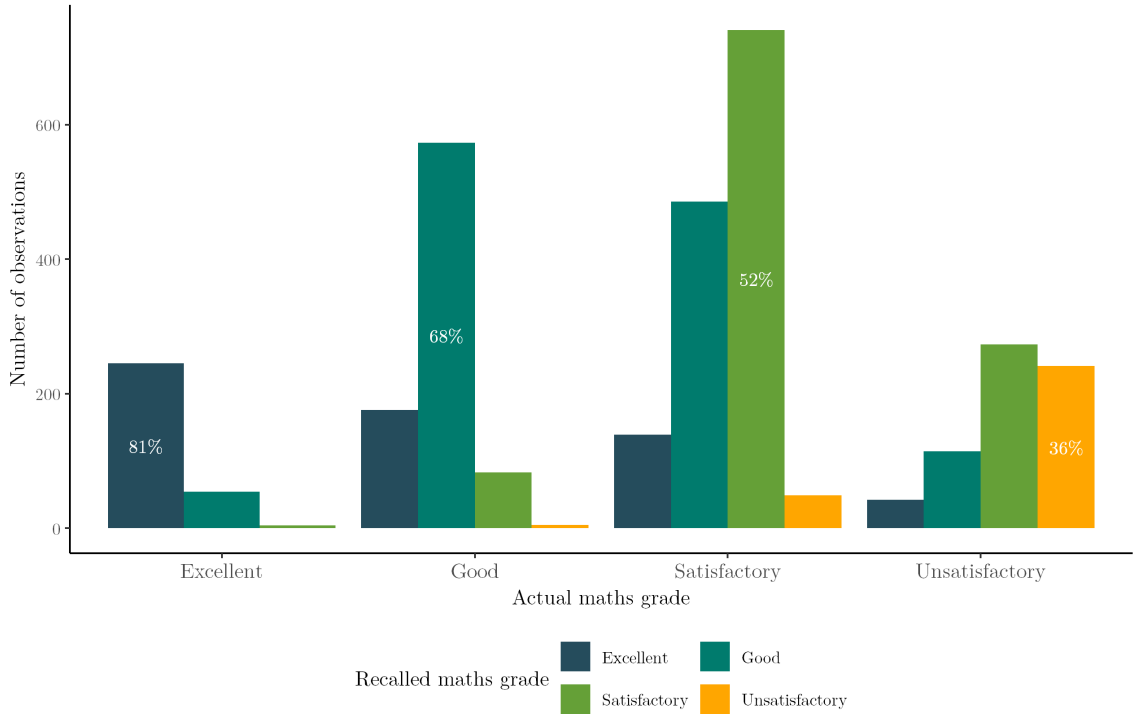
While theories of motivated beliefs differ on the precise motivations for belief distortions, they all predict that unfavourable information should be more unattractive to an individual with some form of ego utility.[6] Before moving to our structural analysis in Section 1.4, which quantifies the trade-off in distorting beliefs, we follow most of the existing empirical literature in simply examining whether the probability of making a recall error differs across favourable and unfavourable information.

[5]Not all schools used the $\{E, G, S, U\}$ grading scale. We use a conversion table in the BSS documentation to map actual grades to that scale. Students were always asked to recall grades on the $\{E, G, S, U\}$ scale, but they were also provided with the percentage score mapping of each grade category in Table 1.3.

[6]For example, in Bénabou and Tirole (2002), overconfidence helps to overcome present bias, so belief distortion has 'instrumental' benefits. In Brunnermeier and Parker (2005), belief distortion has psychological benefits through anticipatory utility (Caplin and Leahy, 2001) but leads to instrumental costs by distorting actions away from their optimum.

In the context of the BSS, the opportunities for information avoidance and suppression began when report cards were transmitted to students, who could ignore their report card if anticipating its poor content via other informal signals, or even avoid fully digesting information while seeing it (Gabaix et al., 2006). Afterwards, biased memory loss may play a role. Both of these mechanisms are likely to be present in measured recall errors. While we examine their combined effects in this section, we demonstrate that biased memory loss plays a distinct and significant role in Section 1.3.2.

Fig. 1.1 Remembered grades vs. actual grades, split by survey edition



Note: Aggregated over survey sweeps. White percentages are accurate recall rates by actual grade.

Figure 1.1 provides a clear indication of biased recall in the BSS data. It plots actual grades received in mathematics and splits them by the grade the student remembered. The first observation is that recall errors are positively skewed, other than for the highest grade (for which upward errors are impossible): conditional on making an error, students generally remembered getting better grades than they actually did. Secondly, errors become increasingly prevalent the lower the actual grade is. Nonetheless, in all cases other than for the worst grade, the modal outcome is to recall the correct grade.

In order to formalise this analysis, we estimate a regression whose dependent variable is the event that a student incorrectly recalls their mathematics grade.[7] The aim is to check whether the relationship between recall errors and the qualitative content of grades is robust to confounding variables. Arguably the most problematic of those are cognitive ability and engagement with schooling, which one might expect to be correlated with both grades and the event that they are correctly recalled. Since the BSS observed students longitudinally, we can make major progress in addressing these concerns by including individual and year fixed effects in the analysis. Intuitively, that means we examine whether the same individuals are more likely to forget lower grades than higher ones. The empirical specification is a linear probability model with robust standard errors.

As in the raw data, the estimated coefficients under column (1) in Table 1.4 provide clear evidence that adolescents are more likely to fail to recall poor grades. This effect is very large: students are almost 50 percentage points (pp) more likely to make recall errors when they receive the lowest grade, *Unsatisfactory* (scores lower than 70%), than when they receive *Excellent* (scores above 90%). The effect of actual grades on recall accuracy is monotonic and large: each progressively worse grade is much less accurately recalled by students. As already indicated, the probability of a recall error increases substantially (by 23.7pp) in Fall periods, which was delivered much longer after report card grades than the Spring edition. The results therefore vindicate the pattern in the raw data (Figure 1.1) and closely match similar patterns in experimental and field research on adults (Eil and Rao, 2011; Huffman et al., 2022).

Column (2) in Table 1.4 omits individual fixed effects. The results undergo little qualitative change other than some small differences in relative magnitudes. This conclusion could provide some reassurance to comparable analyses in field research without longitudinal variation (Huffman et al., 2022), since it indicates that fixed unobserved individual characteristics, such as cognitive ability or engagement with schooling, may not be an important joint determinant of signal valence and the accuracy of recall. However, we will soon see that unobserved factors appear to be a more important joint determinant of biased recall and overconfidence.

[7]This dependent variable also permits negative recall errors. We allow both positive and negative errors because of boundary conditions for the top and bottom grades: it is not possible to remember a better grade than *Excellent* and a worse one than *Unsatisfactory*. We also report the results of an analysis in which we examine whether probability of making a positive recall error depends on the actual grade received, excluding observations where *Excellent* was actually received (Table A.1) The results are very similar, with striking increases in the probability of making a positive recall error when receiving *Satisfactory* and *Unsatisfactory* relative to *Good*.

Table 1.4 Reduced form model — recall errors

	Dependent variable:	
	Incorrect recall	
	(1)	(2)
Actual grade: *Good*	0.116*** (0.040)	0.128*** (0.030)
Actual grade: *Satisfactory*	0.276*** (0.042)	0.251*** (0.029)
Actual grade: *Unsatisfactory*	0.520*** (0.047)	0.461*** (0.032)
Fall	0.237*** (0.020)	0.238*** (0.020)
Individual FE	Yes	No
Academic year FE	Yes	Yes
Observations	2,694	2,694
R^2	0.371	0.129
Adjusted R^2	0.196	0.127

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Linear probability model, robust standard errors in parentheses. All grades are from quarterly report cards for mathematics, and recall of the most recent quarterly grade is elicited in a biannual survey. The omitted actual grade category is *Excellent*.

We also examine sex and racial differences in recall asymmetries (Table A.2). Surprisingly, there are no differences between boys and girls, matching a related analysis in Huffman et al. (2022). On the other hand, Black students appear to display a greater propensity to recall information asymmetrically than others: incorrect recall of *Good* and *Unsatisfactory* is more likely relative to *Excellent* grades than for other students. However, this pattern is uneven: *Satisfactory* grades are no less accurately recalled by Black students.

1.3.2 Biased memory loss

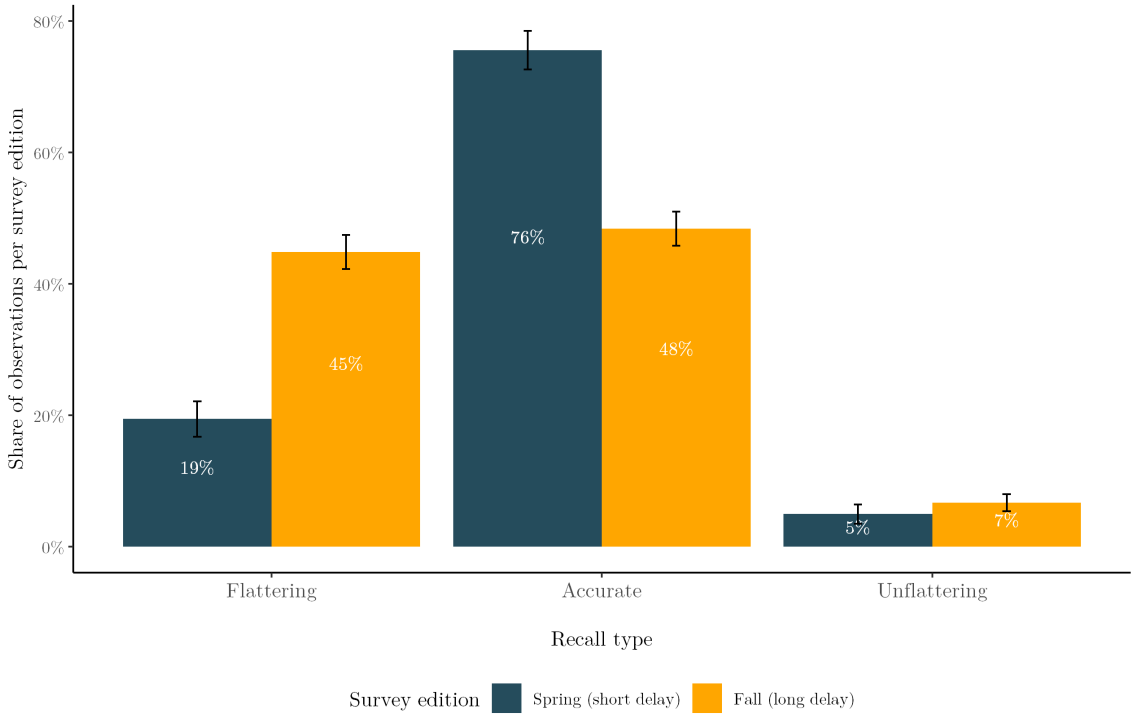
Biases in recall could result from a combination of information avoidance and memory loss. However, the design of the BSS presents a compelling opportunity to isolate the role of the latter.[8] In both versions of the survey within the year, in Spring and Fall, adolescents were asked to remember their most recent grades. In Spring, those grades were likely from just over a month earlier, but in Fall, the most recent grades were likely received more than three months ago. In the raw data, we have already seen sizeable differences in the overall accuracy of recall between Spring and

[8]Our strategy is similar to the main analysis in Zimmermann (2020), which examines how recall of negative signals of IQ deteriorates over time relative to positive signals.

Fall (Table 1.2). The analysis in this section exploits this naturally occurring variation in survey delivery to examine whether recall is more positively biased in Fall than in Spring.

Figure 1.2 plots types of recall by survey edition for the middle two grades, *Good* and *Satisfactory*. For these grades, both positive ('flattering') and negative ('unflattering') recall errors are possible. 'Accurate' recall occurs when the correct grade is remembered. As noted above, recall is less accurate in Fall than in Spring: the

Fig. 1.2 Recall type, split by survey edition



Note: Only *Good* and *Satisfactory* grades included, since it is possible to remember either a better ('flattering') or a worse ('unflattering') grade. Error bars depict 95% confidence intervals.

probability of remembering the correct grade falls from 76% to 48%. However, almost all of the deterioration in recall accuracy is in a flattering direction; the incidence of unflattering errors is barely higher in Fall relative to Spring. By comparing the difference between flattering and unflattering errors in Fall and Spring, we can determine that at least $^2/_3$ of the positive bias in recall in Fall is explained solely by memory loss. Assuming truthful reporting, the remaining $^1/_3$ could be explained by

either memory loss during the first month or information avoidance.[9]

We also study biased memory loss in a simple regression, once again using within-individual variation. Our approach is to examine the event that the student recalls getting an *Excellent* grade — as the dependent variable, controlling for the actual grade they received, g_{it}.[10] The key coefficient of interest is on *Fall*: we want to establish that the probability of recalling the top grade increases in *Fall* compared to *Spring*, for the same student.

Table 1.5 Reduced form model — memory loss

	Dependent variable:	
	Recalled grade: *Excellent*	
	(1)	(2)
Actual grade: *Good*	-0.598^{***} (0.035)	-0.600^{***} (0.030)
Actual grade: *Satisfactory*	-0.680^{***} (0.035)	-0.724^{***} (0.027)
Actual grade: *Unsatisfactory*	-0.691^{***} (0.037)	-0.741^{***} (0.028)
Fall	0.084^{***} (0.013)	0.084^{***} (0.013)
Individual FE	Yes	No
Academic year FE	Yes	Yes
Observations	2,694	2,694
R^2	0.371	0.129
Adjusted R^2	0.196	0.127

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Linear probability model, robust standard errors in parentheses. All grades are from quarterly report cards for mathematics, and recall of the most recent quarterly grade is elicited in a biannual survey. The omitted actual grade category is *Excellent*.

The results are reported in Table 1.5. The coefficient on Fall in column (1) implies that being surveyed in Fall has a large, positive impact on students' propensity to remember getting the highest grade. This formalises the positive drift in memory illustrated in the raw data, in Figures A.2 and 1.2. That is, the same students

[9]Even in the Spring survey, recall is less accurate for the lower grades, *Satisfactory* and especially *Unsatisfactory* (Figure A.2). Given the relatively short delay for the Spring survey, of just over a month, this could support some role for selective attention or information avoidance. That being said, Zimmermann (2020) successfully uses a delay of one month to measure the effects of memory loss in his laboratory test of the same phenomenon.

[10]Another option would be to examine whether the probability of making a positive recall error increases in Fall, conditional on making an error, as in Figure 1.2. However, as in Section 1.3.1, this strategy only permits us to examine cases where the middle two grades, *Good* and *Satisfactory*, were received. It would involve discarding information on how memory loss affects students receiving the top and bottom grades.

are more likely to remember higher grades when surveyed in Fall than in Spring, controlling for actual grades. Thus, we obtain evidence matching the main exercise in Zimmermann (2020): the positive bias in recall becomes more accentuated as time passes, suggesting that memory loss serves the ego. Unsurprisingly, the coefficient on *Fall* does not differ across specifications with and without individual fixed effects in columns (1) and (2), since the sample should be balanced across editions of the survey.

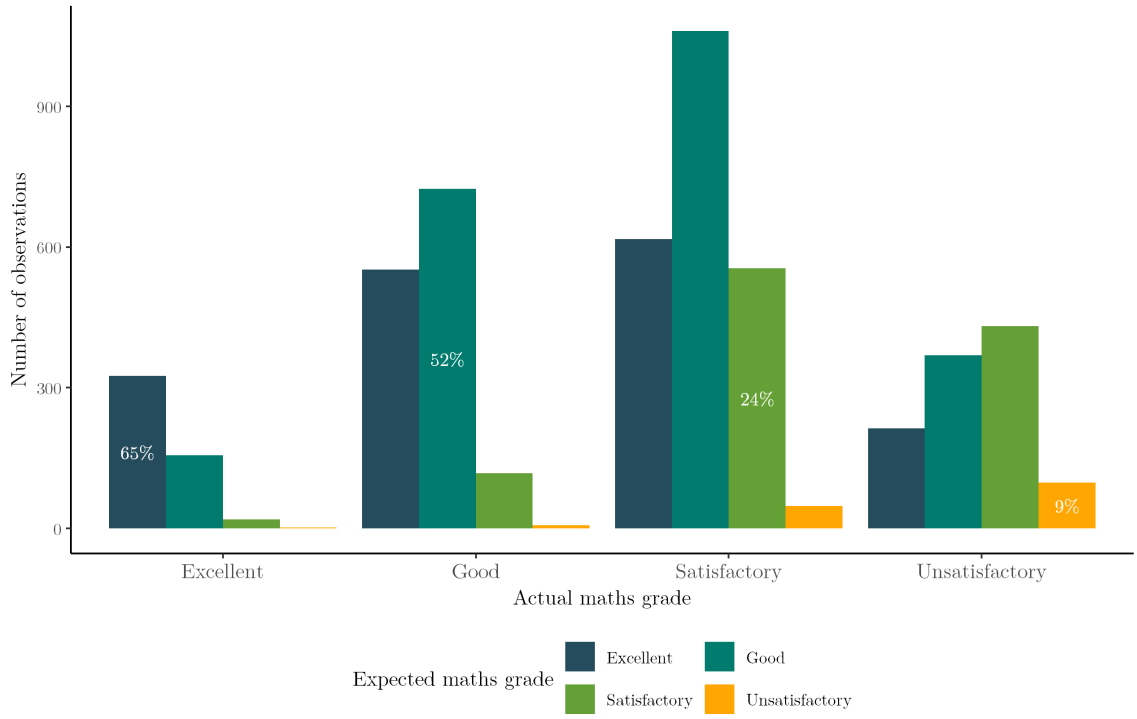
Again, we examine heterogeneity along sex and race dimensions (Table A.3). In this case, we find no evidence of differences in the effect of memory loss across either dimension. That being said, Black students are less likely to recall *Excellent* grades (across both Spring and Fall) having actually achieved either of the lowest two grades.

1.3.3 Overconfidence

In each survey period, the BSS also collected students' expectations of the mathematics grade they would receive in the following quarterly report card. By comparing this variable to actual grades, we can directly measure overconfidence. Our aim is to establish whether biased recall is associated with overconfident expectations of future grades, replicating the main exercise in Huffman et al. (2022) but with the significant benefit of within-individual variation. In doing so, we provide insight on the important matter of whether biased recall allows overconfident beliefs at school to persist despite repeated signals to the contrary.

First, we report on the raw data (Figure 1.3). Students frequently make errors in predicting their next grade. Prediction errors are more common than recall errors. This disparity is particularly stark for the low grades: only 9% of those receiving *Unsatisfactory* expected to do so, compared with the 36% accurately recalling *Unsatisfactory* when it is received. The differences remain stark for *Satisfactory* and *Good* grades: accurate expectations occur at rates of 24% and 52% respectively, compared with accurate recall rates of 52% and 68% respectively. Notably, expectation errors are also more common than recall errors for the upper grade, *Excellent*, suggesting that part of the difference is driven by genuine forecast uncertainty. Nonetheless, students' expectations are strikingly overconfident in general. Other than for students eventually receiving *Good* and *Excellent*, the modal outcome is to expect a higher grade than is actually received. Students also commonly expect a

Fig. 1.3 Expected grades vs actual



Note: Aggregated over survey sweeps. White percentages are accurate recall rates by actual grade.

higher grade than they recall receiving in their last report card (Figure A.1).

To establish a link between biased recall and overconfidence, we estimate a regression which uses longitudinal variation to establish whether a *given student* is more likely to have overconfident expectations of their future grade in a period where they have recalled a higher grade than was actually received.[11] Notably, a within-individual strategy naturally precludes several of the concerns that are addressed individually in Huffman et al. (2022), perhaps most importantly that overconfidence could reflect stable differences in prior beliefs (Benoît and Dubra, 2011) or cognitive ability.

The results are reported in Table 1.6. First, in column (1), we find that a given student is indeed more likely to be overconfident in periods where they they recall an incorrectly high grade. The effect on the probability of overconfidence is very large, at 8.2 percentage points compared to a baseline probability of 28.1% for students who recall the correct grade or one lower than they received. Notably, actual attainment has no impact on the likelihood of overconfidence, suggesting the interpretation that recall is all that matters for beliefs.

In column (2), we add a control variable for Fall survey periods. This specification is of particular interest because, as established in the last section, recall errors are much more likely in Fall than Spring. Interestingly, students are more likely to be overconfident in Fall even though we control for recall errors. This may reflect that the fact that in Fall, expectations become more optimistic in a way that is not captured by recall errors. One example could be that the student could perceive a weaker link between past performance and future performance when more time has passed since the last grade.[12]

In column (3), we omit individual fixed effects from the specification. The most notable effect of this change is to almost double the size of the coefficient on recall errors, relative to the comparable specification in column (2). This suggests that unobserved cross-sectional heterogeneity is an important joint determinant of biased recall and overconfidence. This comparison reveals that students persistently receiving lower actual grades are more likely to make persistent errors in predicting their future grades: those receiving *Satisfactory* grades are 4pp more likely to

[11]This strategy requires us to drop all observations where students receive *Excellent* in the prior or next report card. An alternative specification, which uses the whole sample, considers whether prediction errors are more likely when recall errors have been made. The results are qualitatively very similar (Table A.4).

[12]Since the length of time to the next grade is approximately the same in both the Spring and Fall editions of the survey, the coefficient on Fall cannot be explained by a difference in forecast horizon.

Table 1.6 Reduced form models — overconfidence

	Dependent variable:		
	Overconfident prediction		
	(1)	(2)	(3)
Positively biased recall	0.082*** (0.023)	0.069*** (0.024)	0.129*** (0.020)
Actual grade: *Satisfactory*	−0.022 (0.029)	−0.018 (0.029)	0.040* (0.023)
Actual grade: *Unsatisfactory*	−0.039 (0.037)	−0.030 (0.037)	0.076*** (0.028)
Fall		0.051** (0.024)	0.036 (0.024)
Individual FE	Yes	Yes	No
Academic year FE	Yes	Yes	Yes
Observations	2,473	2,473	2,473
R^2	0.341	0.342	0.041
Adjusted R^2	0.130	0.131	0.038

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Linear probability model, robust standard errors in parentheses. All grades are from quarterly report cards for mathematics, and recalled and expected quarterly grades are elicited in a biannual survey. The sample excludes all observations where the previous or next grade is *Excellent*. The omitted actual grade category is *Good*.

make errors in predicting their next grade than those receiving *Good*, with an even larger difference of 7.6pp for those receiving *Unsatisfactory*. Thus, controlling for cross-sectional unobserved heterogeneity plays an important role in convincingly identifying the link between biased recall and overconfidence.

1.4 Structural analysis

We now outline a structural model that positions recall distortions at the centre of a psychological framework to optimise welfare. It assumes students (subconsciously) maximise lifetime utility by trading off the benefits and costs of distorting their recall of grades in each period. Since the model embodies a dynamic interaction of recall and self-esteem, with links possibly running in both directions, it provides a natural environment in which to test the key mechanisms underpinning fragile self-esteem theory (Koszegi et al., 2022).

A further attraction of the structural approach is that, in an extended version of the model, we can separate students into unobserved types with potentially differing preferences for inflating their recall of grades. In particular, we model unobserved heterogeneity as a finite mixture distribution where key parameters for

both preferences and state variables differ across discrete classes of student. This approach means the model can determine whether some students have stronger preferences for positively biased recall than others, also allowing these differences to be correlated with unobserved heterogeneity in attainment and self-esteem.

1.4.1 Basic model

For ease of computation and interpretation, we redefine our measures of recalled and actual grades as binary rather than categorical. In this updated definition, $r_t = 1$ if the student recalls *Good* or *Excellent* in period t. g_t has the same interpretation as r_t but refers to the relevant grade actually achieved. Also key to the model is y_t, an indicator of self-esteem based on the survey measure in Table 1.7. As is visible from the summary statistics, the majority of students select the middle category. We therefore define our binary indicator such that $y_t = 1$ for values greater than 3; this threshold extracts the maximal amount of information from the survey indicator. This definition implies $y_t = 1$ if a student believes they are above average ability for their school and year group.

Table 1.7 Self-esteem measure

How smart do you think you are compared to other kids in your school this year?	Value	Share of observations
One of the smartest	5	14.7%
Smarter than most kids	4	21.0%
About as smart as everybody else	3	57%
Not as smart as most kids	2	6.5%
Not very smart at all	1	0.1%

In each period, students choose r_t to maximise expected lifetime utility,[13]

$$E \sum_{t=1}^{\infty} \beta^{t-1} u(r_t, y_{t-1}, g_t, f_t, \epsilon_t; \theta, \lambda). \tag{1.1}$$

In (1.1), f_t is an indicator for the Fall survey, and $\epsilon_t = (\epsilon_{0t}, \epsilon_{1t})$ are choice-specific shocks to utility observed by the student in period t but unobserved in the data. We

[13]Since we only observe key variables for an intermediate part of students' school careers, and since the process of interpreting signals about ability likely continues past schooling years, we set up the problem with an infinite horizon.

make the standard assumption that they are drawn from a Type 1 Extreme Value distribution. Periodic utility has the following functional form, where $\Theta = (\theta, \theta_y)$ and $\Lambda = (\lambda, \lambda_f)$,

$$u(r_t, y_t, g_t, \epsilon_t; \Theta, \Lambda) = \begin{cases} \theta + \theta_y y_{t-1} - (\lambda \mathbb{1}(g_t \neq r_t) + \lambda_f \mathbb{1}(g_t \neq r_t) f_t) + \epsilon_{1t} & \text{for } r_t = 1 \\ -(\lambda \mathbb{1}(g_t \neq r_t) + \lambda_f \mathbb{1}(g_t \neq r_t) f_t) + \epsilon_{0t} & \text{for } r_t = 0. \end{cases} \tag{1.2}$$

θ and θ_y capture the hedonic benefit of recalling a higher grade, while λ and λ_f capture the cost of recall distortions. The cost of recall distortions is symmetric, in that recalling an incorrectly high or an incorrectly low grade incurs the same cost.

To understand identification of the key parameters, consider a student and survey period where g_t, y_{t-1} and f_t are all equal to 0. For such a student, the marginal utility of distorting recall is $\theta - \lambda$. Thus, if positive recall errors are more common *relative to* negative recall errors, we will obtain a larger estimate of θ. If instead $g_t = 1$, the marginal utility of a negative recall error $r_t = 0$ is $-\lambda$. Thus, λ is only smaller if both negative and positive recall errors are less common. Identification therefore exploits information from all types of recall: flattering, accurate, and unflattering.

The parameter λ_f also appears directly in utility, capturing the fact that the costs of memory distortion are lower in Fall. We would expect that $\lambda_f < 0$. This specification naturally accommodates the finding that recall becomes more positively biased in Fall: for a student with low prior self-esteem $y_{t-1} = 0$ achieving a low grade $g_t = 0$, the marginal utility of $r_t = 1$ is $\theta - \lambda$ in Spring and $\theta - (\lambda + \lambda_f)$ in Fall. The marginal utility of $r_t = 1$ conditional on $g_t = 1$ is θ in both Spring and Fall. If $\lambda_f < 0$, our model can embody biased memory loss: the prospect of recalling a high grade is equally attractive in both Spring and Fall, but it should be harder to do so in Spring when relatively little time has passed since the report card in question. Finally, θ_y allows preferences for $r_t = 1$ to vary with prior self-esteem. In particular, if $\theta_y > 0$, students with high prior self-esteem have a stronger preference for self-enhancing recall distortions, capturing the key mechanism underpinning fragile self-esteem (Koszegi et al., 2022).

Our specification of hedonic utility imposes no explicit structure on the motivations for biasing recall. An alternative approach would be to do so, perhaps by specifying that the benefits of biased recall accrue through its expected effects on self-esteem and grades. There are two main justifications for the more parsimonious

approach we use. Instead of attempting to explain why students might bias their recall of grades, our model focuses on estimating the strength of preferences for positively biased recall. This has significant benefits: identification is transparent; and our model folds all possible motivations for distorting grades, including unobserved ones, into the essentially reduced form parameters θ and θ_y. Thus, we do not make any assumptions about *what* motivates students to bias their recall, freeing up the model to focus on our main question of interest: how the strength of preference for biased recall relates to the dynamics of beliefs.

The second justification is that it would be very difficult to achieve plausible identification of any indirect marginal benefits of biased recall. For instance, in a model in which periodic utility were a direct function of $\hat{y}_t = E(y_t|r_t, y_{t-1}, g_t, f_t)$ or $\hat{g}_t = E(g_t|r_t, y_{t-1}, g_{t-1}, f_t)$, identification of the marginal utility of both variables would need to exploit heterogeneity in choice probabilities across observations with different marginal effects of r_t on $\hat{y}_t$ and $\hat{g}_t$. We would not have a source of variation in the marginal effect of r_t on $\hat{y}_t$ and $\hat{g}_t$ other than any non-linearity contrived from functional form assumptions, which we do not think would be behaviourally credible.

The two state variables g_t and y_{t-1} are specified to follow first-order Markov processes,

$$P(g_t = 1|g_{t-1}, y_{t-1}, r_{t-1}) = \text{logit}(\mu_1^g + \mu_y^g y_{t-1} + \mu_g^g g_{t-1} + \mu_r^g r_{t-1}), \tag{1.3}$$

and

$$P(y_t = 1|g_{t-1}, y_{t-1}, r_{t-1}) = \text{logit}(\mu_0^y + \mu_y^y y_{t-1} + \mu_y^y g_t + \mu_r^y r_{t-1}). \tag{1.4}$$

Let $x_t = (y_{t-1}, g_t)$. We make the standard assumption that (x_t, ϵ_t) is a stationary controlled first-order Markov process, with transition

$$P(x_{t+1}, \epsilon_{t+1}|x_t, \epsilon_t, r_t) = P(\epsilon_{t+1}|x_{t+1}, r_t)P(x_{t+1}|x_t, r_t), \tag{1.5}$$

so that ϵ_t is serially independent conditional on x_t, and x_{t+1} is independent of ϵ_t conditional on x_t and r_t. The student observes ϵ_t in period t but not before. Under these assumptions, the model is stationary and can be estimated by full information maximum likelihood after computing the expected value function via iteration of the Bellman equation (Rust, 1987). To elaborate on this procedure in our context, let

$\bar{u} = u - \epsilon$ and define the expected value function $\bar{V}(x_t, r_t)$. The value function is

$$V(x_{t+1}, \epsilon_{t+1}) = \max_{r_{t+1} \in \{0,1\}} \{\bar{u}(x_{t+1}, r_{t+1}; \Theta, \Lambda) + \epsilon_{t+1} + \beta \bar{V}(x_{t+1}, r_{t+1})\}. \tag{1.6}$$

By taking expectations over $(x_{t+1}, \epsilon_{t+1})$ conditional on (x_t, r_t), we obtain the Bellman equation for $\bar{V}(x_t, r_t)$, where X is the state space for x_t,

$$\bar{V}(x_t, \epsilon_t) = \sum_{x_{t+1} \in X} \log \left[\sum_{r_{t+1} \in \{0,1\}} \exp \left(\bar{u}(x_{t+1}, r_{t+1}; \Theta, \Lambda) + \beta \bar{V}(x_{t+1}, r_{t+1}) \right) \right] P(x_{t+1} | x_t, r_t). \tag{1.7}$$

The sample log-likelihood for the basic model can be written,

$$l(\Theta, \Lambda, \mu, \omega, \gamma) = \sum_{i=1}^{N} \log \left\{ \prod_{t=1}^{T} P(r_{it} | x_{it}; \Theta, \Lambda, \mu) P(x_{it} | x_{i,t-1}; \mu) \right\}. \tag{1.8}$$

In all of what follows, we assume $\beta = 0.9$ since it is not identified (Rust, 1987). However, we report the results of estimation with a range of values of β in the Appendix. Computation and maximisation of the sample log-likelihood is then straightforward because choice probabilities $P(r_{it} | x_{it}; \Theta, \Lambda, \mu)$ have the well-known logistic closed form, and state transition probabilities stem from the dynamic logistic regressions defined in (1.4) and (1.3).

1.4.2 Unobserved heterogeneity

We also estimate a version of the model with unobserved heterogeneity modelled as a finite mixture distribution (Heckman and Singer, 1984). We use this approach to allow for heterogeneous preferences which may be correlated with unobserved determinants of state variables. We assume students can be one of two unobserved types, $c \in \{1, 2\}$.[14] The crucial feature we would like to capture is heterogeneity in the key parameters driving preferences for self-enhancing recall. As such, θ^c and θ_y^c are now permitted to vary across types, while we impose that the cost parameters are fixed across types — our interest *a priori* is not in heterogeneity in the accuracy of

[14] The model is not identified with more than two types, as is the case in other similar applications (Arcidiacono, 2005).

recall across students. Thus, periodic utility can now be written as:

$$u(r_t, y_t, g_t, c, \epsilon_t; \Theta^c, \Lambda) = \begin{cases} \theta^c + \theta^c_y y_{t-1} - (\lambda \mathbb{1}(g_t \neq r_t) + \lambda_f \mathbb{1}(g_t \neq r_t) f_t) + \epsilon_{1t} & \text{for } r_t = 1 \\ -(\lambda \mathbb{1}(g_t \neq r_t) + \lambda_f \mathbb{1}(g_t \neq r_t) f_t) + \epsilon_{0t} & \text{for } r_t = 0. \end{cases} \tag{1.9}$$

Since heterogeneity in preferences may be correlated with state variables, we control for type-specific fixed effects acting on both actual grades and self-esteem. We also want to be able to say if heterogeneity in preferences for recall is correlated with heterogeneity in the marginal effect of recall on state variables, so we also include the parameters ω^g_r and ω^y_r in the transition equations:

$$P(g_t = 1 | g_{t-1}, y_{t-1}, r_{t-1}, c = 2; \mu_g, \omega^y, \omega^g_r) = P(g_t = 1 | g_{t-1}, y_{t-1}, r_{t-1}; \mu_g) + \omega^g + \omega^g_r r_t, \tag{1.10}$$

and

$$P(y_t = 1 | g_t, y_t, r_t, c = 2; \mu_y, \omega^y, \omega^y_r) = P(y_t = 1 | g_{t-1}, y_{t-1}, r_{t-1}; \mu_y) + \omega^y + \omega^y_r r_t. \tag{1.11}$$

Although unobserved heterogeneity should be well identified by within-individual correlation in choices and state variables over time, we provide further information to the model and aid interpretation by permitting type membership to be predicted by two key covariates from Fall '85, about three years before the first period in which the rest of the variables are observed. A growth mindset is understood to be crucial in developing resilience to negative feedback and improving attainment (Dweck, 2002; Alan et al., 2019; Yeager et al., 2019), and Koszegi et al. (2022) also speculate that mindset could introduce heterogeneity in self-esteem equilibria. We thus specify the following auxiliary model for unobserved type probabilities:

$$P(c = 2 | X; \gamma) = \pi(X; \gamma) = \text{logit}(X'\gamma), \tag{1.12}$$

where X contains a constant term and the following variables:

1. $pfixed$: Parents' responses to the question *'For students to do well in school which is most important?'*. 'Effort/Personality' = 0; 'Ability' = 1.

2. $fixed$: Students' own responses to the question '*What do you think matters most in school, how hard you try, or how smart you are?'*. 'How hard you try' = 0; 'How smart you are' = 1.

Choice probabilities have the same logistic closed form as in the basic model. The only difference is that the sample likelihood is a weighted sum across the two types. Define $\Theta = (\theta^1, \theta^2, \theta^1_y, \theta^2_y)$, $\Lambda = (\lambda, \lambda_f)$, $\mu = (\mu_g, \mu_y)$, and $\Omega = (\omega^g, \omega^g_r, \omega^y, \omega^y_r)$. We can then write the sample log-likelihood $l(\Theta, \Lambda, \mu, \omega, \gamma)$, where we reintroduce the individual index $i \in \{1, 2, ..., N\}$,

$$l(\Theta, \Lambda, \mu, \omega, \gamma) = \sum_{i=1}^{N} \log \left\{ \sum_{c=1}^{2} \pi(X_i; \gamma) \left[\prod_{t=1}^{T} P(r_{it}|x_{it}; \Theta, \Lambda, \mu, c) P(x_{it}|x_{i,t-1}; \mu, \Omega, c) \right] \right\}. \tag{1.13}$$

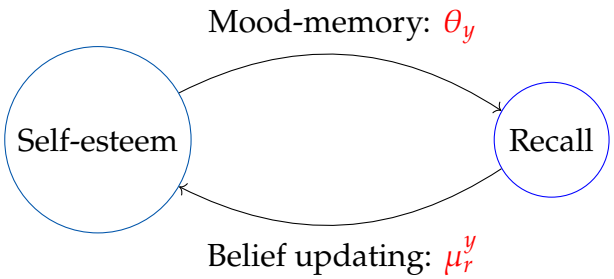
Consistent parameter estimates can be obtained by maximising $l(\Theta, \Lambda, \mu, \omega, \gamma)$, following computation of the value function via iteration of the Bellman equation as before.

1.4.3 Connection to Fragile Self-Esteem

Koszegi et al. (2022)'s theory of 'fragile self-esteem' (henceforth FSE) hinges on a mutual interaction between beliefs and recall, outlined in Figure 1.4. In their framework, much as in other models of biased recall, 'mood' (utility) is determined by an agent's recall of favourable information. The novelty of FSE is in a 'mood-memory' relation that implies a relationship running in the other direction: an individual's recall bias depends directly on their mood. In particular, past outcomes are more likely to be recalled if they are concomitant with one's mood. If an individual feels positively about herself, associative memory allows her to recall memories in support of that belief. Since mood determines memory, and memory determines mood, 'self-esteem personal equilibrium' (SEP) in FSE is characterised by a fixed-point condition. Importantly, multiple equilibria may be possible, depending on the functional form of the mood-memory relationship. The authors' favoured specification is a functional form which gives rise to two SEP. The 'fragility' of each equilibrium is defined as the size of shock required to dislodge self-esteem to the other SEP.

Our data offer a novel opportunity to indirectly test the mechanisms underpinning FSE, since they offer dynamic measures of both self-esteem and recall. Recalling (1.4), μ^y_r captures the well-known impact of recall on self-esteem. The parameter introducing the novel mood-memory link proposed by FSE is θ_y: if mood influences memory, recalling a positive grade should be more likely when self-esteem is already high, meaning $\theta_y > 0$. Separately, the utility parameter θ, capturing students'

Fig. 1.4 Interaction of self-esteem and recall in FSE



preference for recalling favourable grades conditional on $y_{t-1} = 0$, has a close analogue in the FSE parameter k. In that model, k governs the strength of the individual's ego concern and functionally determines the extent of positive bias in the distribution of recall conditional on self-esteem.

Some judgement is required in translating FSE to an empirical setting. One particularly important point is how to characterise SEP empirically. Koszegi et al. (2022) do not specify the speed of adjustment to equilibrium, but in our data, where each time period spans approximately 6 months, movements between self-esteem states ($y_t = 0, 1$) occur within single periods. Neither self-esteem state is absorbing in our data. State and choice variables thus form an ergodic Markov chain and the long-run distribution of self-esteem is invariant to initial conditions. In other words, starting at high or low self-esteem cannot influence the probability of having high self-esteem in the long run; in each consecutive period, it is possible to experience a shock large enough to shift from high self-esteem to low or vice-versa.

Koszegi et al. (2022) define fragility of a self-esteem equilibrium as the smallest shock required to dislodge an individual from one equilibrium to another. On the assumption that $y_t = 0$ and $y_t = 1$ both capture equilibria, we can infer fragility from transition probabilities for y_t. To illustrate the intuition, consider the latent variable representation of y_t, where $y_t = \mathbb{1}(y_t^* > 0)$ and

$$y_t^* = \mu_0^y + \mu_y^y y_{t-1} + \mu_g^y g_t + \mu_r^y r_t + \epsilon_t^y, \tag{1.14}$$

where ϵ_t^y is a random variable whose distribution is logistic with location 0 and scale 1. y_t^* is thus a continuous measure of self-esteem like the one used in FSE. Now suppose that the high self-esteem equilibrium is located at $y_t^* = \bar{y}_1 > 0$. If we make the assumption that students are in equilibrium in each period, a shift to the low self-esteem equilibrium (say, at some $\bar{y}_0 \leq 0$) occurs for any shock large enough to

imply $y_t^* \leq 0$. Suppose for simplicity that $g_t = 1$ and $r_t = 1$. Then, the critical value for ϵ_t^y,

$$\bar{\epsilon}^y = -\mu_0^y - \mu_y^y - \mu_g^y - \mu_r^y, \tag{1.15}$$

is a monotonic transformation of $1 - P(y_t = 1|y_{t-1} = 1, g_t = 1, r_t = 1)$. Thus, a natural quantification of fragility in our setting arises from the transition probabilities $P(y_t = k|y_{t-1} = 1 - k)$ for $k \in \{0, 1\}$: they are monotonically related to the size of shock required to move students from one level of self-esteem to another.

In FSE, self-esteem is a function of an individual's recall of their entire history of ego-relevant information, which is unobserved in our setting. We only have access to recall of the most recent grades. This substantively implies that θ_y only partially identifies the mood-memory relation. If mood-memory acts through recall of any information other than the most recent grade, such a channel would be captured by μ_y^y: y_{t-1} affects y_t independently of its effect on r_t.

1.4.4 Results

Basic model. Parameter estimates for the basic model are reported in Table 1.9. State variables are strongly serially dependent, indicated by μ_g^g and μ_y^y respectively. Most important for our test of fragile self-esteem theory are the results that $\mu_y^y > 0$ and $\mu_r^y > 0$: we have that self-esteem is both serially dependent and affected by recall of ego-relevant information. However, the former significantly outweighs the latter. Notably, self-esteem does not depend on actual grades: $\mu_3^y = 0$.

Actual grades do depend on self-esteem ($\mu_y^g > 0$) and biases in recall ($\mu_r^g > 0$). This finding is particularly notable because it is consistent with belief distortion having instrumental benefits (Bénabou and Tirole, 2002; Compte and Postlewaite, 2004) rather than costs (Brunnermeier and Parker, 2005), at least within the horizon we examine. This has some relevance to a later discussion on the welfare effects of interventions to correct beliefs.

In interpreting the parameters of the utility function, notice that $\theta + \theta_y < \lambda + \lambda_f$, so even in *Fall* and when prior self-esteem is high, recalling the correct grade delivers higher periodic expected utility. This is driven by the fact that empirically, the modal behaviour is to recall the correct grade. However, when recall errors do occur, $\theta > 0$ indicates a (strong) preference for recalling positive grades over negative ones.

Notably, the magnitude of θ_y tells us that the marginal benefit of recalling a high grade increases by over a third if prior self-esteem is high. This result is highly

consistent with the mood-memory mechanism distinctive to FSE. However, it should be noted that the impact of θ_y on the dynamics of y_t is quantitatively small. Both θ_y and μ_r^y are small enough that any impact of y_{t-1} on the distribution of r_t filters through to little impact on the distribution of y_t. By far the most important source of persistence in y_t is through direct serial correlation, captured by μ_{yy}. However, as previously noted, μ_{yy} is likely to capture additional mood-memory effects: if prior self-esteem is high, recall of signals other than r_t (such as for other subjects, or for mathematics in periods prior to t, could be more favourable).

Estimates of λ and λ_f indicate that the cost of making a recall error is almost halved in Fall relative to Spring, indicating a rapid time decay in recall accuracy. As previously illustrated, the main effect of this change is that positively biased recall becomes much more likely in Fall: if $g_t = 0$, the marginal utility of recalling $r_t = 1$ is substantially higher when $f_t = 1$.

Recalling a higher grade also affects continuation values, in that both g_t and y_t are more likely to equal 1 if a higher grade is recalled. However, these feedback effects on state variables are only of second-order importance for decision making.[15]

Unobserved heterogeneity. In Table 1.9, we report estimates for the model which permits students to take one of two unobserved types, $c \in \{1, 2\}$. Recall that in this model, we permit c to influence both preferences for ego-enhancing recall and the evolution of state variables.

Parameter estimates for this specification reveal that the basic model masks substantial heterogeneity in recall preferences across students. $\theta^1 = 2.3$, so type 1 is characterised by very strong preferences for self-enhancing recall of grades. Type 1's preferences for recall are also characteristic of fragile self-esteem: $\theta_y^1 > 0$, meaning positive memories of grades are more attractive when prior self-esteem is higher. On the other hand, type 2 has substantially weaker preferences for self-enhancing recall: $\theta^2 = 0.19$. Type 2 also exhibits fragile self-esteem, although less so than for type 1. Notably, our auxiliary model reveals that both indicators of a growth mindset strongly predict membership of type $c = 1$.[16]

Another substantial difference between the two types is in the unobserved factors

[15] As previously noted, we take the standard practice of fixing the discount factor β since it is not identified in stationary dynamic discrete choice models. However, in Table A.6 we report coefficient estimates for a wide set of values of β. Parameter estimates vary very little across specifications. The only parameter significantly affected by changes to β is θ: if β is smaller, the present benefit of biased recall has more work to do in explaining choice probabilities.

[16] Excluding the auxiliary model for type probabilities leads to similar results.

Table 1.8 Parameter estimates for basic structural model

Utility function	
θ	0.65^{***} (0.01)
θ_y	0.23^{***} (0.02)
λ	2.23^{***} (0.02)
λ_f	-1.00^{***} (0.02)
Markov process: actual grades	
μ_1^g	-1.96^{***} (0.02)
μ_y^g	0.42^{***} (0.02)
μ_g^g	1.62^{***} (0.03)
μ_f^g	-0.08^{***} (0.02)
μ_r^g	0.90^{***} (0.03)
Markov process: self-esteem	
μ_1^y	-1.78^{***} (0.02)
μ_2^y	1.86^{***} (0.02)
μ_3^y	0.01 (0.03)
μ_4^y	0.07^{***} (0.02)
μ_5^y	0.67^{***} (0.03)
Observations	1,176

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. Standard errors in parentheses. Model estimated by full-information maximum likelihood. Discount factor set to $\beta = 0.9$.

affecting attainment. Type 2 achieves much more poorly than type 1: $\omega^g = -0.25$, indicating a 25pp lower conditional probability of receiving $g_t = 1$. They also have lower self-esteem: $\omega^y = -0.04$, indicating the conditional probability of having high self-esteem is about 4pp lower for type 2. Thus, our approach of permitting heterogeneity in recall preferences to be correlated with heterogeneity in state variables is justified by the data.

There are also differences across types in the effects of recalled grades on state variables. Since $\omega_r^y > 0$, the self-esteem of type 1 is *less* vulnerable to recalling low grades. This suggests that in addition to being less likely to inflate recall of poor grades, type 2s are also less likely to defend their self-esteem when they do recall low grades. One interpretation is that there are two layers of self-deception when receiving a poor grade. First, if possible, students recall a higher grade than actually received. However, if it is infeasible or psychologically too costly to recall an incorrect grade, type 1s further inure themselves to low attainment by perceiving both a weaker mapping from grades to ability and a higher unconditional mean for ability. Another possibility is of course that type 1 is right to be less threatened by low attainment because it is objectively less informative. There is some support for this interpretation from the fact that $\omega_r^g > 0$, implying attainment is genuinely less responsive to recall for type 1 than type 2.

While the self-esteem of type 1 is more resilient to low grades than that for type 2, a natural question is whether this conclusion extends to wellbeing. As previously noted, θ in our model is related to the parameter k in Koszegi et al. (2022)'s model, governing the strength of the individual's concern for their ego. In FSE, higher values of k increase the probability of recalling positive outcomes and reduces the probability of recalling negative ones, holding mood constant. However, a larger value of k also crucially implies that mood is more sensitive to self-esteem. For this purpose we re-estimate the model with an auxiliary state variable capturing school-centric welfare, measured by the question 'How much do you like school in general?', and allow y_t to have a heterogeneous marginal effect on welfare by type.[17] Instructively, the marginal effect of self-esteem on welfare is *greater* for type 1. This suggests an interpretation of the difference in θ^c across types; recalling favourable grades is more important to type 1s because self-esteem is more important for their welfare.

One concern with this interpretation may be that θ^c differs across types simply

[17] More detail on this alternative version of the model can be found in Section A.2.

because, when grades are forgotten, recall converges on historical attainment and type 1 does substantially better than type 2 on average. We can directly examine the plausibility of this claim by inspecting the steady-state distributions of r_t and g_t for both types. For type 1, $P(r_t = 1|f_t = 1, c = 1) = 0.90$ in the long run, whereas $P(g_t = 1|f_t = 1, c = 1) = 0.57$. Likewise for Spring periods, $P(r_t = 1|f_t = 0, c = 1) = 0.80$, but $P(g_t = 1|f_t = 0, c = 1) = 0.55$. So, type 1 recalls high grades 90% of the time in Fall, and 80% of the time in Spring, but only achieve them 55–57% of the time in the long run. The difference in long-run recall *bias* relative to type 2 is very stark. In Spring, the long-run probability of type 2 recalling a high grade (30%) is only 9pp higher than the probability of achieving one (21%), much smaller than the 25pp bias for type 1. In Fall, the positive bias in recall is 16pp for type 2s, relative to a 33pp bias for type 1. Thus, differences in recall preferences across types cannot be rationalised by average differences in attainment.

1.4.5 The dynamics of self-esteem

The model outlined in Sections 1.4.1 and 1.4.2 creates a rich dynamic interaction of recall choices, self-esteem and actual attainment. In this section, we explore how the dynamic distributions of these variables depend on initial conditions and across the two unobserved types of student.

As previously noted, shocks to self-esteem are possible in every period. Neither high self-esteem ($y_t = 1$) or low self-esteem ($y_t = 0$) is an absorbing state, so the long-run distribution of self-esteem is only a function of model parameters. Nonetheless, our model can tell us for how long various shocks to initial conditions affect the distribution of self-esteem. As previously outlined, this analysis is related to the notion of fragility used in Koszegi et al. (2022): if a self-esteem equilibrium is more fragile, it may be more easily perturbed by shocks in the short run.

Basic model. Figure 1.5 illustrates the dynamics of self-esteem for the basic model, without unobserved heterogeneity. Since t corresponds to survey periods, each time period is about half a year. For this reason, self-esteem is also higher in even-numbered periods (corresponding to *Fall*, when higher grades are recalled and self-esteem is also higher conditional on recall).

Unsurprisingly, the largest dynamic impact on the distribution of self-esteem is from a shock to initial conditions for self-esteem. An initial condition with $y_0 = 0$

Table 1.9 Parameter estimates for structural model with unobserved heterogeneity

Utility function	
θ^1	2.27^{***} (0.25)
θ^2	0.19^{***} (0.03)
$\theta^2 - \theta^1$	-2.08^{***} (0.23)
θ^1_y	0.54^{*} (0.28)
θ^2_y	0.14^{**} (0.06)
$\theta^2_y - \theta^1_y$	-0.40 (0.42)
λ	2.37^{***} (0.03)
λ_f	-1.20^{***} (0.03)
Markov process: actual grades	
μ^g_1	-0.80^{***} (0.13)
μ^g_y	0.25^{***} (0.02)
μ^g_g	1.55^{***} (0.03)
μ^g_f	0.12^{***} (0.01)
μ^g_r	0.06 (0.15)
ω^g	-0.25^{***} (0.01)
ω^g_r	0.12^{***} (0.01)
Markov process: self-esteem	
μ^y_0	-1.55^{***} (0.13)
μ^y_y	1.82^{***} (0.02)
μ^y_g	0.02 (0.03)
μ^y_f	0.06^{***} (0.02)
μ^y_r	0.42^{***} (0.14)
ω^y	-0.04^{***} (0.00)
ω^y_r	0.06^{***} (0.01)
Auxiliary model for $c = 2$	
γ_1	0.41^{***} (0.13)
γ_2	1.09^{***} (0.34)
γ_3	0.78^{*} (0.44)
Observations	1,176

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. Standard errors in parentheses. Model estimated by full-information maximum likelihood. Discount factor set to $\beta = 0.9$.

rather than $y_0 = 1$ implies just over a 40-percentage-point decrease in the likelihood of having high self-esteem in the following period. Notably, even 2 years later (in $t = 4$), $P(y_t = 1)$ is about 5 percentage points lower in scenarios with $y_0 = 0$. This persistence is largely driven by serial dependence in y_t, captured by μ_y^y, while the mood-memory utility parameter θ_y is quantitatively unimportant. Thus, if mood-memory plays an important role in the dynamics of self-esteem, it must be through other means than just recall of the most recent mathematics grade. As previously noted, this is plausible if y_{t-1} affects recall of unobserved signals in t.

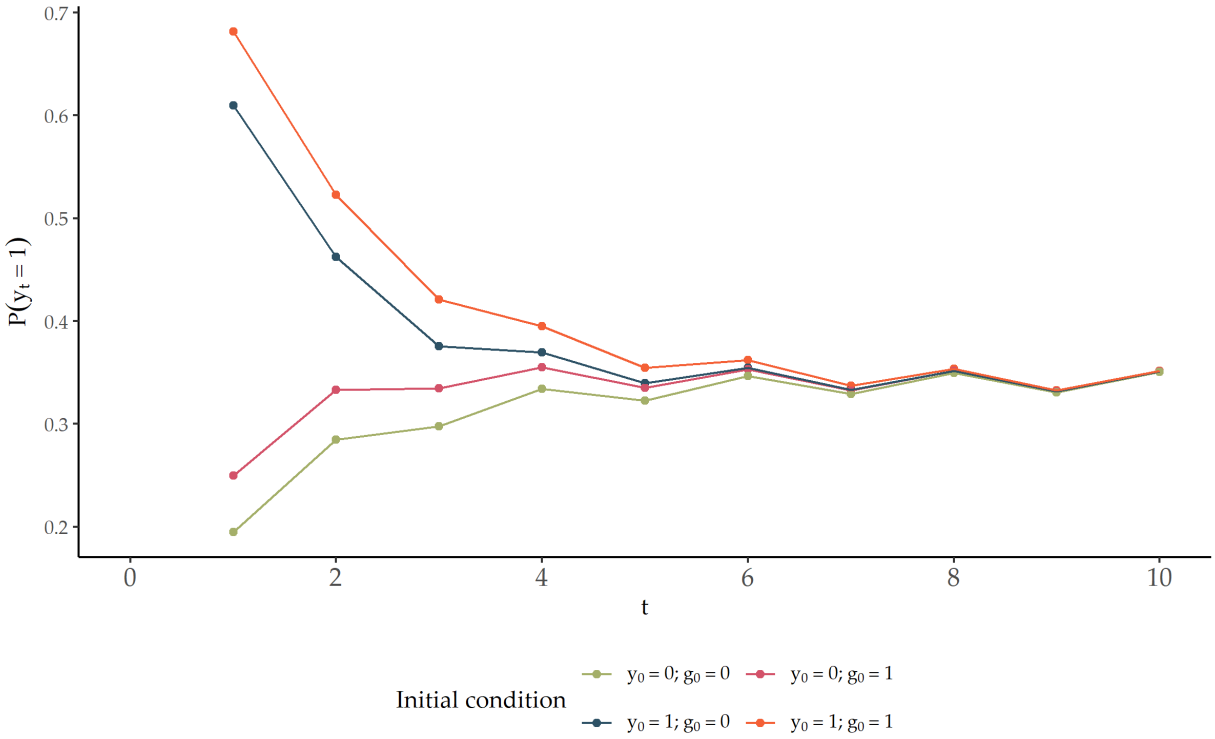
Although actual grades g_t have no direct impact on y_t ($\mu_g^y = 0$), they significantly impact what grade is recalled and therefore indirectly impact self-esteem because $\mu_r^y > 0$. Figure 1.5 indicates a highly persistent impact on self-esteem of receiving a single low grade in a quarterly report card. $P(y_t = 1)$ is around 8 percentage points lower in starting conditions with $g_0 = 0$, and the effect barely dissipates until about 1.5–2 years after the initial shock.

Model with unobserved heterogeneity. Figure 1.6 demonstrates the dynamics of $P(y_t = 1)$ by type in the extended model with unobserved heterogeneity. Types 1 and 2 have quite different long-run equilibria for self-esteem. Around half of the difference in the long-run equilibrium level of $P(y_t = 1)$ is driven by a fixed effect ω^y, and the other half by differences in recall preferences (θ and θ_y). That is, type 1 has higher self-esteem, and around half of this can be explained by the fact that in each period, they have a stronger preference to recall higher grades than type 2.

Recalling our earlier discussion, one measure of the fragility of an SEP at $y_t = k$ is $Pr(y_t = k|y_{t-1} = k)$. This measure is monotonically related to the size of shock required to displace a student from $y_t = k$ to the other equilibrium in the following period. Using this definition, we can see that the high SEP at $y_t = 1$ is equally stable for type 1 and type 2, conditional on receiving a high grade $g_t = 1$. However, conditional on receiving a low grade, the high SEP is more fragile for type 2. This type has only a 60% chance of remaining at $y_t = 1$ in the following period, compared to 70% for type 1. In other words, an equilibrium with unrealistically high self-esteem is more stable for type 1, who places a much higher value on ego.

While both types are affected similarly by direct shocks to self-esteem, they differ significantly in their responses to shocks to actual grades. In particular, type 1 is much more resilient to attainment shocks than type 2, largely because they are more likely to recall a high grade even when they receive a low one. This is illustrated

Fig. 1.5 Dynamics of $P(y_t = 1)$ — basic model



Note: $y_t = 1$ implies the student believes they have above average academic ability for their school and year group. t corresponds to survey periods, so each period is approximately half a year.

Fig. 1.6 Dynamics of $P(y_t = 1)$ — model with unobserved heterogeneity

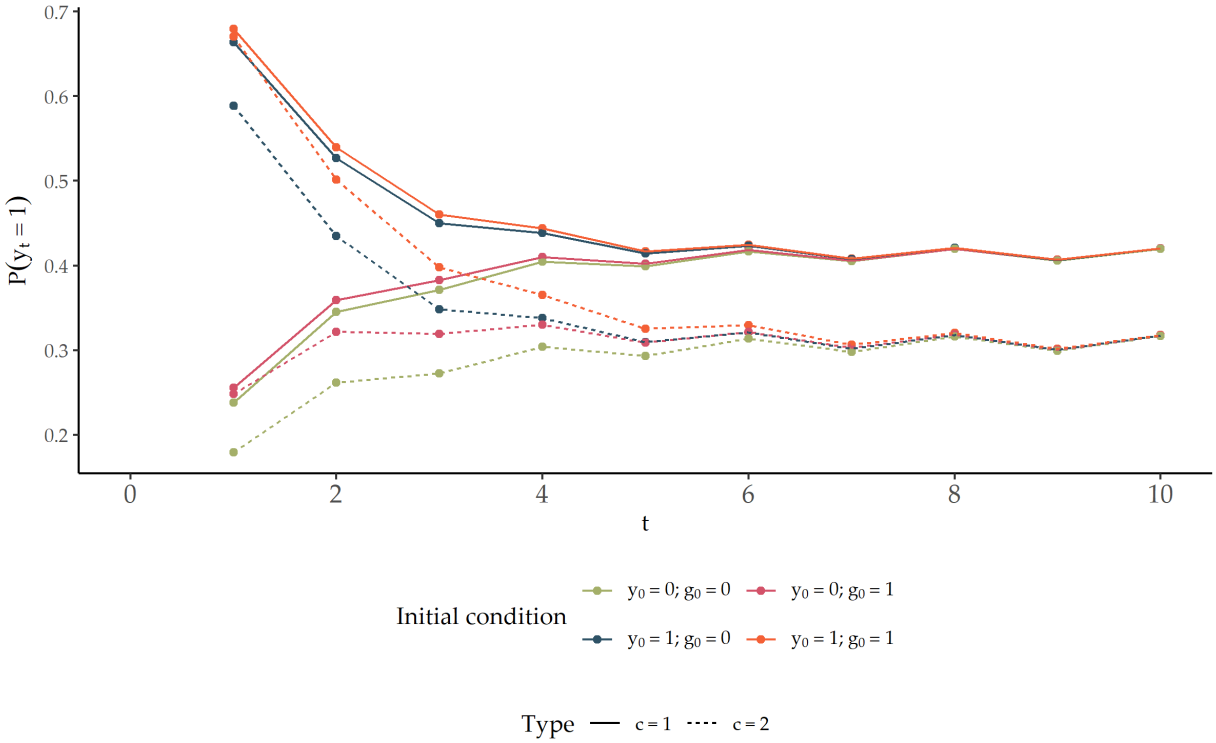
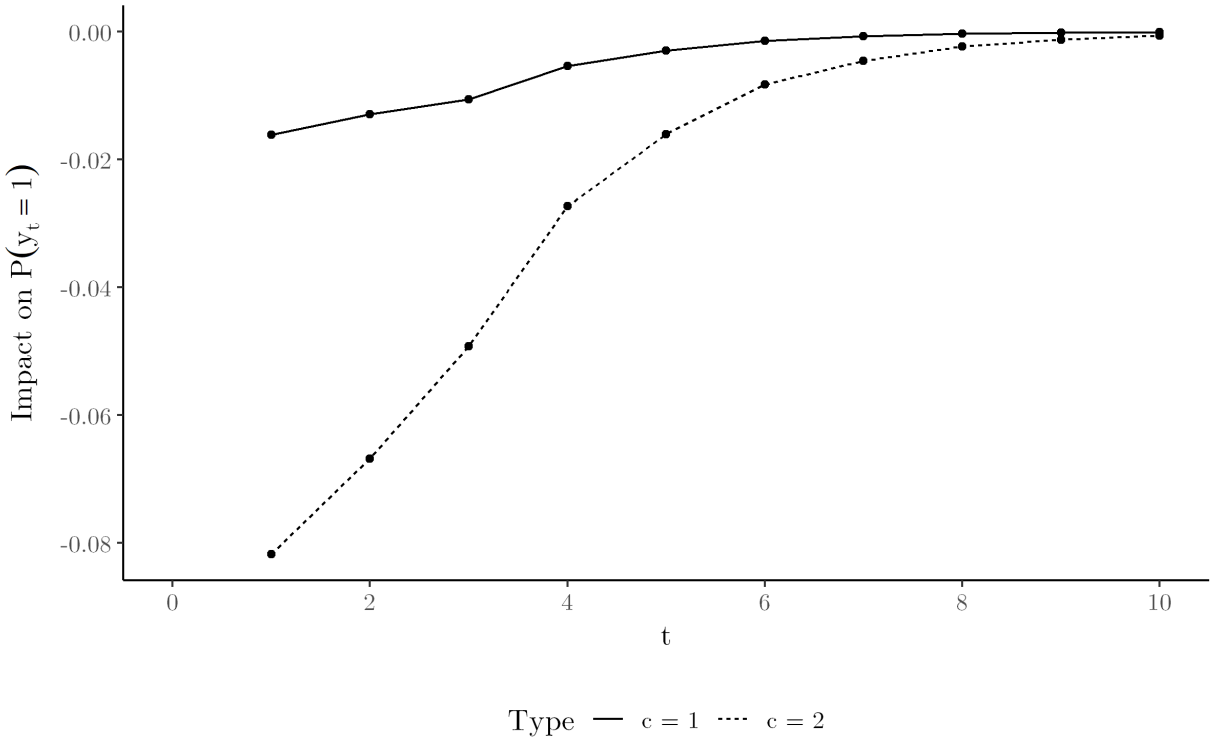


Fig. 1.7 Impact of initial condition $g_0 = 0$ vs. $g_0 = 1$



Note: $y_t = 1$ implies the student believes they have above average academic ability for their school and year group. t corresponds to survey periods, so each period is approximately half a year. Initial condition in both cases has $y_0 = 1$.

more clearly in Figure 1.7: the impact of an initial condition ($y_0 = 1; g_0 = 1$); relative to ($y_0 = 1; g_0 = 1$) is sharply negative for type 2 but negligible for type 1. Just as in FSE, those who place a higher value on ego tend to have a more positively biased recall process, making high self-esteem less vulnerable to shocks.

Information intervention. One category of policy intervention that has attracted particular interest in the literature on education is the provision and reinforcement of feedback information (see e.g. Dizon-Ross, 2019). In our setting, we can model this intervention by setting $r_t = g_t$ and examining how the dynamics of our two key state variables, y_t and g_t, are affected.

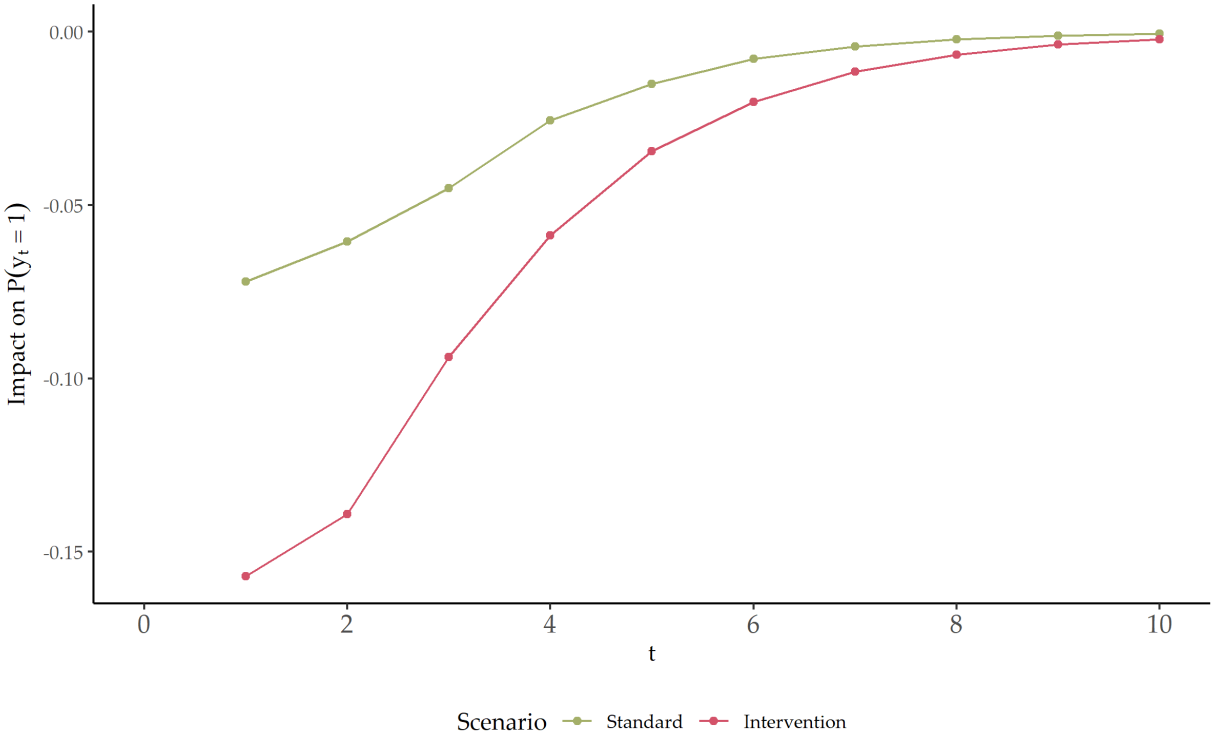
The effects of this intervention are most obvious when a poor grade is received. To explore this case, we plot the impact on $P(y_t = 1)$ of starting in an initial condition ($y_0 = 1; g_0 = 0$) relative to an initial condition ($y_0 = 1; g_0 = 1$), and explore how that impact differs with and without the intervention. To begin with, we do so using the basic model without unobserved heterogeneity (Figure 1.8).

The initial impact of receiving a low grade on self-esteem is almost three times larger under the informational intervention. Two years after the intervention, the expected impact is almost as large as would have been the initial impact without the intervention. Thus, self-deception is crucial for the protection of self-esteem, especially for low achievers. That protective mechanism would precluded by better information transmission by limiting opportunities for biased beliefs.

Clearly, if self-esteem is founded on self-deception, it is likely to be unrealistic. A natural question is whether correcting unrealistically high self-esteem could be beneficial for students. For two reasons, we can be fairly sure that such an effect is implausible in our context. First, we know that higher self-esteem is associated with higher psychological wellbeing while at school, controlling for grades (Section A.2). Second, our basic parameter estimates indicate that high self-esteem (and recalling a higher grade) is associated with higher future attainment; both μ_y^g and μ_r^g are greater than 0.

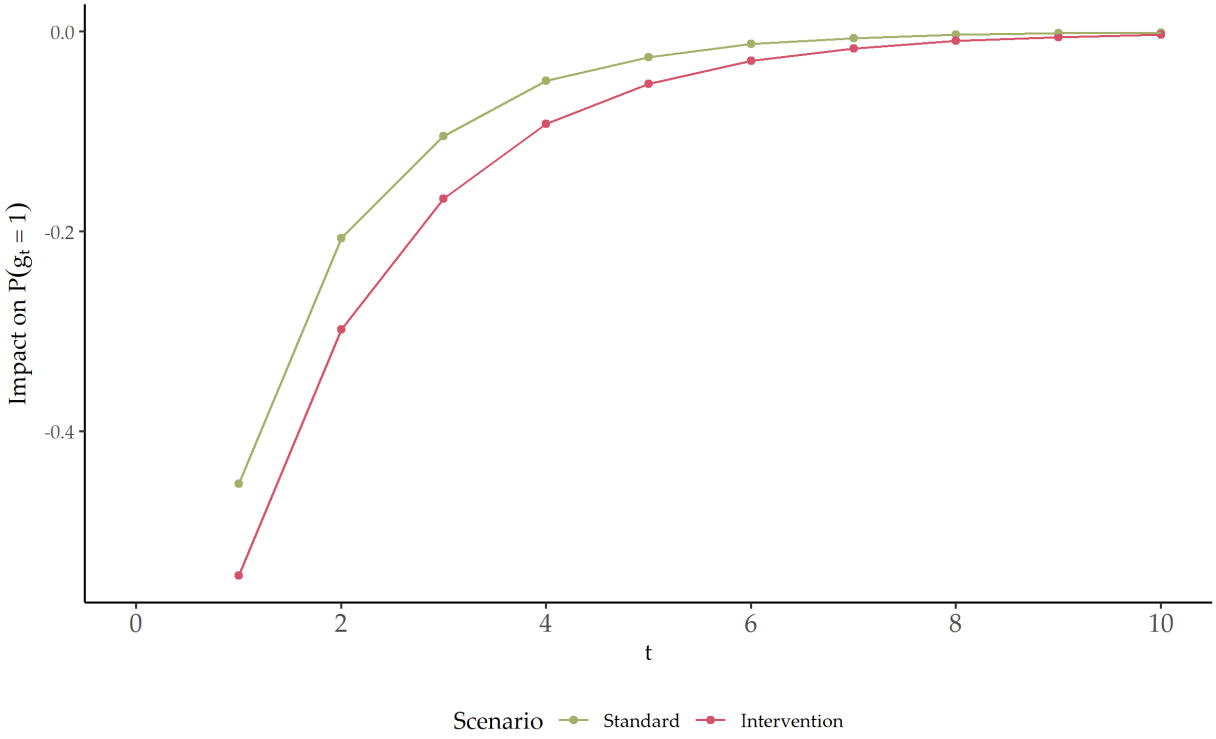
To illustrate this point, we also plot the impact of the informational intervention on how g_t responds after a shock to its own initial condition (Figure 1.9). Under the informational intervention, we see that the impact on g_t is more pronounced: recalling a low grade directly affects attainment in the following period, and this negative impact is inherited in future periods. Strikingly, the impact of the informational intervention in this setting is that students are approximately 10pp (25%) more likely

Fig. 1.8 Impact of poor initial grade on $P(y_t = 1)$ — with and without information intervention



Note: $y_t = 1$ implies the student believes they have above average academic ability for their school and year group. t corresponds to survey periods, so each period is approximately half a year. Initial condition in both cases has $y_0 = 1$.

Fig. 1.9 Impact of poor initial grade on $P(g_t = 1)$ — with and without information intervention

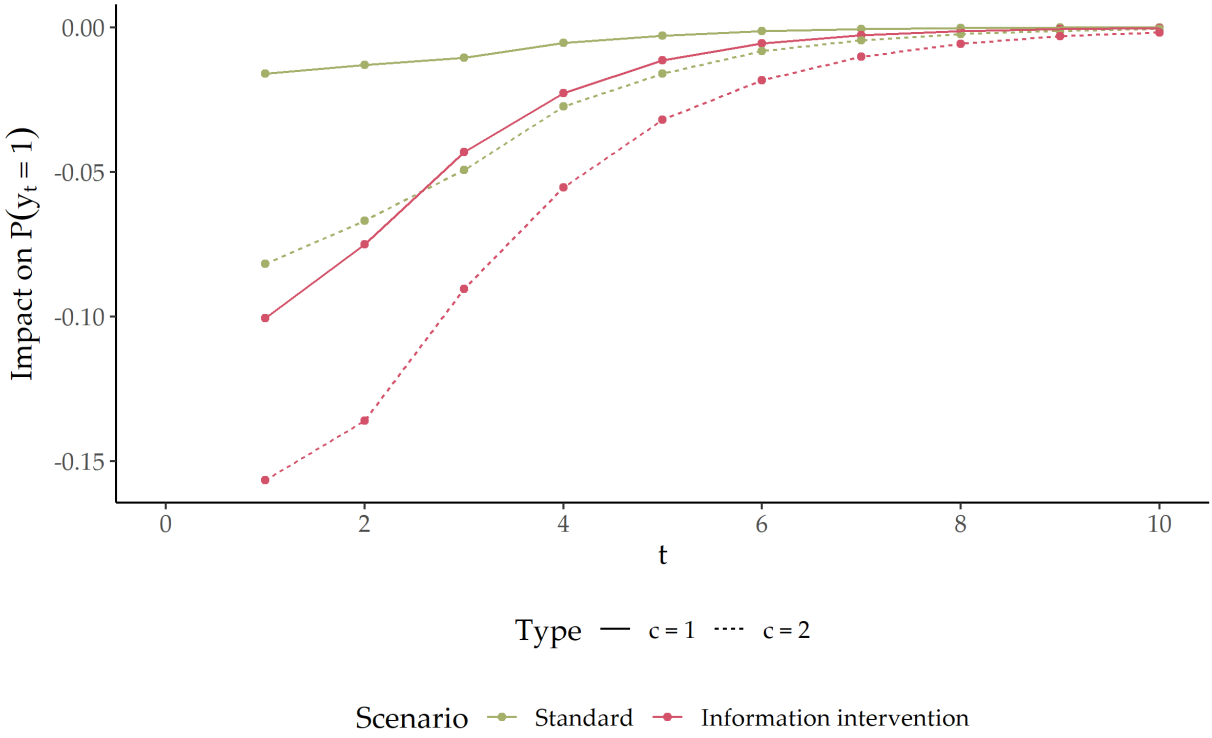


Note: $g_t = 1$ indicates an actual grade of *Good* or *Excellent*. t corresponds to survey periods, so each period is approximately half a year. Initial condition in both cases has $y_0 = 1$.

to achieve another poor grade in the period after a low one. While one should not use this result to claim that unrealistically high self-esteem is beneficial in general, our results do highlight a possible negative side-effect of informational interventions: for students who achieve poorly, constraints on self-deception can result in psychological costs and a negative impact on attainment. These findings is consistent with theories connecting self-confidence to performance (Bénabou and Tirole, 2002; Compte and Postlewaite, 2004), but has received little attention in applied work examining the removal of informational frictions in education.

We also examine how the same intervention affects different types in our extended model with unobserved heterogeneity. As before, the main effect of the intervention is to increase students' sensitivity to achieving poor grades. Starting with the impact on self-esteem, we see that both types are strongly negatively affected by the intervention (Figure 1.10). If the intervention were only implemented on type 1, it would have the effect of completely closing the gap in the impact of receiving a poor

Fig. 1.10 Impact of poor initial grade on $P(y_t = 1)$ — with and without information intervention, 2-type model

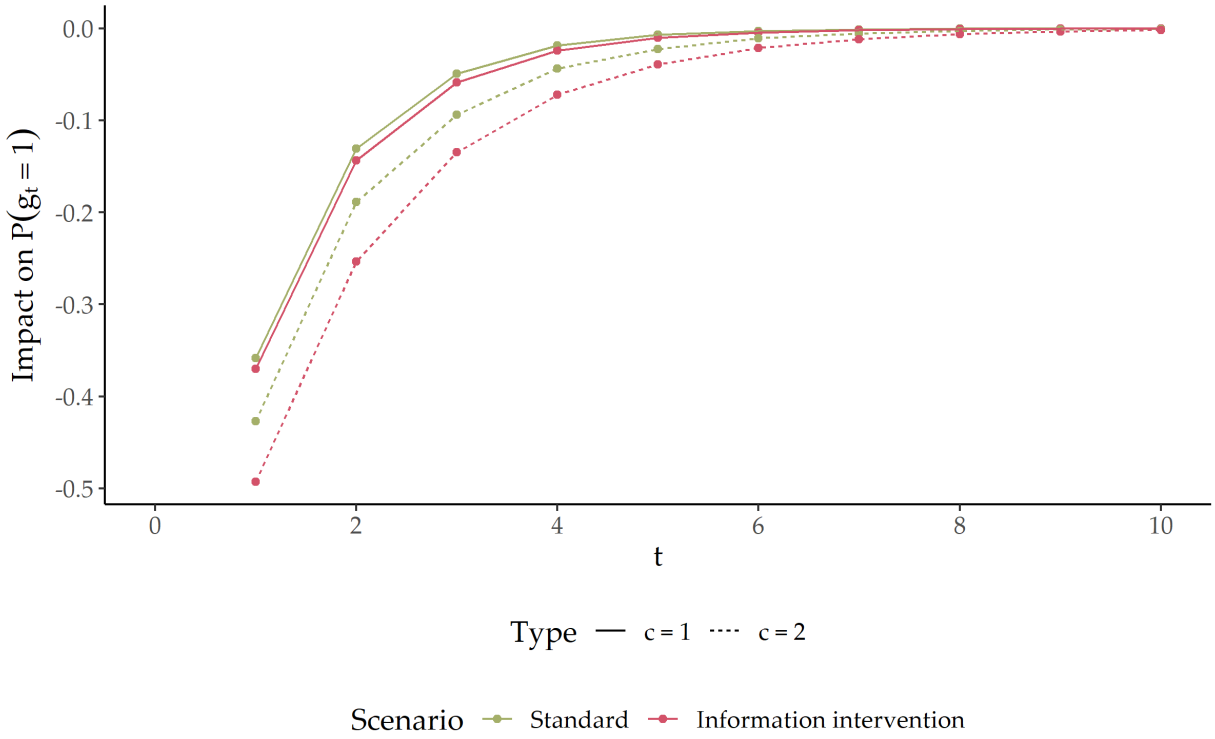


Note: $y_t = 1$ implies the student believes they have above average academic ability for their school and year group. t corresponds to survey periods, so each period is approximately half a year. Initial condition in all cases with $y_0 = 1$.

grade: type 1 relies on biased recall to maintain their self-esteem in the face of poor attainment. Even though type 2 is much less likely to bias their recall of poor grades without the intervention, the information intervention has almost as large an impact on self-esteem: since $\omega_r^y = 0.06$, their self-esteem is more sensitive to what grade is recalled.

We can also examine how the intervention would impact dynamics for g_t (Figure 1.10). Here, the difference between types is significant: type 1 is essentially unaffected by the intervention, since r_{t-1} has little impact on g_t. However, recall is more strongly associated with future attainment for type 2, who is generally low achieving. Thus, an informational intervention could worsen attainment gaps because shocks to self-esteem are more consequential for low achievers.

Fig. 1.11 Impact of poor initial grade on $P(g_t = 1)$ — with and without information intervention, 2-type model



Note: $g_t = 1$ indicates an actual grade of *Good* or *Excellent*. t corresponds to survey periods, so each period is approximately half a year. Initial condition in all cases with $y_0 = 1$.

1.5 Conclusion

This chapter finds support for theories predicting that individuals distort their recall of ego-relevant information to protect their self-esteem. It also demonstrates that these distortions are enabled by biased memory loss. Principally, these findings indicate that in contexts where negative signals convey socially valuable information, additional effort may be required to ensure that they are properly digested and retained. The role we find for biased memory loss, matching Zimmermann (2020) in the laboratory, is particularly notable, since it highlights that beliefs may rapidly diverge from objective reality over time. It also suggests that greater effort may be required to bolster the retention of negative signals as time passes.

Our use of rich longitudinal data facilitates a test of a recently proposed two-way relationship between beliefs and recall (Koszegi et al., 2022). This theory is particularly powerful because it suggests a possible path-dependence for beliefs through biased, associative recall. It also provides a more compelling explanation for biased recall than existing models: unfavourable signals may produce a disproportionate impact on self-esteem, relative to a standard Bayesian model, by stimulating recall of associated negative signals. Our model finds support for the key mechanism, that biases in recall depend on prior self-esteem. Like the theory suggests, the effect of sole signals on self-esteem can be both large and long-lasting.

The welfare implications of biased recall are generally not well studied empirically. In our model, we take it as given that recall distortions serve to optimise psychological welfare, taking cues from existing work on ego utility. This approach seems justified through auxiliary evidence that recall and self-esteem impact students' wellbeing at school. Additionally, it does not appear to be the case that inflated recall negatively affects academic performance; recalling a higher grade (and having higher self-esteem) is in fact associated with improved attainment in our sample, especially for low achievers. This is consistent with theories of belief- or confidence-based motivation (Bénabou and Tirole, 2002; Compte and Postlewaite, 2004) rather than theories in which belief distortion comes at the cost of distorting choices (Brunnermeier and Parker, 2005). Thus, our results highlight that policies aiming to correct recall errors and overconfidence should be combined with measures to protect the psychological welfare of low achievers.

Chapter 2

Recall Distortion and Past Choices

2.1 Introduction

The recall of past information is crucial for belief formation but is often systematically biased. One potentially important source of bias is belief-based utility, the idea that beliefs can affect utility independently of experienced outcomes. One basic testable prediction of this theory is that individuals may distort recall of information that underpins such beliefs. While this prediction has found extensive empirical support in the domain of self-image, little is known about whether it holds for beliefs that are not relevant to the ego.

One possibility of particular interest is that individuals distort beliefs to support optimism about uncertain outcomes, consistent with anticipatory utility theory (Caplin and Leahy, 2001) and related models of optimal expectations (Brunnermeier and Parker, 2005). We call this *consequentialist* belief distortion. This phenomenon would have material implications for the measurement of beliefs in a large span of work in empirical economics: it would suggest that beliefs are often endogenous to past choices. For example, when measuring parents' beliefs on returns to investment in their children, this mechanism means that heterogeneity in beliefs across parents with different investment choices increases after choices have been made; those who have chosen to invest more in their children will distort beliefs to favour their decision.

In order to demonstrate that consequentialist motives cause beliefs to be distorted, we experimentally study recall of a variable which has the same objective distribution across treatment and control groups but also determines an unobserved, favourable outcome in the treatment group. Crucially, since subjects in the treatment group

receive no information on this outcome, any standard model of rational belief formation would predict that average beliefs are exactly the same as in the control group. However, in models where subjects prefer to hold optimistic beliefs about relevant outcomes, such as anticipatory utility, there may be a motive to distort beliefs in the treatment group.

The experiment is divided into two sessions, each featuring a task that involves dragging sliders to predetermined values (Gill and Prowse, 2012). Participants in the control group must complete the tasks in both sessions. In contrast, those in the treatment group choose in session 1 whether they will complete the tasks in session 2. For these participants, the sole benefit of completing the tasks in session 2 is the potential to earn a payment for an assigned partner in the control group. Notably, 81% of the treatment group participants opt to perform the additional tasks.

The crucial element of the experiment design is a *signal* hidden in the task for session 1: the number of times they had to drag sliders to the maximum value. After completing this task, all participants are informed that they were shown the signal and are given its objective distribution. However, treatment group members who chose to complete the session 2 task are also informed that the signal is positively correlated with the likelihood of generating a payment for their partner. Since this structure amounts to an encouragement design where the treatment is only administered to those electing to do the task for their partner, we can safely assume that the motive to distort beliefs should be consistent across all those receiving the treatment. If average beliefs about the signal are higher in the treatment group, beliefs are biased. Only individuals receiving treatment have any motive to adopt a more optimistic view of the task's outcome.

The results confirm our main hypothesis: individuals have systematically higher recall of the signal in the treatment group, even when they can almost double their experimental payment for providing the right guess. In our setting, where recall is cognitively demanding because the signal was hidden in a tedious task, we observe this effect shortly after the signal is transmitted. Our experiment is also set up to explore whether recall becomes more distorted over time through memory loss. We obtain weak evidence in favour of this hypothesis: conditional on recall in session 1, recall in session 2 is higher in the treatment group than the control group, but the average treatment effect on the *change* in recall between sessions is not significant, seemingly because errors in recall across the two sessions are partly independent of each other. That being said, we observe clear evidence of biased memory loss

among those with higher self-reported altruism scores. Belief distortions are also generally larger for this group and for women. These subsets of our sample are likely to have stronger motives to hold the belief that their altruistic action actually yielded positive consequences, supporting the interpretation that the belief distortion we observe is driven by belief-based utility.

If beliefs are persistent through time, models of anticipatory utility imply that the present motive to distort beliefs should be larger when belief distortions can be sustained for longer. To test this prediction, we include a further treatment condition within our treatment group: for half of participants receiving the belief treatment, uncertainty is never resolved. For the other half, they know the true value of the signal will be revealed at the end of the experiment. We find no difference in beliefs across these two groups, demonstrating that the ability to distort beliefs indefinitely has no influence on the size of belief distortions. In our setting, since biases in beliefs are also not eliminated by relatively large financial incentives, our interpretation is that biases in recall appear to be too heavily encoded to be negated. Future work may be able to elucidate whether the same effect is observed in settings where biases in beliefs are more easily eliminated with enough cognitive effort.

Current literature. The literature on memory and beliefs in economics is split into two branches. One of those deals with systematic biases in recall that can be loosely characterised as dealing with automatic cognitive processes, such as associative recall and salience (Bordalo et al., 2020, 2023). This chapter is more closely related to another branch in which systematic distortions of recall result from belief-based utility, the idea that beliefs — not just experienced outcomes — directly affect utility.

Within work on belief-based utility, two major motives for belief distortion have been proposed. The first is the desire to maintain or project positive self-image (Bénabou and Tirole, 2002; Köszegi, 2006; Bénabou, 2015). This idea has received significant attention in empirical work, which shows that individuals systematically distort recall and beliefs about ego-relevant variables in laboratory settings (Eil and Rao, 2011; Di Tella et al., 2015; Zimmermann, 2020) and the field (Huffman et al., 2022; Roy-Chowdhury, 2022). Closest to the present study is Gödker et al. (2021), who consider the role of biased memory in an investment setting. Subjects in their experiment systematically display positive biases in their recall of investment returns, oversampling positive outcomes and undersampling negative ones. They establish that biases appear to be driven by ego preservation; they only arise when investors

actively choose their assets.

Our study shares Gödker et al. (2021)'s objective of exploring motivated distortions to beliefs on the yet unknown consequences of past choices, but the principal point of divergence is that our experiment eliminates any rational link between beliefs and self-image and instead closely examines the role of anticipatory (consequentialist) motives for belief distortion. In our experiment, where subjects make choices before knowing the signal even exists, subjects cannot rationally influence their beliefs about their own preferences or characteristics. Although recent evidence suggests individuals prefer information structures that allow them to maintain optimistic views about future outcomes in the absence of rational ego motives (Masatlioglu et al., 2023), we know relatively little about whether such motives are important in distorting recall and belief formation. Our results provide an affirmative answer to this question.

Since we leverage other-regarding preferences as the main motive for belief distortion in our study, we also contribute to the literature on the role of belief distortion and self-deception in altruism. Other studies find that individuals distort beliefs (Dana et al., 2007; Di Tella et al., 2015; Exley, 2016), avoid information (Roy-Chowdhury, 2024), or implement more creative cognitive distortions (Exley and Kessler, 2024) to rationalise selfish behaviour. Our study takes a slightly different approach in exploring how those who take potentially altruistic actions distort their beliefs to hold more favourable views of the consequences of those outcomes. In working to eliminate self-image concerns, we also demonstrate a purer form of altruism: in our experiment, unlike other existing studies, individuals cannot possibly manipulate their beliefs about their own *rationale* for making an altruistic or selfish choice, but merely the *consequences* of such a choice for a third party.

Outline. This chapter proceeds as follows. In Section 2.2, we outline the design of the experiment. In Section 2.3, we analyse the experimental data. In Section 2.4, we conclude.

2.2 Experiment design

Our experiment is run on over 1000 participants from Prolific, an online survey panel. Participants are residents of the United Kingdom.

The experiment is split into two sessions, taking place five days apart. Subjects are randomised into equally sized treatment and control groups. The key difference

between the two groups is that the consequentialist motive to distort beliefs on a variable x is only introduced in the treatment group, but x is known by participants to be generated by the same objective distribution in both treatments.

At the beginning of session 1, subjects in the treatment group choose whether to complete a task in session 2 and potentially generate a payment made to an assigned *partner* in the control group. To introduce a guilt motive as well as a possible fairness one, subjects are informed that their partner will have to complete mandatory tasks in session 2 and will receive a lower payment by default, and that their partner will find out what the subject chose and whether a payment was (or could have been) made.

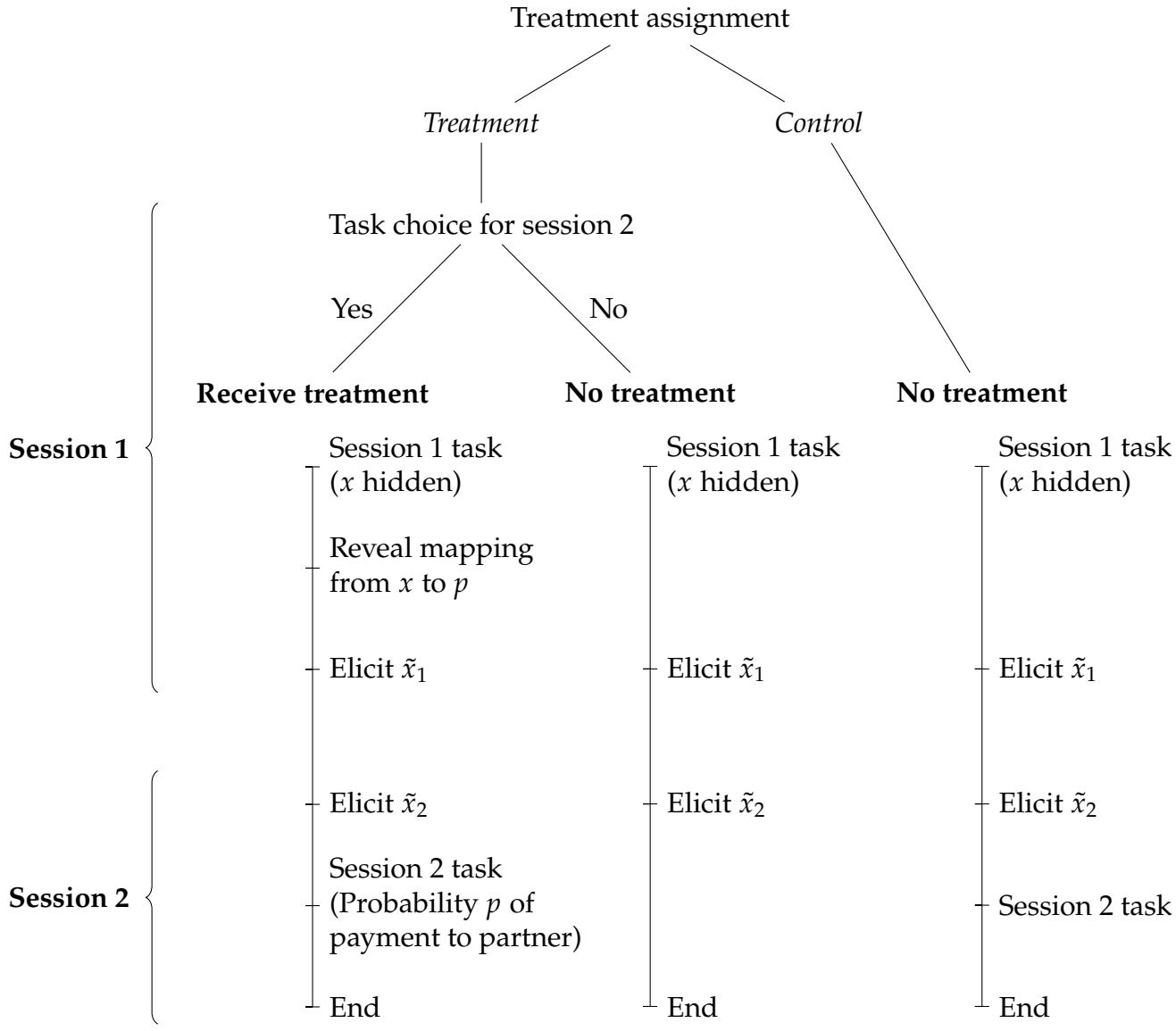
Our experiment uses an encouragement design, in that the belief treatment is only implemented when a subject in the treatment group chooses to do the session 2 tasks. As such, we generally focus on average treatment effects (ATEs) but also report average treatment effects on the treated (ATTs) for key analyses. We only implement the treatment for a subset of the treatment group because the remainder, those who choose not to do the tasks, could prefer either for the payment not to be generated or for it to be generated. Implementing the treatment for the whole treatment group would have been likely to introduce unwanted heterogeneity in the direction of belief distortion.

Screenshots of experiment pages can be seen in the Appendix (Section B.2). The two sessions of the experiment proceed as follows:

Session 1:

- Subjects in the treatment group decide if they want to complete the 2-minute task in session 2. They are told that if they do so, on average, there is a 60% chance of a payment to a randomly assigned *partner* in the control group.
- Subjects complete the task in session 1, involving dragging 20 sliders to specified numbers. The number displayed for each slider can be between 0 and 150. Within the last 10 sliders, a random number of sliders $x \in \{0, 1, ..., 10\}$ are generated to display the maximum value, 150.
- After completing the sliders, subjects receiving treatment are informed that x influences the probability of a payment being made to their partner in session 2 and are shown $p(x)$, the mapping from x to the probability of a payment to their partner.

Fig. 2.1 Experiment overview



Note: Session 2 takes place 5 days after session 1.

- Subjects are shown the distribution of x.
- We elicit point recall of x. Call it $\tilde{x}_1$.

Session 2 takes place a week later:

- We elicit point recall of x again. Call it $\tilde{x}_2$.
- Subjects complete the task if in the control group.
- Subjects in the treatment group complete the task if they chose to do so.
- x is revealed, if applicable.

Signal and payoffs. x is the number of times the participant has to drag sliders to 150 in the task in session 1. We expect x to be relatively salient because participants have to drag the relevant sliders all the way to the end, and because this only ever occurs on the last page of the task.

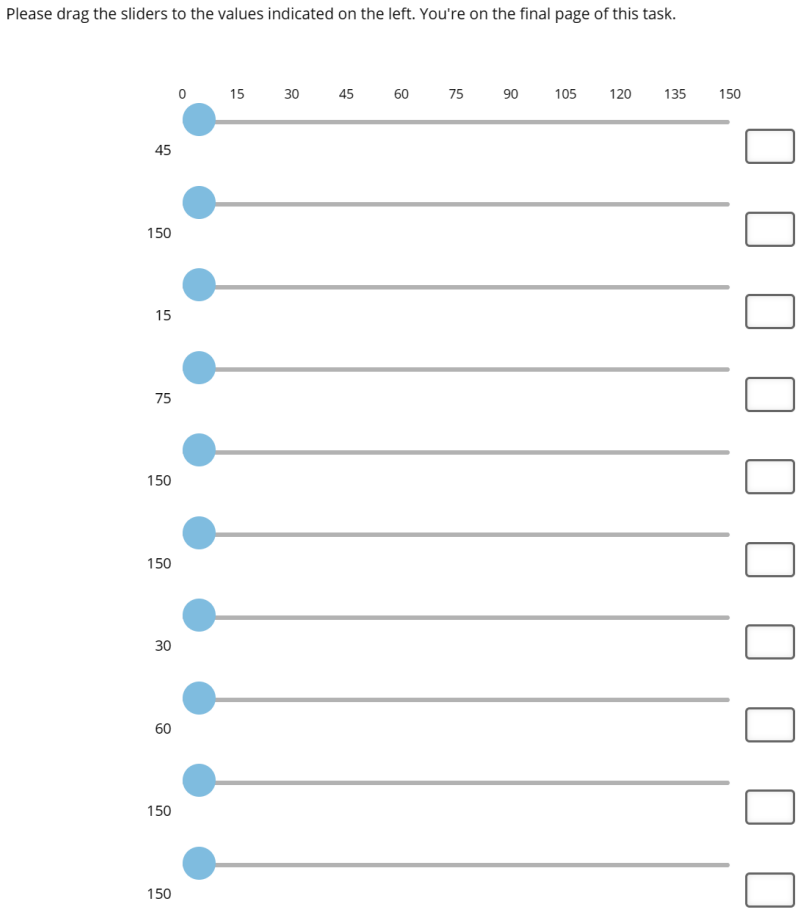
The distribution of x is as follows: with 98% probability, x is drawn from $\{2, 3, 4, ..., 8\}$ with an equal chance of each number in that set. With 2% probability, x is drawn from $\{0, 1, 9, 10\}$, again with equal chance of each number. Thus, the expected value of x is 5 and the distribution is symmetric. Subjects are made fully aware of this distribution each time their beliefs on x are elicited. We use this particular distribution so as to minimise boundary issues: recall is bounded by the support of x, so we want the vast majority of the sample to be situated away from both the upper and lower bounds.

The relevance of x to subjects in the treatment group who have chosen to do the session 2 task is that it influences the probability p that a payment will be generated for their partner. The exact mapping is $p(x) = 0.2 + 0.08x$: if $x = 0$, $p = 0.2$, but if $x = 10$, the payment will certainly be made. The average value of p is thus 0.6, matching the information given to subjects in the treatment group at the time they chose whether to do the session 2 task.

Mapping x to $p(x)$ may introduce additional cognitive demands for those receiving the treatment. To minimise any additional computational burden, $p(x)$ is displayed on-screen in real time as subjects select their belief.

Task choice and implementation of treatment. Subjects in the treatment group receive an additional payment at the end of session 2, almost doubling their pay for that session. Subjects in the control group do not receive the payment and must

Fig. 2.2 Revelation of signal x during task, for $x = 5$



Note: The signal x is the number of sliders on this page taking the maximum value, 150. Participants were unaware that this number had any relevance whilst on this page and it was impossible for them to return to it later.

complete an extra task.[1] In session 1, we give this information to subjects in the treatment group, tell them that they have the option to complete a task in session 2 to generate an equally sized payment for a *partner* in *Control*, and that their partner will find out what they chose and whether a payment was (or could have been) generated.

As previously mentioned, the treatment, where x is revealed to be relevant to p, is only *implemented* for those in the treatment group who choose to complete the tasks. This is because completing the tasks implies that the subject prefers for the payment to be made, creating a motive to distort beliefs on x upwards. However, choosing not to complete the tasks is consistent with both indifference and anti-social preferences. Individuals in this category could theoretically have a motive to upwardly or downwardly distort their belief that a payment would have been made, so we ensure that they do not receive the treatment at all.

Recall elicitation. We elicit point recall of x by simply asking subjects 'How many sliders do you think you dragged to the largest value, 150, on the last page of the task [in session 1]?'.

All subjects are incentivised to recall the correct value of x. In implementing the incentive scheme, we would like to mitigate common concerns around incentive compatibility and hedging. Danz et al. (2022) find that the best-performing incentivisation scheme is the binarised scoring rule without specific information on its implementation. We follow their approach in implementing the binarised scoring rule but only telling participants that if their guess is more accurate, they are more likely to be paid the chosen incentive. Incentives are drawn from $\{£0.25, £0.50, £0.75, £1, £1.25, £5\}$. All incentives other than £5 are drawn with 18% probability, and £5 is drawn with 10% probability. By including a range of possible payments, we can examine whether larger incentives are enough to offset motivated distortions of beliefs (as in Zimmermann, 2020; Gödker et al., 2021).

Our main interest is in comparing recall of x, in sessions 1 and 2, across the treatment and control groups. Across the two groups, participants have exactly the same information set on x, comprising its objective distribution and their best recollection of x as it was displayed. However, the subset of the treatment group

[1] One concern with this design may be that we generate wealth effects across treatment groups. However, the difference in wealth across groups is relatively small (£2), and later, we show that direct incentives for beliefs of up to £5 have no effect on elicited beliefs.

receiving the treatment may have a stake in believing x is higher: if so, there is a higher chance of them not having wasted their time with an irrelevant task, in that the probability p of a payment being generated for their partner is higher.

2.3 Data and Results

First, in Section 2.3.1, we outline some descriptive statistics on the sample. In the rest of this section, we conduct our main analysis on recall of the signal in sessions 1 and 2.

2.3.1 Summary statistics: Sample

1066 participants were recruited from Prolific, an online panel. Participants were only permitted to sign up if 100% of their prior study submissions had been accepted by researchers. The panel included only residents of the United Kingdom and all payments were denominated in pounds sterling.

One major attraction of online recruitment is that participants tend to be more representative of the population than in the typical laboratory. The sample was almost exactly balanced on sex, with 49.8% of respondents being female. The median age of the sample was 39, just below the last recorded UK-wide median of 40.7 in 2022. Participants were more likely to be white than the average UK resident, with a share of 87.8% in our sample compared to 82% across England and Wales as a whole. Unsurprisingly, participants were also less likely to be employed than the average UK resident, with 60.4% of respondents currently in work — lower than the UK's latest employment rate (at the time of writing) of 75%.

Largely to motivate high retention and response quality, participants were compensated at far above Prolific's average hourly pay rate of £9: in total, average pay was £6.55 for approximately 9 minutes of time, implying an hourly pay rate of over £43. The data demonstrate that participants provided high-quality responses: only 5 participants of over 1000 failed either of the two attention checks featured in the experiment; and retention was excellent, with 95.7% of participants returning for session 2 five days after session 1. We omit the 5 participants who failed either attention check from all of the ensuing analysis.

Table 2.1 Summary statistics

	Value	Share	...of eligible observations
Mean age	41	—	1060
Median age	39	—	1060
Std. dev age	13	—	1060
Mean self-reported altruism (0-10)	7.0	—	1020
Mean self-reported risk (0-10)	4.8	—	1020
Mean self-reported patience (0-10)	6.7	—	1020
Mean reward	£6.55	—	890
Session 1 median duration (s)	4m	—	1066
Session 2 median duration (s)	4m 58s	—	1022
Female (N)	528	49.8%	1060
Ethnicity: white (N)	924	87.8%	1052
Employed (N)	644	60.4%	837
Assigned to treatment (N)	507	47.6%	1066
Assigned to disclosure condition (N)	246	49%	507
Completed optional task in treatment group (N)	409	80.7%	507
Attended session 2 (N)	1020	95.7%	1066
Failed either attention check (N)	5	0.0%	1066

2.3.2 Summary statistics: Recall

80.7% of participants in the treatment group elected to do the tasks and therefore received the full belief treatment. After making their choice, if applicable, x was randomly generated for each participant. As a reminder, x was drawn from the set $\{0, 1, 2, ..., 9, 10\}$, with each of $\{0, 1, 9, 10\}$ being generated with 0.5% probability and the remaining 98% being evenly divided over the range $\{2, 3, 4, ..., 8\}$. The expected value of x was therefore 5. Participants were made fully aware (and subsequently reminded) of this distribution immediately before we elicited beliefs on x in sessions 1 and 2.

The sample means for x differed slightly across treatment and control groups, at 5.0 and 5.1 respectively (Table 2.2). The difference was numerically small — only 0.1 more sliders were generated to take the value 150 for subjects in the control group, on average — and not statistically significant.

Overall, recall of x was highly accurate in both sessions, with little variation

Table 2.2 Signal and recall by treatment

Group	x	$\tilde{x}_1$	$\lvert\tilde{x}_1 - x\rvert$	$\tilde{x}_2$	$\lvert\tilde{x}_2 - x\rvert$	N
Control	5.09 (0.09)	4.49 (0.09)	0.83 (0.04)	4.82 (0.09)	1.04 (0.05)	556
Treatment	4.97 (0.09)	4.61 (0.09)	0.81 (0.04)	5.04 (0.1)	1.03 (0.05)	505

Note: x is the hidden signal, the number of times the subject had to drag sliders to 150 in session 1. $\tilde{x}_1$ is recall of that number in session 1, while $\tilde{x}_2$ is recall of that number in session 2, 5 days later.

between subjects. 42% of subjects correctly recalled the exact value of x in session 1 and 36% of participants did so in session 2, with mean absolute numerical errors of 0.8 and 1.0 sliders respectively in each session. Recall of x was slightly negatively biased on average in both the treatment and control groups, especially in session 1: in general, the tendency was to recall dragging fewer sliders to 150 in the session 1 task than in reality. This finding underscores the importance of including a motive-free control group: even in the absence of any motive to distort beliefs, they are biased on average in session 1 and change systematically from session 1 to session 2. This pattern holds relatively consistently across values of x (Figure B.1). However, the power of our design is that we can still retain our interpretation that those not receiving treatment serve as the unmotivated baseline for beliefs. As such, when we refer to a 'bias' in beliefs in the ensuing results, we mean a bias relative to this rational baseline rather than the true value of x.

2.3.3 Results: Short-term recall (session 1)

Our main interest is in whether beliefs on x differ systematically across the treatment and control groups. In the treatment group, those who chose to do the task have a possible anticipatory motive to recall a higher value of x: even though they have the same information set on x as participants in the control group, they know that a higher value of x means a higher chance that their choice to do the optional tasks in session 2 is not pointless. We first examine whether this was the case in session 1, immediately after subjects had completed the task in which x was hidden.

For a comprehensive test of this hypothesis, we run linear regressions where $\tilde{x}_1$, recall of x in session 1, is the dependent variable. To allow for the apparent non-linearity in the relationship between x and $\tilde{x}_1$ (Figure B.1), we add separate dummy variables for each possible integer value of x. These are omitted from

the table for brevity but, as expected, are highly significant: recall is generally very accurate. We also control for each possible value of the belief incentivisation payment.

Table 2.3 OLS regressions for treatment effect on recall in session 1

	$\tilde{x}_1$	
	ATE	ATT
Constant	2.587***	2.612***
	(0.636)	(0.635)
Treatment effect	0.188***	0.232***
	(0.067)	(0.083)
Payment for belief		
£0.50	−0.139	−0.135
	(0.108)	(0.108)
£0.75	−0.054	−0.051
	(0.109)	(0.108)
£1	−0.172	−0.170
	(0.110)	(0.110)
£1.25	−0.091	−0.079
	(0.111)	(0.111)
£5	−0.141	−0.147
	(0.135)	(0.135)
Observations	1,061	1,061
Controls for x	Yes	Yes

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. Standard errors in parentheses. x is the hidden signal, the number of times the subject had to drag sliders to 150 in session 1. $\tilde{x}_1$ is recall of that number in session 1. Control variables for each individual value of x are included in the regression but not reported here for brevity. The omitted dummy variable is $x = 1$. ATT estimated by using treatment assignment as an instrumental variable for actually receiving the treatment (choosing to complete the task in session 2).

While initial beliefs are generally precise across both groups, we obtain a small but highly significant treatment effect. The results in Table 2.3 demonstrate that those assigned to the treatment group remember dragging 0.19 more sliders (out of 10) to 150 than those in the control group. Given that the treatment was only administered to the 80.7% of subjects in the treatment group who chose to do the tasks in session 2, the treatment effect on those who were treated was slightly higher, at 0.23 sliders. This amounts to a small distortion of beliefs on $p(x)$: subjects believed there was around a 1.8 percentage point higher likelihood of the tasks generating a payment to their partner than would have been the case in the absence of a motive to distort beliefs.

In our experiment, incentives for correct beliefs have no influence on average beliefs. Given that this holds even for a payment of £5, which would almost double the total experimental payment for the average participant, it appears that most participants cannot influence the accuracy of recall with more effort. This stands in contrast to existing studies, especially where beliefs are elicited some time after signal transmission (Zimmermann, 2020), but may be less surprising in our context, where errors are already very small in general.[2]

While we have established that average beliefs are higher in the treatment group than the control group, we can enrich our conclusions by examining differences across the treatment and control groups in the variance of recall and recall errors. $\tilde{x}_1$ is no more variable in the treatment group than the control group, but errors $\tilde{x}_1 - x$ are more variable. Since x is generated from the same distribution (and has an identical sample variance) across treatment and control groups, the higher variance of recall errors in the treatment group indicates a weaker correlation between $\tilde{x}_1$ and x. One way of interpreting this result is that our treatment effect is not the result of lower recall effort in the treatment group: instead, there appears to be a shift in the distribution of recall that weakens its correlation with reality.

2.3.4 Results: Medium-term recall (session 2)

In prior studies of motivated beliefs and memory, beliefs have been shown to become more biased as time passes (Zimmermann, 2020; Gödker et al., 2021; Roy-Chowdhury, 2022), indicating that memory loss favours motives to distort beliefs. We examine

[2]This conclusion also holds within the treatment and control groups individually. While the largest incentive we offer is smaller in absolute terms than some existing studies, it would almost double the entire payment for the experiment.

Table 2.4 Standard deviation of x and $\tilde{x}_1$ by group

Group	x	$\tilde{x}_1$	$\tilde{x}_1 - x$	N
Control	2.08	2.07	1.06	556
Treatment	2.09	2.1	1.21	505
p-value	0.926	0.719	0.001	

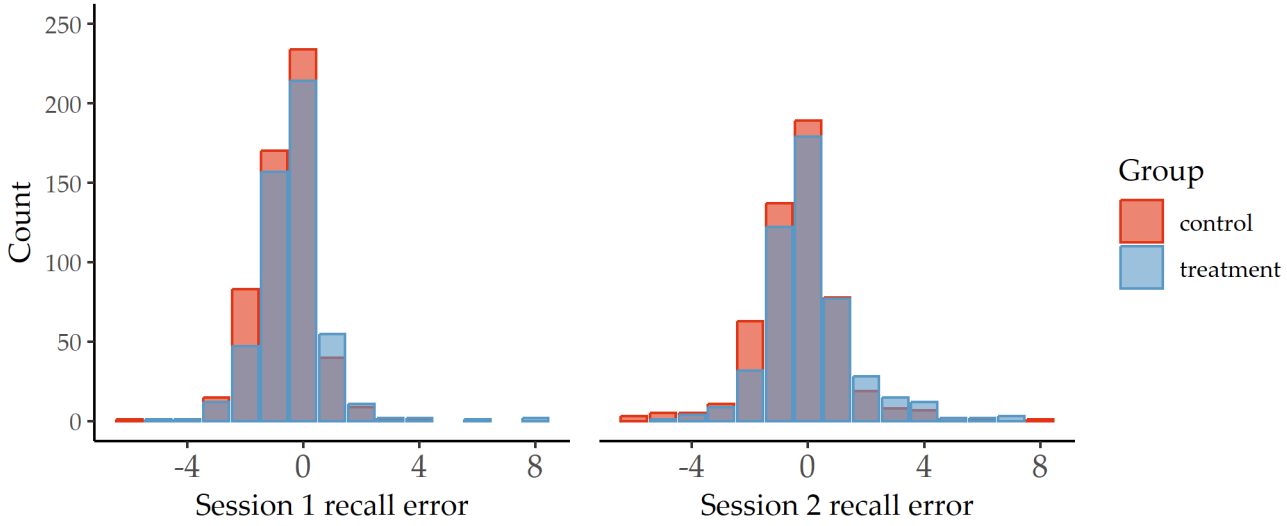
Note: x is the hidden signal, the number of times the subject had to drag sliders to 150 in session 1. $\tilde{x}_1$ is recall of that number in session 1. p-values are for F-tests of equal variances across groups for each variable.

whether the same phenomenon is observed here, albeit on a shorter timeline and with relatively precise initial beliefs. Session 2 took place 5 days after session 1 and beliefs were elicited using exactly the same mechanism as in session 1, before subjects completed the session 2 task (if applicable). As indicated by Table 2.2 and Figure 2.3, recall of x drifted upwards across both control and treatment groups. However, we can only comment on distortions in memory by comparing recall across the treatment and control groups.

We first examine differences across treatment and control groups in beliefs, in an analogous approach to that used for analysis of beliefs in session 1 (columns (1) and (2) of Table 2.5). The average difference in recall across treatment groups is almost twice as high in session 2 as in session 1, at 0.33 sliders. Our estimate of the ATT in column (2) indicates that the causal effect of treatment on beliefs in session 2 was to recall seeing 0.42 more sliders valued 150, amounting to a distortion of 3.3 percentage points in the perceived probability that the task generates a payment for the subject's partner. As in session 1, the magnitude of incentive payments for belief accuracy has no impact on average beliefs.

While the results in columns (1) and (2) suggest an increased distortion of beliefs relative to session 1, we conduct more rigorous statistical tests of this hypothesis in columns (3) to (6). In columns (3) and (4), we control for session 1 beliefs and examine whether treatment assignment and delivery have an independent on beliefs in session 2. To allow for possible differences in patterns of memory across values of $\tilde{x}_1$, we include individual control variables for each possible value of $\tilde{x}_1$. Using this specification, we obtain strong evidence of a treatment effect on session 2 beliefs that is orthogonal to the original treatment effect on beliefs in session 1. The

Fig. 2.3 Distributions of recall errors in sessions 1 and 2



Note: Recall error calculated as recall $\tilde{x}_t$ in session t minus the signal x.

corresponding ATE and ATT estimates are 0.22 and 0.27 sliders respectively.

The final two regressions in columns (5) and (6) estimate treatment effects on the change in recall from session 1 to session 2. In this case, both estimates of the treatment effect just miss the threshold for significance at the 10% level. The likely reason for this is an imperfect correlation between recall in session 1 and session 2: recall errors in session 1 are not entirely inherited in session 2. However, it does weaken support for the conclusion that beliefs systematically change over time, at least on average, due to a consequentialist motive to distort beliefs.

As we did for recall in session 1, we also test whether recall in session 2 is more or less variable in the treatment group than the control group. Similar to session 1, there is no evidence of a difference in variability in beliefs across groups. In fact, perhaps surprisingly, $\tilde{x}_2$ is no more variable than $\tilde{x}_1$ for either the treatment or the control group. However, errors in recall do become more variable in session 2 across both the treatment and control groups, implying that the correlation between recall and the signal is weaker in session 2 than in session 1.

Table 2.5 OLS regressions for treatment effect on recall in session 2

	$\tilde{x}_2$		$\tilde{x}_2$		$\tilde{x}_2 - \tilde{x}_1$	
	ATE	ATT	ATE	ATT	ATE	ATT
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	2.466***	2.509***	0.354***	0.351***	0.434	0.483
	(0.852)	(0.851)	(0.107)	(0.108)	(0.825)	(0.824)
Treatment effect	0.332***	0.410***	0.140	0.173	0.220***	0.272***
	(0.092)	(0.113)	(0.087)	(0.107)	(0.082)	(0.101)
Payment for belief						
£0.50	−0.199	−0.190	−0.044	−0.040	−0.128	−0.121
	(0.147)	(0.147)	(0.139)	(0.139)	(0.131)	(0.131)
£0.75	−0.218	−0.212	−0.170	−0.168	−0.184	−0.180
	(0.149)	(0.149)	(0.141)	(0.141)	(0.132)	(0.132)
£1	−0.220	−0.213	−0.058	−0.055	−0.135	−0.130
	(0.151)	(0.151)	(0.143)	(0.143)	(0.134)	(0.134)
£1.25	−0.087	−0.067	−0.011	−0.002	−0.041	−0.027
	(0.153)	(0.153)	(0.144)	(0.144)	(0.136)	(0.136)
£5	−0.248	−0.256	−0.060	−0.063	−0.174	−0.177
	(0.187)	(0.187)	(0.177)	(0.177)	(0.168)	(0.168)
Observations	1,015	1,015	1,015	1,015	1,015	1,015
Controls for $\tilde{x}_1$	No	No	Yes	Yes	No	No
Controls for x	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.535	0.535	0.053	0.053	0.640	0.640

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. Standard errors in parentheses. x is the hidden signal, the number of times the subject had to drag sliders to 150 in session 1. $\tilde{x}_1$ is recall of that number in session 1, while $\tilde{x}_2$ is recall of that number in session 2, 5 days later. Control variables for each individual value of x, and $\tilde{x}_1$ where applicable, are included in the regression but omitted here for brevity. The omitted dummy variables are for $x = 1$ and $\tilde{x} = 1$. ATT estimated by using treatment assignment as an instrumental variable for actually receiving the treatment (choosing to complete the task in session 2).

Table 2.6 Standard deviation of $\tilde{x}_2$ by group

Group	$\tilde{x}_2$	$\tilde{x}_2 - x$	$\tilde{x}_2 - \tilde{x}_1$	N
Control	2.07	1.53	1.33	556
Treatment	2.15	1.57	1.41	505
p-value	0.387	0.519	0.18	

Note: x is the hidden signal, the number of times the subject had to drag sliders to 150 in session 1. $\tilde{x}_1$ is recall of that number in session 1, while $\tilde{x}_2$ is recall of that number in session 2, 5 days later. p-values are for F-tests of equal variances across groups for each variable.

2.3.5 Results: Disclosure

In order to test the role of anticipatory motives, we included a further treatment condition within the treatment group. Before each belief elicitation, half of participants were informed that the true value of x would be disclosed to them at the end of session 2. The other half never find out the true value of x. If beliefs are persistent through time, a model of anticipatory utility would hold that individuals have a stronger incentive to forgo monetary incentives for accurate beliefs if they expect to be able to sustain belief distortions indefinitely.

Table 2.7 Signal and recall by disclosure treatment within treatment group

Disclosure	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_2 - \tilde{x}_1$	N
Yes	4.6 (0.1)	5 (0.1)	0.4 (0.1)	245
No	4.6 (0.1)	5.1 (0.1)	0.4 (0.1)	260
p-value	0.918	0.909	0.972	

Note: x is the hidden signal, the number of times the subject had to drag sliders to 150 in session 1. $\tilde{x}_1$ is recall of that number in session 1, while $\tilde{x}_2$ is recall of that number in session 2, 5 days later. p-values are for t-tests of differences in means across groups for each variable.

The results in Table 2.7 show definitively that being able to sustain belief distortion indefinitely has no influence on present belief distortion in our experiment.

Participants who anticipated finding out x at the end of session 2 had almost exactly the same beliefs as those who anticipated never finding out.

Contingent on the assumption that biases in recall could have been eliminated with sufficient cognitive effort, this result suggests that any hedonic incentive to distort beliefs must be heavily weighted towards the short term. However, this assumption is unlikely to be valid in our experiment: relatively large monetary incentives, which at the highest level would have almost doubled subjects' pay for the experiment, were not sufficient to offset biases in recall. This suggests that any biases in recall were stubbornly encoded soon after the transmission of information. It may be valuable to repeat this exercise in a context where long-term biases in recall are more strongly driven by the retrieval of information, since retrieval has been shown in prior studies to be relatively sensitive to incentives.

2.3.6 Results: Heterogeneous treatment effects

The results in the previous two subsections focus on average treatment effects on recall in sessions 1 and 2 of the experiment. However, we are also interested in how the treatment affected various segments of the population. We begin by running quantile regressions at the 10^{th}, 30^{th}, 50^{th}, 70^{th}, and 90^{th} conditional percentiles for $\tilde{x}_1$, $\tilde{x}_2$, and $\tilde{x}_2 - \tilde{x}_1$ (Figure 2.4 and Table B.1).[3]

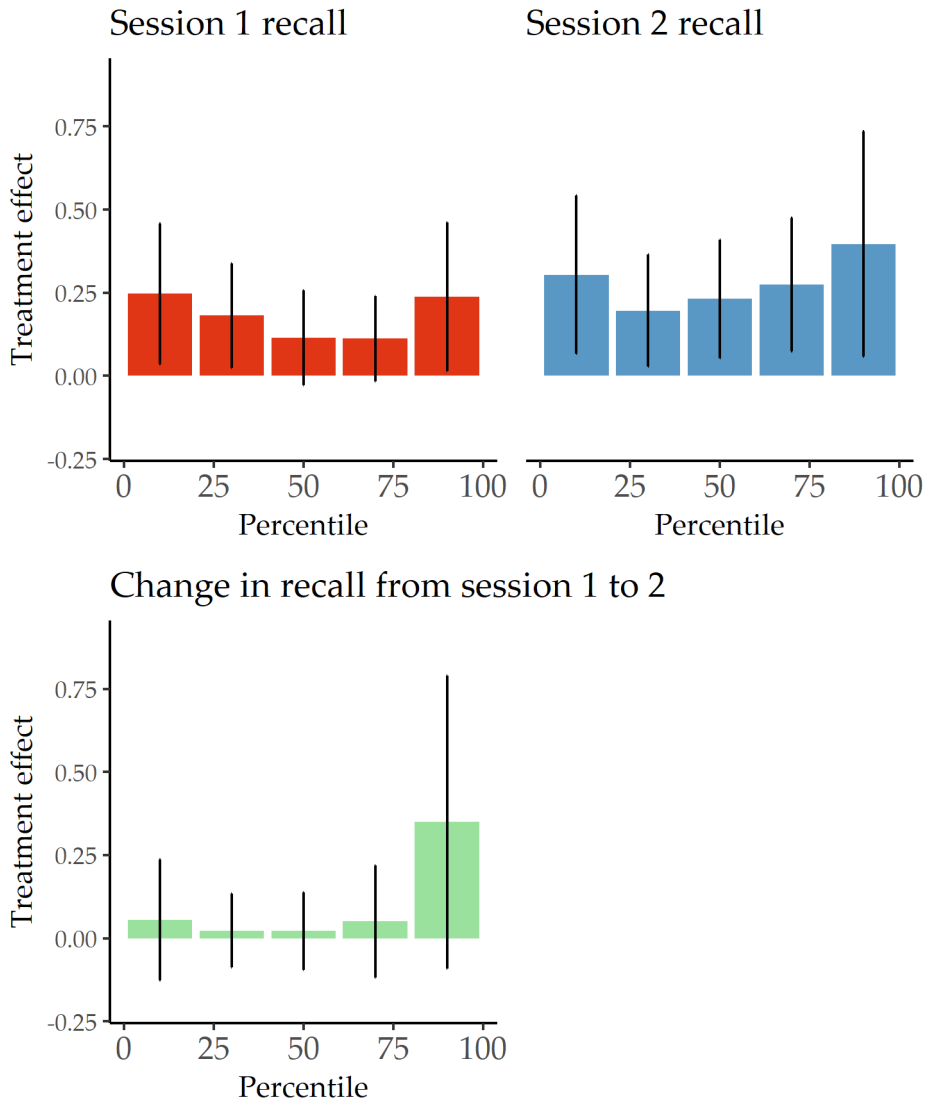
Since we control for the actual value of x in all cases, the estimated treatment effects are for quantiles of beliefs conditional on the signal, so, effectively, quantiles of errors in recall. In sessions 1 and 2, treatment assignment has a statistically significant effect on recall at all tested quantiles of the distribution Figure 2.4. However, there is heterogeneity across the distribution: the effects are numerically larger at both extremes of the distribution.

Thus, our results suggest two conclusions about the parts of the recall distribution in which treatment was relatively potent. First, it was especially likely to curtail the likelihood of relatively large negative errors in recall. Second, it was especially likely to increase the chance of sizeable positive errors in recall. One plausible interpretation is that $p(x)$, the probability of a payment being made to the subjects' partner as a result of doing the task, enters subjects' utility non-linearly. In particular, the marginal disutility of a particularly low belief on $p(x)$, and the marginal utility

[3]Since all recall variables are integers, we take the standard approach of adding random noise to each variable, generated by a uniform distribution with support $[0, 1)$ (Machado and Silva, 2005). We run 2000 such simulations and report mean parameter estimates.

of a particularly high belief on $p(x)$, may be relatively large. However, as is standard in the analysis of heterogeneous treatment effects, we cannot rule out that variation in unobserved characteristics is the driver of differences in treatment effects.

Fig. 2.4 Treatment effects across conditional quantiles for each variable



Note: Results are the average of 2000 simulations jittering recall values with uniformly distributed noise on $[0, 1)$. Error bars depict 95% confidence intervals. Control variables for the signal x and belief incentives are included in the regressions but omitted here for brevity. In the regression for recall in session 2, we do not control for recall in session 1.

Building on other work examining gender differences in altruism and related beliefs (Eckel and Grossman, 1998; Andreoni and Vesterlund, 2001; Fišar et al., 2016;

Brañas-Garza et al., 2018), it is also of interest to estimate heterogeneous treatment effects by sex (Tables B.2 and B.3). In session 1, there is suggestive evidence that being assigned to the treatment group more strongly affected women than men: the ATE for women is positive and statistically significant, while the ATE for men is not significant. The difference in coefficients is not statistically significant, though this test is, of course, more poorly powered. The gap in the size of treatment effects closes slightly by session 2, where statistically significant differences across treatment groups are detected in both subsamples. Finally, matching our finding for the whole sample, the ATE for the change in recall from session 1 to session 2 is not significant for either men or women.

Our suggestive finding that the treatment had a stronger effect on women in session 1 may partly explained by higher altruism among women. We collected a self-reported measure of other-regarding preferences via the question 'How do you assess your willingness to share with others without expecting anything in return?', assessed on a scale from 0-10. As a note of caution, self-reported altruism scores were collected at the end of the experiment so as not to prime subjects, so they may be endogenous to the treatment. However, they do not differ significantly across treatment and control groups. With this caveat in mind, we estimate how treatment effects differ across this measure (Tables B.4 and 2.8).

ATEs are systematically higher for those with above-average self-reported altruism scores (Table B.4). Beliefs are positively biased by 0.3 sliders in session 1 and 0.5 sliders in session 2 for those with above average self-reported altruism. For those with average self-reported altruism or below, assignment to the treatment group did not have a significant effect. Notably, unlike in our other analyses, we also obtain statistically significant evidence of positively biased memory loss within this subsample of participants: recall of x increases by about 0.2 sliders.

Table 2.8 ATTs by self-reported altruism score

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_2 - \tilde{x}_1$
Avg or below altruism	0.165 (0.12)	0.25 (0.172)	0.085 (0.163)
Above avg altruism	0.334*** (0.122)	0.568*** (0.152)	0.234** (0.139)
Difference	-0.169 (0.172)	-0.318 (0.229)	-0.149 (0.214)
N	1015	1015	1015

Since ATEs reflect both choices to do the task and distortions of beliefs resulting

from receiving the treatment, we can focus on the latter factor by examining differences in ATTs across both self-reported altruism and sex (Tables B.3 and 2.8). In both cases, our conclusions are largely sustained. Women and those with higher self-reported altruism distorted their beliefs more strongly than others when receiving the treatment. So, even among those who chose to do the task for their partner, the motive to distort beliefs is greater among those with stronger other-regarding preferences, consistent with the interpretation that belief distortion is driven by belief-based utility.

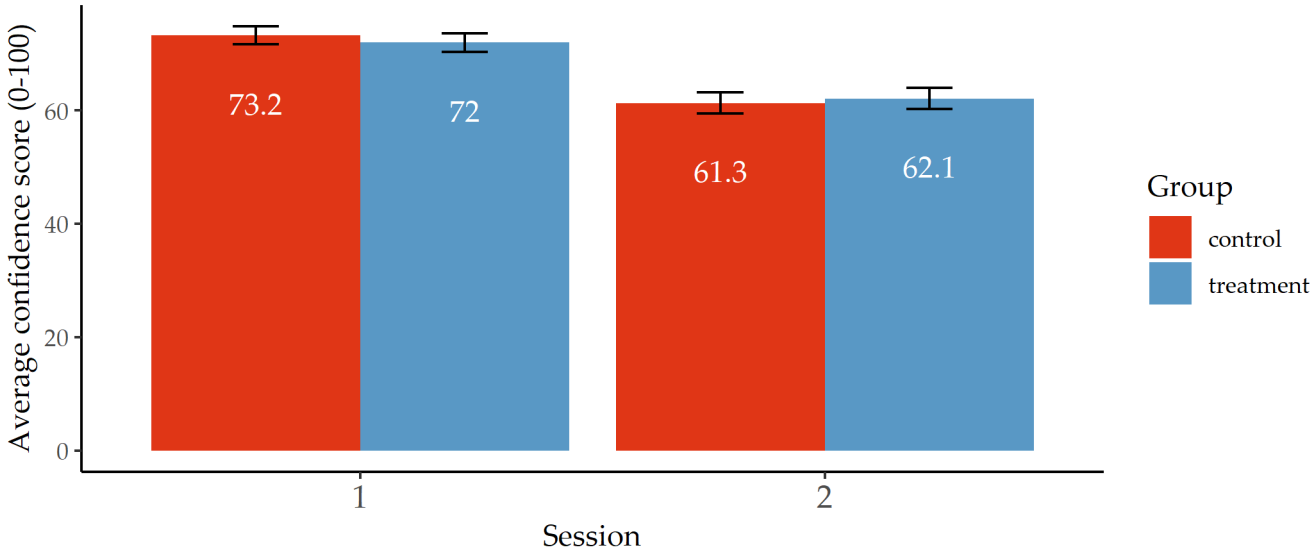
2.3.7 Results: Confidence

We also included unincentivised questions on confidence. Our measurement is somewhat similar to that in recent work on cognitive uncertainty (Enke and Graeber, 2023): after each elicitation of recall, we asked subjects to rate how confident they were in their answer, on a scale of 0-100. Our initial interest is in whether confidence in recall is associated with a lower error, implying subjects are sophisticated about the imprecision of their own belief. Then, we would like to test whether the relationship between confidence scores and recall errors is different across treatment and control groups, as a means of testing whether subjects are more sophisticated about motivated *distortions* of recall than unmotivated errors. If subjects are sophisticated about distortions, we would expect recall errors to have a weaker relationship with confidence scores in the treatment group.

Average confidence scores by session and treatment are plotted in Figure 2.5. Confidence scores are just over 70, on average, in session 1. They decline on average by around 10 points to session 2, consistent with the general decline in recall accuracy across the two sessions. Also visible is the fact that confidence scores are fairly similar across the treatment and control groups. Given that we interpret treatment effects as biases relative to the rational baseline provided by the control group, this may reflect a lack of sophistication about belief distortion in the treatment group. However, it could also reflect that any effect of the treatment on confidence is second-order, since they must run through beliefs in the first instance, and treatment effects on beliefs are numerically small.

For a more precise insight, we run regressions on the confidence scores corresponding to the recall measures $\tilde{x}_1$ and $\tilde{x}_2$, in sessions 1 and 2 respectively. In both cases, confidence has the expected direction of relationship with absolute recall

Fig. 2.5 Confidence scores by session and treatment



Note: Error bars are 95% confidence intervals.

errors $|\tilde{x}_1 - x|$ and $|\tilde{x}_2 - x|$. In session 1, a recall error of 1 slider (out of 10) is associated with a 3.8 decrease in the confidence score (out of 100). The relationship is substantially weaker in session 2: the coefficient on recall errors, -2.3, is 40% smaller. This suggests that the process of memory loss also involves a loss of sophistication on recall accuracy.

In neither session does the relationship between recall errors and confidence scores vary across the treatment and control groups. This suggests subjects are not sophisticated about the distortion of recall causally induced by our treatment. This result is consistent with our prior findings that distortions are not influenced by high incentives for belief accuracy or expected disclosure: in our experiment, distortions of recall appear to be quite stubbornly encoded. In this sense, motivated belief distortion in our experiment seems to be buried further in the subconscious than in other studies (Zimmermann, 2020; Gödker et al., 2021), wherein motivated biases are overturned by sufficient incentives and various treatment conditions.

Table 2.9 Regressions for confidence in recall

	Confidence in $\tilde{x}_1$	Confidence in $\tilde{x}_2$
$\lvert\tilde{x}_1 - x\rvert$	-3.768^{***}	
	(0.886)	
$\lvert\tilde{x}_2 - x\rvert$		-2.262^{***}
		(0.817)
Treatment group	-2.249	-0.050
	(1.529)	(1.839)
Male	2.446^{**}	2.951^{**}
	(1.146)	(1.379)
$\lvert\tilde{x}_1 - x\rvert$*Treatment group	0.921	
	(1.234)	
$\lvert\tilde{x}_2 - x\rvert$*Treatment group		0.627
		(1.182)
Constant	75.206^{***}	62.237^{***}
	(1.227)	(1.445)
Observations	1,048	1,007
R^2	0.033	0.016

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. Standard errors in parentheses. Confidence scores are measured on a scale of 0-100. x is the hidden signal, the number of times the subject had to drag sliders to 150 in session 1. $\tilde{x}_1$ is recall of that number in session 1, while $\tilde{x}_2$ is recall of that number in session 2, 5 days later.

2.4 Conclusion

Our experiment tests whether the recall of information is affected by a consequentialist motive to distort beliefs. We find that this is the case: in our treatment group, where some participants are told that a variable hidden in a task determines whether a tedious task will yield a favourable unobserved outcome, beliefs are systematically distorted in favour of the task being productive. Our experiment is specifically designed to remove self-image concerns: subjects in the treatment group cannot meaningfully distort their beliefs about themselves because the signal had no bearing on the *rationale* for the decision to do the tasks, only its consequences. We also demonstrate that some participants, particularly women and those who rate themselves as more altruistic, hold greater motivated biases in beliefs and, in the latter case, also experience biased memory loss supporting this consequentialist distortion of their beliefs. Overall, this supports the interpretation that subjects distort beliefs about the consequences of their choice because of belief-based utility, which in our context must stem from other-regarding preferences.

Our results therefore reveal that individuals distort recall of objective information simply for the purpose of holding falsely optimistic beliefs about utility-relevant outcomes and, in some cases, that this process is supported by biased memory loss. This finding is consistent with the anticipatory utility model (Caplin and Leahy, 2001) and, especially, its application to belief distortion in a model of optimal expectations (Brunnermeier and Parker, 2005). However, somewhat less consistent with these models is that belief distortions in our study are very resistant to incentives. They cannot be overturned by relatively large financial incentives for belief accuracy, nor are they eliminated when subjects are told to anticipate imminent disclosure of the truth.

One plausible explanation for this finding is that belief distortions in our study are more strongly driven by initially biased encoding of information rather than biased retrieval of information, as is the case in other studies of biased memory (Zimmermann, 2020; Gödker et al., 2021). This narrative is supported by our finding that biases in beliefs occur almost immediately after the transmission of information in our experiment. In our context, this means the finding that anticipated disclosure has no impact on belief distortion is somewhat ambiguous: it could be that subjects still distort beliefs to motivate effort during the tedious task in session 2, or that distortions are simply too hard-wired to be unwound by the treatment. Future

research may be able to clarify whether anticipated disclosure affects biased beliefs in a context where distortions are shallower, in that they can be eliminated by incentives for belief accuracy.

Chapter 3

Temptation to Consume Information

3.1 Introduction

We are faced with an unprecedented supply of information (Hilbert and López, 2011). In many cases, consuming it involves psychological, strategic, financial, or time costs, meaning information avoidance may be rational in a plethora of important settings (Golman et al., 2017). However, a long-standing intuition suggests that curiosity often leads individuals to consume information that does not benefit them.[1] This assertion appears to be consistent with emerging evidence on the consumption of social media (Allcott et al., 2022) and news (Robertson et al., 2023). Do certain kinds of information therefore possess the typical markers of a *temptation* (Gul and Pesendorfer, 2001) — an undesirable good whose presence in the choice set calls for the exertion of costly self-control?

Temptation has potentially material consequences for theoretical and empirical work on preferences for information, both where information does not have decision-making value and where it does. In particular, even if information avoidance is desirable, it may be rendered unfeasible or more costly by temptation, especially when informational commitment devices are unavailable. In this sense, the potential inefficiencies resulting from temptation for information extend beyond just the psychological domain. Additionally, since the availability of a tempting good can be harmful to welfare, temptation for information would have welfare implications across a multitude of contemporary markets in which information is offered at a cost,

[1] An early example of this idea can be seen in Cicero's *De Finibus*: "So great is our innate love of learning and of knowledge, that no one can doubt that man's nature is strongly attracted to these things even without the lure of any profit".

either extrinsic (e.g. financial) or intrinsic (e.g. psychological), to consumers.

In this chapter, we take a first step towards testing formally whether undesired information can be classified as a temptation. First, we reveal the prevalence of temptation for information by collecting individuals' preferences to restrict their own access to potentially discomforting information in the future. Such preferences imply that the mere availability of undesired information is costly and raise the possibility of dynamically inconsistent preferences for information. Our second objective is then to test for dynamic inconsistency in preferences for information, making this study the first to do so in this context.

Our vehicle for these empirical tests is an experiment, guided by canonical theories of self-control. Participants first perform an irrelevant task, and are then told that (i) they can choose to accept or reject a lump-sum payment; (ii) a charitable donation will be made as part of the experiment; and (iii) if they accept the lump-sum payment, the charitable donation will be reduced with a 'small' probability. Those who took the payment then learn they will be offered information on whether the donation to the charity was indeed reduced. This information is the focus of our experiment. We elicit *menu preferences* (Gul and Pesendorfer, 2001; Toussaert, 2018) relating to the information to reveal not only preferences to avoid information but also temptation: individuals are classified as tempted by information if they are willing to pay to restrict their own ability to access it a day later. Then, in a second session, we randomly implement menu preferences. 50% of individuals are offered the information, regardless of whether they had a strict preference for it not to be made available. By examining the behaviour of those tempted by information when it is made available against their wishes, we test whether strict preferences for commitment reflect sophistication about the strength of temptation.

We obtain three main findings. First, almost 75% of individuals who would like to avoid information are also willing to pay not to have the *option* to consume it in a day's time. These individuals, comprising just under half of the whole sample, are therefore classified as being tempted by information. Second, when some of those who wanted to eliminate information from their choice set are offered it, a small share (3%) succumb to temptation and obviously have dynamically inconsistent preferences for information. However, for the remainder, we use a novel empirical strategy to show that those with a higher willingness to pay for commitment actually find it harder to resist temptation when it is offered. This suggests that preferences for information are dynamically inconsistent for those tempted by it, even when consumption choices

are dynamically consistent. Third, using a treatment condition in which partial information is automatically provided in the second session, we obtain tentative evidence that unavoidable partial information erodes ex-ante preferences to avoid full information.

Our main result reveals temptation through ex-ante preferences for commitment, harnessing the experimental paradigm introduced in Toussaert (2018). If choices are not dynamically inconsistent, models of temptation suggest that self-control costs — which can be either avoided or reduced by taking temptations off the table before the period of choice — must be the motive for commitment. However, this interpretation has been difficult to verify in prior empirical work because self-control costs are typically unobservable and alternative rationalisations of commitment are possible. We demonstrate that it is possible to fill this gap by using data on response times. Our argument makes use of the well-established observation that response times are positively correlated with closeness to indifference in binary choices (see Luce, 1986; Krajbich et al., 2014; Alós-Ferrer et al., 2021 for reviews of the literature). In the context of temptation, we use the well-known dual-self model (Fudenberg and Levine, 2006, 2012) to show that response time therefore increases in self-control costs when deciding whether to turn down an extemporaneous, resistible temptation.

Our data are basically unequivocal in their support for models in which agents are sophisticated about their future self-control costs. By analysing response times in session 2 for those who are offered information against their wishes, we find evidence that menu preferences are strongly predictive of experienced self-control costs. For example, those who are willing to pay more for their commitment to avoiding information in session 1 are largely those who spend longer deciding whether or not to consume information in session 2. This finding provides the basis for our conclusion that preferences for information can be dynamically inconsistent even if choices are not: those who have stronger preferences to commit to avoiding information find it *harder* to resist information when offered it, exactly as predicted by a model of temptation under sophistication about future self-control costs. Qualitative measures of preferences for information across the two sessions further corroborate this interpretation. Through this analysis, to the best of our knowledge, this chapter entails the first attempt to find empirical evidence in any context that commitment is motivated by dynamically inconsistent preferences when choices are dynamically consistent.

Partial information is seemingly deployed to stimulate further demand for

information in many contexts — for instance, notifications on mobile devices and social media. Using a treatment condition in which an initial signal is automatically shown in the second session, we find evidence that partial information weakens *ex-ante* preferences to avoid discovering something undesirable. Our interpretation is that partial information leaves less room for wishful thinking than an ambiguous prior.

More broadly, our findings introduce an argument for policies treating ex-ante undesirable information like other well-known temptation goods: sometimes, it may be better for such information not to be manufactured at all. We postulate that this could be an important dimension of policy assessments of some recent developments in social media and consumer access to artificial intelligence. Although many theoretical mechanisms could generate temptation for information, we use a general model to argue that temptation could be natural if individuals are curious about undesired information but recall it less accurately over time. Under intuitive conditions on preferences, this means information intrinsically delivers short-term gratification, as is typical of a temptation.

Related literature. This chapter makes several contributions to the literature on preferences for information. Empirical work on intrinsic preferences for information — preferences for information independent of its decision-making value — has overwhelmingly focused on preferences for the temporal resolution of uncertainty (Eliaz and Schotter, 2007; Ganguly and Tasoff, 2017; Nielsen, 2020; Falk and Zimmermann, 2023). Recent work also demonstrates that individuals are often impatient to acquire even potentially discomforting information (Masatlioglu et al., 2023). However, existing work does not consider whether preferences for information become more intense as the time of its availability approaches. This chapter fills that gap: while relatively few individuals make dynamically inconsistent choices, we emphasise that costly self-control makes preferences for information dynamically inconsistent at the *intensive* margin. Conservatively, this means willingness to pay for information at a given point in time may differ depending on how far in advance the individual is asked. Separately, our results have similar implications for the growing body of empirical work on preferences for information with possible instrumental value.[2]

[2]See Eil and Rao (2011), Oster et al. (2013), Zimmermann (2020), Huffman et al. (2022), and Roy-Chowdhury (2022) for examples of ego-relevant information with possible instrumental value. Saccardo and Serra-Garcia (2023) explore a case in which information has clear instrumental value, finding that individuals often choose an information structure that limits their own ability to be

Here, even if there is a strategic motive to avoid information, costly self-control may make it difficult to do so. Thus, temptation over information could cause inefficiencies in important choice contexts.

In early theoretical work by Kreps and Porteus (1978), intrinsic preferences for information are the mirror image of preferences over the temporal resolution of uncertainty. In their model and subsequent extensions (Epstein and Zin, 1989; Grant et al., 1998), individuals can prefer to expedite or delay the receipt of information on the outcome of a lottery depending on properties of their risk and time preferences. Marking a more significant departure from the canonical expected utility model, a substantial body of recent work suggests individuals avoid information because of its role in anticipatory feelings. This model was pioneered in Caplin and Leahy (2001) and has been extended in several directions (Caplin and Leahy, 2004; Brunnermeier and Parker, 2005; Epstein, 2008; Koszegi and Rabin, 2009; Ely et al., 2015). While there are many ways of expanding these theories to account for our results, we propose one involving imperfect recall, meaning the uncertainty-reducing value of information decays over time. Additionally, our results suggest that in settings where information theoretically has instrumental value, it may be valuable to model both intrinsic and instrumental costs of avoiding information rather than just the latter.

We also contribute to the broader empirical literature on temptation and self-control. The main innovation of the Gul and Pesendorfer (2001) model is that the presence of tempting options in the choice set is costly even if they are not consumed. This hypothesis is supported by many empirical studies of procrastination and effort (Alan and Ertac, 2015; Royer et al., 2015; Toussaert, 2018; Sadoff et al., 2020), which find evidence of demand for commitment does not necessarily predict dynamically consistent choice. Relative to this literature, our use of response time to infer self-control costs is a methodological innovation; our paper is the first to test explicitly whether realised self-control costs are related to preferences for commitment.

Outline. This chapter proceeds as follows. In Section 3.2, we provide details of the experiment and sample. In Section 3.3, we outline a model of temptation, based closely on Fudenberg and Levine (2012), which generates empirical tests connecting menu preferences to dynamic inconsistency and self-control costs. Section 3.4 contains the results of the experiment; Section 3.4.1 considers menu preferences,

selfish.

Sections 3.4.2 and 3.4.3 examine information choices and self-control costs in session 2, Section 3.4.4 compares preferences across our treatment conditions, and Section 3.4.5 explores qualitative measures of preferences for information. In Section 3.5, we return to a theoretical framework to argue why information could be intrinsically tempting. Finally, in Section 3.6, we discuss the implications of our results for theoretical and empirical research on information preferences, as well as policy.

3.2 Experiment design

Participants submit their preferences in session 1 over the feasible set of information menus they could face in session 2. The main objective of the experiment is to measure the frequency of the two menu preferences implying information is tempting. However, by randomly implementing menu preferences in a second session, we also analyse data on choices and behaviour to test whether preferences for choice set restrictions reflect sophistication about experienced temptation.

Information object. Several papers demonstrate that individuals avoid information or distort beliefs to facilitate selfish choices or avoid feeling guilty about them (Dana et al., 2006, 2007; Lazear et al., 2012; Di Tella et al., 2015). We exploit this phenomenon to construct an object of information with no decision-making value but possible *psychological* (affective) value.

In session 1, subjects complete a short task involving moving sliders to specified values. They are then informed they can claim or forgo a lump-sum payment of \$4. The base payment for session 1 is \$1.40, meaning accepting the \$4 substantially increases participants' pay. The only benefit of forgoing the bonus is that a 'small' share of participants (in reality, 15%) have been allocated to be *Donors*, such that taking the bonus results in a charity donation being reduced by \$15. The remainder of participants are not *Donors*, meaning there is no consequence of taking the \$4.

After choosing whether to claim the bonus, participants are informed that they will be able to find out if they were a *Donor* in session 2 (by setting $a = 1$). The information is contained in a virtual envelope, which is generated in session 2 unless the choice menu $\{0\}$ is implemented. To ensure strict preferences for information in session 2, opening the envelope costs money. The price is randomly selected from $\{\$0.25, \$0.50, \$0.75, \$1\}$, and participants are informed of the price they will eventually face when submitting menu preferences in session 1.

Temptation and menu preferences. Gul and Pesendorfer (2001) provide a menu-preference-based definition of temptation which forms the basis of our experimental procedure (Toussaert, 2018). They consider a two-period setting in which a decision maker (DM) makes a choice from a menu in period 2, using a preference relation $\succeq_2$. In period 1, the DM may have preferences $\succeq_1$ over menus to be presented in period 2. The axiom capturing temptation in their model is *Set-Betweenness*, which imposes that for two menus A and B, if $A \succeq_1 B$ then $A \succeq_1 A \cup B \succeq_1 B$. This axiom permits a preference for commitment: a decision maker may strictly prefer to eliminate B from her choice menu if A is strictly preferred to B. In particular, a menu preference with $A \succ_1 A \cup B \succeq B$ indicates B is a temptation.

Our experiment involves a binary choice set $\{0, 1\}$. Consuming information is given by 1 and avoiding it is given by 0. Thus, the relevant menus are $\{\{0\}, \{0, 1\}, \{1\}\}$. Following the definition provided above, there are two menu preference rankings indicating that information is tempting: $\{0\} \succ_1 \{0, 1\} \succ_1 \{1\}$ and $\{0\} \succ_1 \{0, 1\} \sim_1 \{1\}$. In our experiment, participants therefore have menu preferences $\{0\} \succ_1 \{0, 1\} \succeq_1 \{1\}$ if they are tempted to find out whether taking a \$4 bonus resulted in a much larger \$15 being deducted from a charity donation.

The main objective of our experiment is to elicit menu preferences for information and investigate the prevalence of these two preference rankings. However, in Section 3.3, we provide a model that links menu preferences to self-control costs and choices under sophistication, motivating empirical tests of the motives for commitment.

Menu preference elicitation. Toward the end of session 1, participants submit preference rankings over the *menus* of information that will eventually be offered to them in session 2. To aid comprehension, the session 2 information choice is framed as opening a virtual envelope, and commitment is framed as the envelope not being generated. The set of menus in session 1 then has the following presentation:

- Automatically open the envelope. ($\{1\}$)
- Do not generate the envelope at all. ($\{0\}$)
- Ask me if I want to open the envelope next time. ($\{0, 1\}$)

Importantly, participants are informed that session 2 will take 'about the same amount of time' regardless of which menu is implemented.[3] To ensure full comprehension and incentive compatibility, we follow a multi-step elicitation procedure in which participants must repeatedly confirm any strict preferences they submit. The corresponding experiment pages can be viewed in the Appendix, starting with Figure C.5. Initially, participants give each menu a numerical rank between 1 and 3; multiple menus can be given the same rank. They then navigate to a second page where the menus are displayed in a re-randomised order and they cannot proceed until their preference rankings on the two pages match. Then, where any strict preferences have been submitted, participants proceed to the second stage of the procedure.

This second stage provides our preferred measure of menu preferences, used throughout the study. Participants are prompted to confirm that they 'definitely have a clear preference' between each adjacent pair of menus in their ranking and are advised that they will have to complete a short task to confirm their preference. This means our primary outcome measure is somewhat conservative, as participants who are not willing to pay the effort cost to confirm each strict preference are defined as being indifferent. The task is an exact repetition of the one they completed at the start of the experiment, involving dragging sliders to numerical values, and participants are able to amend their menu preferences upon seeing the task.

After the effort task, we elicit participants' willingness to pay to maintain the first adjacent strict preference in their final ranking using the Becker-DeGroot-Marschak (BDM) mechanism. Participants are told that a random bid between \$0 and \$0.50 will be chosen. If the bid is higher than their stated willingness to pay, which is also constrained to be between \$0 and \$0.50, they receive the bid and their first and second choices are swapped. Response values are restricted to multiples of \$0.05, starting from \$0.

Implementation of menu preferences. Session 2 takes place one day after session 1. Participants choose whether or not to open the envelope, depending on the choice menu implemented. The menu they face is influenced by their preference ranking from session 1, although as we mention below, half of participants face the menu $\{0, 1\}$ regardless of their preferences.

After the result of the BDM auction, menu preferences are implemented in session

[3] In practice, this is implemented using buffer tasks when participants receive $\{0, 1\}$ in session 2.

2 according to the following regime, which ensures incentive compatibility: with 50% probability, the menu $\{0, 1\}$ is exogenously implemented so participants must decide whether to access information regardless of their preferences. Otherwise, menu preferences determine implementation: with 40% probability, first choices are implemented; with 10% probability, second choices are implemented. If participants are indifferent between any two options, one of them is randomly selected to be strictly preferred and the scheme is implemented as usual.

Treatment conditions. As previously mentioned, we included two treatment conditions to investigate how anticipating unavoidable signals affects the attractiveness of commitment devices to avoid information. Across conditions, participants taking the bonus face one of two subtly different information offers in session 2.[4] In *FullAvoid*, the envelope contains an initial signal s of whether or not the individual was a *Donor*. In particular, s can take values s_0 or s_{20}, implying either a 0% or a 20% chance respectively that taking the \$4 bonus resulted in the charity donation being reduced by \$15. Conditional on $s = s_{20}$, participants can then choose to costlessly confirm whether the charity donation was reduced by \$15.

In *PartAvoid*, participants know they will automatically learn the value of s at the start of session 2. Instead, the envelope in session 2 is generated conditional on $s = s_{20}$. As illustrated by Figure C.1, the key point is that 'not generating the envelope' in *PartAvoid* involves a participant committing not to learn whether they are a *Donor* when they already know there is a 20% chance they are one. In *FullAvoid*, the same action means learning nothing about the chance of being a *Donor*.

Sample characteristics. The sample recruited from Prolific initially included 673 adults residing in the US and fluent in English. Of those, 669 were invited to session 2 after passing one of two attention checks in session 1 (Figures C.15 and C.16), and 634 completed both sessions. Table C.2 provides a summary of key statistics. We captured a rich range of ages within the sample, with a standard deviation of 12.3 years. Nonetheless, the median age was a little lower than the average in the US population. White respondents were overrepresented, and the employment rate (for the subset with recent data) was lower than in the general population. The sample was balanced on sex.

[4]The 9% of participants who turned down the bonus were all allocated to a condition in which opening the envelope immediately confirms their *Donor* status.

Numerous measures suggest the experimental data were of high quality. Only 11 of 673 participants failed either of the two attention checks inserted in session 1; 4 of those were immediately disqualified because they failed both attention checks. We exclude all 11 from the ensuing analysis. Moreover, 69% of participants chose to answer optional free-text questions in session 1 rationalising their submitted menu preferences. Finally, as already noted, 95% of participants recruited for session 1 returned for session 2 a day later. Compensation rates were relatively generous: the mean reward for the experiment was $5.80, implying an average hourly pay rate of around $34.80.

91% of participants accepted the bonus. Of the 9% who turned it down, 6 participants were allocated to be *Donors*. Accordingly, $90 was donated to the charity after the experiment concluded, and all participants were contacted as promised after the experiment to inform them of the total donation amount (but nothing else).

3.3 Model

We now present an application of the well-known dual-self model of self-control that produces testable predictions connecting menu preferences to self-control costs. To focus attention on these key hypotheses, we initially present the model without any structure on information preferences, except to suppose that one self has a preference to consume information and the other has a preference to avoid it. Later in the chapter, we embed a model of information preferences in the dual-self framework to provide intuition for our results (Section 3.5).

3.3.1 Dual selves and temptation

Fudenberg and Levine (2006, 2012) propose a model of self-control involving an intra-individual conflict between a long-run self and a short-run self. The only disparity across the two selves is their time preference: the long-run self discounts the future by a factor δ, whereas the short-run self uses a smaller discount factor $\mu\delta$, where $0 < \mu, \delta < 1$.

Suppose that consuming the information in period t results in a sequence of payoffs $(g_t, g_{t+1}, g_{t+2}, ...)$ where $g_t \in \mathbb{R}$ for all z. We write the normalised lifetime

utility of consuming information in the current period as

$$\overline{U}_l = (1-\delta)\sum_{z=0}^{\infty}\delta^z g_t, \tag{3.1}$$

for the long-run self, and

$$\overline{U}_s = (1-\delta\mu)\sum_{z=0}^{\infty}(\delta\mu)^z g_t, \tag{3.2}$$

for the short-run self. We normalise the payoff to avoiding information to 0 in perpetuity. Thus, the model has traction if $\overline{U}_l < 0$ and $\overline{U}_s > 0$: the long-run self wants to avoid information but the short-run self wants to consume it.

In Fudenberg and Levine (2006, 2012), the short-run self selfishly optimises her welfare. The long-run self may have the option of restricting the short-run self's choice set or influencing the state of the world. In empirical applications, it is impossible to determine which choices are made by the short-run and long-run selves. To simplify matters, we assume that the long-run self is the one to provide menu preferences in $t = 1$ and make choices from menus in $t = 2$ — the short-run self does not have any role in decision making. However, the long-run self suffers a self-control cost from any actions it takes, defined by the utility forgone by the short-run self. Thus, the only role of the short-run self is to introduce self-control costs via forgone utility.[5]

3.3.2 Self-control costs and commitment

For our purposes, since we consider a two-period problem with a binary choice, Fudenberg and Levine (2012)'s linear self-control cost specification is sufficient. In particular, let Δ_t be the discounted utility forgone by the short-run self from any action taken in period t by the long-run self. Then, $\lambda\Delta_t$ is the self-control cost experienced in period t. The objective function for the long-run self consists of her normalised lifetime utility net of self-control costs experienced in each period.

A crucial feature of the Fudenberg and Levine (2006, 2012) framework is that self-control costs are experienced in the period when the self-control action is taken. In our application, if the long-run self makes a commitment in $t = 1$ to avoid

[5]This implementation is redolent of Gul and Pesendorfer (2001).

information in $t = 2$, the self-control cost experienced in $t = 1$ is $\lambda\delta\mu\overline{U}_s$. Conversely, if information is avoided in $t = 2$ itself, the self-control cost experienced in $t = 2$ is $\lambda\overline{U}_s$. In this case, the discounted self-control cost from the $t = 1$ perspective is $\delta\lambda\overline{U}_s$. The short-run self cares less about future flows of utility than the long-run self, meaning it is easier to avoid temptations the further in the future they are.

This specification thus gives rise to a motive for commitment, with $\{0\} \succ_1 \{0, 1\}$. When facing $\{0, 1\}$ in $t = 2$, the condition for avoiding information is

$$\lambda\overline{U}_s < -\overline{U}_l. \tag{3.3}$$

Consider a case in which $\lambda\overline{U}_s > |\overline{U}_l|$, so the individual consumes information in $t = 2$ when $\{0, 1\}$ is implemented. From the $t = 1$ perspective, the condition for strictly preferring the menu $\{0\}$ relative to $\{0, 1\}$ is

$$\mu\lambda\overline{U}_s < -\overline{U}_l. \tag{3.4}$$

Thus, if μ is sufficiently small, dynamically inconsistent choice implies the possibility of a strict preference for commitment.

If self-control costs are small enough that $\lambda\overline{U}_s \leq |\overline{U}_l|$, so information is avoided in under $\{0, 1\}$ period 2, a strict preference $\{0\} \succ_1 \{0, 1\}$ is automatic. That is because self-control costs are experienced at the time of taking the self-control action; if that action is taken in advance of the temptation, self-control costs become cheap relative to the long-run benefit of avoiding the temptation. To see why, note that the condition for $\{0\} \succ_1 \{0, 1\}$ is

$$(1 - \mu)\overline{U}_s > 0, \tag{3.5}$$

which is certainly satisfied. Thus, if self-control costs are sufficiently low that the temptation would be avoided when $\{0, 1\}$ is implemented, implementing $\{0\}$ in $t = 1$ is strictly preferred to taking the same action in $t = 2$. It is easier for the long-run self to 'persuade' the short-run self to forgo a future temptation than a current one because the latter discounts future utility flows at a faster rate.

3.3.3 Types and self-control costs

There are two menu preferences implying temptation for the information we offer: $\{0\} \succ_1 \{0, 1\} \succ_1 \{1\}$ and $\{0\} \succ_1 \{0, 1\} \sim_1 \{1\}$. These two preference rankings reflect varying motives for commitment, presented below as a Proposition and Corollary

for ease of reference. Importantly, both account for the cost of effort required to submit strict preferences in our experiment. Proofs can be found in Section C.3.

Proposition 1: *If submitting strict preferences is costly,*

1. *Tempted types* ($\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$) *have relatively small self-control costs,* $\lambda\overline{U}_s$*, compared with their long-run preference for avoiding information,* $|\overline{U}_l|$.

2. *Strongly Tempted types* ($\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$) *have large* $\lambda\overline{U}_s$ *relative to* $|\overline{U}_l|$.

A corollary captures predictions on willingness to pay for menu preferences:

Corollary 1: *Under menu preferences* $\{0\} \succ_1 \{0,1\} \succeq_1 \{1\}$,

1. *Willingness to pay for the preference* $\{0\} \succ_1 \{0,1\}$ *increases in* $\lambda\overline{U}_s$*. If* $\lambda\overline{U}_s \geq -\overline{U}_l$*, it also increases in* $|\overline{U}_l|$.

2. *Willingness to pay for the preference* $\{0,1\} \succ_1 \{1\}$ *decreases in* $\lambda\overline{U}_s$ *and increases in* $|\overline{U}_l|$.

Thus, $\{0,1\} \succ_1 \{1\}$ implies relatively small self-control costs $\lambda\overline{U}_s$ and relatively strong long-run preferences to avoid information $\overline{U}_l$. Intuitively, $\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$ captures those who expect to experience self-control costs so high as to make them (almost) indifferent between $\{0,1\}$ and $\{1\}$.

3.3.4 Self-control costs and deliberation time

Proposition 1 and Corollary 1 provide theoretical predictions on the relationship between menu preferences, willingness to pay, and self-control costs. In other studies of temptation and commitment, a role has been inferred for self-control costs through the absence of dynamically inconsistent choice. In this section, we demonstrate that response times allow us to construct a behavioural measure of self-control costs when a temptation is successfully resisted.

A considerable body of evidence from economics, psychology, and neuroscience finds that decision times increase when preferences approach indifference (see Luce (1986), Krajbich et al. (2014), and Alós-Ferrer et al. (2021) for overviews and Fudenberg et al. (2018) for a theoretical argument using the drift diffusion model).

In our context, the implication is that higher self-control costs in $t = 2$ result in higher response times for a resistible temptation, conditional on the long-run

preference to avoid information. To see why, note that when information will be avoided in period 2 marginal utility of consuming information is $\overline{U}_l + \lambda\overline{U}_s$. As long as $\lambda\overline{U}_s < |\overline{U}_l|$, this quantity approaches 0 as $\lambda\overline{U}_s$ grows relative to $|\overline{U}_l|$. Thus, higher self-control costs increase the closeness of the choice alternatives in $\{0, 1\}$, conditional on resisting the temptation. That is, if self-control costs are higher, it takes individuals longer to determine that resisting information is the right decision.

Strikingly, our model implies that self-control costs have quite different implications for information preferences in $t = 1$. That is, they strengthen the strict preference $\{0\} \succ_1 \{0, 1\}$ and weaken the preference $\{0, 1\} \succ_1 \{1\}$. This leads to two key predictions on the relationship between deliberation times in session 2 and menu preferences in session 1. First, 'Strongly Tempted' types with the preference $\{0\} \succ_1 \{0, 1\} \sim_1 \{1\}$ should have higher deliberation times than 'Tempted' types with the preference $\{0\} \succ_1 \{0, 1\} \succ_1 \{1\}$; high self-control costs diminish the value of $\{0, 1\}$ relative to $\{1\}$, meaning some agents who have only a marginal preference for $\{0, 1\}$ will be unwilling to pay the corresponding effort cost. Second, for a resistible temptation, willingness to pay for the preference $\{0\} \succ_1 \{0, 1\}$ should be positively correlated with deliberation time: early commitment becomes more valuable if self-control costs are expected to be high.

In empirical analysis of deliberation times, we also need to consider the role of $|\overline{U}_l|$. In particular, deliberation time should decrease in $|\overline{U}_l|$ conditional on the temptation being resisted. For our analysis of whether 'Strongly Tempted' types $\{0\} \succ_1 \{0, 1\} \sim_1 \{1\}$ have higher self-control costs than 'Tempted' types $\{0\} \succ_1 \{0, 1\} \succ_1 \{1\}$, differences in deliberation times across types could also reflect differences in $|\overline{U}_l|$. On the other hand, $|\overline{U}_l|$ should be irrelevant for *WTP* when the temptation is resisted: the only factor which determines the preference $\{0\} \succ_1 \{0, 1\}$ is the cost of self-control $\lambda\overline{U}_s$.[6]

Separately, empirical findings from psychology and economics indicate that longer response times are associated with the use of deliberative decision-making processes rather than intuitive processes, which could be considered analogous to the two selves in our model of temptation (Kessler et al., 2023). Through this lens, another interpretation of the argument presented above is that it takes longer for deliberative processes to win over intuitive processes when the appeal of a temptation

[6]If individuals believe they have some chance of succumbing to temptation, as in Dekel and Lipman (2012), a larger $|\overline{U}_l|$ could increase willingness to pay for $\{0\} \succ_1 \{0, 1\}$. However, it has the opposite effect on deliberation times. Thus, our model dictates that a positive relationship between willingness to pay and deliberation times must reflect self-control costs $\lambda\overline{U}_s$.

is stronger.

3.4 Results

Our first set of results concerns participants' menu preferences from session 1, which reveal whether information is seen as tempting. To support the interpretation that commitment is driven by temptation, we then ask whether there is any evidence of the two motives for commitment outlined in Section 3.3: dynamically inconsistent choice, and costly self-control. For this purpose, we analyse whether participants succumb to temptation, or appear to suffer self-control costs in resisting it, when offered information against their wishes. In doing so, we test the model's precise predictions on heterogeneity in the strength of temptation across individuals. We then compare preferences across our treatment conditions, investigating whether anticipating partial information makes further information more or less tempting *ex-ante*. Finally, we offer some qualitative evidence supporting the interpretation that information is tempting.

3.4.1 Menu preferences from session 1

Result 1: *49% of all individuals, and 73% of those who want to avoid information, have menu preferences revealing temptation to consume information.*

A menu preference ranking with $\{0\} \succ_1 \{0,1\} \succeq_1 \{1\}$ implies information is seen as tempting: it encompasses a strict preference both to avoid information and to remove it from the later choice set, relative to being offered it. There are two types for whom this holds: the 'Tempted' type $\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$, and the 'Strongly Tempted' type $\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$. However, 13 menu preference rankings are possible in total.

The main results of the experiment are in Table 3.1, which reports the frequency of each full menu preference. Using our primary measurement, information is tempting for 49% of individuals: not only do they strictly prefer not to see information ($\{0\} \succ_1 \{1\}$), but they would explicitly rather not be *offered* information than be offered it ($\{0\} \succ_1 \{0,1\}$).[7] Within this set, the large majority (38% of the sample) are the 'Tempted' type $\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$. Recalling Proposition 1, this type is tempted

[7] In all, just over two-thirds of the sample strictly prefers to avoid the information offered than see it (Table C.1).

Table 3.1 Menu preferences from session 1

Menu preference	Type	N	Share
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	Tempted	254	38.4% (1.9)
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	80	12.1% (1.3)
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	Strongly Tempted	71	10.7% (1.2)
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	Standard info. averse	54	8.2% (1.1)
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	50	7.6% (1)
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	34	5.1% (0.9)
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	Flex	33	5% (0.8)
$\{1\} \succ_1 \{0,1\} \succ_1 \{0\}$	Tempted (info. loving)	29	4.4% (0.8)
$\{0,1\} \sim_1 \{1\} \succ_1 \{0\}$	Other	17	2.6% (0.6)
$\{0\} \succ_1 \{1\} \succ_1 \{0,1\}$	Flexibility averse	14	2.1% (0.6)
$\{1\} \succ_1 \{0\} \sim_1 \{0,1\}$	Other	13	2% (0.5)
$\{0\} \sim_1 \{1\} \succ_1 \{0,1\}$	Commitment loving	9	1.4% (0.5)
$\{1\} \succ_1 \{0\} \succ_1 \{0,1\}$	Flexibility averse	4	0.6% (0.3)
Total		662	100%

Note: In this, our primary measure, participants must complete an effort task for each strict preference submitted in their ranking. If present, the first strict preference additionally requires participants to have a strictly positive willingness to pay to avoid a swap of their first and second choices. If $WTP = 0$, the first strict preference is replaced with indifference. Standard errors in parentheses, in percentage points. 'Flex' types are those with a strict preference for $\{0,1\}$ over both $\{0\}$ and $\{1\}$.

by information but has strong *ex-ante* preferences to avoid information relative to self-control costs.

All 13 possible preference rankings are represented in the sample. This is even true in the first stage of elicitation, before participants are prompted to reconsider any strict preferences they submit (and pay a small effort cost to confirm them). However, the 'Tempted' type is by far the most common of the 13 in the sample, followed by those who are indifferent between all three options, and then 'Strongly Tempted' types with the ranking $\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$. This last type also finds information tempting but expects either high self-control costs or a high chance of succumbing to temptation, such that the marginal value of being offered $\{0,1\}$ relative to just $\{1\}$ is relatively small. As we will confirm when analysing session 2 behaviour, as well as our naturally occurring measure of self-control costs during session 2, costly

self-control is an important motive for commitment.

Setting aside those who can be classified as finding information tempting, 18% of individuals in the sample have a strict preference for flexibility over either choice set restriction, suggesting uncertain preferences for information.[8] Within this set, binary preferences over {0} and {1} are approximately evenly distributed over the three possibilities.

One concern may be that the information we offer is used instrumentally by some participants, for example to motivate a donation of their own after the experiment's conclusion. A priori, this strategy would work in the opposite direction of our results: individuals are more likely to be tempted not to donate to charity (Andreoni and Serra-Garcia, 2021), making commitments to consuming information more attractive and reducing the prevalence of temptation to consume information. An even closer examination of the data confirms that we can rule out this concern: the distribution of menu preferences for the small subset of participants who did *not* take the bonus is remarkably similar to that for the whole sample (Table C.4). For these participants, the information in the envelope reveals whether they could have taken the bonus without affecting the charity donation, but is tempting in precisely the same way.

This last analysis also encourages some optimism about the external validity of our results. Information on *Donor* status has quite different, although possibly still negative, psychological implications for those who did not take the bonus: rather than guilt, an individual finding out they could have taken the bonus without consequence would likely feel regret. Nonetheless, temptation occurs with exactly the same frequency. As such, it does not appear that temptation for information is an idiosyncratic feature of guilt aversion.

Willingness to pay. Since effort costs are mandatory for strict preferences in our main measure, it is possible that we measure some genuine strict preferences as indifferences where participants value a strict preference less than the cost of effort. Overall, 29% of participants with initial strict preferences decline to pay at least one unit of the effort costs in proceeding to the next stage. In this sense, this measure of preferences — our preregistered primary outcome — can be seen as conservative on balance.[9]

After participants completed any tasks to confirm strict preferences, we also

[8]Type categories are condensed in Table C.5.

[9]More detail on how preferences evolve through the elicitation procedure can be found in Section C.2.

collected monetary *WTP* for the first strict preference in their ranking. Some other studies of commitment devices, such as Augenblick et al. (2015), find that demand drops to nearly 0 when participants are asked if they are willing to pay money. In our case, about 43% of Tempted types are also willing to pay strictly positive sums of money in order for $\{0\}$ and $\{0, 1\}$ not to be swapped in their preference rankings, even though they have already completed two effort tasks confirming their strict preferences (Table C.6). This leaves 16% of individuals willing to pay both effort costs and small sums of money in order for information not to be generated and offered to them. Following this stage, the remaining 57% of Tempted types would be classified as standard information averse. However, given that all of these individuals already paid an effort cost to justify their strict preference by this stage, this classification is very conservative and we do not emphasise it. Instead, it can be seen as a sort of lower bound on the prevalence of strict preferences for commitment, particularly for readers who are not convinced that effort costs are sufficient to extract genuinely strict preferences for commitment over flexibility.

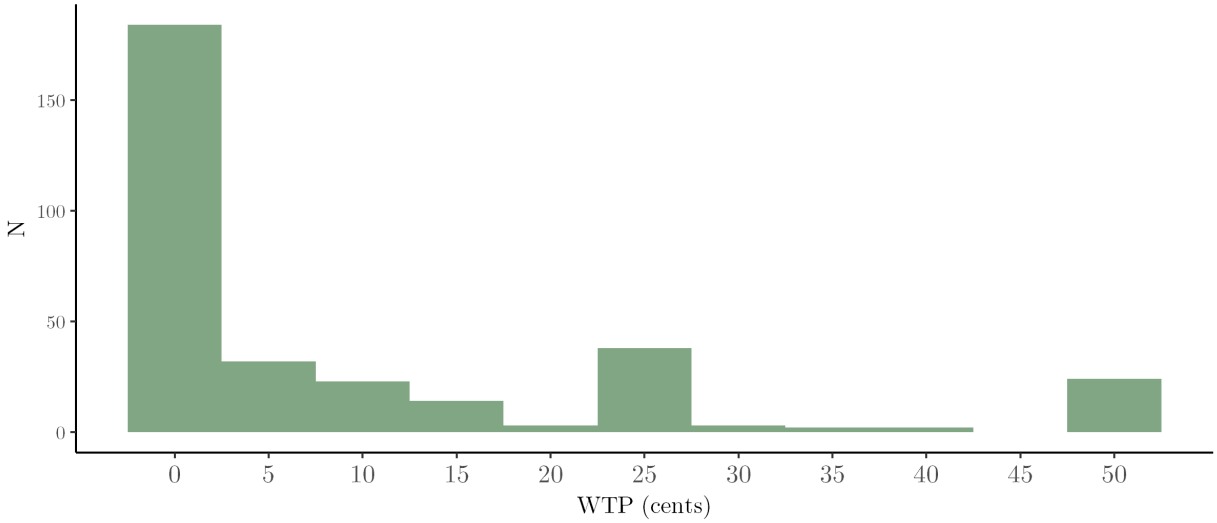
The average *WTP* among Tempted types is \$0.09, reflecting the high share with a *WTP* of 0 (Table C.6).[10] Among those with a positive *WTP*, the mean is \$0.20. Strongly Tempted types express similar valuations, with 45% having $WTP > 0$ and an average *WTP* of \$0.26 among those with $WTP > 0$. There is some evidence of bunching at \$0.50 across both types, suggesting that we underestimate average *WTP*: 17% of tempted and Strongly Tempted types with a positive *WTP* submit the maximum *WTP* of \$0.50. Otherwise, WTP is somewhat evenly distributed across the range of permitted values (Figure 3.1).

Heterogeneity. Our data also include a set of participant characteristics that we can use to examine determinants of preference heterogeneity. Participants' age, ethnicity, and sex were automatically provided by the panel, whereas we manually elicited qualitative measures of risk preferences, patience, self-control, and altruism at the end of session 2.[11] Since the range of preference rankings is broad, we report two regressions focusing on cases of particular interest (Table C.7). The first predicts the event of an implied binary preference to avoid information, $\{0\} \succ_1 \{1\}$. In the first case, the likelihood of submitting a strict preference to avoid information is

[10] *WTP* is bounded above at \$0.50, and participants could only submit multiples of 5 cents (\$0, \$0.05, \$0.10,..., \$0.50).

[11] Since these measures were collected after bonus and information choices, they could be endogenous to both. As such, this analysis is presented descriptively.

Fig. 3.1 *WTP* histogram: tempted and Strongly Tempted types



Note: Willingness to pay for the menu preference $\{0\} \succ_1 \{0,1\}$, elicited via the BDM procedure. If the participant lost the auction, their new menu preference ranking would begin with $\{0,1\} \succ_1 \{0\}$.

lower among some older age categories and participants of Asian ethnic origin. Other variables are not strong predictors of preferences to avoid information, notably including the price of information and self-reported scores relating to altruism, patience, and self-control. The exception is that participants with higher subjective risk tolerance are less likely to submit preferences to avoid information. This is likely a reflection of the fact that the decision to seek information is inherently risky, given that it can contain either good or bad news. The model of preferences for information provided in Section 3.5 formalises this idea.

The second regression is for the probability of a strict preference for commitment ($\{0\} \succ_1 \{0,1\}$) conditional on a strict preference to avoid information ($\{0\} \succ_1 \{1\}$). Very few of the individual characteristics we examine have any predictive power. The exception is a weakly positive correlation of commitment demand with age.

3.4.2 Information choices from session 2

Result 2: *3% of individuals tempted by undesired information pay to access it in session 2.*

In session 2, taking place a day after session 1, participants' preferences are randomly implemented according to the scheme previously described. With 50%

probability, menu preferences were ignored and $\{0, 1\}$ is implemented. With 40% probability, the top menu from $\succeq_1$ was implemented. With 10% probability, the second menu from $\succeq_1$ is implemented. Of the segment facing $\{0, 1\}$ exogenously, 16% elect to pay to open the envelope overall.

Table 3.2 splits session 2 information choices by implied session 1 binary ranking of $\{0\}$ and $\{1\}$. Information access is markedly more common for participants who submitted rankings with $\{1\} \succ_1 \{0\}$ than those with $\{1\} \sim_1 \{0\}$. Matching experimental studies of dynamic consistency in other choice contexts, succumbing to temptation is relatively uncommon (Toussaert, 2018). This may be particularly unsurprising in our context, given that doing so would involve paying a small sum of money to access information having borne at least a cost of effort not to be offered it just a day earlier.

For those individuals who submitted $\{1\} \succ_1 \{0\}$, 65.6% follow through on their session 1 preference by consuming information when asked to choose again. However, it should be noted that a majority (56%) of participants in this category weakly preferred flexibility to committing to see the information. For these participants, either $0 \succeq_2 1$ or $1 \succeq_2 0$ could be consistent with the menu preference $\succeq_1$.

Table 3.2 Session 2 information choice by session 1 information preference

Session 1 information preference	% $1 \succ_2 0$	...of N
$\{0\} \succ_1 \{1\}$	5.4% (1.7)	184
$\{0\} \sim_1 \{1\}$	21.6% (5.9)	51
$\{1\} \succ_1 \{0\}$	65.6% (8.6)	32
Total	15.5% (2.2)	265

Note: Standard errors in parentheses, in percentage points. Only includes participants who were randomly allocated to receive $\{0, 1\}$ regardless of their menu preferences.

Table 3.3 provides a more granular separation of session 2 choices by session 1 preferences. We first focus on the two main types of interest: tempted and Strongly Tempted types, both of whom view information as tempting and strictly prefer not to be offered it as a result. Preference reversals are relatively infrequent across both types, at rates of 2% for Tempted types and 7% for Strongly Tempted types. As we will see later, individuals who are more likely to succumb to temptation are less willing to pay an effort cost for the strict preference $\{0, 1\} \succ_1 \{1\}$, meaning they

Table 3.3 Session 2 information choice by session 1 menu preference

Session 1 menu preference	Type	% $1 \succ_2 0$	...of N
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	Tempted	1.9% (1.3)	105
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	23.5% (7.3)	34
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	Dynamic inconsistency	6.9% (4.7)	29
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	Standard info. averse	8% (5.4)	25
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	15.8% (8.4)	19
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	Flex	7.7% (7.4)	13
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	80% (12.6)	10
$\{0,1\} \sim_1 \{1\} \succ_1 \{0\}$	Other	42.9% (18.7)	7
$\{1\} \succ_1 \{0,1\} \succ_1 \{0\}$	Tempted (info. loving)	83.3% (15.2)	6
$\{0\} \succ_1 \{1\} \succ_1 \{0,1\}$	Flexibility averse	16.7% (15.2)	6
$\{1\} \succ_1 \{0\} \sim_1 \{0,1\}$	Other	60% (21.9)	5
$\{0\} \sim_1 \{1\} \succ_1 \{0,1\}$	Commitment loving	66.7% (27.2)	3
$\{1\} \succ_1 \{0\} \succ_1 \{0,1\}$	Flexibility averse	33.3% (27.2)	3
Total			265

Note: Standard errors in parentheses, in percentage points. Only includes participants who were randomly allocated to receive $\{0,1\}$ regardless of their menu preferences.

are classified as Strongly Tempted types in the effort cost stage of our elicitation procedure.

Among types with $\{0\} \succ_1 \{1\}$, the highest probability of accessing information in a group of non-trivial size is within the type $\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$. Although these individuals prefer $\{0\}$ to $\{1\}$ under $\succeq_1$, they strictly prefer flexibility to either form of commitment. In that sense, either $1 \succ_2 0$ or $0 \succ_2 1$ is consistent with $\succeq_1$ for these individuals, unlike those for whom $\{0\} \succ_1 \{0,1\} \succeq_1 \{1\}$.

3.4.3 Self-control costs

Result 3: *Higher willingness to pay for commitment is significantly associated with longer deliberation periods, other than at exceptionally long periods. Strongly Tempted types spend significantly longer deciding whether to access information than Tempted types.*

So far, we have established that a substantial proportion of individuals in our experiment believe they will suffer from the mere presence of information in their

choice set. 49% of individuals have strict preferences for restricting their choice set, and 43% of that subset are also willing to pay small sums of money as well as costs of effort to maintain those preferences. As in Toussaert (2018), dynamically inconsistent choice or random self-indulgence (Dekel and Lipman, 2012) are unlikely to explain all of the demand for commitment devices; individuals rarely succumb to temptation in session 2. Our model suggests that a major role must therefore be played by $\lambda\overline{U}_s$, the self-control cost from having to turn down information when it is offered in session 2.

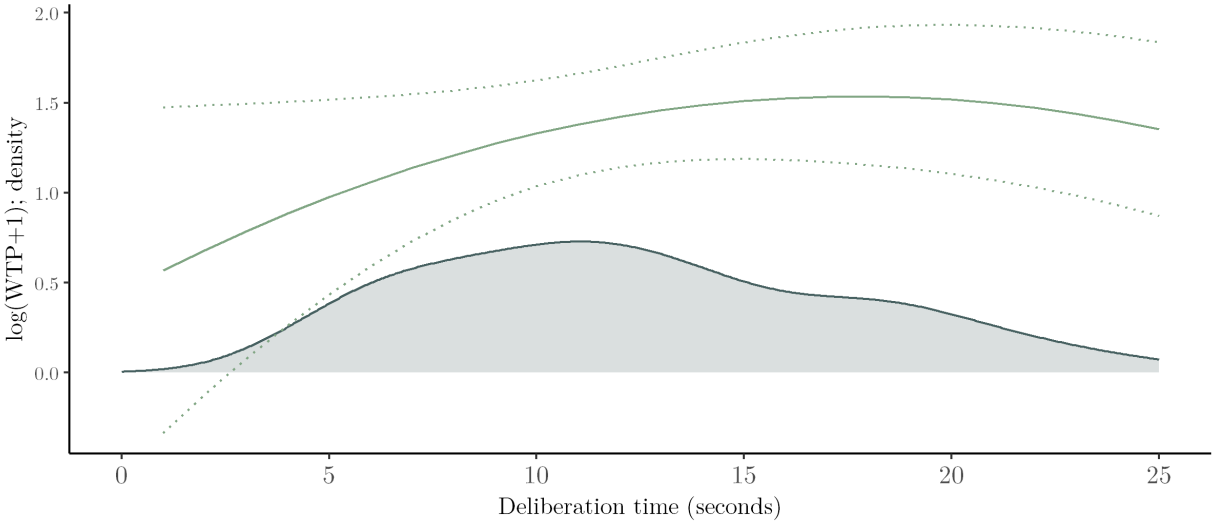
As argued in Section 3.3, deliberation times for those tempted by information should increase in self-control costs, conditional on resisting the temptation. Since the vast majority of tempted and Strongly Tempted types resist the temptation to consume information in session 2, the effect of increasing $\lambda\overline{U}_s$ must be to bring the marginal utility of consuming information closer to 0 as required.

Participants spend an average of 16 seconds on the page when $\{0,1\}$ was exogenously implemented. As a benchmark on our measure, notice that average response times for unincentivised questions on wellbeing in online studies like Liu and Netzer (2023) range from 2–4 seconds. As such, deciding whether to access information clearly involves careful deliberation from our participants. However, our main interest is in how self-control costs differ across session 1 menu preferences, which were elicited a day earlier, as well as across individuals with different willingness to pay effort costs and money for their preferences for commitment.

Willingness to pay and self-control costs. By examining the *intensity* of individuals' preferences for commitment ($\{0\} \succ_1 \{0,1\}$), we obtain a direct test of whether sophistication about self-control costs motivates commitment (Corollary 1). The most notable prediction of our model is that willingness to pay for the strict preference $\{0\} \succ_1 \{0,1\}$ should increase in self-control costs $\lambda\overline{U}_s$, assuming that the temptation is resisted.

Our main emphasis is on a test of whether experienced self-control costs (deliberation times) explain monetary *WTP* to retain $\{0\} \succ_1 \{0,1\}$. However, we also analyse a binary measure of whether an individual is willing to bear a cost of effort to retain any strict preferences submitted before that stage. Accordingly, we examine whether experienced self-control costs explain willingness to complete the effort task for both $\{0\} \succ_1 \{0,1\}$ and $\{0,1\} \succ_1 \{1\}$, for those who submit the preference

Fig. 3.2 Period 2 deliberation time vs. predicted period 1 *WTP* to maintain $\{0\} \succ_1 \{0,1\}$ for Tempted types



Note: Scaled density of decision times plotted underneath predicted *WTP*. 95% prediction interval plotted around predicted *WTP*.

$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$ in the unincentivised elicitation stage.[12]

In our model, the long-run preference for information $|\overline{U}_I|$ does not determine *WTP* for commitment when temptation is resistible: the individual knows they will not consume information when offered it, so self-control costs $\lambda\overline{U}_s$ are all that matter. However, some agents may believe that they have a chance of succumbing to temptation even if they eventually do not (Dekel and Lipman, 2012). This could make $|\overline{U}_I|$ an unobserved variable that is positively correlated with *WTP* in practice. If so, our estimate of the effect of self-control costs on *WTP* may be biased towards 0. To help to address this possibility, we add a qualitative measure of interest in accessing the information, at the time of session 1, as a control variable for $|\overline{U}_I|$. Further, we also control for whether the information was actually accessed in session 2; deliberation time measures $\lambda\overline{U}_s$ only while $\lambda\overline{U}_s < |\overline{U}_I|$ and the information is not consumed. However, given that most participants do not consume information, we expect neither variable to play a significant role.

The results are reported in Table 3.4. We first consider willingness to pay effort costs and money for $\{0\} \succ_1 \{0,1\}$, which our model predicts to increase in self-control

[12]Since the latter was the second adjacent strict preference for Tempted types, we did not elicit monetary *WTP* for it.

costs. Column (1) contains the results for monetary *WTP*, while (2) examines the decision to complete the real-effort task after submitting a strict preference in the unincentivised stage. Since we do not know have precise theoretical predictions on the functional form of the empirical relationship, we allow for response time to have a non-linear marginal effect by including a quadratic term.

Table 3.4 Regressions for preference intensity

	$\{0\} \succ_1 \{0,1\}$		$\{0,1\} \succ_1 \{1\}$
	OLS	*logistic*	*logistic*
	$log(WTP+1)$	Paid effort cost	Paid effort cost
	(1)	(2)	(3)
Not at all interested ($\lvert\overline{U}_I\rvert$)	−0.28 (0.30)	0.84 (0.53)	−0.06 (0.45)
Response time ($\lambda\overline{U}_s$)	0.12 (0.06)	0.12 (0.09)	−0.04 (0.07)
Response time^2 ($\lambda\overline{U}_s$)	−0.00 (0.00)	−0.00 (0.00)	−0.00 (0.00)
Opened envelope	0.54 (0.77)	−0.15 (1.23)	−1.57 (0.97)
Constant	0.64 (0.56)	0.60 (0.91)	1.98 (0.80)
Observations	133	149	149

Note: Coefficients are log-odds, other than for column (2) which is OLS. Sample restricted to those randomised to face $\{0,1\}$ exogenously in session 2. Columns (1) and (3) capture all those tempted by information in the raw elicitation stage, while (2) includes only those classified as tempted or Strongly Tempted in the main stage.

The results in column (1) indicate that within the set of those tempted by information, those willing to pay larger amounts of money to maintain $\{0\} \succ_1 \{0,1\}$ generally appear to find it harder to resist information a day later. This is precisely as predicted by our model of costly self-control. The effect can be more easily observed visually, as in Figure 3.2. Those with higher self-control costs in period 2 are willing to pay more money to maintain their preference for commitment over flexibility, $\{0\} \succ_1 \{0,1\}$. Roughly, each additional second of response time corresponds to an additional \$0.01 of *WTP*.[13] As we approach uncommonly high deliberation times (above the 90$^{\text{th}}$ percentile) the effect of deliberation time on *WTP* begins to level

[13]To address the potential concern that we are capturing general response time or attention effects, we run a placebo test using total completion time for session 1 (Table C.10). There is no relationship between this variable and *WTP*.

off and decrease gently. These very long deliberation periods approach the levels measured for individuals with a strict preference for flexibility in period 1 (Table C.8), suggesting that preference uncertainty begins to reduce the value of commitment for this small set of individuals.

Columns (2) and (3) concern our coarser measures of willingness to pay effort for the preferences $\{0\} \succ_1 \{0,1\}$ and $\{0,1\} \succ_1 \{1\}$. Here, we do not find a significant relationship between self-control costs and willingness to pay. However, individuals who do eventually succumb to temptation in period 2 are significantly less willing to pay effort costs for this strict preference. Intuitively, if successful self-control is unlikely, it is not worth bearing effort costs to have the ability to avoid information.

Types and deliberation times. Our model of temptation and self-control also predicts a relationship between menu preferences and deliberation times. From Proposition 1, within the set of individuals who find information tempting, 'Strongly Tempted' types ($\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$) anticipate higher costs of self-control, $\lambda\overline{U}_s$, and have weaker long-run preferences to avoid information, $|\overline{U}_l|$, than 'Tempted' types ($\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$). Both factors should increase Strongly Tempted types' deliberation time when facing the menu $\{0,1\}$.

Mean deliberation time for Tempted types is about 13 seconds (Table 3.5).[14] On the other hand, Strongly Tempted types spend around 5 seconds longer on average deciding whether or not to open the envelope. Thus, deliberation times closely match the predictions of Proposition 1: Strongly Tempted types should be closer to indifferent between accessing and avoiding information in period 2.[15] Notably, the longest deliberation times are among types who strictly preferred flexibility in session 1, strongly consistent with the interpretation that deliberation times are explained by the drift-diffusion model of decision speed: Flexibility types were the most uncertain of their preference for information in session 1.

An open question is whether a longer period of deliberation predicts failure of self-control. Our model does not provide clear guidance on this matter, since it only generates a strong prediction on the relationship between self-control costs and deliberation time when the temptation is *resisted*.

We estimate a logistic regression for the event that the envelope is opened in

[14]Deliberation times for all types are reported in the Appendix, in Table C.8.

[15]The difference in deliberation times across types is retained after controlling for our coarse measure of $|\overline{U}_l|$ (a qualitative measure of interest in seeing the information in session 1), suggesting that anticipated self-control costs may play a significant role in type classification.

Table 3.5 Measures of deliberation under exogenous $\{0,1\}$ in session 2 by session 1 preference ranking

Menu preference	Type	Time (seconds)	N
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	Tempted	13.1 (0.5)	105
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	17.5 (1.9)	33
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	Strongly Tempted	18 (1.9)	28
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	Standard info. averse	15.6 (2.3)	25
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	21.5 (2.9)	19
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	Flex	25.9 (3.2)	13
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	16.6 (1.7)	10

Note: Only the most common types included, for brevity. Full table in Appendix.

session 2 under exogenously implemented $\{0,1\}$, with quadratic deliberation time as the regressor of interest, for those whose session 1 preferences imply they find information tempting (Table C.9). The log-odds of failed self-control increase in deliberation time at low levels, before levelling off and decreasing slightly at very long deliberation periods.[16]

3.4.4 Comparison of treatment conditions

Result 4: *Ex-ante preferences to avoid information are weaker when partial information is unavoidable.*

Since the total content and average price of information are identical across treatment conditions, our prior analysis considered menu preferences and information choices aggregated across the whole sample. In reality, individuals who take the bonus are randomly allocated to one of two treatment conditions: *PartAvoid* and *FullAvoid*. To recall, there is only a consequence of taking the bonus if the individual is a *Donor*. Participants in both treatments are able to confirm if they were a *Donor* during session 2. What differs is how much information participants can choose to avoid in session 2, and correspondingly what they can commit to avoiding in session 1.

In *PartAvoid*, participants cannot avoid receiving an initial signal of the consequences of their decision to take the bonus. The signal s takes values s_0 or s_{20},

[16]The probability of accessing information reaches its peak at just over 15 seconds of deliberation and decreases after that point only slightly.

where $P(Donor|s = s_0) = 0$ and $P(Donor|s = s_{20}) = 0.2$. If $s = s_{20}$, an envelope letting the participant learn the value of *Donor* is generated. Recalling Figure C.1, the participant then chooses to open or not open the envelope.[17] In *FullAvoid*, the signal s can be avoided, so participants always have an opportunity to obtain strictly less information than those in *PartAvoid*. In *FullAvoid*, if the envelope is opened and $s = s_{20}$, the participant must then choose whether or not to confirm their *Donor* status.

The only difference across the two treatments in session 1 consists of a small variation in the text explaining the information offer sequence and the contents of the envelope (Figures C.5 and C.6). The text describing the choice alternatives and the subsequent stages of the elicitation procedure only make reference to the envelope, so they are identical across the two treatments. Differences in menu preferences across the treatment conditions therefore measure whether it is easier or harder to commit to avoiding information when more of it must be avoided. If commitment demand is higher in *FullAvoid* than *PartAvoid*, we can conclude that avoiding more information makes commitment less attractive. If commitment demand is higher in *FullAvoid*, we can conclude that committing to avoid information is harder when partial information cannot be avoided.

Menu preferences. Table 3.6 splits session 1 menu preferences by treatment group. There is no significant difference across treatment groups in the prevalence of Tempted types. However, the share of Strongly Tempted types, with the menu preference $\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$, is marginally greater in *FullAvoid* than *PartAvoid*. This type strictly prefers commitment but is indifferent between being offered the information and automatically seeing it. Relative to Tempted types, of which there is no difference in prevalence across treatment arms, Strongly Tempted types rank the menus $\{0,1\}$ and $\{1\}$ equally. As noted in Proposition 1, one possibility is that Strongly Tempted types have a weaker *ex-ante* preference for avoiding information. Since in *PartAvoid* the proposition of avoiding information ($\{0\}$) involves having already learned there was a 20% chance of being a *Donor*, it appears to become less attractive to do so. This also explains the finding that the increased prevalence of

[17] Note that all the menus in *PartAvoid* are only offered if $s = s_{20}$, in order to avoid introducing noise by asking subjects for multiple preference rankings involving contingent reasoning. While this could reduce participants' willingness to pay for their strict preferences, we find no evidence of this phenomenon: there is little difference in participants' willingness to pay effort costs to confirm their strict preferences.

Table 3.6 Menu preferences: treatment comparison

		N		Share			
Menu preference	Type	*PartAvoid*	*FullAvoid*	*PartAvoid*	*FullAvoid*	Diff.	p-value
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	**Tempted**	118	113	**37.7%** (2.7)	**39.1%** (2.9)	-1.4	0.79
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	**Strongly Tempted**	26	38	**8.3%** (1.6)	**13.1%** (2)	-4.84	0.07
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	43	30	13.7% (1.9)	10.4% (1.8)	3.36	0.26
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	**Standard info. averse**	21	27	**6.7%** (1.4)	**9.3%** (1.7)	-2.63	0.3
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	23	25	7.3% (1.5)	8.7% (1.7)	-1.3	0.66
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	13	16	4.2% (1.1)	5.5% (1.3)	-1.38	0.55
$\{1\} \succ_1 \{0,1\} \succ_1 \{0\}$	Strongly Tempted (info. loving)	15	11	4.8% (1.2)	3.8% (1.1)	0.99	0.69
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	**Flex**	23	8	**7.3%** (1.5)	**2.8%** (1)	4.58	0.02
$\{0,1\} \sim_1 \{1\} \succ_1 \{0\}$	Other	9	6	2.9% (0.9)	2.1% (0.8)	0.8	0.71
$\{1\} \succ_1 \{0\} \sim_1 \{0,1\}$	Other	8	5	2.6% (0.9)	1.7% (0.8)	0.83	0.68
$\{0\} \succ_1 \{1\} \succ_1 \{0,1\}$	Flexibility averse	7	5	2.2% (0.8)	1.7% (0.8)	0.51	0.88
$\{0\} \sim_1 \{1\} \succ_1 \{0,1\}$	Commitment loving	5	3	1.6% (0.7)	1% (0.6)	0.56	0.81
$\{1\} \succ_1 \{0\} \succ_1 \{0,1\}$	Flexibility averse	2	2	0.6% (0.5)	0.7% (0.5)	-0.05	1
Total		313	289	100%	100%		

Note: The 9% of participants who did not take the bonus are excluded from this table, since they were all allocated to the same condition (in which all of the information was accessed at the same time). Difference in percentage points.

Strongly Tempted types in *FullAvoid* appears to be at the expense of flexibility types with $\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$, who have the same preference except with $\{0\}$ and $\{0,1\}$ switched. This type is consistent with preference uncertainty: they are indifferent between $\{0\}$ and $\{1\}$ under $\succeq_1$, and so would like to wait to make their decision. These conclusions are also borne out when condensing menu preferences into binary information preferences, as in Table C.12.

To sum up, information choices in *PartAvoid* are made only after unfavourable partial information has already been received. It is striking that individuals nonetheless find it harder to resist acquiring further information *ex-ante*. This suggests the provision of partial information could be an effective strategy in tempting access to further information: it is harder for individuals to plan to avoid even potentially upsetting information when they expect to be given an initial indication of its content.

Choices and deliberation times. Since the distribution of menu preferences differs across treatment groups, we cannot compare choices and deliberation times within types. However, we can analyse how behaviour differs across treatment groups for all individuals facing $\{0,1\}$ exogenously.[18]

[18]Fewer than 50% of participants in *PartAvoid* face $\{0\}$, since it is only offered conditional on

17.2% of individuals in *FullAvoid* access the information, compared with 13% in *PartAvoid*. The corresponding standard errors are about 3 percentage points, so analysis of differences is somewhat underpowered. However, there is no clear evidence that the slightly weaker session 1 preference to avoid information among individuals in *PartAvoid* is reflected in choices. This finding suggests the short-term motive to access information is less strongly affected by our treatment than the long-term one.

This conclusion is tentatively corroborated by differences in deliberation times across treatment groups. Participants in *PartAvoid* spend 2.3 seconds less time deciding whether or not to access information. This difference is significant at the 10% level. In interpreting this finding, note that across all types, deliberation times measure closeness to indifference in ranking 0 and 1 in the menu $\{0, 1\}$. Since most participants resist information, this result tentatively suggests that information may even be marginally *less* attractive in session 2 in *PartAvoid* than *FullAvoid*, reversing the conclusion from session 1. In Section 3.5, we outline a model of information preferences that can accommodate these findings.

3.4.5 Qualitative measures of information preferences and self-control costs

We collected a set of qualitative measures in both sessions, which we now use to further interrogate the motives for choice set restrictions.

We begin by more closely examining our curiosity measure 'How interested are you to know if the charity donation was reduced by \$15 as a result of your choice?', which we previously introduced as a coarse measure of $|\overline{U}_l|$ in our analysis of *WTP*.[19] Average scores are reported by type and session in Table C.13. In session 1, self-control, weak self-control, and standard information-averse types are those with the lowest measured curiosity, which mirrors the idea from our model that the commitment evaluation of avoiding information $|\overline{U}_l|$ should be the highest among these types. Curiosity in session 1 is much higher among flexibility types and those indifferent between all menus.

Since we asked this question in both sessions, within-individual comparisons reveal how $\overline{U}_s$, the short-run value of information, differs from the long-run value

receiving the signal $s = s_{20}$.

[19]For those who did not take the bonus, the question was 'How interested are you to know if you could have taken the \$4 without affecting the charity donation?'.

$\overline{U}_l$. Noting that we only include participants who successfully resisted information in period 2, Tempted types experience a significant increase in the subjective attractiveness of information from session 1 to session 2. This supports the prediction of our model, based on Fudenberg and Levine (2012), that it is easier to resist the temptation to consume information in advance. The increase in interest in information just misses the margin for statistical significance for Strongly Tempted types, although the test is more poorly powered. Like Tempted types, standard information-averse types also experience a significant increase in the attractiveness of information. Since these types were unwilling to pay effort costs to commit to avoiding information, the increase in the attractiveness of information could suggest some naivete about self-control costs at the time of submitting menu preferences. Finally, flexibility types generally did not experience increases in the attractiveness of information. Indeed, those with the ranking $\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$ experience a sizeable *decrease* in the attractiveness of information conditional on deciding not to access it.

Participants were also given the option to provide written rationales for their session 1 menu preferences. In all, 69% of participants chose to answer the corresponding free-text question. Most respondents chose to focus on their binary information preference $\{0\} \succ_1 \{1\}$, reflecting the strength of sentiments about the \$4 choice and the desire to avoid information about its consequences. Overwhelmingly, responses within this set appear to indicate a desire to engage in wishful thinking about the consequences of their action or simply avoid thinking about it at all.

Perhaps more interesting is the set of responses for Tempted types which provide rationales for submitting $\{0\} \succ_1 \{0,1\}$. Although it is often difficult to distinguish clearly between these two interpretations, about half allude to self-control costs (e.g. 'I would like to remove the option because it will help me forget feeling bad about my choice.'; 'I'd rather not know and not have to make the decision.') and half seem to perceive a risk of failed self-control (e.g. 'I would rather not be given the option so I don't feel tempted to open it and find out.'; 'Because that way I don't even try to open it at all'). This is consistent with our previous findings: at least some individuals correctly believe they have a chance of later accessing information in spite of their preferences, in line with a model of random self-indulgence (Dekel and Lipman, 2012), but many others seem to be sure they will not open the envelope even if it is offered next time, suggesting costly self-control must be the key motive. In the latter category, several participants mention that being offered the envelope would

prolong the guilt they feel about their decision to take the bonus. This suggests that one benefit of the menu $\{0\}$ relative to $\{0, 1\}$ is that it allows attention to be diverted away from the contents of the envelope after session 1.

3.5 Why is information tempting?

To provide some intuition for our results, we now show how temptation can arise naturally when intrinsic preferences for information are introduced into models of self-control. Critical to our argument is the role of timing in temptation: in order for information to be attractive to the short-run self and unattractive to the long-run self, it must deliver benefits in the short run and costs in the long run. By using existing results from the literature on intrinsic preferences for information, we provide general conditions in which temptation might occur if agents recall information increasingly poorly over time. Then, to provide further intuition, we illustrate a stylised model of preferences satisfying those conditions.

Recall that temptations in the dual-self framework, like in other representations of dynamically inconsistent preferences, are characterised by a conflict between short-run and long-run evaluations. In order to build our argument, we first define a simple temptation in a similar sense to Fudenberg and Levine (2012):

Definition 1: *A simple temptation is initially positively valued, but its value monotonically decreases and becomes negative over time.*

In the dual-self model, note that temptation is controlled by the size of the short-run self's relative discount factor μ. That is, for small enough values of μ, Definition 1 is sufficient for temptation to actually occur: the short-run self and long-run self will be in conflict.

Mode primitives. Let $\theta \in \Theta$ capture the state of the world, and consider an agent with a prior belief $\mu \in \Delta\Theta$. Posterior beliefs in period t are given by $\nu_t \in \Delta\Theta$. In our context, preferences for information are from its psychological value, meaning ν_t, rather than θ, enters preferences. Suppose the agent has a utility function over belief distributions, $u : \Delta\Theta \to \mathbb{R}$ and maximises expected utility over *information prospects* I_t, defined as in Lipnowski and Mathevet (2018) as an element of $\mathcal{R}(\mu) = \left\{I_t \in \Delta\Delta\Theta : E_{\nu_t \sim I_t}(\nu_t) = E(\mu)\right\}$.

An information prospect I_t is evaluated by the agent by taking expectations over $u(\nu_t)$ at her prior belief μ. In order to introduce temptation into this framework, we permit dynamic preferences over information prospects in each period. If the agent has a discount factor δ, her (normalised) expected lifetime utility from a sequence of information prospects $(I_t, I_{t+1}, I_{t+2}, ...)$ can be written $(1-\delta)\sum_{n=1}^{\infty}\delta^n \int_{\Delta\Theta} u(\nu_{t+n})dI_{t+n}$.

General conditions for temptation. It is well established (see e.g. Grant et al., 1998; Lipnowski and Mathevet, 2018) that the agent's intrinsic preference for information is defined by the convexity of $u(\nu_t)$. That is, the agent is averse to information if and only if $u(\nu_t)$ is concave, and information-loving if and only if $u(\nu_t)$ is convex. The intuition for this result is that information induces a lottery over posterior beliefs, which the decision maker evaluates using her prior beliefs; more information implies a mean-preserving spread over the lottery of belief distributions.

The only novel feature we introduce here is imperfect recall. Let $Var(\nu_t) = \sigma_t^2$. Imperfect recall is captured by the assumption that for any posterior belief ν_t, $\sigma_{t+1}^2 > \sigma_t^2$. In addition, we impose the natural constraint that for any two distributions with σ_{1t}^2 and σ_{2t}^2 such that $\sigma_{1t}^2 < \sigma_{2t}^2$, $\sigma_{1z}^2 < \sigma_{2z}^2$ for all $z > t$: memory loss never causes a more precise initial belief to become more diffuse than a belief which began less precise. Finally, to simplify our results, we assume that the variance of the prior belief μ is fixed over time, which implies that the growing variances of the posterior beliefs are bounded above by the constant variance of the prior belief.

Definition 2: *If $\frac{\partial u}{\partial \sigma^2} < 0$, the decision maker is uncertainty averse.*

We also restrict attention to uncertainty-averse decision makers. This rules out decision-makers who, for instance, prefer less precision in beliefs with unfavourable expected values. In our setting, this restriction means we focus on those with both an initial motive to acquire information and a decay in the value of information over time. Temptation for information is possible for other classes of decision-makers but would require different restrictions on preferences.

We now consider an agent's evaluation of consuming *full information* — that is, information which fully reveals θ. The proposition below defines conditions for full information to be a simple temptation.

Proposition 2: *Under imperfect recall, full information is a simple temptation for an uncertainty-averse decision maker only if there exist thresholds $x, z : 0 < x \leq z$ such that u is locally convex for $\sigma_t^2 \in [0, x]$ and locally concave for $\sigma_t^2 > z$.*

Under the condition in Proposition 2, the initial prospect of information is attractive: in every state of the world, it implements a posterior belief with 0 variance. However, over time, posterior beliefs become more variable, meaning the evaluation of information in later periods eventually occurs in the concave region of utility. At this stage, the value of information has decayed sufficiently for the agent to prefer not to have acquired it; maintaining the prior belief is preferred when the alternative involves a lottery over a set of only slightly more informative posterior distributions.

A diverse range of preferences for information could be compatible with Proposition 2. However, the conditions are likely to be satisfied by some prominent models of preferences for information proposed in the literature. A notable example is if individuals incur a fixed cost when shifting from their prior belief, such as from confirmation bias (Rabin and Schrag, 1999). This condition on preferences also happens to be a stronger version of that implied by Rabin's calibration theorem for money, except of course in this case for preferences over belief distributions.[20]

To illustrate the intuition, we set out an example below in the context of our experiment. The model of preferences we outline embodies a trade-off between risk aversion over expectations of *Donor* status and a motive to satisfy curiosity.

Stylised example. In our experiment, individuals form beliefs about the event that they were a *Donor*, say $\theta \in \{0, 1\}$. Denote an agent's subjective belief that they were *not* a *Donor* in period t, $P(\theta = 0)$, as $p \in [0, 1]$. Since the prior distribution of θ is in general ambiguous from participants' perspectives, they have posterior beliefs $p \sim \nu_t$.

Suppose $E_{\nu_t}(p) = \omega$ and $Var_{\nu_t}(p) = \sigma_t^2$. We first impose that agents are averse to the bad states of the world; they are happier if ω is higher. As previously noted, it is well known that utility must be non-linear in ω for this element of preferences to provide a motive to consume or avoid new information. Thus, we assume utility is a function of ω^ρ for some $\rho > 0$. If $\rho < 1$, this element results in an aversion to information because it results in more dispersed posterior beliefs. If $\rho = 1$, so utility is linear in expectations over p, agents are indifferent between acquiring and avoiding information.

We then assume agents have a conflicting motive to consume information because

[20]Locally (weakly) convex and then concave utility functions have an early tradition in economics (Friedman and Savage, 1948) and have found some support in recent empirical evidence (Bombardini and Trebbi, 2012).

they dislike increases in $Var(p) = \sigma_t^2$.[21] This second component of utility can be seen as capturing discomfort with uncertainty, and the strength of this motive is captured by the coefficient κ. A simple additive representation of preferences under these two motives is given below:

$$u(\nu_t) = \omega^\rho - \kappa\sigma_t^2. \tag{3.6}$$

u_t satisfies the conditions provided in Proposition 2. To see why, note that the first component of preferences implies concavity over posterior beliefs: the agent is averse to dispersion of $E(p)$ across possible belief distributions. The second implies convexity: the agent prefers lower belief variance *at* each belief. The long-run prospect of information involves evaluating a set of posterior beliefs with progressively higher variance over time but a fixed dispersion of expectations over p. Thus, over time, the first motive increasingly dominates the second.

To formalise imperfect recall, we assume that the agent's prior on p (without acquiring information) has variance σ_0^2, and

$$\sigma_t^2 = \sigma_{t-1}^2 + m(\sigma_0^2 - \sigma_{t-1}^2), \tag{3.7}$$

with $0 < m < 1$, so that any posterior belief distribution with $\sigma_t^2 < \sigma_0^2$ is affected. Purely to simplify exposition, this formulation imposes that imperfect recall plays no role in the case where no information is consumed.

Denote the mean of the prior distribution for p as ω_0. In this case, the (normalised) lifetime utility resulting from avoiding information discount factor δ is

$$\omega_0^\rho - \kappa\sigma_0^2 \tag{3.8}$$

The expected lifetime utility from learning the value of $\theta \in \{0, 1\}$, is

$$(1-\delta)\left[\omega_0\left\{\sum_{t=0}^{\infty}\delta^t\left[1-\kappa(1-(1-m)^t)\sigma_0^2\right]\right\} + (1-\omega_0)\left\{\sum_{t=0}^{\infty}\delta^t\left[-\kappa(1-(1-m)^t)\sigma_0^2\right]\right\}\right]. \tag{3.9}$$

After some rearrangement, the expected marginal utility of consuming information is

$$\overline{U}_l = \omega_0 - \omega_0^\rho + \frac{(1-\delta)\kappa\sigma_0^2}{1-\delta(1-m)}. \tag{3.10}$$

We now bring this model into the dual-self framework by introducing a short-run

[21] We assume the agent is unconcerned with uncertainty in θ conditional on p.

self with a smaller discount factor, $\delta\mu$. In the example above, we can write the normalised marginal utility of information to the short-run self as

$$\overline{U}_s = \omega_0 - \omega_0^{\rho} + \frac{(1-\delta\mu)\kappa\sigma_0^2}{1-\delta\mu(1-m)}. \tag{3.11}$$

To see how temptation arises, first note that the marginal utility of information increases monotonically in impatience. If $m = 0$, there would be no conflict in evaluation between the short- and long-run selves. However, because information is increasingly difficult to recall over time, it necessarily provides more utility to a psychological agent in the short run. Thus, it is straightforward to construct an example in which $\overline{U}_s > 0$ and $\overline{U}_l < 0$. All that is required is for acquiring information to result in a cost that does not diminish as fast as the benefit diminishes.

In the representation provided by (3.6), for a small enough value of μ, $\rho < 1$ is sufficient for information to be tempting. If the individual is risk averse over the probability of being a *Donor*, she must trade off the prospect of bad news with her desire to minimise uncertainty. The short-run self is more willing to accept the risk of bad news because she is more preoccupied with the short-term benefit of reducing her uncertainty over p. A further implication of $\rho < 1$ is that the motive to access information diminishes in ω_0. The more optimistic the individual's prior belief on θ, the weaker the incentive to access information.

Treatment conditions. Our stylised example also produces a useful interpretation of the differences in menu preferences we observe across treatment conditions. Recall that in *FullAvoid*, opening the envelope occurs after learning nothing else about the chance of being a *Donor*, relative to $t = 1$. In *PartAvoid*, opening the envelope occurs after learning there was a 20% chance of being a *Donor*.

In our model, the provision of partial information has two, possibly opposing, effects on the demand for complete information. The first is that the uncertainty-reducing motive is weaker: more information has been provided, making $\{0\}$ more attractive relative to $\{0,1\}$ and $\{1\}$. However, the second is that the expectation ω implemented under $\{0\}$ is different in *PartAvoid* than in *FullAvoid*. In *FullAvoid*, $\{0\}$ involves maintaining the prior belief with an expectation ω_0. In *PartAvoid*, it involves implementing a posterior distribution with $\eta = 0.8$. If $\omega_0 > 0.8$ and $\rho < 1$, an individual's desire to believe that they were not a *Donor* raises the possibility that their preference for $\{1\}$ relative to $\{0\}$ is stronger under *PartAvoid*. In the

opposite direction, more information is provided under $\{0\}$ in *PartAvoid*, meaning the uncertainty-reducing motive provided by the $-\kappa\sigma_t^2$ term in utility is weaker.

In our experimental data, $\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$ is less common in *PartAvoid* than in *FullAvoid*, while $\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$ is more common. Through the lens of our model, our interpretation is that individuals know they will be unable to retain optimistic prior beliefs about θ in a case where $\eta = 0.8$. This makes it more attractive to gamble on learning the value of θ in *PartAvoid*.[22]

Notably, our model can also accommodate our tentative finding that information is weakly less attractive in $t = 2$ in *PartAvoid* than *FullAvoid*, despite the reverse having been true in $t = 1$. Concretely, the arguments stated above demonstrate how partial information could decrease $\overline{U}_s$ as well as $|\overline{U}_l|$: $\overline{U}_s$ is more sensitive than $|\overline{U}_l|$ to the effect of additional information on σ_t^2. $\overline{U}_s$ is more important for information preferences in $t = 2$ than in $t = 1$, since the preferences of the short-run self become more important in $t = 2$. Thus, our treatment could have an inverted effect on preferences in $t = 1$ and $t = 2$ respectively, as tentatively suggested by our data.

3.6 Discussion

Our experimental results provide novel evidence that undesired information is often intrinsically tempting. In doing so, they have possible implications for three separate areas of research on the economics of information: intrinsic preferences for information; strategic information avoidance; and consumer markets for information. We now discuss these implications in more detail.

3.6.1 Intrinsic preferences for information

Our analysis of self-control costs implies that intrinsic (psychological) preferences for information are dynamically inconsistent for many individuals. A small share of individuals directly demonstrate dynamic inconsistency by paying to access information after having strictly preferred not to be offered it. However, preferences also appear dynamically inconsistent at the intensive margin: using response times, we find evidence that for Tempted types, session 2 preferences for information are negatively correlated with session 1 preferences for information at the individual

[22]Strictly speaking, our model would require some uncertainty in $\succeq_1$ to explain why the observed shift is from $\{0\} \succ_1 \{1\}$ to $\{0\} \sim_1 \{1\}$.

level. Qualitative measures also indicate that participants classified as tempted by information are more interested in seeing it in session 2, when it is available, than in session 1.

Most existing theories of intrinsic preferences for information would require extension to accommodate dynamic inconsistency. In Section 3.5, we proposed a natural explanation for temptation and dynamic inconsistency in a parsimonious model of 'psychological' preferences for information: since recall tends to deteriorate over time, the psychological benefit of information is stronger in the short term than the long term. Further research could test whether this exact mechanism operates in practice.

Our results also have some implications for empirical measurements of preferences for information: individuals may have different preferences when asked in advance than instantaneously. The fact that some preference reversals were observed in our experiment is fairly striking given individuals had to pay to access information and made their choices in a highly controlled environment. It seems plausible that preference reversals could be more common in other contexts, especially when information can be freely accessed. Given the literature's focus on temporal preferences for information, this seems a pertinent issue.

Existing empirical work measures participants' *WTP* for a wide range of information structures. In Masatlioglu et al. (2023), participants are willing to pay on average \$0.08 to receive unavoidable, non-instrumental information (on the outcome of a lottery whose expected value is \$5) 30 minutes earlier. *WTP* for receiving positively skewed information early rather than negatively skewed information early is about three or four times higher. In Ganguly and Tasoff (2017), 83.4% of participants are willing to pay \$0.50 to affect the timing of information receipt in a similar context.

It is notable that individuals appear to have a comparable *WTP* for committing not to see information (in our experiment) and seeing it (in others), especially considering that commitments are implemented only probabilistically in our setting. Both behaviours are consistent with a model in which individuals are tempted by information, but they result in contradictory conclusions about preferences. Indeed, *WTP* positively predicts self-control costs at the individual level when commitments are not implemented: those who find it more difficult to resist information in session 2 are willing to pay *more* money to commit to avoiding it in session 1. In other settings, it cannot be ruled out that those who are especially impatient to instantaneously acquire information when it is unavoidable are actually those with the strongest

preference to avoid it ahead of time.

3.6.2 Strategic information avoidance

While the classical view is that the provision of instrumentally useful information is welfare-improving, an innovation of recent research has been to demonstrate that information avoidance can be welfare-improving in the presence of certain behavioural biases (Lipnowski and Mathevet, 2018). For example, avoiding unfavourable information about one's ability may be a crucial tool to overcome motivation problems resulting from present bias (Bénabou and Tirole, 2002). Avoidance of ego-relevant information has been documented in a large and expanding set of empirical studies (Eil and Rao, 2011; Oster et al., 2013; Zimmermann, 2020; Huffman et al., 2022), many of which establish a link to overconfident beliefs. Roy-Chowdhury (2022) finds that selective recall of attainment by adolescents is associated with higher self-esteem and short-run attainment outcomes. Saccardo and Serra-Garcia (2023) demonstrate that individuals manipulate information demand to strategically influence their own decision-making. Thus, information avoidance is well-documented empirically, and could be welfare-improving, in contexts where information has more than just psychological value.

Our results highlight the possibility that in circumstances where an individual would like to avoid information for strategic reasons, being offered instant access to it makes it much harder to resist. This raises questions about the theoretical plausibility of such strategies and suggests that they may be more feasible when individuals are offered devices ahead of time to avoid information.

Our results also confirm the conjecture of many authors that avoiding information involves *intrinsic* psychological costs. This has implications for the theoretical analysis of information avoidance, including in settings where information has instrumental value. For example, the core models in Bénabou and Tirole (2002) and Brunnermeier and Parker (2005) treat information avoidance as intrinsically costless: the only consequence of avoiding information is on decision making. Our results suggest that a different approach may be warranted: avoiding information seems to be intrinsically costly. A further insight from the dual-self model is that self-control costs become exponentially less important if commitment actions can be taken in advance of the temptation. This model and our results suggest that strategic information avoidance may be more feasible if informational commitment devices

are offered.

3.6.3 Markets for information

Intrinsic preferences for information are of major policy relevance, especially in view of the dramatic growth of information markets in recent decades (Hilbert and López, 2011). Demonstrating that information is tempting raises a stronger case for managing the manufacture and creation of unwanted information. In our experiment, many individuals strictly prefer not to be offered the information we generate, and are willing to pay both effort costs and money for that preference. In our setting, for many individuals, one interpretation is that the optimal outcome would have been for the information not to be created at all. Alternatively, as implied by both the dual-self model of self-control and our results, it may be better to offer individuals devices to avoid potentially undesirable information ahead of it becoming available.

We highlight applications to two markets that have been of particular interest to economists. The first is the market for news. A large literature explores the demand for news, with a particular focus on its role in driving political polarisation (Mullainathan and Shleifer, 2005; DellaVigna and Kaplan, 2007). While little formal work has been done to understand whether news is tempting, recent work alludes to the possibility: Robertson et al. (2023) find that negatively framed news headlines attract greater interest than positively framed ones with the same content. Temptation for information could be an important consideration in markets for news, given that their primary purpose is to convey information. If so, the availability of some kinds of news may have negative welfare implications.

The welfare costs of social media are somewhat better understood and have been robustly linked to limited self-control (Allcott et al., 2020; Mosquera et al., 2020; Allcott et al., 2022; Braghieri et al., 2022). Our findings suggest a deeper behavioural foundation for this phenomenon: one major function of social media is to supply information that could very well be tempting: undesired but hard to resist. Our treatment conditions also shed light on the potential role of notifications, now a fundamental feature of information supply on digital platforms. Our results suggest that the supply of partial information vitiates *ex-ante* preferences to avoid undesirable information. This suggests the widespread use of notifications may be an effective strategy in eroding individuals' willingness to commit to reducing their use of digital media, and that this could negatively impact their welfare. Further research could

be informative on this and other related policy matters.

3.7 Conclusion

We conduct the first empirical investigation of temptation and dynamic inconsistency in preferences for information. We find evidence of widespread temptation for undesired information: around half of all individuals, and the majority of those who would like to avoid information, strictly prefer not to be given the option to access information in a day's time. Some individuals are offered information against their wishes on day 2, and we find that only a small share pay to consume it. However, we show that dynamic inconsistency in preferences is suggested by an analysis of the relationship between menu preferences on day 1 and response times on day 2. The relationship between menu preferences and response times is as predicted by a model of temptation under sophistication, suggesting that preferences for information are dynamically inconsistent. Using treatment conditions, we find weak evidence that information is harder to resist ex-ante when an unavoidable, unfavourable signal is anticipated. Finally, we show how temptation for information can arise in a dual-self model of psychological preferences for information.

References

Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, *134*(3), 1121–1162.

Alan, S., & Ertac, S. (2015). Patience, self-control and the demand for commitment: Evidence from a large-scale field experiment. *Journal of Economic Behavior & Organization*, *115*, 111–122.

Alexander, K., & Entwisle, D. (2003). *The beginning school study, 1982-2002*. https://doi.org/10.7910/DVN/NYYXIO

Alexander, K., Entwisle, D., & Olson, L. (2014). *The long shadow: Family background, disadvantaged urban youth, and the transition to adulthood*. Russell Sage Foundation. http://www.jstor.org/stable/10.7758/9781610448239

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–76.

Allcott, H., Gentzkow, M., & Song, L. (2022). Digital addiction. *American Economic Review*, *112*(7), 2424–63.

Alós-Ferrer, C., Fehr, E., & Netzer, N. (2021). Time will tell: Recovering preferences when choices are noisy. *Journal of Political Economy*, *129*(6), 1828–1877.

Andreoni, J., & Serra-Garcia, M. (2021). Time inconsistent charitable giving. *Journal of Public Economics*, *198*, 104391.

Andreoni, J., & Vesterlund, L. (2001). Which is the fair sex? gender differences in altruism. *The Quarterly Journal of Economics*, *116*(1), 293–312.

Arcidiacono, P. (2005). Affirmative action in higher education: How do admission and financial aid rules affect future earnings? *Econometrica*, *73*(5), 1477–1524.

Augenblick, N., Niederle, M., & Sprenger, C. (2015). Working over time: Dynamic inconsistency in real effort tasks. *The Quarterly Journal of Economics*, *130*(3), 1067–1115.

Bénabou, R. (2015). The economics of motivated beliefs. *Revue d'économie politique*, *125*(5), 665–685.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, *117*(3), 871–915.

Benoît, J.-P., & Dubra, J. (2011). Apparent overconfidence. *Econometrica*, *79*(5), 1591–1625.

Bombardini, M., & Trebbi, F. (2012). Risk aversion and expected utility theory: An experiment with large and small stakes. *Journal of the European Economic Association*, *10*(6), 1348–1399.

Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., & Shleifer, A. (2023). Memory and probability. *The Quarterly Journal of Economics*, *138*(1), 265–311.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2020). Memory, attention, and choice. *The Quarterly Journal of Economics*, *135*(3), 1399–1442.

Braghieri, L., Levy, R., & Makarin, A. (2022). Social media and mental health. *American Economic Review*, *112*(11), 3660–3693.

Brañas-Garza, P., Capraro, V., & Rascon-Ramirez, E. (2018). Gender differences in altruism on Mechanical Turk: Expectations and actual behaviour. *Economics Letters*, *170*, 19–23.

Brunnermeier, M. K., & Parker, J. A. (2005). Optimal expectations. *American Economic Review*, *95*(4), 1092–1118.

Caplin, A., & Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *The Quarterly Journal of Economics*, *116*(1), 55–79.

Caplin, A., & Leahy, J. (2004). The supply of information by a concerned expert. *The Economic Journal*, *114*(497), 487–505.

Compte, O., & Postlewaite, A. (2004). Confidence-enhanced performance. *American Economic Review*, *94*(5), 1536–1557.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47.

Cunha, F., Heckman, J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*(3), 883–931.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, *100*(2), 193–201.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*, 67–80.

Danz, D., Vesterlund, L., & Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, *112*(9), 2851–2883.

De Bondt, W. F., & Thaler, R. H. (1995). Financial decision-making in markets and firms: A behavioral perspective. *Handbooks in Operations Research and Management Science*, *9*, 385–410.

Dekel, E., & Lipman, B. L. (2012). Costly self-control and random self-indulgence. *Econometrica*, *80*(3), 1271–1302.

DellaVigna, S., & Kaplan, E. (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, *122*(3), 1187–1234.

Di Tella, R., Perez-Truglia, R., Babino, A., & Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, *105*(11), 3416–3442.

Dizon-Ross, R. (2019). Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review*, *109*(8), 2728–65.

Dweck, C. S. (2002). The development of ability conceptions. In *Development of achievement motivation* (pp. 57–88). Elsevier.

Eckel, C. C., & Grossman, P. J. (1998). Are women less selfish than men?: Evidence from dictator experiments. *The Economic Journal*, *108*(448), 726–735.

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–38.

Eliaz, K., & Schotter, A. (2007). Experimental testing of intrinsic preferences for noninstrumental information. *American Economic Review*, *97*(2), 166–169.
Ely, J., Frankel, A., & Kamenica, E. (2015). Suspense and surprise. *Journal of Political Economy*, *123*(1), 215–260.
Enke, B., & Graeber, T. (2023). Cognitive uncertainty. *The Quarterly Journal of Economics*, *138*(4), 2021–2067.
Epstein, L. G. (2008). Living with risk. *The Review of Economic Studies*, *75*(4), 1121–1141.
Epstein, L. G., & Zin, S. E. (1989). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica*, *57*(4), 937–969.
Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, *83*(2), 587–628.
Exley, C. L., & Kessler, J. B. (2024). Motivated errors. *American Economic Review*, *114*(4), 961–987.
Falk, A., & Zimmermann, F. (2023). Attention and dread: Experimental evidence on preferences for information. *Management Science*.
Fišar, M., Kubák, M., Špalek, J., & Tremewan, J. (2016). Gender differences in beliefs and actions in a framed corruption experiment. *Journal of Behavioral and Experimental Economics*, *63*, 69–82.
Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, *56*(4), 279–304.
Fudenberg, D., & Levine, D. K. (2006). A dual-self model of impulse control. *American Economic Review*, *96*(5), 1449–1476.
Fudenberg, D., & Levine, D. K. (2012). Timing and self-control. *Econometrica*, *80*(1), 1–42.
Fudenberg, D., Strack, P., & Strzalecki, T. (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review*, *108*(12), 3651–3684.
Gabaix, X., Laibson, D., Moloche, G., & Weinberg, S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, *96*(4), 1043–1068.
Ganguly, A., & Tasoff, J. (2017). Fantasy and dread: The demand for information and the consumption utility of the future. *Management Science*, *63*(12), 4037–4060.
Gill, D., & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, *102*(1), 469–503.
Gödker, K., Jiao, P., & Smeets, P. (2021). Investor memory. *Available at SSRN 3348315*.
Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, *55*(1), 96–135.
Grant, S., Kajii, A., & Polak, B. (1998). Intrinsic preference for information. *Journal of Economic Theory*, *83*(2), 233–259.
Gul, F., & Pesendorfer, W. (2001). Temptation and self-control. *Econometrica*, *69*(6), 1403–1435.
Heckman, J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 271–320.
Heckman, J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*(3), 411–482.
Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, *332*(6025), 60–65.

Huffman, D., Raymond, C., & Shvets, J. (2022). Persistent overconfidence and biased memory: Evidence from managers. *American Economic Review*, *112*(10), 3141–3175.
Kessler, J. B., Kivimaki, H., Litwin, A., & Niederle, M. (2023). Please take a minute: How prosocial preferences change with deliberation.
Koszegi, B., Loewenstein, G., & Murooka, T. (2022). Fragile self-esteem. *The Review of Economic Studies*, *89*(4), 2026–2060.
Koszegi, B., & Rabin, M. (2009). Reference-dependent consumption plans. *American Economic Review*, *99*(3), 909–36.
Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, *4*(4), 673–707.
Krajbich, I., Oud, B., & Fehr, E. (2014). Benefits of neuroeconomic modeling: New policy interventions and predictors of preference. *American Economic Review*, *104*(5), 501–506.
Kreps, D. M., & Porteus, E. L. (1978). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*, 185–200.
Lazear, E. P., Malmendier, U., & Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics*, *4*(1), 136–163.
Lipnowski, E., & Mathevet, L. (2018). Disclosure to a psychological audience. *American Economic Journal: Microeconomics*, *10*(4), 67–93.
Liu, S., & Netzer, N. (2023). Happy times: Measuring happiness using response times.
Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press, USA.
Machado, J. A. F., & Silva, J. S. (2005). Quantiles for counts. *Journal of the American Statistical Association*, *100*(472), 1226–1237.
Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment. *The Journal of Finance*, *60*(6), 2661–2700.
Masatlioglu, Y., Orhun, Y., & Raymond, C. (2023). Intrinsic information preferences and skewness. *American Economic Review*, *113*(10), 2615–2644.
Mosquera, R., Odunowo, M., McNamara, T., Guo, X., & Petrie, R. (2020). The economic effects of facebook. *Experimental Economics*, *23*, 575–602.
Mullainathan, S., & Shleifer, A. (2005). The market for news. *American Economic Review*, *95*(4), 1031–1053.
Nielsen, K. (2020). Preferences for the resolution of uncertainty and the timing of information. *Journal of Economic Theory*, *189*, 105090.
Ortoleva, P., & Snowberg, E. (2015). Overconfidence in political behavior. *American Economic Review*, *105*(2), 504–535.
Oster, E., Shoulson, I., & Dorsey, E. (2013). Optimal expectations and limited medical testing: Evidence from Huntington disease. *American Economic Review*, *103*(2), 804–30.
Rabin, M., & Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, *114*(1), 37–82.
Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature Human Behaviour*, *7*(5), 812–822.

Roy-Chowdhury, V. (2022). Self-confidence and motivated memory loss: Evidence from schools.

Roy-Chowdhury, V. (2024). Temptation to consume information.

Royer, H., Stehr, M., & Sydnor, J. (2015). Incentives, commitments, and habit formation in exercise: Evidence from a field experiment with workers at a fortune-500 company. *American Economic Journal: Applied Economics*, *7*(3), 51–84.

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 999–1033.

Saccardo, S., & Serra-Garcia, M. (2023). Enabling or limiting cognitive flexibility? evidence of demand for moral commitment. *American Economic Review*, *113*(2), 396–429.

Sadoff, S., Samek, A., & Sprenger, C. (2020). Dynamic inconsistency in food choice: Experimental evidence from two food deserts. *The Review of Economic Studies*, *87*(4), 1954–1988.

Toussaert, S. (2018). Eliciting temptation and self-control through menu choices: A lab experiment. *Econometrica*, *86*(3), 859–889.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., et al. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review*, *110*(2), 337–61.

Appendix A

Appendices to Chapter 1

A.1 Additional reduced form analyses

Table A.1 Reduced form models — biased recall

	Dependent variable:	
	Flattering recall error	
	(1)	(2)
Actual grade: *Satisfactory*	0.257^{***} (0.025)	0.193^{***} (0.020)
Actual grade: *Unsatisfactory*	0.543^{***} (0.032)	0.434^{***} (0.024)
Fall	0.234^{***} (0.020)	0.240^{***} (0.020)
Individual FE	Yes	No
Academic year FE	Yes	Yes
Observations	2,448	2,448
R^2	0.431	0.158
Adjusted R^2	0.258	0.156

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Linear probability model, robust standard errors in parentheses. All grades are from quarterly report cards for mathematics, and recall of the most recent quarterly grade is elicited in a biannual survey. Observations with an actual grade of *Excellent* are omitted since flattering recall errors are impossible; the omitted actual grade category in this case is *Good*.

Table A.2 Reduced form models — recall errors, heterogeneity by sex and race

	Dependent variable:	
	Incorrect recall	
	(1)	(2)
Actual grade: *Good*	0.071 (0.081)	0.113* (0.062)
Actual grade: *Satisfactory*	0.232*** (0.086)	0.245*** (0.059)
Actual grade: *Unsatisfactory*	0.403*** (0.095)	0.391*** (0.063)
Female	—	0.025 (0.051)
Black	—	−0.037 (0.051)
Fall	0.237*** (0.020)	0.238*** (0.020)
Actual grade: *Good**Female	−0.093 (0.080)	−0.066 (0.061)
Actual grade: *Satisfactory**Female	−0.008 (0.082)	0.015 (0.058)
Actual grade: *Unsatisfactory**Female	−0.067 (0.093)	−0.045 (0.064)
Actual grade: *Good**Black	0.153* (0.083)	0.085 (0.062)
Actual grade: *Satisfactory**Black	0.087 (0.088)	0.011 (0.060)
Actual grade: *Unsatisfactory**Black	0.227** (0.098)	0.144** (0.066)
Individual FE	Yes	No
Academic year FE	Yes	Yes
Observations	2,694	2,694
R^2	0.375	0.134
Adjusted R^2	0.199	0.128

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Linear probability model, robust standard errors in parentheses. All grades are from quarterly report cards for mathematics, and recall of the most recent quarterly grade is elicited in a biannual survey. The omitted actual grade category is *Excellent*.

Table A.3 Reduced form model — memory loss, heterogeneity by sex and race

	Dependent variable:	
	Recalled grade: *Excellent*	
	(1)	(2)
Actual grade: *Good*	0.074 (0.082)	0.115* (0.062)
Actual grade: *Satisfactory*	0.235*** (0.086)	0.247*** (0.059)
Actual grade: *Unsatisfactory*	0.408*** (0.095)	0.393*** (0.063)
Female	—	0.017 (0.054)
Black	—	−0.021 (0.056)
Fall	0.252*** (0.039)	0.248*** (0.038)
Actual grade: *Good**Female	−0.094 (0.081)	−0.064 (0.061)
Actual grade: *Satisfactory**Female	−0.008 (0.083)	0.017 (0.058)
Actual grade: *Unsatisfactory**Female	−0.069 (0.092)	−0.043 (0.063)
Actual grade: *Good**Black	0.150* (0.084)	0.081 (0.063)
Actual grade: *Satisfactory**Black	0.084 (0.088)	0.007 (0.060)
Actual grade: *Unsatisfactory**Black	0.223** (0.099)	0.140** (0.065)
Female*Fall	−0.010 (0.035)	0.011 (0.035)
Black*Fall	−0.013 (0.039)	−0.022 (0.038)
Individual FE	Yes	No
Academic year FE	Yes	Yes
Observations	2,694	2,694
R^2	0.375	0.134
Adjusted R^2	0.198	0.128

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Linear probability model, robust standard errors in parentheses. All grades are from quarterly report cards for mathematics, and recall of the most recent quarterly grade is elicited in a biannual survey. The omitted actual grade category is *Excellent*.

Table A.4 Reduced form models — grade prediction errors

	Dependent variable:		
	Incorrect prediction		
	(1)	(2)	(3)
Recall error	0.062*** (0.021)	0.042** (0.022)	0.089*** (0.019)
Actual grade: *Good*	−0.008 (0.044)	0.010 (0.044)	0.085** (0.035)
Actual grade: *Satisfactory*	0.016 (0.045)	0.037 (0.045)	0.164*** (0.033)
Actual grade: *Unsatisfactory*	0.015 (0.050)	0.043 (0.050)	0.201*** (0.036)
Fall		0.081*** (0.022)	0.078*** (0.022)
Individual FE	Yes	Yes	No
Academic year FE	Yes	Yes	Yes
Observations	2,883	2,883	2,883
R^2	0.306	0.310	0.045
Adjusted R^2	0.110	0.115	0.042

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Linear probability model, robust standard errors in parentheses. All grades are from quarterly report cards for mathematics, and recalled and expected quarterly grades are elicited in a biannual survey. The omitted actual grade category is *Excellent*.

A.2 Structural model: Extensions and robustness checks

Here, we briefly outline an extended structural model with an additional state variable capturing school-centric welfare. The purpose of this model is to investigate whether different types of student place different weights on academic self-esteem in utility. As mentioned in the main body, the measure of welfare we use is based on the question 'How much do you like school in general', and we code responses $z_t = 1$ if the student's response was 'A lot' and $z_t = 0$ if they responded with 'A little' or 'Not at all'. We then specify the same model as in Section 1.4.2, but with z_t as an additional state variable which appears nowhere else. As with the other state variables, z_t follows a conditional Markov process,

$$P(z_t = 1|z_{t-1}, g_t, y_t, c = 1) = \text{logit}(\mu_1^z + \mu_z^z z_{t-1} + \mu_g^z g_t + \mu_y^z y_t), \tag{A.1}$$

for $c = 1$, and

$$P(z_t = 1|z_{t-1}, g_{t-1}, y_{t-1}, c = 2) = P(z_t = 1|z_{t-1}, g_t, y_t, c = 1) + \omega^z + \omega_y^z(y_t = 1), \tag{A.2}$$

for $c = 2$. Crucial is the coefficient ω_y^z, which lets the marginal effect of self-esteem vary by type. The coefficient estimates are reported in Table A.5. The coefficients are qualitatively the same as in the baseline 2-type model, although the fragile self-esteem parameter θ_y^1 is no longer statistically significant because its magnitude is slightly reduced and standard errors are inflated. The main coefficient of interest in this case is ω_y^z, capturing heterogeneity in the impact of self-esteem on welfare. Self-esteem has a substantially smaller impact on the welfare of type 2: the marginal effect of $y_t = 1$ is 9pp smaller for type 2 than for type 1. This evidence is consistent with the interpretation that the much larger θ^c for type 1 reflects a larger relevance of self-esteem for welfare.

Table A.5 Parameter estimates for structural model with unobserved heterogeneity and auxiliary state variable for welfare

Utility function	
θ^1	2.44^{***} (0.31)
θ^2	0.18^{***} (0.03)
θ^1_y	0.39 (0.36)
θ^2_y	0.17^{***} (0.06)
λ	2.35^{***} (0.03)
λ_f	-1.24^{***} (0.03)
Markov process: actual grades	
μ^g_1	-0.82^{***} (0.18)
μ^g_y	0.25^{***} (0.02)
μ^g_g	1.54^{***} (0.03)
μ^g_f	0.15^{***} (0.01)
μ^g_r	0.23 (0.17)
ω^g	-0.24^{***} (0.01)
ω^g_r	0.06^{***} (0.01)
Markov process: self-esteem	
μ^y_1	-1.61^{***} (0.15)
μ^y_y	1.84^{***} (0.03)
μ^y_g	-0.00 (0.03)
μ^y_f	0.08^{***} (0.02)
μ^y_r	0.49^{***} (0.16)
ω^y	-0.03^{***} (0.00)
ω^y_r	0.04^{***} (0.01)
Markov process: welfare	
μ^z_1	-1.84^{***} (0.09)
μ^z_z	1.69^{***} (0.03)
μ^z_g	-0.18^{***} (0.03)
μ^z_f	0.07^{***} (0.02)
μ^z_y	0.84^{***} (0.14)
ω^z	-0.01^{***} (0.00)
ω^z_y	-0.09^{***} (0.01)
Auxiliary model for $c = 2$	
γ_1	0.52^{***} (0.10)
γ_2	1.25^{***} (0.41)
γ_g	0.83^{**} (0.40)
Observations	1,158

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. Standard errors in parentheses. Model estimated by full-information maximum likelihood. Discount factor set to $\beta = 0.9$.

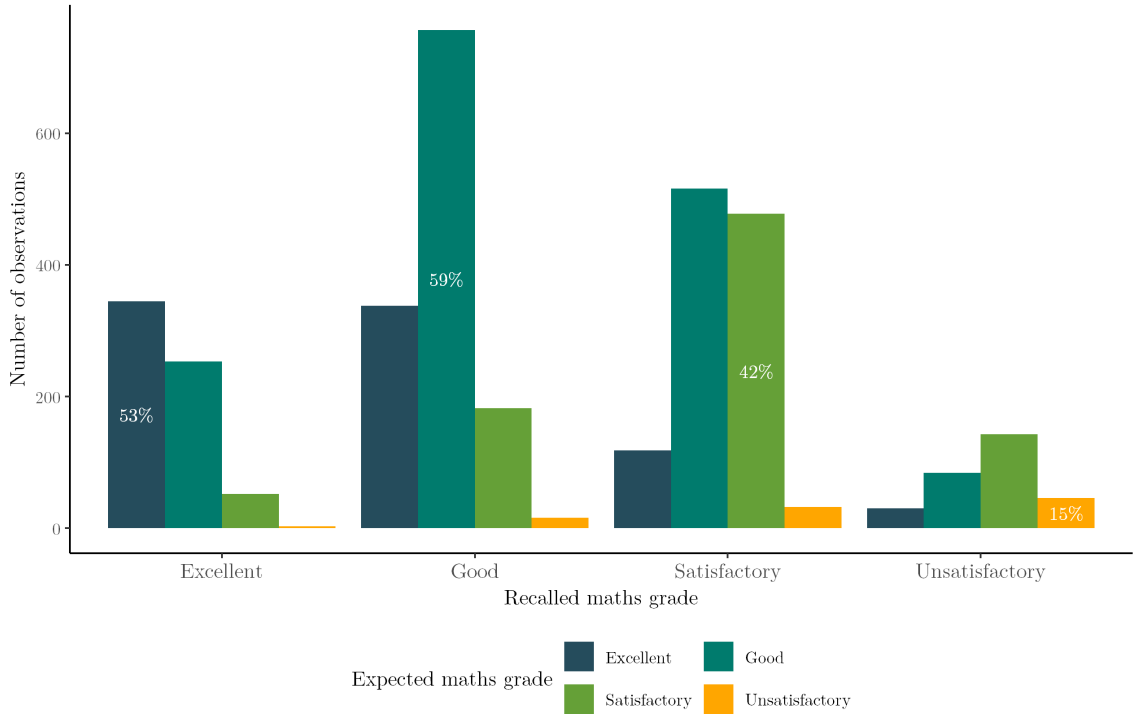
Table A.6 Structural model: variations on β

	β				
	0.1	0.25	0.5	0.75	0.9
Utility function					
θ	0.79^{***}	0.77^{***}	0.73^{***}	0.68^{***}	0.65^{***}
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
θ_y	0.27^{***}	0.26^{***}	0.25^{***}	0.24^{***}	0.23^{***}
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
λ	-2.26^{***}	-2.25^{***}	-2.24^{***}	-2.23^{***}	-2.23^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
λ_f	1.00^{***}	1.00^{***}	1.00^{***}	1.00^{***}	1.00^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Markov process: actual grades					
μ_1^g	-1.78^{***}	-1.78^{***}	-1.78^{***}	-1.78^{***}	-1.78^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
μ_y^g	1.86^{***}	1.86^{***}	1.86^{***}	1.86^{***}	1.86^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
μ_g^g	0.00	0.00	0.01	0.01	0.01
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
μ_f^g	0.07^{***}	0.07^{***}	0.07^{***}	0.07^{***}	0.07^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
μ_r^g	0.67^{***}	0.67^{***}	0.67^{***}	0.67^{***}	0.67^{***}
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Markov process: self-esteem					
μ_1^y	-1.96^{***}	-1.96^{***}	-1.96^{***}	-1.96^{***}	-1.96^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
μ_y^y	0.42^{***}	0.42^{***}	0.42^{***}	0.42^{***}	0.42^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
μ_g^y	1.62^{***}	1.62^{***}	1.62^{***}	1.62^{***}	1.62^{***}
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
μ_f^y	-0.08^{***}	-0.08^{***}	-0.08^{***}	-0.08^{***}	-0.08^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
μ_r^y	0.89^{***}	0.89^{***}	0.89^{***}	0.89^{***}	0.90^{***}
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Observations			1,176		

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. Standard errors in parentheses. Model estimated by full-information maximum likelihood.

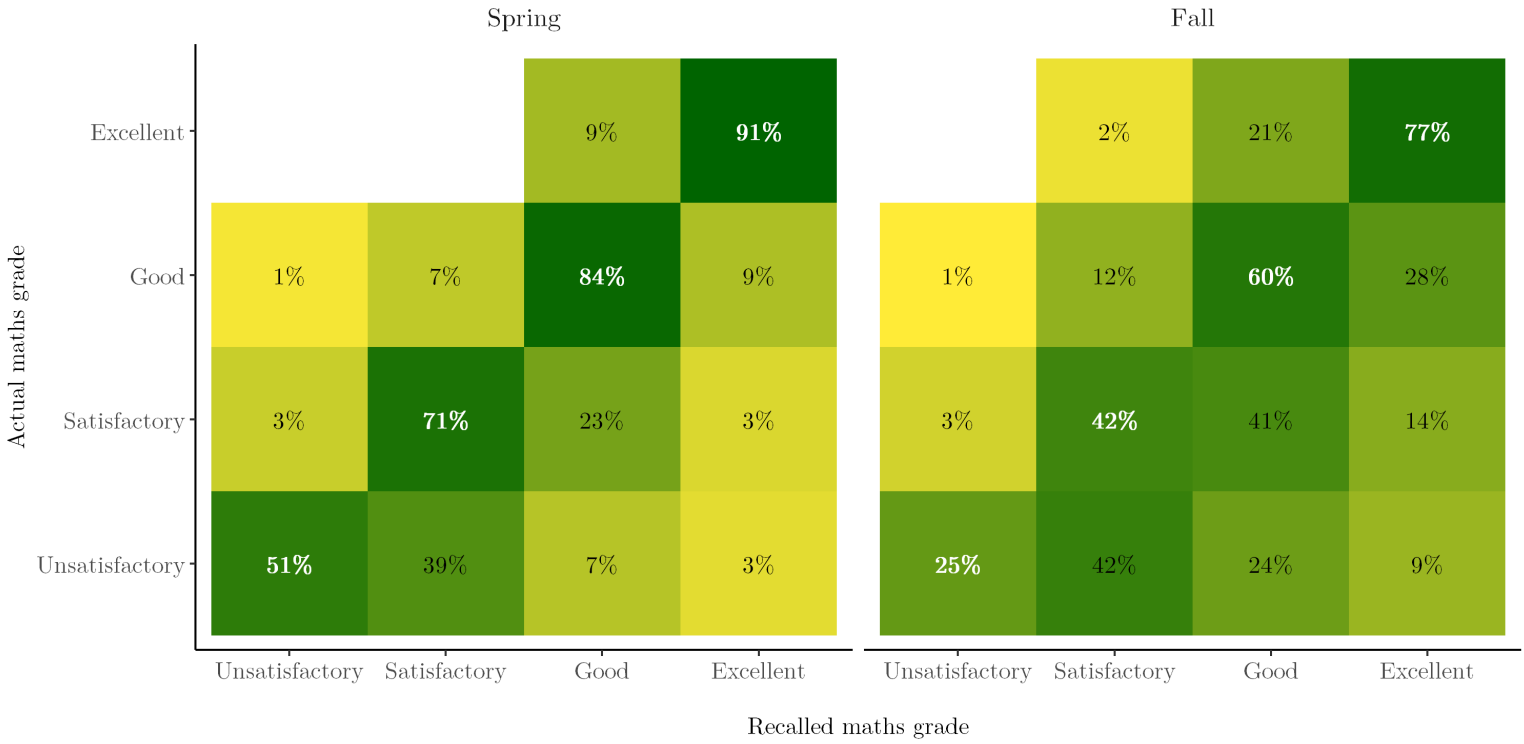
A.3 Additional figures

Fig. A.1 Expected grades vs recalled



Note: Aggregated over survey sweeps. White percentages are the share of observations with expected grades equal to recalled grades, by recalled grade.

Fig. A.2 Actual and remembered grades, split by survey edition



Note: Aggregated over survey sweeps. Percentages indicate shares of recalled mathematics grades by actual grade (row-wise).

Appendix B

Appendices to Chapter 2

B.1 Additional figures and tables

Fig. B.1 Average values of $\tilde{x}_1$ and $\tilde{x}_2$ by x

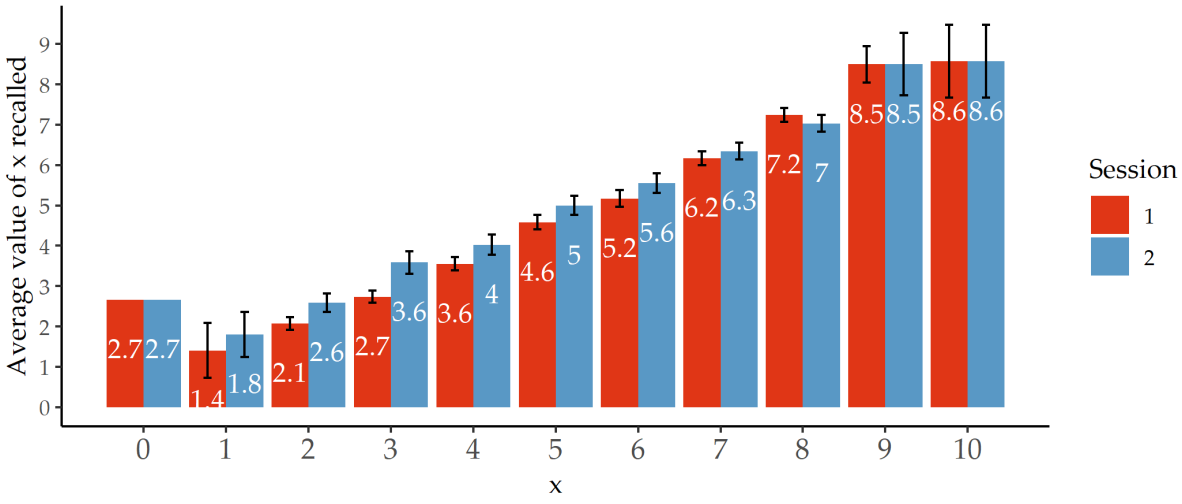


Table B.1 Treatment effects across conditional quantiles for each variable

Percentile	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_2 - \tilde{x}_1$
10^{th}	0.247** (0.108)	0.304** (0.122)	0.056 (0.092)
30^{th}	0.181** (0.08)	0.196** (0.086)	0.023 (0.057)
50^{th}	0.115** (0.073)	0.232** (0.091)	0.021 (0.06)
70^{th}	0.112** (0.065)	0.275** (0.102)	0.05 (0.086)
90^{th}	0.239** (0.114)	0.397** (0.172)	0.349 (0.225)
N	1061	1015	1015

Note: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. Results are the average of 2000 simulations jittering x_1 and x_2 with uniformly distributed noise on $[0, 1)$. Standard errors in parentheses. x is the hidden signal, the number of times the subject had to drag sliders to 150 in session 1. $\tilde{x}_1$ is recall of that number in session 1, while $\tilde{x}_2$ is recall of that number in session 2, 5 days later. Control variables for x and incentives for beliefs are included in the regressions but omitted here for brevity. We do not control for $\tilde{x}_1$.

Table B.2 ATEs by sex

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_2 - \tilde{x}_1$
Female	0.29*** (0.098)	0.393*** (0.129)	0.086 (0.125)
Male	0.095 (0.09)	0.253** (0.129)	0.16 (0.116)
Difference	0.195 (0.133)	0.14 (0.182)	-0.074 (0.17)
N	1055	1014	1014

Table B.3 ATTs by sex

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_2 - \tilde{x}_1$
Female	0.349*** (0.118)	0.471*** (0.155)	0.103 (0.149)
Male	0.122 (0.114)	0.323** (0.164)	0.204 (0.148)
Difference	0.228 (0.164)	0.148 (0.226)	-0.102 (0.211)
N	1055	1014	1014

Table B.4 ATEs by self-reported altruism score

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_2 - \tilde{x}_1$
Avg or below altruism	0.112 (0.087)	0.176 (0.125)	0.064 (0.118)
Above avg altruism	0.288*** (0.11)	0.512*** (0.138)	0.223** (0.126)
Difference	-0.176 (0.139)	-0.336** (0.186)	-0.159 (0.173)
N	1015	1015	1015

B.2 Experiment pages

Fig. B.2 Initial instructions

This study is split into two sessions. It is part of academic research conducted at the University of Cambridge. This first session should take under 5 minutes and involves completing a short task. Session 2 will take around 4-8 minutes and will take place on **Tuesday 14th May**. **Please only start this session if you are sure you can attend session 2.**

You will earn at least £0.75 for session 1. You will earn at least £3.35 for session 2. **Bonus payments are possible** in both cases.

Please do not attempt to restart the survey. If you do, we will ask you to return your submission.

Fig. B.3 Initial instructions, page 2

Please note that we will ask you to return your submission if you do not attend session 2.

You do not have to show up at a particular time for session 2, but **you must complete it within 24 hours of being invited on Tuesday 14th May.**

Your participation in this study is voluntary. You have the right to withdraw at any point during the survey. The survey is anonymous and the responses will be used for research purposes only. If you wish to participate in the study, please continue to the next screen.

Fig. B.4 Initial instructions, page 3

Thank you very much for agreeing to participate in this study. Please do not navigate away from this page in your browser at any point, as you will be asked to return your submission if you do.

It is vital for our study that you devote your full attention to all of the questions. Please read the instructions carefully throughout.

All of the information you will see throughout is true and many of your choices will affect the final payment made to you and other participants.

Fig. B.5 Initial instructions, page 4

Please indicate your degree of agreement with the following statement: 3 plus 7 is equal to 12.

- Strongly disagree ○
- Somewhat disagree ○
- Somewhat agree ○
- Strongly agree ○

Fig. B.6 Recall instructions, standard

On the last page of sliders you just completed, you had to drag a certain number to **150**.

This number was **randomly generated**: there's a 98% chance you were given **2-8 sliders to drag to 150**, with an equal chance of each possibility in this range. It is very unlikely that you were given 0-1 or 9-10 sliders of this kind: there's a 0.5% chance of each of these numbers.

Fig. B.7 Recall instructions, treatment

Your partner will be notified of your choice to do the tasks and whether the £2 payment has been generated.

You have just seen a clue on whether the payment will be generated. There is a higher chance of generating the payment for your partner if you had to drag **more sliders to 150 on the page you just completed**. You cannot return to that page now.

Fig. B.8 Recall instructions, treatment

Below is a table showing how the chance of a £2 payment to your partner from the session 2 tasks depends on the **number of sliders you dragged to 150 today**.

Number of times you dragged sliders to 150	**Probability of £2 payment to your partner for the tasks**
0	20%
1	28%
2	36%
3	44%
4	52%
5	60%
6	68%
7	76%
8	84%
9	92%
10	100%

This number was **randomly generated**: there's a 98% chance you were given **2-8 sliders to drag to 150**, with an equal chance of each possibility in this range. It is very unlikely that you were given 0-1 or 9-10 sliders of this kind: there's a 0.5% chance of each of these numbers.

Fig. B.9 Recall page, treatment

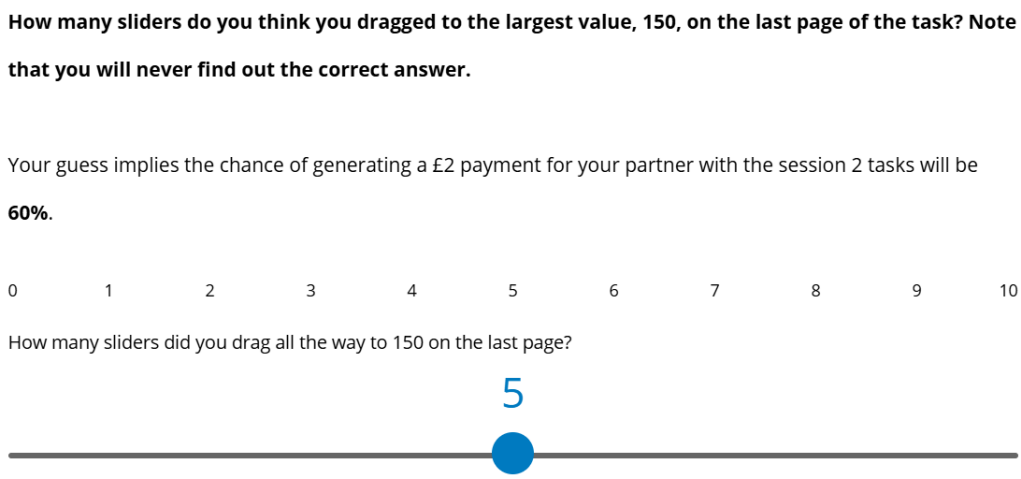


Fig. B.10 Recall confidence elicitation

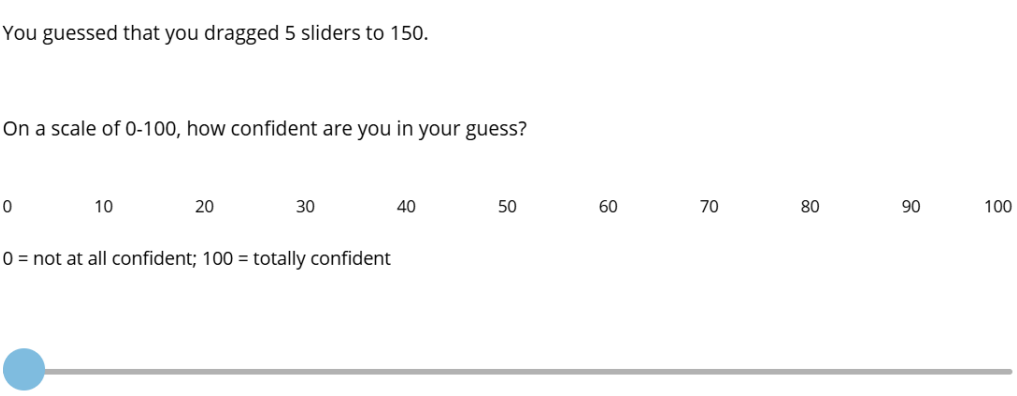


Fig. B.11 Disclosure

We can now reveal that you dragged 5 sliders to 150 in session 1. This means the chance that a £2 payment will be generated for your partner is **60%**.

Fig. B.12 Self-reported risk preference elicitation

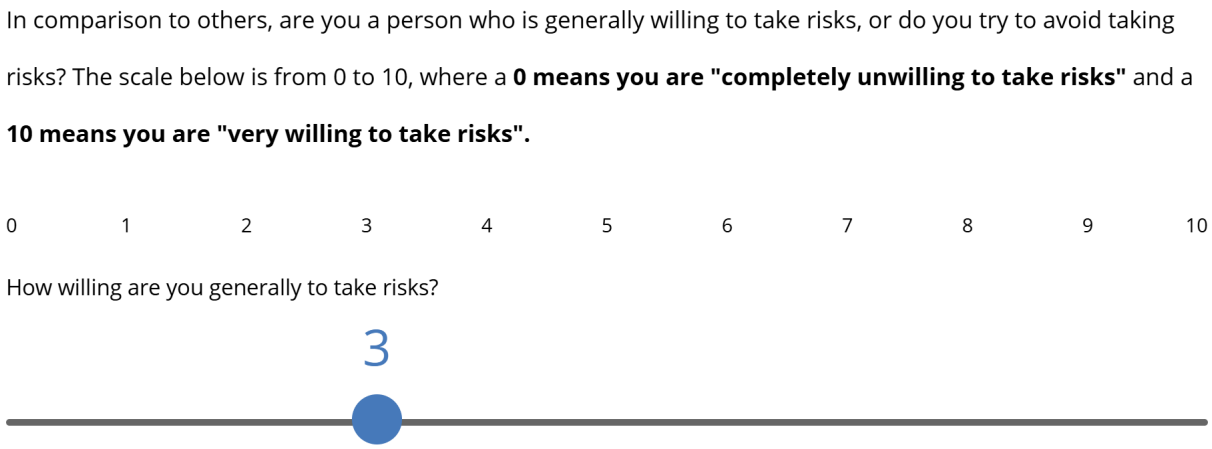


Fig. B.13 Self-reported time preference elicitation

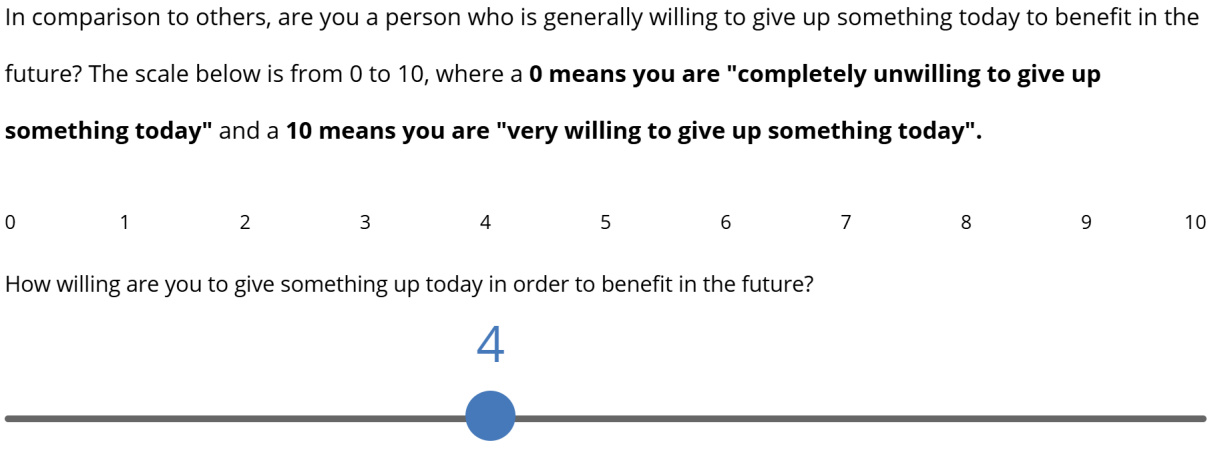
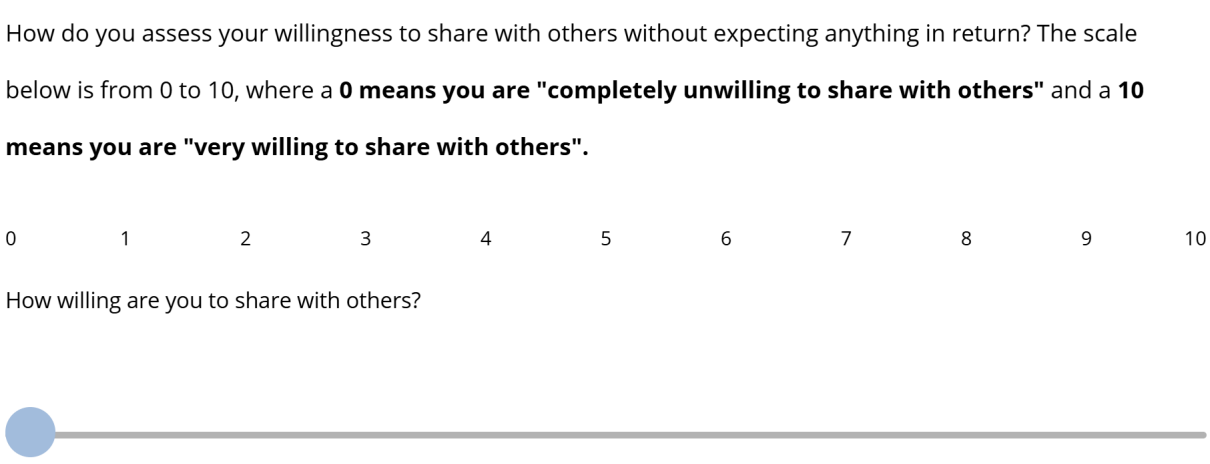


Fig. B.14 Self-reported other-regarding preference elicitation



Appendix C

Appendices to Chapter 3

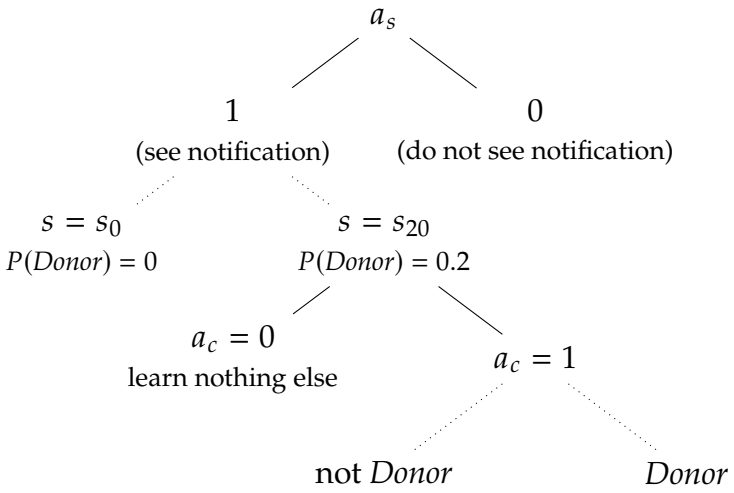
C.1 Additional tables and figures

Table C.1 Implied binary information preferences from session 1

Binary preference	N	Share
$\{0\} \succ_1 \{1\}$	443	66.9% (1.8)
$\{0\} \sim_1 \{1\}$	122	18.4% (1.5)
$\{1\} \succ_1 \{0\}$	97	14.7% (1.4)
Total	662	100%

Note: Standard errors in parentheses, in percentage points.

Fig. C.1 Illustration of treatment conditions



Note: s is the signal. a_s is the action to see the signal. a_c is the action to confirm *Donor* status conditional on $s = s_{20}$. In *PartAvoid*, s is automatically displayed to participants at the beginning of session 2 and menu preferences correspond to a_c. In *FullAvoid*, menu preferences correspond to a_s.

Table C.2 Summary statistics

	Value	...of *N*
Mean age	36.7	662
Standard deviation of age	12.3	662
Male	50%	662
Ethnicity: white	72.5%	661
Ethnicity: black	5.9%	661
Ethnicity: mixed/other	11.4%	661
Mean altruism (0-10)	7.0	632
Mean risk (0-10)	4.7	632
Mean patience (0-10)	6.6	632
Mean self-control (0-10)	5.6	632
Took bonus	91.0%	669
Treatment: *FullAvoid*	47.9%	609
Attended session 2	94.8%	669
Employed	45.1%	478
Mean reward	$5.80	634
Session 1 median duration	7m, 40s	669
Session 2 median duration	2m, 4s	634
Failed attention checks	11	673

Note: Qualitative, self-reported measures elicited at the end of session 2. Treatment allocation was random only for participants who took the bonus. Reported only for participants who completed both sessions. The few participants who did not complete session 2 were just paid the base pay for session 1, $1.40.

Table C.3 Menu preferences from session 1 — all stages

		Raw (no cost)		**Main (effort cost)**		$WTP > 0$	
Menu preference	Type	N	Share	N	Share	N	Share
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	Tempted	316	47.7% (1.9)	254	38.4% (1.9)	109	16.5% (1.4)
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	13	2% (0.5)	80	12.1% (1.3)	144	21.8% (1.6)
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	Strongly Tempted	36	5.4% (0.9)	71	10.7% (1.2)	32	4.8% (0.8)
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	Standard info. averse	20	3% (0.7)	54	8.2% (1.1)	198	29.9% (1.8)
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	82	12.4% (1.3)	50	7.6% (1)	34	5.1% (0.9)
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	55	8.3% (1.1)	34	5.1% (0.9)	28	4.2% (0.8)
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	Flex	16	2.4% (0.6)	33	5% (0.8)	31	4.7% (0.8)
$\{1\} \succ_1 \{0,1\} \succ_1 \{0\}$	Strongly Tempted (info. loving)	51	7.7% (1)	29	4.4% (0.8)	21	3.2% (0.7)
$\{0,1\} \sim_1 \{1\} \succ_1 \{0\}$	Other	13	2% (0.5)	17	2.6% (0.6)	28	4.2% (0.8)
$\{0\} \succ_1 \{1\} \succ_1 \{0,1\}$	Flexibility averse	35	5.3% (0.9)	14	2.1% (0.6)	11	1.7% (0.5)
$\{1\} \succ_1 \{0\} \sim_1 \{0,1\}$	Other	6	0.9% (0.4)	13	2% (0.5)	13	2% (0.5)
$\{0\} \sim_1 \{1\} \succ_1 \{0,1\}$	Commitment loving	8	1.2% (0.4)	9	1.4% (0.5)	9	1.4% (0.5)
$\{1\} \succ_1 \{0\} \succ_1 \{0,1\}$	Flexibility averse	11	1.7% (0.5)	4	0.6% (0.3)	4	0.6% (0.3)
Total		662	100%	662	100%	662	100%

Note: Participants must complete a short task for each strict preference submitted in their ranking. For *WTP* measure: if present, the first strict preference additionally requires participants to have a strictly positive willingness to pay to avoid a swap of their first and second choices. If $WTP = 0$, the first strict preference is replaced with indifference. Standard errors in parentheses, in percentage points.

Table C.4 Menu preferences from session 1 — subjects who turned down bonus

		Raw (no cost)		**Main (effort cost)**		$WTP > 0$	
Menu preference	Type	N	Share	N	Share	N	Share
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	Tempted	27	45% (2)	23	38.3% (2)	13	21.7% (1.7)
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	2	3.3% (0.7)	7	11.7% (1.3)	11	18.3% (1.6)
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	Strongly Tempted	5	8.3% (1.1)	7	11.7% (1.3)	4	6.7% (1)
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	Standard info. averse	3	5% (0.9)	6	10% (1.2)	16	26.7% (1.8)
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	6	10% (1.2)	5	8.3% (1.1)	4	6.7% (1)
$\{1\} \succ_1 \{0,1\} \succ_1 \{0\}$	Self-control (info. loving)	5	8.3% (1.1)	3	5% (0.9)	2	3.3% (0.7)
$\{0\} \succ_1 \{1\} \succ_1 \{0,1\}$	Flexibility averse	5	8.3% (1.1)	2	3.3% (0.7)	2	3.3% (0.7)
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	4	6.7% (1)	2	3.3% (0.7)	1	1.7% (0.5)
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	Flex	2	3.3% (0.7)	2	3.3% (0.7)	2	3.3% (0.7)
$\{0,1\} \sim_1 \{1\} \succ_1 \{0\}$	Other	1	1.7% (0.5)	2	3.3% (0.7)	4	6.7% (1)
$\{0\} \sim_1 \{1\} \succ_1 \{0,1\}$	Commitment loving		0% (—)	1	1.7% (0.5)	1	1.7% (0.5)
Total		60	100%	60	100%	60	100%

Note: Participants must complete a short task for each strict preference submitted in their ranking. For *WTP* measure: if present, the first strict preference additionally requires participants to have a strictly positive willingness to pay to avoid a swap of their first and second choices. If $WTP = 0$, the first strict preference is replaced with indifference. Standard errors in parentheses, in percentage points.

Table C.5 Condensed types from session 1

		Raw (no cost)		**Main (effort cost)**		$WTP > 0$	
Type	Unique rankings	N	Share	N	Share	N	Share
Tempted	1	316	47.7% (1.9)	**254**	**38.4%** (1.9)	109	16.5% (1.4)
Flexibility	3	153	23.1% (1.6)	117	17.7% (1.5)	93	14% (1.4)
Other	6	124	18.7% (1.5)	86	13% (1.3)	86	13% (1.3)
Indifferent	1	13	2% (0.5)	80	12.1% (1.3)	144	21.8% (1.6)
Strongly Tempted	1	36	5.4% (0.9)	**71**	**10.7%** (1.2)	32	4.8% (0.8)
Standard info. averse	1	20	3% (0.7)	54	8.2% (1.1)	198	29.9% (1.8)
Total	13	662	100%	662	100%	662	100%

Note: In main measure, participants must complete a short task for each strict preference submitted in their ranking. For *WTP* measure: if present, the first strict preference additionally requires participants to have a strictly positive willingness to pay to avoid a swap of their first and second choices. If $WTP = 0$, the first strict preference is replaced with indifference. Standard errors in parentheses, in percentage points.

Table C.6 *WTP* by preference ranking

Ranking	Type	$E(WTP)$	% $WTP > 0$	$E(WTP\|WTP > 0)$	...of N
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	Tempted	\$0.09 (0.01)	42.9% (3.11)	\$0.20 (0.01)	254
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	—	—	—	80
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	Strongly Tempted	\$0.12 (0.02)	45.1% (5.95)	\$0.26 (0.03)	71
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	Standard info. averse	\$0.2 (0.03)	67.9% (6.41)	\$0.29 (0.03)	54
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	\$0.12 (0.02)	67.3% (6.7)	\$0.18 (0.02)	50
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	\$0.19 (0.02)	82.4% (6.64)	\$0.23 (0.02)	34
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	Flex	\$0.29 (0.03)	93.9% (4.22)	\$0.31 (0.03)	33
$\{1\} \succ_1 \{0,1\} \succ_1 \{0\}$	Strongly Tempted (info. loving)	\$0.23 (0.04)	72.4% (8.45)	\$0.32 (0.03)	29
$\{0,1\} \sim_1 \{1\} \succ_1 \{0\}$	Other	\$0.27 (0.05)	82.4% (9.53)	\$0.32 (0.04)	17
$\{0\} \succ_1 \{1\} \succ_1 \{0,1\}$	Flexibility averse	\$0.23 (0.05)	78.6% (11.38)	\$0.3 (0.05)	14
$\{1\} \succ_1 \{0\} \sim_1 \{0,1\}$	Other	\$0.26 (0.05)	100% (0)	\$0.26 (0.05)	13
$\{0\} \sim_1 \{1\} \succ_1 \{0,1\}$	Commitment loving	\$0.25 (0.07)	66.7% (16.67)	\$0.38 (0.04)	9
$\{1\} \succ_1 \{0\} \succ_1 \{0,1\}$	Flexibility averse	\$0.15 (0.05)	100% (0)	\$0.15 (0.05)	4
Total					662

Note: Standard errors in parentheses, in \$ and percentage points as applicable.

Table C.7 Logistic regression for preference categories

	$\mathbb{1}(\{0\} \succ_1 \{1\})$	$\mathbb{1}(\{0\} \succ_1 \{0,1\} \succeq_1 \{1\} \mid \{0\} \succ_1 \{1\})$
	(1)	(2)
Age 25-34	−0.31 (0.30)	0.43 (0.34)
Age 35:44	−0.46 (0.32)	0.69 (0.37)
Age 45:54	−0.06 (0.37)	0.79 (0.43)
Age 55:64	−0.98 (0.42)	0.88 (0.55)
Age 65+	−0.65 (0.51)	1.06 (0.72)
Male	−0.05 (0.19)	0.04 (0.24)
Asian	−0.97 (0.29)	0.40 (0.44)
Black	−0.17 (0.39)	0.19 (0.51)
Mixed ethn.	−0.33 (0.37)	0.56 (0.50)
Other ethn.	−0.34 (0.41)	0.20 (0.55)
PartAvoid	−0.44 (0.18)	0.23 (0.23)
Risk tol.	−0.10 (0.04)	0.00 (0.05)
Altruism	−0.05 (0.04)	0.02 (0.05)
Patience	0.02 (0.05)	0.04 (0.06)
Strongly Tempted	0.02 (0.04)	0.03 (0.06)
$p_a = 0.5$	0.07 (0.25)	−0.13 (0.33)
$p_a = 0.75$	0.35 (0.26)	0.06 (0.33)
$p_a = 1$	0.20 (0.26)	−0.05 (0.33)
Constant	1.93 (0.54)	−0.27 (0.63)
Observations	619	420
Log likelihood	−373.03	−242.70
Akaike inf. crit.	782.06	521.40

Note: Coefficients are log-odds.

Table C.8 Deliberation time under exogenous $\{0,1\}$ in session 2 by session 1 preference ranking

Menu preference	Type	Time (seconds)	N
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	Tempted	13.1 (0.5)	105
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	17.5 (1.9)	33
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	Strongly Tempted	18 (1.9)	28
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	Standard info. averse	15.6 (2.3)	25
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	21.5 (2.9)	19
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	Flex	25.9 (3.2)	13
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	16.6 (1.7)	10
$\{0,1\} \sim_1 \{1\} \succ_1 \{0\}$	Other	11.7 (2.3)	7
$\{1\} \succ_1 \{0,1\} \succ_1 \{0\}$	Strongly Tempted (info. loving)	16.3 (3.4)	6
$\{0\} \succ_1 \{1\} \succ_1 \{0,1\}$	Flexibility averse	11.5 (1.8)	6
$\{1\} \succ_1 \{0\} \sim_1 \{0,1\}$	Other	17.3 (3.2)	5
$\{0\} \sim_1 \{1\} \succ_1 \{0,1\}$	Commitment loving	12.5 (2.4)	3
$\{1\} \succ_1 \{0\} \succ_1 \{0,1\}$	Flexibility averse	25.2 (11.8)	3
Total			263

Table C.9 Probability of opening the envelope vs deliberation time under exogenous {0,1}

	$a = 1$
	(1)
Time	3.13 (2.35)
$(\text{Time})^2$	−0.12 (0.09)
Constant	−22.84 (15.14)
Observations	133

Note: Coefficients are log-odds. Time is measured as duration on page where envelope decision is made, measured in seconds.

Table C.10 Placebo test with session 1 time

	$\{0\} \succ_1 \{0,1\}$
	$log(WTP+1)$
Not at all interested	−0.30 (0.32)
Session 1 total time	−0.00 (0.00)
Session 1 total time^2	0.00 (0.00)
Opened envelope	0.63 (0.80)
Constant	1.82 (0.65)
Observations	133

Note: Sample restricted to those randomised to face $\{0,1\}$ exogenously in session 2, those classified as tempted or Strongly Tempted in the main stage.

Table C.11 Placebo test with session 1 time

	$\{0\} \succ_1 \{0,1\}$
	$log(WTP+1)$
Not at all interested ($\lvert\overline{U}_l\rvert$)	−0.30 (0.32)
Session 1 total time	−0.00 (0.00)
Session 1 total time^2	0.00 (0.00)
Opened envelope	0.63 (0.80)
Constant	1.82 (0.65)
Observations	133

Note: Sample restricted to those randomised to face $\{0,1\}$ exogenously in session 2, those classified as tempted or Strongly Tempted in the main stage.

Table C.12 Binary information preferences from session 1: treatment comparison

	N		Share			
Binary preference	*PartAvoid*	*FullAvoid*	*PartAvoid*	*FullAvoid*	Diff.	p-value
$\{0\} \succ_1 \{1\}$	195	208	62.3% (2.7)	72% (2.6)	-9.67	0.01
$\{0\} \sim_1 \{1\}$	71	41	22.7% (2.4)	14.2% (2.1)	8.5	0.01
$\{1\} \succ_1 \{0\}$	47	40	15% (2)	13.8% (2)	1.18	0.77
Total	313	289	100%	100%		

Note: The 9% of participants who did not take the bonus are excluded from this table, since they were all allocated to the same condition (in which all of the information was accessed at the same time). Differences in percentage points.

Table C.13 Interest in information by preference ranking and session, for participants choosing 0 from exogenous $\{0, 1\}$ in session 2

Ranking	Type	Interest (s1)	Interest (s2)	p-value	N
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	Tempted	1.3 (0.1)	1.5 (0.1)	0	103
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	Strongly Tempted	1.2 (0.1)	1.4 (0.2)	0.13	26
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	1.8 (0.1)	2 (0.2)	0.06	26
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	Standard info. averse	1.3 (0.1)	1.7 (0.2)	0.03	23
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	2.2 (0.2)	2.3 (0.2)	0.43	17
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	Flex	2.7 (0.2)	2.1 (0.3)	0.03	12
$\{0\} \succ_1 \{1\} \succ_1 \{0,1\}$	Flexibility averse	1.6 (0.4)	1.6 (0.4)	—	5
$\{0,1\} \sim_1 \{1\} \succ_1 \{0\}$	Other	1.8 (0.2)	2.5 (0.3)	0.22	4
$\{1\} \succ_1 \{0\} \sim_1 \{0,1\}$	Other	3 (0.6)	2.3 (0.3)	0.18	3
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	2.5 (1.5)	3 (0)	0.8	2
$\{1\} \succ_1 \{0\} \succ_1 \{0,1\}$	Flexibility averse	2.5 (1.5)	2 (1)	0.5	2
$\{0\} \sim_1 \{1\} \succ_1 \{0,1\}$	Commitment loving	1 (—)	2 (—)	—	1
$\{1\} \succ_1 \{0,1\} \succ_1 \{0\}$	Strongly Tempted (info. loving)	2 (—)	2 (—)	—	1
Total					225

Note: p-values correspond to two-sided paired t-tests of the difference in interest across sessions within each preference ranking. Responses to the question 'How interested are you to know if the charity donation was reduced by \$15 as a result of your choice?', numerically coded: 'Not at all interested' = 1; 'A little interested' = 2; 'Fairly interested' = 3; 'Very interested' = 4. Values are means, with standard errors in parentheses.

Table C.14 Happiness with $4 choice by preference ranking

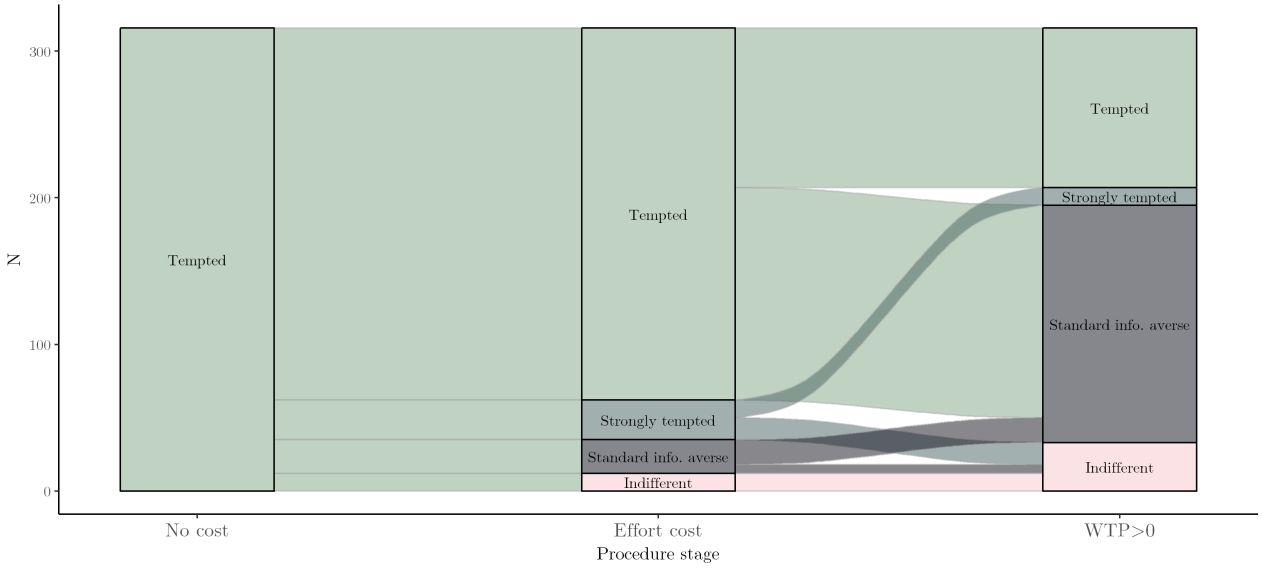
Ranking	Type	Happiness with $4 choice	*N*
$\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$	**Tempted**	**4.1 (0.1)**	254
$\{0\} \sim_1 \{0,1\} \sim_1 \{1\}$	Indifferent	3.9 (0.1)	80
$\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$	**Strongly Tempted**	**3.8 (0.1)**	71
$\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$	**Standard info. averse**	**3.8 (0.1)**	54
$\{0,1\} \succ_1 \{0\} \succ_1 \{1\}$	Flex	4.1 (0.1)	50
$\{0,1\} \succ_1 \{1\} \succ_1 \{0\}$	Flex	3.8 (0.2)	34
$\{0,1\} \succ_1 \{0\} \sim_1 \{1\}$	Flex	3.6 (0.1)	33
$\{1\} \succ_1 \{0,1\} \succ_1 \{0\}$	Strongly Tempted (info. loving)	4.3 (0.2)	29
$\{0,1\} \sim_1 \{1\} \succ_1 \{0\}$	Other	3.6 (0.3)	17
$\{0\} \succ_1 \{1\} \succ_1 \{0,1\}$	Flexibility averse	4.1 (0.2)	14
$\{1\} \succ_1 \{0\} \sim_1 \{0,1\}$	Other	4.3 (0.3)	13
$\{0\} \sim_1 \{1\} \succ_1 \{0,1\}$	Commitment loving	4.4 (0.3)	9
$\{1\} \succ_1 \{0\} \succ_1 \{0,1\}$	Flexibility averse	4.2 (0.2)	4
Total			662

Note: Responses to the question 'How happy are you with your decision to take the $4 bonus?', numerically coded: 'Extremely unhappy' = 1; 'Somewhat unhappy' = 2; 'Neither happy nor unhappy' = 3; 'Somewhat happy' = 4; 'Extremely happy' = 5. Means, with standard errors in parentheses.

C.2 Additional analysis of preference elicitation procedure

Unsurprisingly, the prevalence of indifference increases substantially from our initial 'raw' measurement of menu preferences to the main one, in which participants must complete a short effort task to confirm each strict preference (Table C.3). To understand the importance of this step, take an individual whose true preference $\succeq_1$ implies $\{0,1\} \sim_1 \{1\}$. Submitting the preference $\{0,1\} \sim_1 \{1\}$ produces the same expected payoff as either $\{0,1\} \succ_1 \{1\}$ or $\{1\} \succ_1 \{0,1\}$. The second stage of elicitation breaks these ties: it is now strictly optimal for participants who are genuinely indifferent between two or more options to say so.

Fig. C.2 Preference transitions over the elicitation procedure: types classified as 'Tempted' in the no cost stage.



Note: 'No cost' refers to the stage where participants simply enter their preference rankings into a grid. 'Effort cost' is our main measure of preferences, and is taken after participants must complete a short task to confirm each adjacent strict preference in their ranking. $WTP > 0$ captures those who are willing to pay money for their first adjacent strict preference.

The effect of the multi-stage elicitation procedure on one category of respondent is illustrated in Figure C.2. We see that a large majority (80%) of those with the preference ordering $\{0\} \succ_1 \{0,1\} \succ_1 \{1\}$ before the effort cost stage are willing to pay effort costs to maintain both strict preferences. A roughly equal number of those

unwilling to pay the effort cost are subsequently reclassified as Strongly Tempted ($\{0\} \succ_1 \{0,1\} \sim_1 \{1\}$) or standard information averse ($\{0\} \sim_1 \{0,1\} \succ_1 \{1\}$) types. In the first case, participants state that they are willing to complete a short task to confirm their preference for 'not generating the envelope at all' over 'deciding whether to open the envelope next time', but unwilling to do the same to confirm their preference for 'deciding whether to open the envelope next time' over 'automatically opening the envelope next time'. In contrast, those reclassified as standard information-averse types are unwilling in the first case but willing in the second case. A smaller share are unwilling to pay either effort cost and are thus classified as indifferent.

As previously emphasised, there may be some concern that the initial procedure, involving asking participants to select ranks 1, 2, and 3 for each alternative, naturally encourages participants to submit strict preferences. Our main measure addresses this issue by prompting participants to reconsider the adjacent *pairs* of options for which they initially submitted strict preferences. Any participant with a strict preference for commitment over flexibility in our main measure must have explicitly confirmed they had a clear preference for 'not generating the envelope' over 'deciding whether to open the envelope next time' (Figure C.7), and completed a short task to do so.

C.3 Proofs

Proposition 1 and Corollary 1. Suppose the utility cost in $t = 1$ of submitting a strict preference is c. First, we consider the strict preference $\{0\} \succ_1 \{0,1\}$. The marginal value of $\{0\}$ relative to $\{0,1\}$ when $\lambda\overline{U}_s < -\overline{U}_l$, meaning 0 is chosen in $t = 2$ from $\{0,1\}$, is

$$-\lambda\delta\mu\overline{U}_s + \lambda\delta\overline{U}_s - c. \tag{C.1}$$

Thus, willingness to pay for the preference $\{0\} \succ_1 \{0,1\}$ increases in self-control costs $\overline{U}_s$. Conversely, if self-control costs are sufficiently high, then $\lambda\overline{U}_s \geq -\overline{U}_l$ and the information is consumed in $t = 2$ under $\{0,1\}$. Then, the condition for submitting $\{0\} \succ_1 \{0,1\}$ is

$$-\lambda\delta\mu\overline{U}_s - \delta\overline{U}_l - c > 0, \tag{C.2}$$

and it is important that $\overline{U}_l$ is large relative to $\lambda\overline{U}_s$.

We now consider the strict preference $\{0,1\} \succ_1 \{1\}$ to complete the proof. If $\lambda\overline{U}_s < -\overline{U}_l$, when submitting strict preferences is costless, $\{0,1\} \succ_1 \{1\}$. As

previously argued, that is because it is cheaper to make a commitment in $t = 1$ than in $t = 2$. However, with costs of submitting strict preferences, the requirement on $\{0, 1\} \succ_1 \{1\}$ is more stringent:

$$-\lambda\delta\mu\overline{U}_s - \delta\overline{U}_l - c > 0. \tag{C.3}$$

Proof of the Proposition and Corollary are immediate.

□

Proposition 2. Let $\mathcal{I} = (I_1, I_2, I_3, ...)$ be the sequence of information prospects involving acquiring information in period 1, with $\mathcal{I} \subset \Delta\Theta$. Avoiding information just implements the (imperfectly recalled) prior belief distribution μ. What is required for full information to be a simple temptation is that i) $\int_{\Delta\Theta} u(\nu_t)dI_t$ weakly decreases in t, ii) for some $s > 0$, $\int_{\Delta\Theta} u(\nu_t)dI_t > u(\mu)$ for all $t \in [0, s)$, and iii) $\int_{\Delta\Theta} u(\nu_t)dI_t < u(\mu)$ for all $t \geq s$.

First, note that all posterior distributions implemented by prospects in $\mathcal{I}$ are more informative than μ: they are more correlated than μ with the unobserved state. Thus, if u is concave over the region of $\Delta\Theta$ with $\sigma^2 > y$ for some $y > 0$, we know that after some period, the value of information will always be negative. This demonstrates point iii).

Then, note in order for information to be valued in the first period, it must be the case that u is convex for some region $\sigma^2 \in [0, x]$ where $x < y$. Otherwise, u would be concave over $\Delta\Theta$ in its entirety, so $\int_{\Delta\Theta} u(\nu_t)dI_t > u(\mu)$ for all t. This demonstrates point ii).

Finally, we know that the value of I_t declines monotonically in t simply from the fact that u is strictly increasing in σ_t^2, and I_t always involves a lottery over a set of posterior distributions where each has a strictly higher variance than those implemented by I_{t-1}.

□

C.4 Experiment pages

Fig. C.3 Bonus choice page

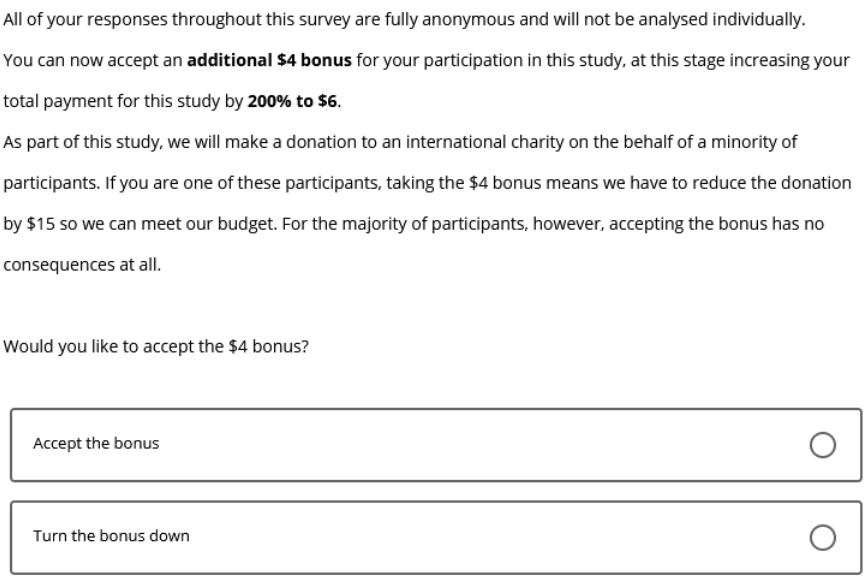


Fig. C.4 Charity information page

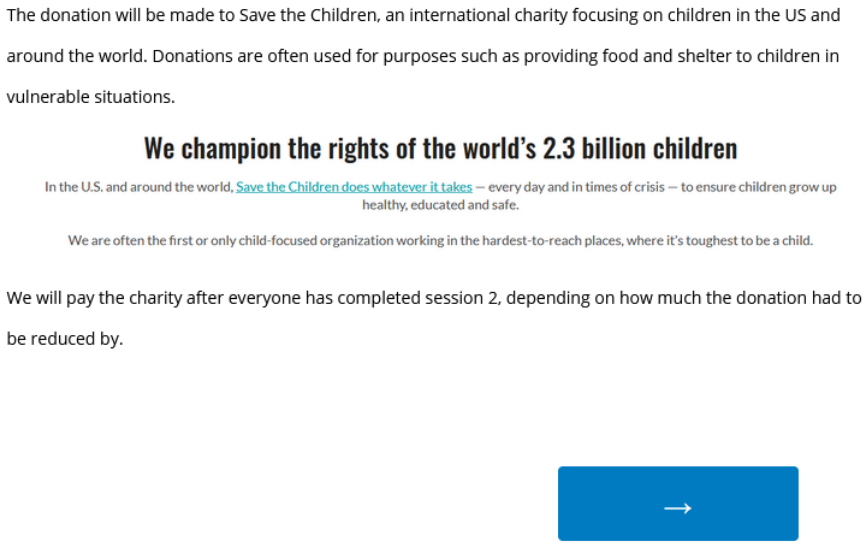


Fig. C.5 Preference elicitation — step 1 (*FullAvoid*)

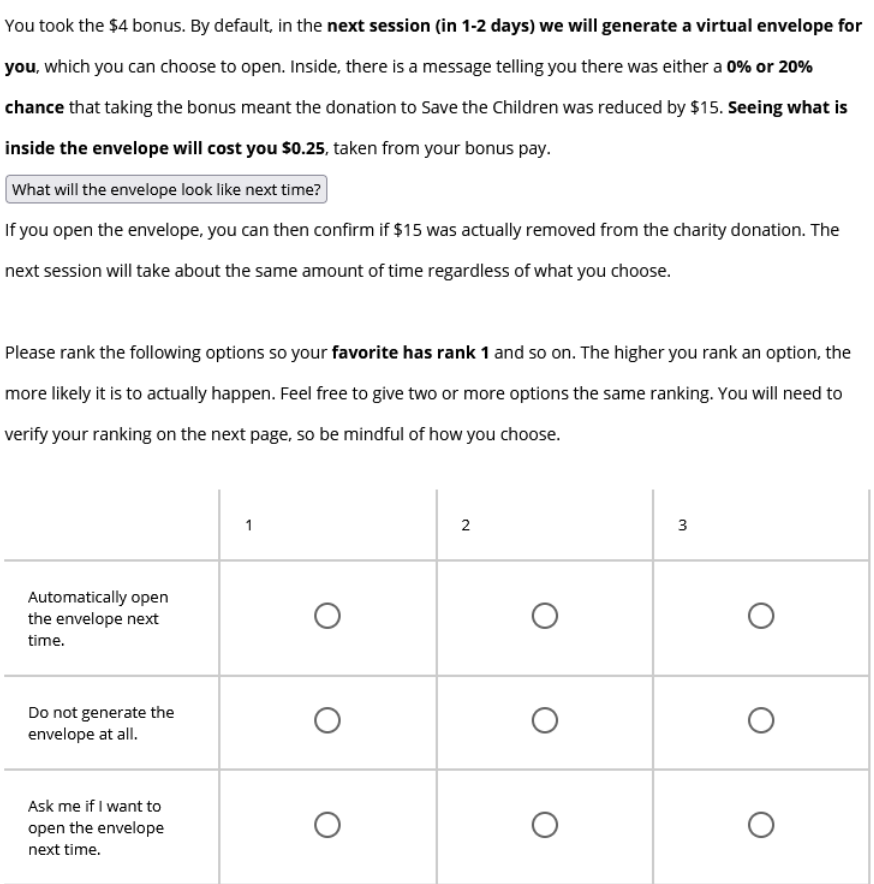


Fig. C.6 Preference elicitation — step 1 (*PartAvoid*)

You took the $4 bonus. At the start of the **next session (in 1-2 days)**, you will first be notified that there was either a 0% or 20% chance that taking the bonus meant the donation to Save the Children was reduced by $15.

If you find out there was a 20% chance of the charity donation being reduced, by default **we will generate a virtual envelope for you**, which you can choose to open. Inside will be a message confirming whether $15 was removed from the charity donation. The next session will take about the same amount of time regardless of what you choose. **Seeing what is inside the envelope will cost you $0.50**, taken from your bonus pay.

What will the envelope look like next time?

Please rank the following options so your **favorite has rank 1** and so on. The higher you rank an option, the more likely it is to actually happen. Feel free to give two or more options the same ranking. You will need to verify your ranking on the next page, so be mindful of how you choose.

	1	2	3
Do not generate the envelope at all.	○	○	○
Ask me if I want to open the envelope next time.	○	○	○
Automatically open the envelope next time.	○	○	○

Fig. C.7 Preference elicitation — step 2

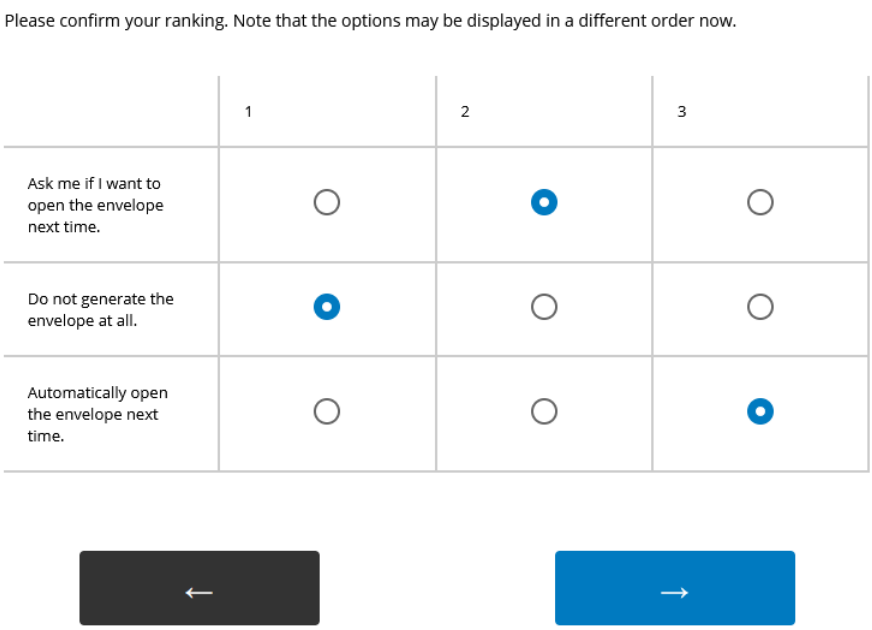


Fig. C.8 Preference elicitation — step 3

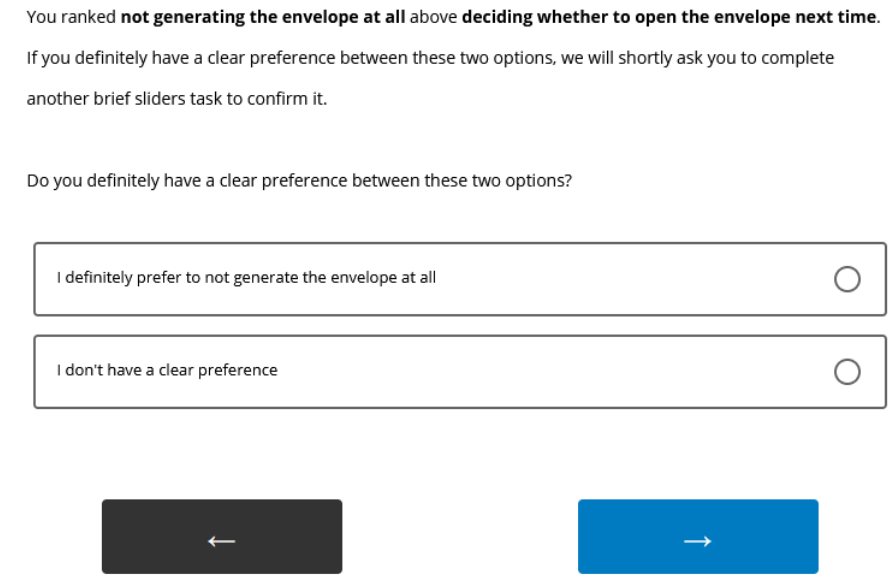


Fig. C.9 Preference elicitation — effort confirmation

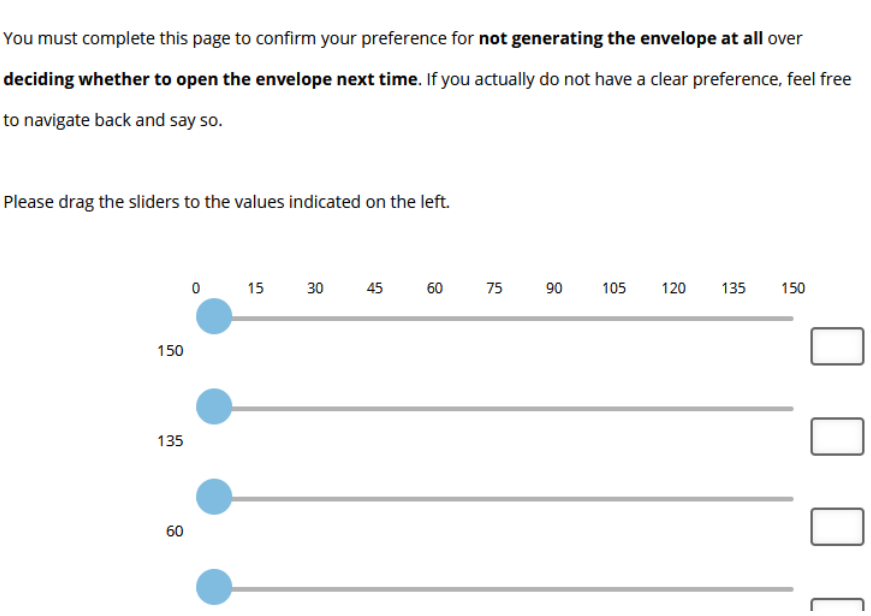


Fig. C.10 Preference elicitation — summary

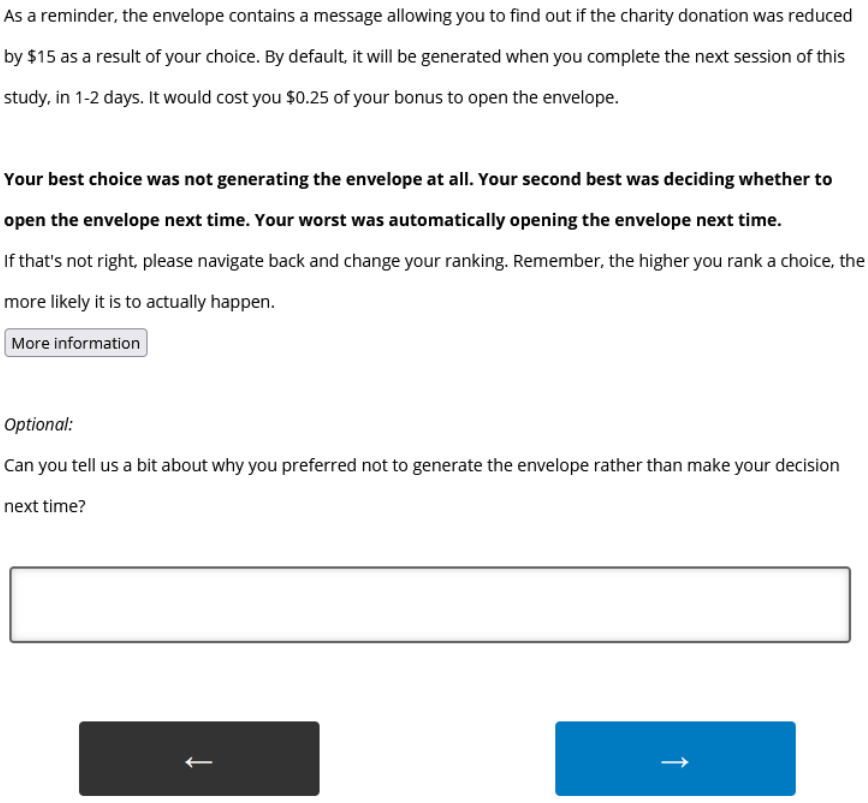


Fig. C.11 WTP elicitation

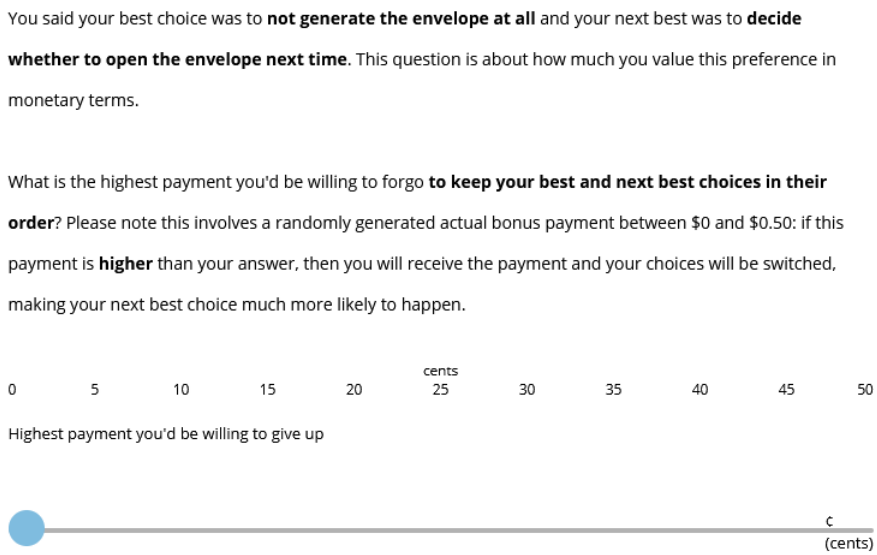


Fig. C.12 Envelope — *FullAvoid*

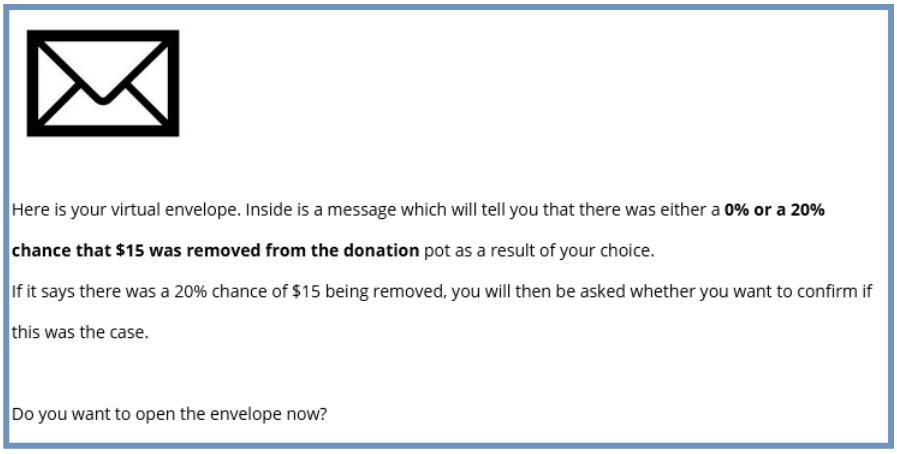


Fig. C.13 Envelope — *PartAvoid*

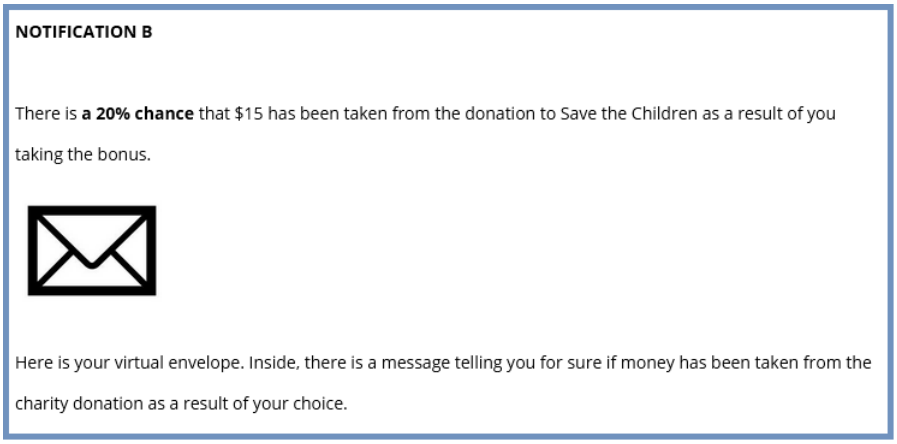


Fig. C.14 Envelope — bonus rejected

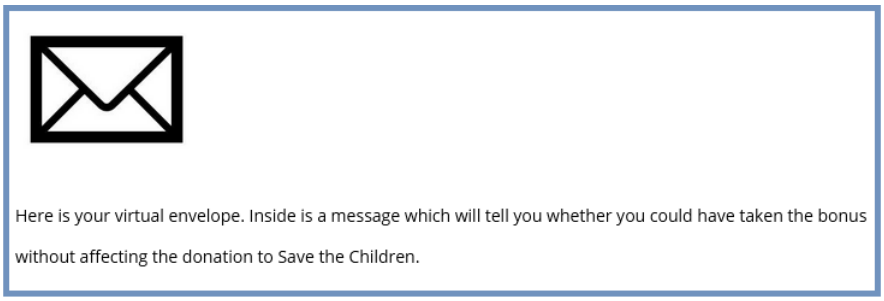


Fig. C.15 First attention check

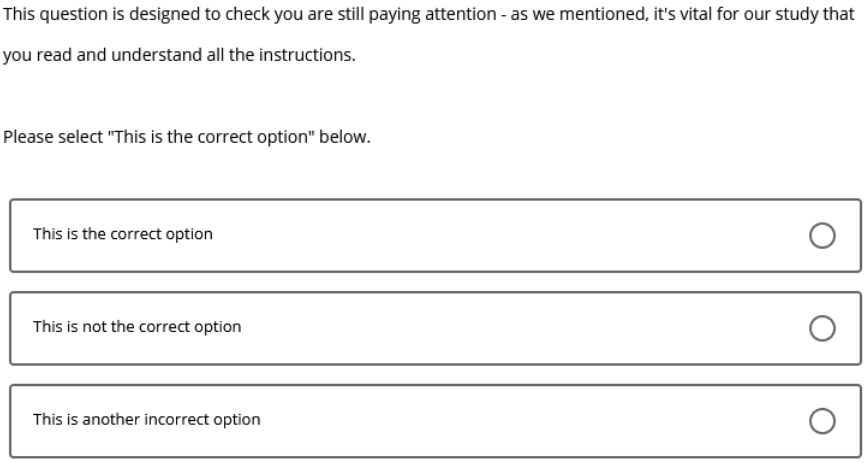


Fig. C.16 First attention check