University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 190

Informational Requirements of Nudging

Jean-Michel Benkert and Nick Netzer

Revised version, August 2016

Informational Requirements of Nudging[‡]

Jean-Michel Benkert[*] and Nick Netzer[**]

This version: August 2016
First version: November 2014

Abstract

A nudge is a paternalistic government intervention that attempts to improve choices by changing the framing of a decision problem. We propose a welfare-theoretic foundation for nudging similar in spirit to the classical revealed preference approach, by investigating a model where preferences and mistakes of an agent can be elicited from her choices under different frames. We provide characterizations of the classes of behavioral models for which nudging is possible or impossible, and we derive results on the required quantity of information. We also study an extended application to a savings problem.

Keywords: nudge, framing, behavioral welfare economics, revealed preference

JEL Classification: D03, D04, D60, D82

[‡]We are grateful for very helpful comments by Sandro Ambühl, Sandeep Baliga, Eddie Dekel, Kfir Eliaz, Jeff Ely, Samuel Häfner, Igor Letina, Konrad Mierendorff, Georg Nöldeke, Ariel Rubinstein, Yuval Salant, Armin Schmutzler, Ron Siegel, Ran Spiegler, Georg Weizsäcker, seminar audiences at DICE Düsseldorf, European University Institute, Goethe University Frankfurt, HECER Helsinki, Northwestern University, NYU Abu Dhabi, Tel Aviv University, UCL, the Universities of Basel, Bonn, Konstanz, Michigan, Surrey, and Zurich, and participants at CESifo Area Conference on Behavioural Economics 2014, BBE Workshop 2015, Midwest Economic Theory Meeting Fall 2015, Swiss Economists Abroad Meeting 2015, Verein für Socialpolitik Theoretischer Ausschuss 2015, and BERA Micro Workshop 2016. All errors are our own.

[*]University of Zurich, Department of Economics, Bluemlisalpstrasse 10, 8006 Zurich, Switzerland, and UBS International Center of Economics in Society at the University of Zurich. Email: jean-michel.benkert@econ.uzh.ch. The author would like to thank the UBS International Center of Economics in Society at the University of Zurich and the Swiss National Science Foundation (Doc.Mobility Grant P1ZHP1_161810) for financial support.

[**]University of Zurich, Department of Economics, Bluemlisalpstrasse 10, 8006 Zurich, Switzerland. Email: nick.netzer@econ.uzh.ch.

1 Introduction

A nudge (Thaler and Sunstein, 2008) is a regulatory intervention that is characterized by two properties. First, it is paternalistic in nature, because "it is selected with the goal of influencing the choices of affected parties in a way that will make those parties better off" (Thaler and Sunstein, 2003, p. 175). Second, it is not coercive but instead manipulates the framing of a decision problem, which makes it more easily acceptable than conventional paternalistic measures. Among the best-known examples already dicussed in Thaler and Sunstein (2003) is retirement saving in 401(k) savings plans, which can be encouraged tremendously by setting the default to automatic enrollment. Another example is the order in which food is presented in a cafeteria, which can be used to promote a more healthy diet.

The intriguing idea that choices can be improved by framing has made the concept of nudging also politically attractive. Governments of numerous countries have set up so-called "nudge units", which develop and implement nudge-based policies. The UK spearheaded this development in 2010 with the foundation of the Behavioral Insights Team.[1] More recently, U.S. President Barack Obama issued an executive order establishing the Social and Behavioral Sciences Team.[2] The executive order encourages all government agencies to "carefully consider how the presentation and structure of [...] choices, including the order, number, and arrangement of options, can most effectively promote public welfare".

This paper addresses the problem of how to define and measure welfare. What does it mean that a frame improves choices? How can we be sure that it is in the employee's own best interest to save more or to eat more healthily? The ordinary revealed preference approach is not suitable to answer these questions, due to the behavioral inconsistencies caused by framing. Instead, the applied nudging literature often takes criteria such as increased savings or improved health for granted (see e.g. Goldin, 2015, for a discussion). Other authors have entirely dismissed the idea of nudging based on the welfare problem (see e.g. Grüne-Yanoff, 2012). We take a different, choice-theoretic approach. We investigate a framework where the welfare preference of an agent can be (partially) inferred from her choices under different frames, and the success of a nudge is evaluated on this basis. We thus attempt to develop a welfare-theoretic foundation for nudging in a revealed preference spirit, but appropriately modified. The twist is that, once we accept that "in certain contexts, people are prone to error" (Sunstein, 2014, p. 4), we may be able to learn about these errors from choice data.[3]

[1] See http://www.behaviouralinsights.co.uk.

[2] See http://go.wh.gov/MKURtv.

[3] Kőszegi and Rabin (2008b) first emphasized the possibility of recovering both welfare preferences and implementation mistakes from choice data, for a given behavioral model. Several contributions have studied this problem for specific models. Recent examples include Masatlioglu et al. (2012) for a model

Our formal framework is a variant of Rubinstein and Salant (2012), henceforth RS, who formulate a generalized approach for eliciting an agent's preferences from choice data. In this framework, which we formally introduce in Section 2, a regulator has a conjecture about the behavioral model d, which relates each pair of a *welfare preference* $\succeq$ and a *frame* f to a *behavioral preference* $d(\succeq, f)$. The interpretation is that an agent with welfare preference $\succeq$ acts as if maximizing $d(\succeq, f)$ if the situation is framed according to f. The welfare preference represents the normatively relevant well-being of the agent but is not observable. Behavioral preferences may be different from the welfare preference but are in principle observable in the usual revealed preference sense. RS investigate the problem of learning about the welfare preference from a data set that contains observations of behavior and, possibly, frames. We follow their approach in a first step, by verifying which welfare preferences could have generated a given data set. In a second step, we evaluate the frames based on the acquired information.

The framework's generality enables us to accommodate many different behavioral models. Among others, we will study well-known models such as choice from lists, default biases, satisficing, priming, and limited search. Our goal is not to take a stand on what the correct behavioral model is, or to argue in favor of any one of these models. Rather, the objective of our analysis is to understand the general properties of decision-making processes that make it possible or impossible to improve choices by framing.

A first contribution of our paper is to provide a choice-theoretic definition of a nudge. After identifying the welfare preferences that are consistent with a given data set and a behavioral model, in Section 3 we evaluate the frames on the basis of each of these preferences. Comparing frames pairwise, we say that a frame f is a weakly successful nudge over frame f' if the induced choices under f are at least as good as under f', irrespective of which of the consistent preferences is the actual welfare preference. This definition captures the above-mentioned idea that the regulator aims at improving the agent's choices by her own standards, i.e., the regulator tries to help the agent do what she really wants to do. It also shares with the literature (e.g. Masatlioglu et al., 2012) the cautious approach of requiring agreement among all possible welfare preferences, thereby ensuring that the regulator does not accidentally make the agent worse off.

Having formalized the concept of a successful nudge, we can formulate notions of global optimality. Ideally, we may be able to identify a frame that is a successful nudge over all the other frames. We show that the ability to identify such an optimal frame coincides with the ability to identify the welfare preference. An optimal frame is revealed by some sufficiently rich data set if and only if the welfare preference is fully revealed by some

of limited attention, and Kőszegi and Szeidl (2013) for a model of focusing. Caplin and Martin (2012) provide conditions under which welfare preferences can be recovered from choice data in a setting where frames contain payoff-relevant information, such that framing effects are fully rational.

sufficiently rich data set. This does not mean that the welfare preference has to be fully elicited for successful nudging, as we will show by example, but it allows us to consider two polar cases: models in which the welfare preference can never be identified completely, and models in which the welfare preference can be identified completely. There are interesting examples for either class of models, such as a satisficing model that has non-identifiable preferences and a limited search model that has identifiable preferences. We also ask how many models belong to each of the two classes and show that the share of models with identifiable preferences converges to 1 as the set of alternatives grows.

In Section 4 we investigate models with non-identifiable preferences more thoroughly. Finding an optimal frame is out of reach for these models, but we can still pursue the more modest goal of identifying frames which are dominated by others. Put differently, even though it is impossible to find a frame that improves upon all other frames, it may still be the case that some frames can be improved upon. Such dominated frames can indeed exist, as we show by example. However, if the behavioral model satisfies a property that we term the *frame cancellation property*, then all frames are always undominated, irrespective of the data set's richness. With the frame cancellation property, observation of choices never reveals the information required to improve these choices. Several important models have the frame cancellation property. A first example is the satisficing model in its different versions. A second example is the much-discussed case where the agent chooses the one alternative out of two that is marked as a default. We also present a decision-making procedure with limited sensitivity that nests all these (and more) behavioral models.

If, by contrast, the welfare preference can ultimately be learned, then questions of complexity arise. How many, and which, observations are necessary to determine the optimal frame? In Section 5 we define an *elicitation procedure* as a rule that specifies the order in which we impose different frames on the agent during an observation phase, contingent on the history of previous observations. This captures the idea that a data set may not be given randomly but can be collected deliberately with the purpose of finding an optimal nudge as quickly as possible. Holding fixed the unknown welfare preference of the agent, an elicitation procedure generates a sequence of expanding data sets. We define the complexity n of the nudging problem as the minimum over all elicitation procedures of the number of observations after which the optimal frame is guaranteed to be known. This number can sometimes be surprisingly small. For instance, we construct an optimal elicitation procedure for the limited search model and show that $n \leq 3$. We then establish a tight bound on n for arbitrary behavioral models. The bound, which is for instance reached by a behavioral model of priming, corresponds to the number of possible welfare preferences and thus grows more than exponentially in the number of alternatives. This implies that the informational requirements of nudging can in general become prohibitively large even with identifiable welfare preferences.

In Section 6 we allow for the possibility that the regulator has additional, non-choice-based prior information about the agent's welfare preference. We study such information in the form of restricted domains and of probabilistic beliefs over the set of preferences. For instance, the introduction of probabilistic beliefs allows us to generalize our notion of complexity in different ways. We investigate the expected running time of an elicitation procedure, and we relax our requirement of optimality and require that a frame is optimal only with a sufficiently large probability (or for a sufficiently large share of a population for which the agent is representative). As a consequence, nudging becomes easier, and sometimes substantially so.

In Section 7 we take the opposite direction and limit the regulator's exogenous information relative to the main model. We in turn relax the assumptions that the regulator has a unique conjecture about the correct behavioral model, thereby allowing for model uncertainty, and that the regulator can perfectly observe (and control) the frame under which the agent chooses. In the case of model uncertainty, for instance, the regulator needs to learn from choice data about both the welfare preference and the behavioral model. A fundamental new difficulty then arises when there are multiple model-preference pairs that are behaviorally equivalent but have different normative implications.

We present an extended application of our model to a savings problem in Section 8. Different frames induce more or less patient behavior in intertemporal choice problems. We argue that it is not a priori clear whether future-oriented or present-oriented behavior corresponds better with an agent's unobservable welfare preference. We model this in a two-period setting with one short-run frame and one long-run frame. Each frame focusses the agent on one of the two time periods. The welfare discount factor and the degree of present- or future-bias induced by the frames can then be elicited from the agent's behavior. We characterize when an optimal nudge exists and, if so, whether the agent should be nudged towards more or less patient behavior. We then take this application to the data. We conducted an experiment on Amazon Mechanical Turk to measure subjects' behavioral discount factors under present- and future-biased frames. We estimate the subjects' welfare discount factors and determine their nudgeability. We find substantial heterogeneity in discount factors, but, contrary to conventional wisdom, for a large share of subjects our model predicts that the optimal nudge is a frame that induces present-oriented behavior.

As noted before, the existing literature on nudging has focussed more on documenting the behavioral effects of framing, taking the welfare criterion for granted. We believe that our choice-theoretic approach adds a valuable new perspective. Several of our results imply strong informational limitations for a regulator who attempts to base the selection of nudges on a welfare-theoretic foundation. At the same time, our analysis reveals that seemingly minor differences between behavioral models – such as whether an agent's

failure to optimize is due to a low aspiration level as in the satisficing model, or due to a restricted number of considered alternatives as in the limited seach model – can have profoundly different consequences for the ability to improve well-being by framing.

Goldin and Reck (2015) also study the problem of identifying welfare preferences when choices are distorted by frames, focussing mostly on binary choice problems with defaults. They estimate the preference shares among fully rational agents by the shares of agents who choose each alternative when it is not the default. The preference shares among the inconsistent agents are then deduced under identifying assumptions, for instance the assumption that they are identical to the rational agents after controlling for observable differences. It is then possible to identify the default that induces the best choice for a majority of the population. Informational requirements are not the only obstacle that a libertarian paternalist has to overcome. Spiegler (2015) emphasizes that equilibrium reactions by firms must be taken into account when assessing the consequences of a nudge-based policy. Even abstracting from informational problems, these reactions can wipe out the intended benefits of a policy. Finally, frames are often not chosen by a benevolent regulator but by profit-maximizing actors in markets, which also gives rise to questions about welfare. Siegel and Salant (2015) study contracts when a seller is able to temporarily influence the buyers' willingness to pay by framing. They provide conditions under which optimal contracts make use of strategic framing, show how framing interacts with market regulation, and discuss the welfare implications.

2 Model and Examples

We begin by introducing the formal framework, which is a variant of RS, and we illustrate it with the help of two examples. Let X be a finite set of alternatives, with $m_X = |X|$. Denote by P the set of linear orders (reflexive, complete, transitive, antisymmetric) on X. A strict preference is a linear order $\succeq \in P$. Let F be a finite set of frames, with $m_F = |F|$. By definition, frames capture all dimensions of the environment that can affect decisions but are not considered welfare-relevant.[4] The agent's behavior is summarized by a distortion function $d : P \times F \to P$, which assigns a distorted preference $d(\succeq, f) \in P$ to each combination of $\succeq \in P$ and $f \in F$. The interpretation is that an agent with true welfare preference $\succeq$ acts as if maximizing the behavioral preference $d(\succeq, f)$ if the choice situation is framed by f.[5] To fix ideas, we formally introduce two possible models.

[4]For specific applications, the modeller has to judge which dimensions are welfare-relevant and which are not. For instance, it may be uncontroversial that an agent's well-being with some level of old age savings is independent of whether this level was chosen by default or by opt-in, but analogous statements would not be true if a default entails substantial switching costs, or if a "frame" actually provides novel information about the decision problem.

[5]This assumes that, given any frame, choices are consistent and can be represented by a preference. Salant and Rubinstein (2008) refer to (extended) choice functions with this property as "salient consider-

Model 1 (Perfect-Recall Satisficing). This model is taken from RS. The agent is satisfied with any of the top k alternatives in her welfare preference, so $k \in \{2, \ldots, m_X\}$ represents her aspiration level. The frame f describes the order in which the alternatives are presented to the agent. Whenever the agent chooses from some non-empty subset $S \subseteq X$ (e.g. the budget set), she considers the alternatives in S sequentially in their order as prescribed by $f \in F = P$. She chooses the first alternative that exceeds her aspiration level, i.e., she picks from S whichever satisfactory alternative is presented first. If S turns out not to contain any satisfactory alternative, the agent recalls all alternatives in S and chooses the welfare-optimal one. Choices between satisfactory alternatives will thus always be in line with the order of presentation, while all other choices are in line with the welfare preference. Hence we can obtain $d(\succeq, f)$ from $\succeq$ by rearranging the top k elements according to their order in f.[6]

Model 2 (Limited Search). This model formalizes a choice heuristic similar to one described in Masatlioglu et al. (2012). When the agent looks for a product online, all alternatives in X are displayed by a search engine, but only k of them on the first result page and $m_X - k$ of them on the second result page. The frame f here is the set of $k \in \{1, \ldots, m_X - 1\}$ alternatives on the first page, such that F is the set of all size k subsets of X. The agent again chooses from non-empty subsets $S \subseteq X$ (e.g. not all displayed alternatives may be affordable to the agent or in stock with the retailer). Whenever the first result page contains at least one of the alternatives from S, then the agent does not even look at the second page but chooses from $S \cap f$ according to her welfare preference. Only if none of the elements of S is displayed on the first page, then the agent moves to the second page and chooses there according to her welfare preference. Choices between alternatives on the same page will thus always be in line with the welfare preference, but any available alternative on the first page is chosen over any alternative on the second page. Hence $d(\succeq, f)$ preserves $\succeq$ among all first and among all second page alternatives, but takes the first page to the top.[7]

The function d should be thought of as representing the regulator's conjecture about the relation between welfare, frames and choice. We consider the case of uncertainty about the behavioral model in Section 7, but for now we assume that the conjecture d is unique and given (keeping in mind that this assumption works in favor of nudging). Such a conjecture will typically rely on insights about the decision-making process and thus

ation functions" (p. 1291). The assumption rules out behavioral models in which choices violate standard axioms already when a frame is fixed. De Clippel and Rozen (2014) investigate the problem of learning from incomplete data sets without such an assumption.

[6] In contrast to RS, we explicitly treat the order of presentation as a variable frame. We also assume that the aspiration level k is fixed, which implies that the distortion function is single-valued.

[7] This model is also similar to the gradual accessibility model in Salant and Rubinstein (2008), but the eventual choice rule is different.

originates from non-choice data.[8] For instance, eye-tracking or the monitoring of browsing behaviors can provide the type of information necessary to substantiate a model like limited search (see the discussion in Masatlioglu et al., 2012), and methods from neuroscience may confirm decision-processes such as perfect-recall satisficing. As noted before, it is not our goal here to argue that a specific model is correct. Hence the only minor assumption that we impose on the behavioral model in general is that for each $\succeq \in P$ there exists an $f \in F$ such that $d(\succeq, f) = \succeq$. This rules out that some preferences are distorted by all possible frames and allows us to focus on the informational requirements of nudging, without having to deal with exogenously unavoidable distortions. The assumption does not imply the existence of a neutral frame that is non-distorting for all preferences.[9] In the satisficing model, all frames which present the k satisfying alternatives in their actual welfare order are non-distorting for that welfare preference. In the limited search model, the non-distorting frame places the k welfare-best alternatives on the first page.

Holding fixed a frame, the regulator now observes the agent's choices from sufficiently many different subsets $S \subseteq X$ to deduce her behavioral preference, in the usual revealed preference sense. Here the only difference to the usual approach is that the behavioral preference is not automatically equated with the welfare preference, and that the procedure generates potentially different revealed behavioral preferences when repeated for different frames. Formally, a data set is a subset $\Lambda \subseteq P \times F$, where $(\succeq', f') \in \Lambda$ means that the agent has been observed under frame f' and her choice behavior revealed the behavioral preference $\succeq'$. Further following RS, we say that $\succeq$ is consistent with data set Λ if for each $(\succeq', f') \in \Lambda$ it holds that $\succeq' = d(\succeq, f')$. In that case, $\succeq$ is a possible welfare preference because the data set could have been generated by an agent with that preference.[10] We illustrate the elicitation of the welfare preference, and also some first implications for nudging, using two examples.

Example 1. Consider an agent whose decision process is described by the perfect-recall satisficing model with aspiration level $k = 2$. The set of alternatives is given by

[8] Arguably, non-choice-based conjectures about the relation between choice and welfare always have to be invoked, even in standard welfare economics, see Kőszegi and Rabin (2007, 2008a) and Rubinstein and Salant (2008). For an opposing perspective and a critical discussion of the ability to identify the decision process, see Bernheim (2009).

[9] Sometimes a neutral or "revelatory" frame (Goldin, 2015, p. 9) may indeed exist, for example when the default can be removed from a choice problem. The existence of such a frame makes the welfare elicitation problem and also the nudging problem straightforward. Often, however, this solution is not available, e.g. defaults are unavoidable for organ donations, alternatives must always be presented in some order or arrangement, and questions must be phrased in one way or another.

[10] Formally, this framework corresponds to the extension in RS where behavioral data sets contain information about frames. It simplifies their setup by assuming that any pair of a welfare preference and a frame generates a unique distorted behavioral preference. This is not overly restrictive, as the different contingencies that generate a multiplicity of distorted preferences can always be written as different frames. It is restrictive in the sense that observability and controllability of these frames might not always be given. See Section 7.2 for the respective generalization.

$X = \{a, b, c, d\}$. The agent has the welfare preference $\succeq_1$ given by $c \succ_1 a \succ_1 b \succ_1 d$, so that alternatives c and a are satisfactory. Denote the frame which presents the alternatives in the alphabetical order by f. Thus, when choosing from some subset $S \subseteq X$, the agent will consider the alternatives in S in alphabetical order and choose the first which is satisfactory. Consequently, because a is presented before c, the agent will choose a whenever $a \in S$, even if also $c \in S$, in which case this is a mistake. She will choose c when $c \in S$ but $a \notin S$, and otherwise she will choose b over d by the perfect-recall assumption. Taken together, these choices look as if the agent was maximizing the preference $\succeq_2$ given by $a \succ_2 c \succ_2 b \succ_2 d$. Formally, we have $d(\succeq_1, f) = \succeq_2$. Suppose the behavioral preference $\succeq_2$ is observed in the standard revealed preference sense, by observing the agent's choices from different subsets $S \subseteq X$ but under the fixed frame of alphabetical presentation. Formally, the regulator obtains the data set $\Lambda = \{(\succeq_2, f)\}$. Given the perfect-recall satisficing conjecture, he can then conclude that the agent's welfare preference must be either $c \succ_1 a \succ_1 b \succ_1 d$ or $a \succ_2 c \succ_2 b \succ_2 d$; these two but no other welfare preferences generate the observed behavior under frame f. Formally, the set of preferences that are consistent with the data set is given by $\{\succeq_1, \succeq_2\}$. Therefore, with as little information as observing behavior under a single frame, the set of possible welfare preferences can be reduced from initially 24 to only 2.

We now illustrate some first implications for nudging, which here amounts to fixing an optimal order of presentation. Any order that presents a before c would be optimal if the agent's welfare preference was $a \succ_2 c \succ_2 b \succ_2 d$, but induces the above described decision mistake between a and c if the welfare preference is $c \succ_1 a \succ_1 b \succ_1 d$. The exact opposite is true for any order that presents c before a. Hence our knowledge is not yet enough to favor any one frame over another. Unfortunately, the problem cannot be solved by observing the agent under additional frames. The order of presentation fully determines choices among the alternatives a and c, so we can never learn about the welfare preference between the two. Since precisely this knowledge would be necessary to determine the optimal order, nudging here runs into irresolvable information problems.

Example 2. Consider an agent whose decision process is described by the limited search model, and $k = 2$ alternatives are presented on the first result page. As in the previous example, the set of alternatives is $X = \{a, b, c, d\}$ and the agent has the welfare preference $\succeq_1$ given by $c \succ_1 a \succ_1 b \succ_1 d$. Let $f = \{a, b\}$ denote the frame which puts the alternatives a and b on the first page. Thus, whenever the agent's choice set $S \subseteq X$ contains either a or b (or both), she will remain on the first page and make her choice there. Consequently, she chooses a whenever $a \in S$, even if also $c \in S$, because c is displayed only on the second page. This is again a mistake. She will choose b when $b \in S$ but $a \notin S$, and otherwise she will choose c over d. Taken together, these choices look as if the agent was maximizing the preference $\succeq_3$ given by $a \succ_3 b \succ_3 c \succ_3 d$. Formally, we have $d(\succeq_1, f) = \succeq_3$. Suppose

again that this behavioral preference is revealed, i.e., the regulator obtains the data set $\Lambda = \{(\succeq_3, f)\}$. Reversing the distortion process now unveils that the agent truly prefers a over b and c over d, which leaves the six possible welfare preferences marked in the first column of Table 1. The set of preferences consistent with the observed behavior is therefore given by $\{\succeq_1, \succeq_2, \succeq_3, \succeq_4, \succeq_5, \succeq_6\}$, meaning that the single observation reduces the set of possible welfare preferences from 24 to 6.

Here, an optimal nudge should place the two welfare-best alternatives on the first page, thus helping the agent avoid decision mistakes like the one between a and c under frame f above. Unfortunately, each of the four alternatives still belongs to the top two for at least one of the consistent welfare preferences, but none of them for all of the consistent welfare preferences. Hence no frame guarantees fewer mistakes than any other. In contrast to the satisficing example, however, gathering more information helps. Observing choices under frame $f' = \{a, d\}$ reveals the behavioral preference $\succeq_7$ given by $a \succ_7 d \succ_7 c \succ_7 b$, from which the welfare candidates marked in the second column of Table 1 can be deduced. Formally, adding this observation to the data set yields $\Lambda' = \{(\succeq_3, f), (\succeq_7, f')\}$, and the set of consistent welfare preferences shrinks to $\{\succeq_1, \succeq_2, \succeq_4, \succeq_5\}$. Note that these preferences all agree that a and c are the two best alternatives. Hence we know that $f'' = \{a, c\}$ is the optimal nudge. The actual welfare preference is still not known, so the example also shows that identifying a nudge is not the same problem as identifying the welfare preference.

Table 1: Reversing Limited Search

	$f = \{a, b\}$: $a \succ_3 b \succ_3 c \succ_3 d$	$f' = \{a, d\}$: $a \succ_7 d \succ_7 c \succ_7 b$
$c \succ_1 a \succ_1 b \succ_1 d$	✓	✓
$a \succ_2 c \succ_2 b \succ_2 d$	✓	✓
$a \succ_3 b \succ_3 c \succ_3 d$	✓	
$a \succ_4 c \succ_4 d \succ_4 b$	✓	✓
$c \succ_5 a \succ_5 d \succ_5 b$	✓	✓
$c \succ_6 d \succ_6 a \succ_6 b$	✓	
$a \succ_7 d \succ_7 c \succ_7 b$		✓
$c \succ_8 b \succ_8 a \succ_8 d$		✓

3 Nudgeability

3.1 Weakly Successful Nudge

In this section, we will provide a formal definition of a nudge. To capture the first step of preference elicitation due to RS in a concise way, let

$$\bar{\Lambda}(\succeq) = \{(d(\succeq, f), f) \mid f \in F\}$$

be the maximal data set that could be observed if the agent's welfare preference was $\succeq$, i.e., the data set that contains an observation for each possible frame. Then the set of all welfare preferences that are consistent with an arbitrary data set Λ can be written as

$$P(\Lambda) = \{\succeq \mid \Lambda \subseteq \bar{\Lambda}(\succeq)\}.$$

Without further mention, we consider only data sets Λ for which $P(\Lambda)$ is non-empty, i.e., for which there exists $\succeq$ such that $\Lambda \subseteq \bar{\Lambda}(\succeq)$. Otherwise, the behavioral model would be falsified by the data.[11] Observe that a frame f cannot appear more than once in such data sets. Observe also that $P(\varnothing) = P$ holds, and that $P(\Lambda) \subseteq P(\Lambda')$ whenever $\Lambda' \subseteq \Lambda$.

We are interested in evaluating the frames after having observed some data set Λ and having narrowed down the set of possible welfare preferences to $P(\Lambda)$. Since previously different frames may now have become behaviorally equivalent, let

$$[f]_\Lambda = \{f' \mid d(\succeq, f') = d(\succeq, f), \forall \succeq \in P(\Lambda)\}$$

be the equivalence class of frames for frame f, i.e., the elements of $[f]_\Lambda$ induce the same behavior as f for all of the remaining possible welfare preferences. We denote by $F(\Lambda) = \{[f]_\Lambda \mid f \in F\}$ the quotient set of all equivalence classes. Our central definition compares the elements of $F(\Lambda)$ pairwise from the perspective of the possible welfare preferences. For any $\succeq$ and any non-empty $S \subseteq X$, let $c(\succeq, S)$ be the element of S that would be chosen from S by an agent who maximizes $\succeq$.

Definition 1 *For any f, f' and Λ, $[f]_\Lambda$ is a weakly successful nudge over $[f']_\Lambda$, written*

$$[f]_\Lambda \; N(\Lambda) \; [f']_\Lambda,$$

if for each $\succeq \in P(\Lambda)$ it holds that $c(d(\succeq, f), S) \succeq c(d(\succeq, f'), S)$, for all non-empty $S \subseteq X$.

[11] RS derive conditions under which data sets do or do not falsify a model conjecture. A falsified model is of no use for the purpose of nudging and would have to be replaced by a conjecture for which $P(\Lambda)$ is non-empty.

The statement $[f]_\Lambda N(\Lambda)[f']_\Lambda$ means that the agent's choice under frame f (and all equivalent ones) is at least as good as under f' (and all equivalent ones), no matter which of the remaining welfare preferences is the true one. The welfare preferences enter the definition not only for the evaluation of choices, but also because agents with different welfare preferences react differently to frames. The binary nudging relation $N(\Lambda)$ shares with other approaches in behavioral welfare economics the property of requiring agreement among multiple preferences (see, for instance, the multiself Pareto interpretation of the unambiguous choice relation by Bernheim and Rangel, 2009), but the multiplicity of preferences here simply reflects lack of information (as in Masatlioglu et al., 2012). Thus, adding observations to a data set can only make the partition $F(\Lambda)$ coarser and the nudging relation more complete, because it can only reduce the set of possible welfare preferences for which improved choices have to be guaranteed. In fact, the only way in which the data set Λ matters for the binary nudging relation is via the set $P(\Lambda)$.

The following Lemma 1 summarizes additional properties of $N(\Lambda)$ that will be useful. It relies on the sets of ordered pairs $B(\succeq, f) = d(\succeq, f) \setminus \succeq$ which record all binary comparisons that are reversed from $\succeq$ by f.[12] For instance, in the satisficing example in the preceding section, where the welfare preference was given by $c \succ_1 a \succ_1 b \succ_1 d$ and alphabetical order of presentation f resulted in the behavioral preference $a \succ_2 c \succ_2 b \succ_2 d$, we would obtain $B(\succeq_1, f) = \succeq_2 \setminus \succeq_1 = \{(a, c)\}$. For the limited search example where frame $f = \{a, b\}$ distorted the same welfare preference to $a \succ_3 b \succ_3 c \succ_3 d$, we would obtain $B(\succeq_1, f) = \succeq_3 \setminus \succeq_1 = \{(a, c), (b, c)\}$.

Lemma 1 *(i)* $[f]_\Lambda N(\Lambda)[f']_\Lambda$ *if and only if* $B(\succeq, f) \subseteq B(\succeq, f')$ *for each* $\succeq \in P(\Lambda)$. *(ii)* $N(\Lambda)$ *is a partial order (reflexive, transitive, antisymmetric) on* $F(\Lambda)$.

The proof of the lemma (and all further results) can be found in Appendix A. Since $B(\succeq, f)$ describes all the mistakes in binary choice that frame f causes for welfare preference $\succeq$, statement (i) of the lemma formalizes the intuition that a successful nudge is a frame that guarantees fewer mistakes. Statement (ii) implies that the binary relation is sufficiently well-behaved to consider different notions of optimality.

3.2 Optimal Nudge

A benevolent regulator would ideally like to choose a frame that is a weakly successful nudge over all other frames and thus guarantees the best possible choices. We call such a frame an *optimal nudge*. Given a data set Λ, let

$$G(\Lambda) = \{f \mid [f]_\Lambda N(\Lambda)[f']_\Lambda, \forall f' \in F\}$$

[12] Even though we often represent preferences as rankings like $c \succ a \succ b \succ d$, we remind ourselves that technically both $d(\succeq, f)$ and $\succeq$ are subsets of the set of ordered pairs $X \times X$.

be the set of frames which have been identified as optimal. Formally, $G(\Lambda)$ coincides with the greatest element of the partially ordered set $F(\Lambda)$, and it might be empty due to incompleteness of the binary nudging relation. Since the nudging relation becomes more complete as we collect additional observations, it follows that optimal nudges are more likely to exist for larger data sets. Therefore, the following result provides a necessary and sufficient condition for the existence of an optimal nudge for maximal data sets. The result is relatively straightforward but important, as it will allow us to classify behavioral models according to whether the search for an optimal nudge is promising or hopeless.

Definition 2 *Preference $\succeq$ is identifiable if for each $\succeq' \in P$ with $\succeq' \neq \succeq$, there exists $f \in F$ such that $d(\succeq, f) \neq d(\succeq', f)$.*

Proposition 1 *$G(\bar{\Lambda}(\succeq))$ is non-empty if and only if $\succeq$ is identifiable.*

The if-statement is immediate: an identifiable welfare preference is known once the maximal data set has been collected, and all the non-distorting frames are optimal with that knowledge. It is worth emphasizing again, however, that the result does not imply that the welfare preference actually has to be learned perfectly for successful nudging. It only tells us that, if $\succeq$ is the true and identifiable welfare preference, then for some sufficiently large data set Λ we will be able to identify an optimal nudge. The set $P(\Lambda)$ might still contain more than one element at that point. The only-if-statement tells us that there is no hope to ever identify an optimal nudge if the welfare preference cannot be identified, i.e., if there exists another welfare preference $\succeq'$ that is behaviorally equivalent to $\succeq$ under all frames. In this case we say that $\succeq$ and $\succeq'$ are *indistinguishable*. A frame could then only be optimal if it does not distort any of the two, but this is impossible as such a frame would generate different observations for $\succeq$ and $\succeq'$ and hence would empirically discriminate between them.

In the following, we will make use of the result in Proposition 1 and consider the two polar classes of behavioral models where all welfare preferences are identifiable or non-identifiable, respectively. Before turning to a detailed analysis of these two classes, we address the question of how plausible each of them is. Our prime example for non-identifiable preferences is the perfect-recall satisficing model. Any two welfare preferences that are identical except that they rank the same best k alternatives differently are mapped into the same distorted preference by any frame, and hence are indistinguishable. Our prime example for identifiable preferences is the limited search model (for $m_X \geq 3$). There, we learn the welfare preference among all alternatives on the same page, and thus we can identify the complete welfare preference by observing behavior under sufficiently many different frames. The decision processes formalized by these two models are both plausible, implying that both classes are important. Another way of looking at the question of

plausibility is to ask how many models belong to each of the classes. We can provide an answer to this question for the limiting case as the number of alternatives grows large.[13] With m_X alternatives, there are $m_P(m_X) = m_X!$ strict preferences. The number of models also depends on how many frames $m_F(m_X)$ we allow, as a function of the number of alternatives. This number should typically be increasing in m_X, but for the following result we only need to assume that $m_F(m_X) \geq 4$ for sufficiently large values of m_X.[14]

Proposition 2 *The share of models with identifiable preferences goes to 1 as* $m_X \to \infty$.

The proof exploits the fact that the number of models with identifiable preferences is given by the number of different ways to assign distinct maximal data sets to the welfare preferences, satisfying the requirement that there must exist a non-distorting frame for each preference. It is difficult to determine this number exactly, but we find a lower bound that is tractable and suffices to show that the share of models with identifiable preferences converges to 1 as the number of alternatives grows. If one accepts that the genericity notion formalized by Proposition 2 captures model plausibility in a meaningful way, the result is good news for the nudging project. If the number of alternatives is large, an optimal nudge can generically be identified. However, we need to add that the complexity of finding this optimal nudge may become prohibitive if m_X is large, a problem to which we will return in Section 5.

4 Non-Identifiable Preferences

We now investigate behavioral models with non-identifiable preferences more thoroughly. From Proposition 1 we know that an optimal nudge cannot be found for these models. However, our previous notion of optimality was strong, requiring an optimal frame to outperform all other frames. Even if such a frame does not exist, we might still be able to exclude some frames that are dominated by others. We now weaken optimality to the requirement that a reasonable frame should not be dominated. Let

$$M(\Lambda) = \{f \mid [f']_\Lambda N(\Lambda)[f]_\Lambda \text{ only if } f' \in [f]_\Lambda\}$$

be the (always non-empty) set of frames which are undominated, based on our knowledge from the data set Λ. Formally, $M(\Lambda)$ is the union of all elements that are maximal in

[13] This approach of quantifying plausibility is similar to Kalai et al. (2002), who are interested in the number of preferences that are necessary to rationalize an arbitrary choice function. They show that the share of choice functions which can be rationalized by less than the maximal conceivable number of preferences goes to 0 as the number of alternatives grows large.

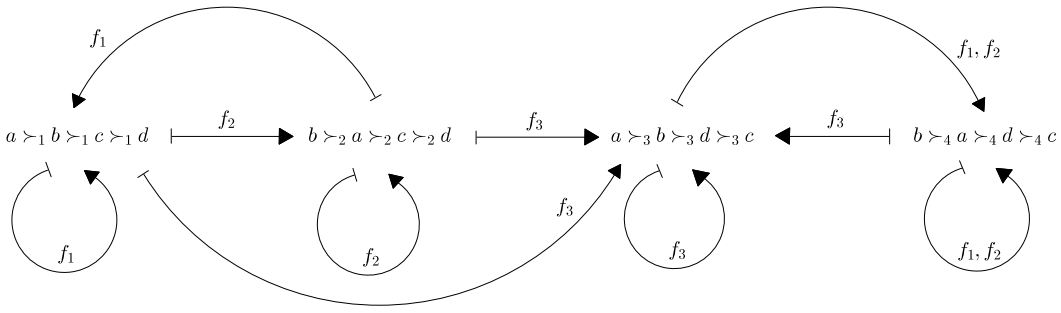
[14] If we restricted attention to models where frames are orders of presentation, we would already obtain $m_F(m_X) = m_X!$. In general, the number of frames can be arbitrarily large. However, there can never be more than $m_F(m_X) = m_X!^{m_X!}$ different non-equivalent frames, the number of mappings from P to P.

the partially ordered set $F(\Lambda)$. To provide an analogy, we can think of $M(\Lambda)$ as the set of Pareto efficient policies, because moving away from any $f \in M(\Lambda)$ makes the agent better off with respect to some $\succeq \in P(\Lambda)$ only at the cost of making her worse off with respect to some other $\succeq' \in P(\Lambda)$. By the same token, a frame which is not in $M(\Lambda)$ can be safely excluded, as there exists a nudge that guarantees an improvement over it.

Dominated frames can exist already ex ante with no knowledge of the agent's welfare preference. For instance, certain informational arrangements could be interpreted as being dominant over others, because they objectively clarify the available information and improve the decision quality (e.g. Camerer et al., 2003). In the following example we show that ex ante undominated frames can become dominated for richer knowledge, too.

Example 3. Assume that $X = \{a, b, c, d\}$ and consider the distortion function for the four preferences and three frames depicted in Figure 1.[15] The two preferences $\succeq_1$ and $\succeq_2$ are indistinguishable, as each frame maps them into the same distorted preference, and the same holds for $\succeq_3$ and $\succeq_4$. Note also that none of the frames is dominated before any data has been collected, $M(\varnothing) = \{f_1, f_2, f_3\}$, because each one is the unique non-distorting frame for one possible welfare preference. Now suppose we observe $\Lambda = \{(\succeq_2, f_2)\}$, so that $P(\Lambda) = \{\succeq_1, \succeq_2\}$. It follows immediately that none of the potentially non-distorting frames f_1 and f_2 is dominated. The frame f_3, however, is now dominated by f_1. If the welfare preference is $\succeq_2$, then f_1 induces a mistake between a and b, but so does f_3, which induces an additional mistake between c and d. Hence we obtain $M(\Lambda) = \{f_1, f_2\}$. We have learned enough to identify a nudge over f_3, but no additional observation will ever allow us to compare f_1 and f_2.

Figure 1: Dominated Frame f_3



The sometimes dominated frame f_3 in Example 3 has a particular property. It maps the indistinguishable set of preferences $\{\succeq_1, \succeq_2\}$ outside of itself. This is the reason why

[15] The example focusses on only four welfare preferences, but it can be expanded to encompass the set of all possible preferences. We can also add additional frames without changing its insight.

the example violates the following property.

Definition 3 *A distortion function d has the frame-cancellation property if*

$$d(d(\succeq, f_1), f_2) = d(\succeq, f_2)$$

holds for all $\succeq \in P$ and all $f_1, f_2 \in F$.

With the frame-cancellation property, the impact of any frame f_1 disappears once a new frame f_2 is applied. Starting from any welfare preference $\succeq$, the preference $d(\succeq, f)$ obtained by applying any frame $f \in F$ is then always observationally equivalent to $\succeq$, and thus is itself an indistinguishable welfare preference. Hence, for any given frame, all maximal indistinguishable sets of preferences are closed under the distortion function, in contrast to Example 3.

A variety of interesting behavioral models has the frame-cancellation property. One extreme example, where frames never have an effect on behavior and $d(\succeq, f) = \succeq$ always holds, is the rational choice model.[16] The opposite extreme case of frame-cancellation arises when $d(\succeq, f)$ is independent of $\succeq$, so that frames override the preference entirely. This is true, for instance, when there are only two alternatives and the agent always chooses the one that is marked as the default. The perfect-recall satisficing model has the frame-cancellation property, too, even though the welfare preference retains a substantial impact on behavior. In this model, the effect of the order of presentation is to overwrite the welfare preference among the top k alternatives, which leaves no trace of previous frames when done successively. We can also establish a connection to the analysis of choice from lists by Rubinstein and Salant (2006). They allow for the possibility that agents choose from lists instead of sets, i.e., the choice from a given set of alternatives can be different when the alternatives are listed differently. Their results imply that we can capture choice from list behavior in reduced form of a distortion function whenever the axiom of "partition independence" is satisfied by the agent's choices for all possible welfare preferences.[17] An example in which this holds is satisficing without recall. In contrast to the perfect-recall version, the agent here chooses the last alternative on a list when no alternative on the list exceeds her aspiration level. Formally, $d(\succeq, f)$ is obtained

[16]Note that all welfare preferences are identifiable in the rational choice model, which constitutes of course the basis for the standard revealed preference approach. The rational choice model is indeed the only model which has both identifiable preferences and the frame-cancellation property. To see why, suppose d is not fully rational, i.e., there exist $\succeq'$ and f' such that $d(\succeq', f') = \succeq'' \neq \succeq'$. If d has the frame-cancellation property we then obtain $d(\succeq'', f) = d(d(\succeq', f'), f) = d(\succeq', f)$ for all $f \in F$, hence $\succeq'$ and $\succeq''$ are indistinguishable and not identifiable.

[17]Partition independence requires that the choice from two concatenated sublists is the same as the choice from the list that concatenates the two elements chosen from the sublists (Rubinstein and Salant, 2006, p. 7). Such behavior can be modelled as the maximization of some non-strict preference that is turned strict by ordering its indifference sets in or against the list order (Proposition 2, p. 8).

from $\succeq$ by rearranging the top k elements in the order of f and the bottom $m_X - k$ elements in the opposite order of f (see RS). It is easy to verify that this model also has the frame-cancellation property. The following general class of decision processes nests all these models with the frame-cancellation property.

Model 3 (Limited Sensitivity). The agent displays limited sensitivity in the sense that she can sometimes not tell whether an alternative is actually better than another. Degree and allocation of sensitivity are described by a vector $(k_1, k_2, \ldots, k_s)$ of positive integers with $\sum_{i=1}^{s} k_i = m_X$. A welfare preference $\succeq$ induces a partition of X, where block X_1 contains the k_1 welfare-best alternatives, X_2 contains the k_2 next best alternatives, and so on. The agent can distinguish alternatives across but not within blocks. When choosing from $S \subseteq X$, she therefore only identifies the smallest i for which $S \cap X_i$ is non-empty, and the frame then fully determines the choice from this set. Thus $d(\succeq, f)$ is obtained from $\succeq$ by rearranging the alternatives within each block of the partition in a way that does not depend on their actual welfare ranking. Formally, let $P_{\succeq}$ be the set of welfare preferences that induce the same partition of X as $\succeq$, for any $\succeq \in P$. Then $d(\succeq', f) = d(\succeq'', f) \in P_{\succeq}$ must hold whenever $\succeq', \succeq'' \in P_{\succeq}$, for all $f \in F$. Any such function satisfies the frame-cancellation property.[18] When f is an order of presentation and the alternatives within each block of the partition are rearranged in or against this order – because the agent chooses the first or the last among seemingly equivalent alternatives – then the process is a successive choice from list model (see Rubinstein and Salant, 2006, for a definition). Special cases include rational choice for the vector $(k_1, k_2, \ldots, k_s) = (1, 1, \ldots, 1)$, perfect-recall satisficing for $(k, 1, \ldots, 1)$, no-recall satisficing for $(k, m_X - k)$, and situations where the welfare preference has no impact on behavior for $k_1 = m_X$.

The following result shows that there are never any dominated frames for models with the frame-cancellation property.

Proposition 3 *If d has the frame-cancellation property, then $M(\Lambda) = F$ for all Λ.*

If the frame-cancellation property holds, then irrespective of how many data points we have collected, we will never know enough to improve upon any given frame. According to our earlier analogy, all frames are always Pareto efficient. If we want to select between them, we need to resort to approaches that can be used to compare Pareto efficient allocations, involving stronger assumptions such as probabilistic beliefs (see Section 6.2).

[18] For any $\succeq \in P$, since $\succeq \in P_{\succeq}$ holds, we have $d(\succeq, f_1) \in P_{\succeq}$ for any $f_1 \in F$. Then we also obtain $d(d(\succeq, f_1), f_2) = d(\succeq, f_2)$ for any $f_2 \in F$, which is the frame-cancellation property. We note that there are models with the frame-cancellation property that do not belong to the class of limited sensitivity models. Any model with frame-cancellation allows us to partition P into maximal indistinguishable sets of preferences, very similar to the sets $P_{\succeq}$ in the limited sensitivity model, but these sets will not in general be generated by some vector $(k_1, k_2, \ldots, k_s)$ as required by the limited sensitivity model.

5 Identifiable Preferences

We now turn to models with identifiable welfare preferences, which guarantee knowledge of an optimal nudge once a maximal data set has been observed. Collecting a maximal data set requires observing the agent under all m_F frames, however, which might be beyond our means. We are thus interested in optimal data gathering procedures and the required quantity of information. The idea is that a regulator, who ultimately seeks to impose the optimal nudge, is also able to impose a specific sequence of frames on the agent, with the goal of eliciting the welfare preference.

For each $s \in \{0, 1, \ldots, m_F\}$, let

$$L_s = \{\Lambda | P(\Lambda) \neq \varnothing \text{ and } |\Lambda| = s\}$$

be the collection of data sets that do not falsify the behavioral model and contain exactly s observations, i.e., observations for s different frames. In particular, $L_0 = \{\varnothing\}$, and L_{m_F} consists of all maximal data sets. Then $L = L_0 \cup L_1 \cup \ldots \cup L_{m_F-1}$ is the collection of all possible data sets except the maximal ones. An elicitation procedure dictates for each of these data sets a yet unobserved frame, under which the agent is to be observed next.

Definition 4 *An elicitation procedure is a mapping $e : L \to F$ with the property that, for each $\Lambda \in L$, there does not exist $(\succeq, f) \in \Lambda$ such that $e(\Lambda) = f$.*

A procedure e starts with the frame $e(\varnothing)$ and, if the welfare preference is $\succeq$, generates the first data set $\Lambda_1(e, \succeq) = \{(d(\succeq, e(\varnothing)), e(\varnothing))\}$. It then dictates the different frame $e(\Lambda_1(e, \succeq))$ and generates a larger data set $\Lambda_2(e, \succeq)$ by adding the resulting observation. This yields a sequence of expanding data sets described recursively by $\Lambda_0(e, \succeq) = \varnothing$ and

$$\Lambda_{s+1}(e, \succeq) = \Lambda_s(e, \succeq) \cup \{(d(\succeq, e(\Lambda_s(e, \succeq))), e(\Lambda_s(e, \succeq)))\},$$

until the maximal data set $\Lambda_{m_F}(e, \succeq) = \bar{\Lambda}(\succeq)$ is reached. Hence all elicitation procedures deliver the same outcome after m_F steps, but typically differ at earlier stages. A procedure does not use any exogenous information about the welfare preference, but the frame to be dictated next can depend on the information generated endogenously by the growing data set.[19]

We now define the complexity n of the nudging problem as the number of steps that the quickest elicitation procedure requires until it identifies an optimal nudge for sure.

[19]Notice that an elicitation procedure dictates frames also for pre-collected data sets that itself never generates. We tolerate this redundancy because otherwise definitions and proofs would become substantially more complicated, at no gain.

Formally, let

$$n(e, \succeq) = \min\{s \mid G(\Lambda_s(e, \succeq)) \neq \varnothing\}$$

denote the first step at which e identifies an optimal nudge if the welfare preference is $\succeq$. Since this preference is unknown, e guarantees a result for sure only after $\max_{\succeq \in P} n(e, \succeq)$ steps. With E denoting the set of all elicitation procedures, we have to be prepared to gather

$$n = \min_{e \in E} \max_{\succeq \in P} n(e, \succeq)$$

data points before we can nudge successfully.

To illustrate the concepts, we first consider the limited search model (assuming $m_X \geq 3$ to make all preferences identifiable). The following result shows that learning and nudging are relatively simple in this model.

Proposition 4 *For any $m_X \geq 3$, the limited search model satisfies*

$$n = \begin{cases} 3 & \text{if } k = m_X/2 \text{ and } k \text{ is odd,} \\ 2 & \text{otherwise.} \end{cases}$$

To understand our construction of an optimal elicitation procedure for the limited search model, consider again Example 2. The procedure starts with an arbitrary frame, $f = \{a, b\}$, and generates the behavioral preference $a \succ_3 b \succ_3 c \succ_3 d$. We now know that the welfare preference satisfies $a \succ b$ and $c \succ d$. The second frame is constructed by taking the top element from f and the bottom element from $X \backslash f$, which yields $f' = \{a, d\}$. From the induced behavioral preference $a \succ_7 d \succ_7 c \succ_7 b$ we learn that $a \succ d$ and $c \succ b$. This information is enough to deduce that a and c are the two welfare-optimal alternatives, because both b and d are worse than each of them. If instead at the second step we had learned that $a \succ d$ and $b \succ c$, we could have concluded that a and b are optimal. If we had learned that $d \succ a$, we could have concluded that c and d are optimal. This argument can be generalized. If $k = m_X/2$ and k is even, for instance, the second frame is constructed to contain the $k/2$ best alternatives from the previous first result page and the $k/2$ worst alternatives from the previous second result page. It can be shown that the k welfare-best alternatives can always be deduced from the resulting data set.

The nudging complexity is surprisingly small for the limited search model. This begs the question to what extent it is representative for more general models. It obviously always holds that $n \leq m_F$ if all welfare preferences are identifiable, but the number of frames m_F can be extremely large (see footnote 14). We therefore derive a tighter bound on n next. The result rests on the insight that there is always an elicitation procedure that

guarantees a reduction of the set of possible welfare preferences at each step. Since there are $m_P(m_X) = m_X!$ different welfare preferences that the agent might have ex ante, an elicitation procedure that reduces the set of possible preferences at each step guarantees identification of the preference and the optimal nudge after at most $m_X! - 1$ steps.

Proposition 5 *Any behavioral model with identifiable preferences satisfies* $n \leq m_X! - 1$.

It turns out that the bound presented in Proposition 5 is tight, because there are models for which it is reached. We illustrate this with the following model, which describes an, admittedly, strong effect of framing.

Model 4 (Strong Priming). The framing of the decision problem suggests that there is a unique proper way of deciding (e.g. priming, persuasion, demand effects). Formally, a frame $f \in F = P$ is identified with the preference that it conveys as being the proper behavior. The effect of the frame is strong, in the sense that the agent can be manipulated to behave in the suggested way whenever there is at least some agreement between the suggestion and the welfare preference. Manipulation fails only when the agent's welfare preference is exactly opposite of the suggestion. In this case the agent behaves in an arbitrarily distorted way that uniquely identifies him. For any $\succeq \in P$, let $o(\succeq)$ denote the opposite order of $\succeq$. Assume $m_X \geq 3$ and let $b : P \to P$ be a bijective mapping such that $b(\succeq) \notin \{\succeq, o(\succeq)\}$, for all $\succeq \in P$. Then

$$d(\succeq, f) = \begin{cases} f & \text{if } f \neq o(\succeq), \\ b(\succeq) & \text{if } f = o(\succeq). \end{cases}$$

Proposition 6 *The strong priming model satisfies* $n = m_X! - 1$.

In the strong priming model, identification of the optimal nudge actually requires identification of the welfare preference, because each frame is optimal only for exactly one welfare preference. This takes all $m_X! - 1$ steps, because observation of behavior under a frame either reveals a specific welfare preference to be the true one, or it excludes it from the set of possible welfare preferences. No matter in which order frames are dictated by the elicitation procedure, it is always possible that the agent's welfare preference is the one not revealed until the end. Hence, learning is particularly slow in this model.

Taken together, Propositions 5 and 6 are bad news for nudging. The tight bound on n grows more than exponentially in the number of alternatives. Thus, nudging may quickly become infeasible despite the general identifiability of preferences.

6 Nudging with Additional Information

6.1 Restricted Domains

Throughout the previous analysis we have maintained the assumption that the regulator considers all preferences over the set of alternatives feasible. In some situations, however, the regulator may be able to rule out certain preferences beforehand using non-choice information. For instance, criteria such as non-satiation or an agreement with some objective dimension of the alternatives may sometimes be uncontroversial and reduce the set of plausible preferences. We can model situations in which some welfare preferences are excluded from the outset by restricting the domain of preferences to some non-empty $\tilde{P} \subseteq P$. We then replace the set $P(\Lambda)$ of welfare preferences that are consistent with data set Λ by $\tilde{P}(\Lambda) = P(\Lambda) \cap \tilde{P}$. Based on this modified definition, all further concepts remain unchanged. We will explore two different implications of such domain restrictions. First, models with non-identifiable preferences may become identifiable. Second, the complexity of the elicitation procedure can be reduced.

We extend Definition 2 by saying that a preference $\succeq \in \tilde{P}$ is identifiable on $\tilde{P}$ if for each $\succeq' \in \tilde{P}$ with $\succeq' \neq \succeq$, there exists $f \in F$ such that $d(\succeq, f) \neq d(\succeq', f)$. It then follows exactly as for Proposition 1 that $G(\bar{\Lambda}(\succeq))$ is non-empty if and only if $\succeq$ is identifiable on $\tilde{P}$. Hence we will call $\tilde{P}$ a *nudging domain* if all its elements are identifiable on $\tilde{P}$. The universal domain P is a nudging domain if and only if all welfare preferences are identifiable as defined previously. To characterize nudging domains more generally, let

$$P_{\succeq} = \{\succeq' \in P \,|\, d(\succeq', f) = d(\succeq, f), \forall f \in F\}$$

be the equivalence class of welfare preferences that are indistinguishable from $\succeq$, and denote by $\bar{P} = \{P_{\succeq} \mid \succeq \in P\}$ the set of all these equivalence classes, which form a partition of P. Then it follows that $\tilde{P}$ is a nudging domain if and only if $|\tilde{P} \cap P_{\succeq}| \leq 1$ for all $\succeq \in P$, i.e., the domain $\tilde{P}$ can contain at most one element from each of the behaviorally equivalent classes of preferences.

Unfortunately, this may not be a particularly plausible or easily justifiable requirement. Consider the perfect-recall satisficing model. The set $P_{\succeq}$ contains all preferences which agree with $\succeq$ with respect to the bottom $m_X - k$ alternatives and their ranking. Hence, the restriction necessary to obtain identifiable preferences is that for each selection and ordering of the bottom $m_X - k$ alternatives, there exists at most one preference in $\tilde{P}$. Put differently, the preference over the bottom alternatives must fully determine the preference over all alternatives. This is very different from often studied domain restrictions such as single-peaked preferences (which do not constitute a nudging domain for the satisficing model or any of the other models with non-identifiable preferences studied in this paper).

The extent to which the domain needs to be restricted can still be interpreted as a measure of the degree of non-identifiability of a model. Different models may be unambiguously comparable by their demand for exogenous information. We show this for the satisficing models with perfect recall and no recall.

Proposition 7 *Any nudging domain for no-recall satisficing is also a nudging domain for perfect-recall satisficing, while the converse is not true whenever* $k < m_X - 1$.

A satisficer with no recall is harder to nudge than a satisficer with perfect recall (whenever the two are behaviorally different), because more knowledge about the welfare preference is necessary and less can be learned from behavior. As a general rule, a model comparison as in Proposition 7 is possible whenever the partition $\bar{P}$ is finer for one model than for another.

Let us now consider the effect of domain restrictions on the complexity of nudging. Whenever $\tilde{P}$ is a nudging domain for model d, we can adapt our previous definition of complexity to

$$\tilde{n} = \min_{e \in E} \max_{\succeq \in \tilde{P}} n(e, \succeq),$$

no matter whether or not d has identifiable preferences on the universal domain P. We obtain the following generalization of Propositions 5 and 6, which shows that the modified complexity bound mirrors the amount of non-choice information we put in.

Proposition 8 *Any behavioral model on a nudging domain* $\tilde{P}$ *satisfies* $\tilde{n} \leq |\tilde{P}| - 1$. *The strong priming model satisfies* $\tilde{n} = |\tilde{P}| - 1$.

6.2 Probabilistic Beliefs

We now explore the possibility to introduce non-choice-based prior information in the form of probabilistic beliefs. We assume that the regulator has a prior belief p over the set of welfare preferences P. We number the preferences in the order of their prior probabilities, so that $P = \{\succeq_1, \succeq_2, \ldots, \succeq_{m_X!}\}$ with $p_1 \geq p_2 \geq \ldots \geq p_{m_X!} > 0$.[20] Beliefs can be utilized in different ways. A first possibility is to replace our previous notion of complexity n by the expected complexity

$$\bar{n} = \min_{e \in E} \sum_{i=1}^{m_X!} p_i n(e, \succeq_i).$$

[20]In contrast to the previous subsection, here we make the full support assumption $p_{m_X!} > 0$. This is for simplicity and allows us to circumvent technical issues with Bayesian updating which would otherwise require a redefinition of elicitation procedures.

While n was the minimal number of observations necessary to guarantee identification of an optimal nudge, $\bar{n}$ can be thought of as the average running time of the quickest elicitation procedure. Note that different procedures may be required to achieve n or $\bar{n}$, respectively. The following result provides the expected complexity for the strong priming model, which we used before to illustrate the potentially large complexity of the elicitation problem.

Proposition 9 *The strong priming model satisfies*

$$\bar{n} = \sum_{i=1}^{m_X!-1} p_i i + p_{m_X!}(m_X! - 1).$$

At any given step, the elicitation procedure that has not yet identified the optimal nudge should always try to verify or exclude the remaining welfare preference with the *highest* belief probability, by prescribing the frame that corresponds to the opposite of this preference. The elicitation process then concludes with highest possible probability at every step, which gives rise to the formula in the proposition.

The complexity $\bar{n}$ and its behavior for large m_X depend on the shape of prior beliefs. As an example of a relatively informative prior, consider a (truncated) geometric distribution where the prior probabilities are given by

$$p_i = \rho^{i-1}\left(\frac{1-\rho}{1-\rho^{m_X!}}\right)$$

for some parameter $\rho \in (0,1)$. In Appendix B we show that $\lim_{m_X \to \infty} \bar{n} = 1/(1-\rho)$ holds for this distribution. The expected complexity thus remains bounded as the number of alternatives grows, and it may be small if ρ is small. On the other hand, for a uniform prior, where $p_i = 1/m_X!$, we show that $\bar{n}$ is still of the same order of magnitude as the previous $n = m_X! - 1$ and thus grows more than exponentially in the number of alternatives.

Let us therefore consider a second way in which belief-dependent complexity could be defined. In particular, we introduce a probabilistic notion of optimality of a nudge. Let $\pi_\Lambda(\succeq)$ denote the updated belief probability that the regulator attaches to welfare preference $\succeq$ when the data set Λ has been collected. We thus have $\pi_\varnothing(\succeq_i) = p_i$ and can apply Bayesian updating to obtain

$$\pi_\Lambda(\succeq) = \begin{cases} \pi_\varnothing(\succeq) / \left(\sum_{\succeq' \in P(\Lambda)} \pi_\varnothing(\succeq')\right) & \text{if } \succeq \in P(\Lambda), \\ 0 & \text{otherwise,} \end{cases}$$

for all data sets Λ with $P(\Lambda) \neq \varnothing$. In addition to narrowing down the set of possible welfare preferences, collecting data magnifies differences in prior beliefs on $P(\Lambda)$. Now

let $\varphi_\Lambda(f)$ denote the probability that frame f is an optimal nudge. With the definition of $P(\Lambda, f) = \{\succeq \in P(\Lambda) \mid d(\succeq, f) = \succeq\}$ as the set of remaining preferences for which f is non-distorting, we can calculate

$$\varphi_\Lambda(f) = \sum_{\succeq \in P(\Lambda, f)} \pi_\Lambda(\succeq).$$

We will denote by $\bar{\varphi}_\Lambda = \max_{f \in F} \varphi_\Lambda(f)$ the confidence that an optimally chosen frame induces non-distorted behavior.

From our previous arguments we obtain that $\varphi_\Lambda(f) = 1$ if and only if $f \in G(\Lambda)$. Hence the complexity n was based on the requirement that we want to ensure complete confidence, $\bar{\varphi}_\Lambda = 1$. We may now content ourselves with identifying a frame that is optimal with a sufficiently large probability $q \in (0, 1]$. The optimal elicitation procedure then is the one that guarantees a level $\bar{\varphi}_\Lambda \geq q$ as quickly as possible. This is captured by the generalized definition $n(q, e, \succeq) = \min\{s \mid \bar{\varphi}_{\Lambda_s(e,\succeq)} \geq q\}$ and

$$n(q) = \min_{e \in E} \max_{\succeq \in P} n(q, e, \succeq).$$

The following result provides the generalized complexity for the strong priming model.

Proposition 10 *The strong priming model satisfies that $n(q)$ is the smallest integer $k \geq 0$ for which*

$$\sum_{j=1+k}^{m_X!-1} p_{j+1} \leq p_1 \left(\frac{1-q}{q} \right).$$

At any given step, a generalized optimal procedure that has not yet identified the optimal nudge should always try to verify or exclude the remaining welfare preference with the *second-highest* belief probability, by prescribing the frame that corresponds to the opposite of this preference. It can always occur that the procedure still does not identify the welfare preference, but in that case it guarantees maximal posterior beliefs.

The result implies our previous result for n when we consider the limit as $q \to 1$, i.e., for large enough q we always obtain $n(q) = m_X! - 1$. With a uniform prior, for instance, we can rearrange the condition in Proposition 10 to obtain the ceiling

$$n(q) = \max \left\{ \left\lceil (m_X! - 1) - \left(\frac{1-q}{q} \right) \right\rceil, 0 \right\},$$

as shown in Appendix B. This implies $n(q) = m_X! - 1$ whenever $q > 1/2$. The uniform prior can be interpreted as the criterion of counting the welfare preferences for which a given frame is optimal. The result thus shows that the previous complexity bound remains

the same as long as we require optimality for a strict majority of welfare preferences. On the other hand, the combination of an informative prior belief and a low degree of required confidence can reduce the complexity of the nudging problem substantially. An extreme case is $p_1 \geq q$, so that the prior belief already provides sufficient confidence and $n(q) = 0$ follows. For the geometric distribution, we show that the generalized complexity remains bounded as the number of alternatives grows whenever $q < 1$.

The complexity $n(q)$ is also interesting for models with non-identifiable preferences. Our previous results imply that we can never achieve $\bar{\varphi}_\Lambda = 1$ for such models, but we may be able to achieve $\bar{\varphi}_\Lambda \geq q$ when q is sufficiently small. We can easily extend our definition of $n(q)$ to the case of non-identifiable preferences, by defining $n(q, e, \succeq) = \infty$ whenever elicitation procedure e never achieves a confidence of q or larger when $\succeq$ is the true welfare preference, i.e., when

$$q > \overline{q}(e, \succeq) = \max_{s \in \{0,\ldots,m_F\}} \bar{\varphi}_{\Lambda_s(e,\succeq)}.$$

Let $\underline{q} = \bar{\varphi}_\varnothing$ denote the prior confidence and $\overline{q} = \max_{e \in E} \min_{\succeq \in P} \overline{q}(e, \succeq)$ the maximal confidence that an optimal procedure can guarantee. We thus have $0 < \underline{q} \leq \overline{q} < 1$ and

$$n(q) \in \begin{cases} \{0\} & \text{if } q \leq \underline{q}, \\ \{1, \ldots, m_X! - 1\} & \text{if } \underline{q} < q \leq \overline{q}, \\ \{\infty\} & \text{if } \overline{q} < q. \end{cases}^{21}$$

For models with the frame-cancellation property, we can make the following more precise statement.

Proposition 11 *If d has the frame-cancellation property, then*

$$n(q) = \begin{cases} 0 & \text{if } q \leq \underline{q}, \\ 1 & \text{if } \underline{q} < q \leq \overline{q}, \\ \infty & \text{if } \overline{q} < q. \end{cases}$$

As shown previously, models with the frame-cancellation property stand out because the scope of learning about optimal nudges is particularly limited. Proposition 11 shows that, at the same time, the speed of learning is particularly fast for these models. All that can be learned under the frame-cancellation property is learned already after a single observation. Hence if we are willing to reduce our confidence aspiration, models with the frame-cancellation property stand out because of the simplicity of learning.

[21] The fact that $n(q) \leq m_X! - 1$ when $\underline{q} < q \leq \overline{q}$ follows as for Proposition 5. There is always an elicitation procedure that guarantees a reduction of the set of possible welfare preferences at each of the initial steps, until a further reduction is no longer possible at all. Thus the entire range of achievable confidence levels can be achieved during the first $m_X! - 1$ steps.

We conclude by pointing out two potential pitfalls of this result. First, significant learning will only be possible if there is significant information already in the prior beliefs. Second, achieving a satisfactory level of confidence is not tantamount to having a good guidance in the choice between frames. We illustrate this by studying the case of a uniform prior distribution. Let $\bar{s}$ denote the average size of the elements of partition $\bar{P}$, i.e., the average number of preferences in an indistinguishable equivalence class.

Proposition 12 *If d has the frame-cancellation property and prior beliefs are uniform, then $\underline{q} = \overline{q} = 1/\overline{s}$. Furthermore, $\varphi_\Lambda(f) = \varphi_\Lambda(f')$ for all $f, f' \in F$ and all Λ.*

With a uniform prior, no procedure guarantees to increase the prior confidence. In addition, all frames are always equally likely to be optimal. A regulator can thus never do better than by choosing a frame at random.

7 Nudging with Limited Information

7.1 Model Uncertainty

We have so far assumed that there is a unique conjecture about the behavioral model, while it may be more appropriate to assume that the regulator considers a number of different models possible. We can replace the assumption of a unique behavioral model by the assumption that the regulator considers any distortion function $d \in D$ possible, where D is a given set of conjectures. For instance, there could be uncertainty about the aspiration level of a satisficer, or one of the models in D could be the rational agent.[22] As a consequence, we no longer have to learn about the welfare preference only, but about the pair $(d, \succeq) \in D \times P$ of the distortion function and the welfare preference.[23]

Let $\bar{\Lambda}(d, \succeq) = \{(d(\succeq, f), f) \mid f \in F\}$ denote the maximal data set generated by the pair $(d, \succeq)$. Then the set of pairs $(d, \succeq)$ that are consistent with data set Λ is $DP(\Lambda) = \{(d, \succeq) \mid \Lambda \subseteq \bar{\Lambda}(d, \succeq)\}$. We again assume that $DP(\Lambda)$ is non-empty, i.e., at least one conjecture is not falsified by the data. Once we have narrowed down the set of model-preference pairs to $DP(\Lambda)$, we obtain the equivalence class of frame f by $[f]_\Lambda = \{f' \mid d(\succeq, f) = d(\succeq, f'),\ \forall (d, \succeq) \in DP(\Lambda)\}$. We can then modify our definition of the binary nudging relation in a natural way, taking into account that both model and welfare preference are unknown. In particular, we define $[f]_\Lambda \ N(\Lambda) \ [f']_\Lambda$ if for each $(d, \succeq) \in DP(\Lambda)$ it holds that $c(d(\succeq, f), S) \succeq c(d(\succeq, f'), S)$ for all non-empty $S \subseteq X$, so that for each remaining behavioral model the agent's choice under frame f is at least as

[22] It is central to the idea of asymmetric paternalism (Camerer et al., 2003) that there are different types of agents, some of which are rational and should not be restricted by regulation.

[23] We continue to assume that there is a non-distorting frame for each pair $(d, \succeq)$, which will typically depend both on the model and on the welfare preference.

good as under f', no matter which of the welfare preferences that are consistent with the behavioral model and the data set is the true one.

We are again interested in the existence of an optimal nudge. By the same reasoning as in Section 3, we consider maximal data sets only. An immediate extension of Definition 2 could require identifiability of $\succeq$ in d, for a given pair $(d, \succeq)$. This property is in fact necessary but no longer sufficient for the existence of an optimal nudge. It rules out that the maximal data set $\bar{\Lambda}(d, \succeq)$ could have been generated by a different welfare preference $\succeq'$ and the same model d, but it does not rule out that it could have been generated by a different welfare preference $\succeq'$ and a different model d'. Since two behaviorally equivalent model-preference pairs $(d, \succeq)$ and $(d', \succeq')$ can have different normative implications (see e.g. Kőszegi and Rabin, 2008b; Bernheim, 2009; Masatlioglu et al., 2012), identifiability in the extended setting must aim at all aspects of the pair $(d, \succeq)$ that are normatively relevant.

Definition 5 *Pair $(d, \succeq)$ is virtually identifiable if for each $(d', \succeq') \in D \times P$ with $\succeq' \neq \succeq$, there exists $f \in F$ such that $d(\succeq, f) \neq d'(\succeq', f)$.*

Virtual identifiability implies that the welfare preference $\succeq$ is known for sure once the maximal data set has been collected. It still allows for some uncertainty about the behavioral model, but only to the extent that we may not be able to predict the behavior of an agent with a different welfare preference $\succeq' \neq \succeq$.

Proposition 13 *With model uncertainty, $G(\bar{\Lambda}(d, \succeq))$ is non-empty if and only if $(d, \succeq)$ is virtually identifiable.*

We can have multiple models with identifiable preferences each, that, if considered jointly, do not have virtually identifiable model-preference pairs. Model uncertainty of this type poses a fundamental new problem to nudging. On the other hand, adding a rational agent to any given behavioral model with identifiable preferences preserves the property of virtually identifiable model-preference pairs. Thus the possibility of agents being rational has no substantial impact on our previous results. The analysis in Sections 4 and 5 could also be adapted to the case of model uncertainty. For instance, if each distortion function $d \in D$ satisfies the frame-cancellation property, then it follows immediately that no data set allows us to exclude any dominated frame. Applications include the uncertainty about a satisficer's aspiration level. With virtually identifiable model-preference pairs, on the other hand, elicitation procedures now generate sequences of expanding data sets with the goal of learning about both preferences and models.

We could go one step further and dispense with any model conjecture. Instead of following our model-based approach to behavioral welfare economics, we could work with

the purely choice-based approach by Bernheim and Rangel (2009).[24] In fact, we can easily adapt our definition of the binary nudging relation and evaluate the frame-induced choices based on the weak unambiguous choice relation R' (Bernheim and Rangel, 2009, p. 60), rather than on a set of welfare preferences. Formally, a generalized choice situation (GCS) consists of a set of alternatives $S \subseteq X$ and a frame $f \in F$, and a choice correspondence describes the chosen alternatives for each GCS that we have observed. Let us assume that the observed choice has always been a unique alternative $C(S, f) \in S$. To eliminate all traces of non-choice-based theories about mistakes, let us also assume that all the observed GCSs are welfare-relevant. Now consider two frames f and f' of which we know that they have a differential impact on behavior, i.e., we have observed two GCSs $(\bar{S}, f)$ and $(\bar{S}, f')$ with $C(\bar{S}, f) = x \neq y = C(\bar{S}, f')$. In line with our previous analysis, we could say that f is a weak unambiguous nudge over f' if $C(S, f)\, R'\, C(S, f')$ holds for all matching pairs (S, f) and (S, f') that we have observed. It follows immediately from the definition of R' that such a ranking is impossible. The mere fact that $C(\bar{S}, f) = x \neq y = C(\bar{S}, f')$ implies that neither $xR'y$ nor $yR'x$ holds, and hence neither of the two frames can be a weak unambiguous nudge over the other. Nudging is impossible without assumptions about decision mistakes (as already pointed out by Bernheim and Rangel, 2009, p. 62).

7.2 Imperfectly Observable Frames

So far we have assumed that frames are perfectly observable and controllable by the regulator. Since a frame can be very complex, this assumption deserves to be relaxed. The generalization also allows us to model fluctuating internal states of the agent that affect her choices. For instance, consider a modified satisficing model in which the aspiration level k fluctuates in a non-systematic and unobservable way, as in the original RS model. We can capture this by including the aspiration level into the description of the frame (k affects choice but not welfare), but the extended frame cannot be fully observable and controllable for an outsider.

Imperfect observability can be modelled as a structure $\Phi \subseteq 2^F$ with the property that for each $f \in F$ there exists $\phi \in \Phi$ with $f \in \phi$. The interpretation is that the regulator observes only sets of frames $\phi \in \Phi$ and does not know under which of the frames $f \in \phi$ the agent was acting. The example with a fluctuating aspiration level can be modelled as $F = P \times \{2, \ldots, m_X\}$ and $\Phi = \{\phi_p \mid p \in P\}$ for $\phi_p = \{(p, k) \mid k \in \{2, \ldots, m_X\}\}$. A behavioral data set is a subset $\Lambda \subseteq P \times \Phi$, where $(\succeq', \phi') \in \Lambda$ means that the agent has been observed behaving according to $\succeq'$ when the frame must have been one of the

[24] Another interesting choice-based approach is due to Apesteguia and Ballester (2015), who propose using as a welfare benchmark the preference that is closest to a given behavior, measured by their "swaps" criterion. Their framework does not allow for frames, but it would be interesting to develop the respective generalization and derive the implications for nudging.

elements of ϕ'. Thus a welfare preference $\succeq$ is consistent with Λ if for each $(\succeq', \phi') \in \Lambda$ we have $\succeq' = d(\succeq, f')$ for some $f' \in \phi'$, so that $\succeq$ might have generated the data set from the regulator's perspective. The set of welfare preferences that are consistent with Λ is $P(\Lambda) = \{\succeq \mid \Lambda \subseteq \bar{\Lambda}(\succeq)\}$, where $\bar{\Lambda}(\succeq) = \{(d(\succeq, f), \phi) \mid f \in \phi \in \Phi\}$ is again the maximal data set for $\succeq$. Note that a non-singleton set of frames ϕ can appear more than once in a maximal data set, combined with different behavioral preferences. This also implies that the cardinality of $\bar{\Lambda}(\succeq)$ is no longer the same for all $\succeq \in P$, because two different frames $f, f' \in \phi$ might generate two different observations for some preference but only one observation for another preference.

In many applications, such as a satisficing model with fluctuating aspiration level, it is reasonable to assume that the same Φ applies to observing and nudging, i.e., the frame dimensions that the regulator can observe are identical to those that he can control. We allow for the more general case where a set of frames can be chosen as a nudge from a potentially different structure Φ_N.[25] When comparing two elements $\phi, \phi' \in \Phi_N$, we will not necessarily want to compare the agents' choices under each $f \in \phi$ with her choices under each $f' \in \phi'$. For instance, we want to compare orders of presentation for each aspiration level separately, not across aspiration levels. To this end, we introduce a set H of selection functions, which are functions $h : \Phi_N \to F$ with the property that $h(\phi) \in \phi$. The elements of H capture the comparisons that we need to make: when comparing ϕ with ϕ' we compare only the choices under the frames $h(\phi)$ and $h(\phi')$, for each $h \in H$. In the satisficing model we would have one $h_k \in H$ for each aspiration level $k \in \{2, \ldots, m_X\}$, defined by $h_k(\phi_p) = (p, k)$. The only assumption that we impose on H is that for each $f \in \phi \in \Phi_N$ there exist $h \in H$ such that $h(\phi) = f$. We can then define the equivalence class $[\phi]_\Lambda = \{\phi' \mid d(\succeq, h(\phi')) = d(\succeq, h(\phi)), \forall (h, \succeq) \in H \times P(\Lambda)\}$ for any Λ and ϕ. As before, let $[\phi]_\Lambda N(\Lambda)[\phi']_\Lambda$ if for each $(h, \succeq) \in H \times P(\Lambda)$ it holds that $c(d(\succeq, h(\phi)), S) \succeq c(d(\succeq, h(\phi')), S)$, for all non-empty $S \subseteq X$.

Let $G(\Lambda) = \{\phi \mid [\phi]_\Lambda N(\Lambda)[\phi']_\Lambda, \forall \phi' \in \Phi_N\}$ be the set of optimal nudges. We again consider maximal data sets. An immediate extension of identifiability of $\succeq$ (Definition 2) could require that for each $\succeq' \neq \succeq$ there exists $f \in \phi \in \Phi$ such that $d(\succeq, f) \neq d(\succeq', f)$. This property turns out to be necessary but not sufficient for $G(\bar{\Lambda}(\succeq))$ to be non-empty. It implies that the maximal data set for $\succeq$ is different from the maximal data set for every other preference, so that $\succeq$ is identified once $\bar{\Lambda}(\succeq)$ has been collected and once it is known that this set is indeed maximal. Unfortunately, the cardinality of $\bar{\Lambda}(\succeq)$ no longer

[25] In continuation of our previous approach, we assume that for each $\succeq \in P$ there exists $\phi \in \Phi_N$ such that $d(\succeq, f) = \succeq$ for all $f \in \phi$. This implies that nudging is not per se impeded by the lack of control over frames. The assumption is clearly much stronger here than before. For instance, it holds in the described satisficing application when there is perfect recall (because the order of presentation that coincides with the welfare preference is non-distorting for all possible aspiration levels) but would not hold with no recall (because the non-distorting order of presentation then depends on the aspiration level).

carries that kind of information, as we could have $\bar{\Lambda}(\succeq) \subset \bar{\Lambda}(\succeq')$ for some $\succeq' \neq \succeq$. Upon observing $\bar{\Lambda}(\succeq)$ we then never know if we have already arrived at the maximal data set for $\succeq$, or if there is an additional observation yet to be made. Our notion of identifiability in the setting with imperfectly observable frames must therefore ensure that the maximal data set reveals itself as maximal.

Definition 6 *Preference $\succeq$ is potentially identifiable if for each $\succeq' \in P$ with $\succeq' \neq \succeq$, there exist $f \in \phi \in \Phi$ such that $d(\succeq, f) \neq d(\succeq', f')$ for all $f' \in \phi$.*

When frames are not directly observed, identifiability requires more than the existence of a frame $f \in \phi \in \Phi$ that distinguishes between $\succeq$ and $\succeq'$. We can exclude welfare preference $\succeq'$ as a candidate only if the observed distorted preference $d(\succeq, f)$ could not as well have been generated by $\succeq'$ for any other $f' \in \phi$. For instance, no preference is potentially identifiable in the perfect-recall satisficing model with fluctuating aspiration level.[26]

Proposition 14 *With imperfectly observable frames, $G(\bar{\Lambda}(\succeq))$ is non-empty if and only if $\succeq$ is potentially identifiable.*

We use the term potential identifiability because there is no guarantee that we will ever arrive at $\bar{\Lambda}(\succeq)$. An appropriately redefined elicitation procedure might impose a set of frames ϕ multiple times on the agent, but a specific element $f \in \phi$ still does not materialize. This is in contrast to the case of observable frames, where a maximal data set can always be collected in exactly m_F steps.

8 Savings Application

We now study an extended application of our approach to a savings problem. The question how to encourage savings has received much attention in the nudging literature from the beginning.[27] The application also extends our previous setting in various directions, illustrating the flexibility and portability of our approach.

We consider a two-period environment with alternatives $x = (x_1, x_2) \in X = \mathbb{R}^2_+$ that specify a present payment of x_1 and a future payment of x_2. In line with much of the literature that estimates discount rates from behavioral data (see e.g. Cohen et al., 2016), we assume that the agent is risk-neutral. This is an acceptable approximation when

[26] To see why, note that two preferences which coincide except for the ranking of the two top alternatives are behaviorally equivalent for every order of presentation and every aspiration level $k \geq 2$. This was different if we allowed the agent to be sometimes rational ($k = 1$) as in the original RS model, in which case all preferences are potentially identifiable.

[27] For a recent contribution see Bernheim et al. (2015), who derive weak generalized Pareto optimal 401(k) defaults in the sense of Bernheim and Rangel (2009), with and without pruning.

the agent's background consumption is large relative to the payoff consequences of the considered choices. We thus focus on the restricted domain $\tilde{P}$ of welfare preferences that can be represented by a utility function of the form

$$u(x_1, x_2) = x_1 + \delta x_2,$$

for some unknown discount factor $0 < \delta \leq 1$. This environment extends our previous setting by allowing an infinite set of alternatives and non-strict preferences.

We aim at modelling that the framing of the decision problem necessarily prompts the agent to take a more *short-run* or a more *long-run* perspective. Examples of frames that induce more or less patient choices include the timing of the choice, the default allocation, the status quo, the phrasing of the question, or whether the agent is hungry or sated when making the choice (see e.g. Loewenstein and Prelec, 1992). The literature has described various channels through which these frames operate; they may focus the agent's attention on a particular time period, activate hot or cold states, moderate visceral influences, rouse different behavioral selves, or set reference points. We capture these effects by defining two frames, a short-run frame f_S and a long-run frame f_L. We do not advocate the simple view that the choices under one of the two frames are always welfare-maximizing. Frame f_S may induce "lapses of self-control" but frame f_L may cause future benefits to be "excessively intellectualized at arm's length" (Bernheim and Rangel, 2009, p. 58). Hence we believe it is most reasonable to assume that, if anything, the choices under f_S are present-biased while the choices under f_L are future-biased. We model this by assuming that an agent with true discount factor δ acts as if maximizing the preference represented by the utility function

$$u_S(x_1, x_2) = \gamma x_1 + \delta x_2$$

under frame f_S, and the preference represented by the utility function

$$u_L(x_1, x_2) = x_1 + \gamma\delta x_2$$

under frame f_L. The parameter $\gamma \geq 1$ captures the extra focus that each frame puts on one of the two time periods, and its magnitude measures the agent's susceptibility to framing. We assume that γ is an unknown parameter of the behavioral model and treat (γ, δ) as the model-preference pair that has to be elicited from the behavioral data set. Since both frames are distorting whenever $\gamma > 1$, we are relaxing the previous assumption that a non-distorting frame must exist for each model-preference pair.

We first turn to the problem of eliciting (γ, δ). After a normalization of u_S it follows that the agent applies the behavioral discount factor $\delta_S = \delta/\gamma$ under frame f_S. Holding

the frame fixed, this discount factor can be measured by observing the agent's choices from sufficiently many different subsets $S \subseteq X$. To that end, the experimental literature that we will discuss below typically uses paradigms such as multiple price lists or matching (Cohen et al., 2016). There are still many model-preference pairs that are consistent with some measured δ_S (except if $\delta_S = 1$), but the procedure can be repeated to also obtain a measure of the behavioral discount factor $\delta_L = \delta\gamma$ under frame f_L. The maximal data set $\Lambda = \{(\delta_S, f_S), (\delta_L, f_L)\}$ then reveals that (γ, δ) is given by $\gamma = \sqrt{\delta_L/\delta_S}$ and $\delta = \sqrt{\delta_L \delta_S}$. Hence each model-preference pair is (fully) identifiable.

We now turn to the problem of nudging an agent who is characterized by (γ, δ). If we required one frame to outperform the other frame for all (compact) choice sets $S \subseteq X$, as we did in our previous analysis where a non-distorting frame was always available, we would immediately find that none of the frames is a successful nudge over the other.[28] We therefore work with the weaker but reasonable requirement that only the choices in a prespecified range of plausible market conditions have to be improved by the optimal nudge. A market condition is described by the interest rate r and the present value y of the agent's investment opportunities, which jointly generate a budget line $X(r, y) = \{x \in X \mid x_1 + x_2/(1+r) = y\}$. Let C be a set of market conditions (r, y) for which we want to ensure optimal choices. We then say that frame f_S is a weakly successful nudge over frame f_L if each u_S-optimal element in choice set S is weakly u-better than each u_L-optimal element in S, and this holds for all compact subsets $S \subseteq X(r, y)$ for all $(r, y) \in C$. Intuitively, we ensure that the agent's choices under frame f_S are welfare-better than her choices under frame f_L for all admissible market conditions. We consider subsets $S \subseteq X(r, y)$ instead of only the entire budget lines $X(r, y)$ to reflect the possibility that there are floors or caps on investment, or that savings rates must be selected from a finite set. The definition of f_L being a successful nudge over f_S is analogous.

The set C may be generated by an interval of interest rates that we deem plausible, and a range of money amounts potentially available for investment to the agent (e.g. shares of current labor income). In general we only assume that C is compact and connected. Let $\underline{r}$ denote the smallest interest rate and $\overline{r}$ the largest interest rate in C. We allow interest rates to be negative but assume that $\underline{r} > -1$. Then we obtain the following result on the possibility of nudging.

Proposition 15 *Frame f_S is an optimal nudge if $\delta \leq 1/(1+\overline{r})$. Frame f_L is an optimal nudge if $1/(1+\underline{r}) \leq \delta$. Both frames are undominated if $1/(1+\overline{r}) < \delta < 1/(1+\underline{r})$.*

Impatient agents are nudged optimally by the short-run frame, and patient agents are

[28] Just consider the binary choice set $S = \{x, x'\}$ where $x = (x_1, x_2)$ and $x' = (x_1 - \epsilon_1, x_2 + \epsilon_2)$. When $\epsilon_2\delta/\gamma < \epsilon_1 < \epsilon_2\delta$, the welfare optimal choice and the choice under frame f_L is x', while frame f_S induces the wrong choice x. When $\epsilon_2\delta < \epsilon_1 < \epsilon_2\delta\gamma$, the welfare optimal choice and the choice under frame f_S is x, while frame f_L induces the wrong choice x'.

nudged optimally by the long-run frame. The precise thresholds for δ depend on the range of market interest rates for which the nudge is supposed to work. The thresholds do not depend on the size of the agent's susceptibility to framing γ, and therefore the behavior of the nudgeable types is far from being unambiguous. For instance, an impatient agent with $\delta < 1/(1+\overline{r})$ may behave very patiently and choose the highest available savings rate under frame f_L. This happens for sufficiently high interest rates whenever $\gamma\delta > 1/(1+\overline{r})$, and it even happens for all interest rates when $\gamma\delta > 1/(1+\underline{r})$. Our analysis then recommends to overrule these seemingly cold and rational long-run choices, by nudging the agent in a way that induces more impatient behavior. Conversely, our analysis recommends to correct lapses of self-control of patient agents, by nudging them to take the long-run perspective.

We now use empirical estimates of behavioral discount factors to obtain quantitative predictions from our model. As discussed above, several of the behavioral anomalies in intertemporal choice could be mapped into the model, because they can be understood as a frame-driven conflict between the short-run and the long-run perspective. A particularly lucid effect is the asymmetry between *delay* and *speed-up* framing (Loewenstein, 1988; Benzion et al., 1989; Shelley, 1993; Weber et al., 2007). If the choice problem is framed as a problem of delaying immediate rewards, behavior often reveals greater impatience than if it is framed as a problem of speeding up future rewards. Whether a problem is perceived as involving delay or speed-up is typically determined by the phrasing of the question, but it could also be induced by a default (i.e., opting into a 401(k) plan is a delay of immediate claims while opting out is a speed-up of later claims). We will use delay framing as an instance of f_S and speed-up framing as an instance of f_L. The literature has proposed different *positive* models to predict the asymmetry between delay and speed-up framing, including the added compensation hypothesis (Benzion et al., 1989), reference-dependence and gain-loss asymmetry (Loewenstein, 1988; Shelley, 1993), or query theory (Weber et al., 2007). Our *normative* analysis only assumes that the delay frame generates present-biased and the speed-up frame generates future-biased choices with respect to welfare.[29]

We conducted an experiment on Amazon Mechanical Turk to obtain individual-level data for a large number of diverse subjects.[30] The experiment took place in August

[29]In fact, query theory (Weber et al., 2007), which is prominent in psychology, rests on an explicit description of the internal decision-making process, exactly as required for our normative analysis. In our setting, it postulates that there are reasons for early consumption and reasons for late consumption, which we could capture by a welfare utility function of the form $u(x_1, x_2) = r_1x_1 + r_2x_2$. When facing a choice, the decision-maker has to access these reasons from memory. Access happens serially, and the frame determines whether information about the short-run or the long-run is accessed first. Due to "output interference", access is "less successful for later queries than for earlier queries" (p. 517). Denoting by $0 < s \leq 1$ the relative share of reasons retrieved successfully in the second query, we would obtain the behavioral utility functions $u_S(x_1, x_2) = r_1x_1 + sr_2x_2$ and $u_L(x_1, x_2) = sr_1x_1 + r_2x_2$, which corresponds exactly to our model with $\delta = r_2/r_1$ and $\gamma = 1/s$.

[30]The experiments in Loewenstein (1988), Benzion et al. (1989), Shelley (1993), and Weber et al. (2007)

2016. Our design follows the earlier literature closely; the instructions can be found in Appendix B.2. After reporting demographic information (gender, age, education), each participant had to answer two pairs of questions about intertemporal choice. Each pair of questions implemented a different frame. In the short-run frame, subjects were asked about their willingness to pay v_0 for an Amazon gift card of given value to be received on the same day. Then they were asked about the minimal discount v_D for which they would accept delaying receipt of the gift card by one year. The answers to this pair of questions reveal the behavioral discount factor $\delta_S = (v_0 - v_D)/v_0$. In the long-run frame, subjects were asked about their willingness to pay v_1 for an Amazon gift card of given value to be received in one year. Then they were asked about the maximal additional fee v_F for which they would accept speeding up receipt of the gift card to the same day. The answers to this pair of questions reveal the behavioral discount factor $\delta_L = v_1/(v_1 + v_F)$. The value of the gift card was \$75 in one frame and \$85 in the other frame, but the assignment of values and the order of the questions were randomized.[31] Responses were not incentivized, but the subjects obtained a compensation of \$0.75 for participation, yielding an average hourly wage of about \$19.[32]

Overall, 1059 subjects completed our survey. We dropped 218 subjects who did not obey our instructions or who responded in a way inconsistent with the model.[33] This leaves us with 841 independent observations. About half of the subjects (44.8%) are female. Ages range between 18 and 77, with a mean of 35 and a median of 32. Subjects' educational backgrounds are diverse, including high school (32.5%), undergraduate degree (48.6%), and graduate degree (18.4%) as the highest completed level of education.

The average of the discount factors δ_S revealed by the subjects in the delay frame

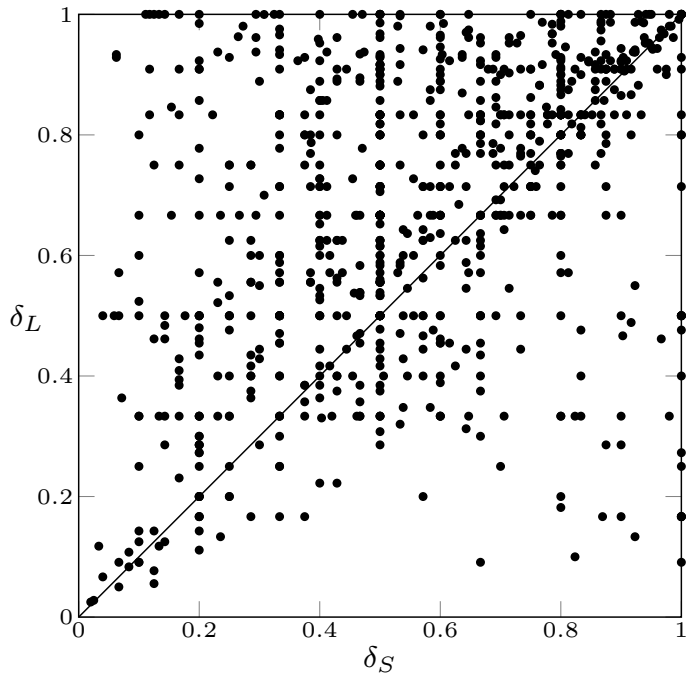
feature between 66 and 208 subjects but individual-level results are not reported.

[31]We chose different values of the gift card in the two frames to avoid suggesting that there was an objectively correct answer to our questions, thereby generating a demand effect for consistency. We still chose the values to be similar to each other because the earlier literature has documented an effect of the stake size on discount rates (e.g. Benzion et al., 1989; Shelley, 1993).

[32]The experiments by Benzion et al. (1989) and Shelley (1993) were not incentivized either. Loewenstein (1988) reports on three different experiments, one of which had real monetary incentives. All three experiments by Weber et al. (2007) were incentivized. A significant framing effect is found in all these studies, and, as we will argue below, our quantitative results are also well in the range of their findings. More generally, Cohen et al. (2016) review the literature on the measurement of time preferences and conclude that there are no significant differences between the results of experiments with and without monetary incentives (p. 32f). We add that the primary goal of our experiment is to illustrate the applicability of our theoretical approach, and one may want to replicate the results in an incentivized experiment to increase the confidence in our findings.

[33]From the beginning, we restricted participation eligibility to U.S. subjects with an experience of at least 500 approved MTurk HITs and an approval rate of at least 95%. In the short-run frame, the subjects were instructed to report a discount v_D between 0 and the value v_0 that they had stated earlier. One hundred subjects did not obey this instruction but reported values such that $v_D > v_0$. We deliberately allowed for such responses as a test of (in)attentiveness, following a suggestion by Paolacci et al. (2010) for experiments on MTurk. Furthermore, 118 subjects responded in a way that implies either $\delta_S = 0$ or $\delta_L = 0$, which is ruled out in our model.

Figure 2: Revealed Behavioral Discount Factors



is 0.56 ($s = 0.009$). The average of the discount factors δ_L revealed in the speed-up frame is 0.67 ($s = 0.008$). A t-test clearly rejects the null hypothesis that these averages are identical ($p = 0.000$, one-sided). Hence we replicate the earlier finding that average impatience is greater in the delay frame than in the speed-up frame. Our results are also quantitatively within the range of the previous findings.[34] Figure 2 is a scatterplot of the individual subjects' behavioral discount factors. The correlation between δ_S and δ_L is positive ($\rho = 0.41$) and significant ($p = 0.000$). The share of fully rational subjects (for whom $\delta_S = \delta_L$) is 6.3%. About one quarter of the subjects (26.3%) exhibit a framing effect opposite to the one conjectured above ($\delta_S > \delta_L$). Column (1) in Table 2 reports a linear regression of δ_S on the demographic variables, and column (2) reports the analogous regression with δ_L as the dependent variable. The regressions show that only education has a significant effect on behavioral discount rates, with higher levels of education implying weakly higher discount factors and thus more patient behavior under both frames.[35]

[34]We report here the average one-year discount factors (δ_S, δ_L) obtained in the previous studies, for the respective treatments that are most similar to ours. Applying our formulas to the average responses in the VCR treatment of Loewenstein (1988) yields $(0.54, 0.80)$. The discount rates reported in Benzion et al. (1989) for the treatment with a \$40 receipt and a one-year time horizon can be translated into the discount factors $(0.72, 0.80)$. Similarly, the pooled results in Shelley (1993) for receipts of \$40 and \$200 and time horizons of 6 months and one year translate into the one-year discount factors $(0.78, 0.83)$. While these values from Benzion et al. (1989) and Shelley (1993) are systematically larger than ours, those from Weber et al. (2007) are lower: the average one-year discount factors in Experiment 1 for gift certificates of values \$50 and \$75 and a time horizon of 3 months are $(0.34, 0.57)$.

[35]The dummy variables "High school", "Undergraduate", and "Graduate" code the highest level of education that a subject has completed. The coefficient "Undergraduate" is significantly larger than the

Table 2: Regression Analysis

	(1)	(2)	(3)	(4)
Dependent variable	δ_S	δ_L	δ	bias
Constant	0.463***	0.466***	0.443***	0.293***
	(0.051)	(0.074)	(0.046)	(0.078)
Female	-0.002	-0.134	-0.016	0.052**
	(0.017)	(0.017)	(0.015)	(0.025)
Age	-0.001	-0.001	-0.001	−0.002**
	(0.001)	(0.001)	(0.001)	(0.001)
High school	0.077**	0.222***	0.141***	0.054
	(0.036)	(0.065)	(0.032)	(0.063)
Undergraduate	0.111***	0.300***	0.201***	0.022
	(0.035)	(0.064)	(0.031)	(0.062)
Graduate	0.141***	0.274***	0.204***	0.007
	(0.039)	(0.066)	(0.034)	(0.064)
Other controls	Yes	Yes	Yes	Yes
R-squared	0.039	0.036	0.031	0.031
No. of observations	841	841	841	841

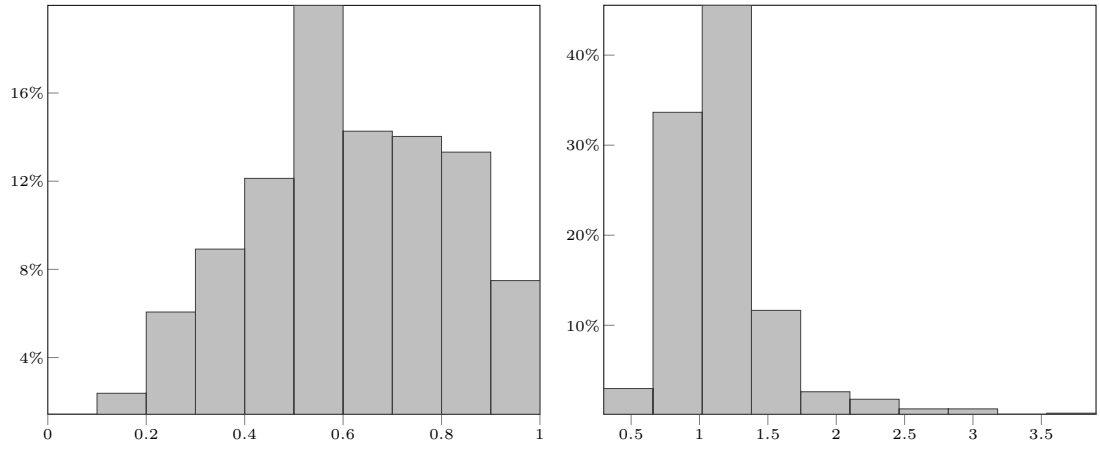
Notes: The table reports linear regressions. Robust standard errors are indicated in parentheses. The omitted education category is "None of the others". "Other controls" are response time and dummy variables for the randomization. The symbols *, **, *** denote significance at the 10%, 5% and 1% levels.

We next examine the welfare discount factors δ and the susceptibility to framing parameters γ that can be deduced from the subjects' behavioral discount rates. The average of δ across subjects is 0.60 ($s = 0.007$), which means that \$1.00 in one year is worth \$0.60 today from an average welfare perspective. Panel (a) of Figure 3 shows the entire distribution of δ and reveals considerable heterogeneity across subjects. The average of γ across subjects is 1.17 ($s = 0.013$), and panel (b) of Figure 3 shows its distribution in the population. The correlation between δ and the framing bias, which we define as $|\gamma-1|$ to take account of subjects with opposite framing effect, is negative ($\rho = -0.45$) and significant ($p = 0.000$), indicating that the subjects who are more susceptible to framing are those who are less patient from a welfare perspective. Regression (3) in Table 2 shows that only education affects the welfare discount factor, in the expected direction.[36] Maybe surprisingly, the framing bias $|\gamma - 1|$ is not significantly affected by education, as can be

coefficient "High school" in both regressions (1) and (2), but only at marginal significance level in the former (Wald-test, (1) $p = 0.085$, (2) $p = 0.000$). The coefficient "Graduate" is not significantly different from the coefficient "Undergraduate" in both regressions (Wald-test, (1) $p = 0.190$, (2) $p = 0.242$). Hence the effect of education on behavioral patience is only weakly monotonic.

[36]The coefficient "Undergraduate" is significantly larger than the coefficient "High school" (Wald-test, $p = 0.001$), while "Graduate" is not significantly different from "Undergraduate" (Wald-test, $p = 0.890$).

Figure 3: Welfare Discount Factors and Susceptibility Parameters



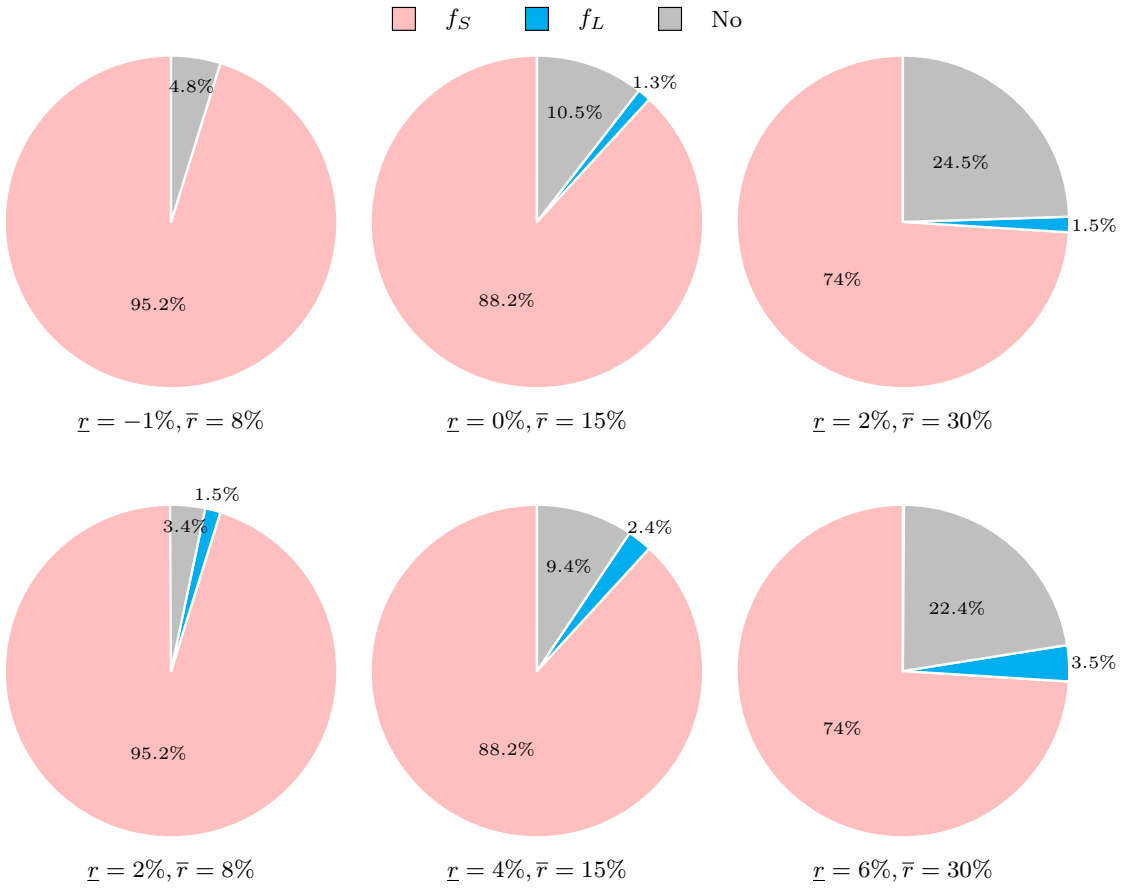
(a) discount factor δ (b) susceptibility γ

seen from column (4) in Table 2. Hence the effect of education on behavior operates by changing the welfare preference rather than the behavioral bias. The bias is significantly affected by gender, with women exhibiting larger biases, and by age, with smaller biases for older subjects.

We can now address the question of optimal nudging, by combining the theoretical result in Proposition 15 with our empirical findings. For transparency, we restrict attention to those subjects for whom we estimated a framing effect $\gamma \geq 1$ (but the analysis could easily be extended to the entire subject population). Figure 4 shows how many of the subjects should be nudged by one of the two frames, for six different intervals of potential market interest rates. Maybe contrary to conventional wisdom, the short-run frame f_S is an optimal nudge for a substantial share of the subjects across all interest rate conditions (between 74.0% and 95.2%). The share of subjects for whom f_L is optimal is very small (never exceeding 3.5%) and the share of non-nudgeable subjects is limited (between 3.4% and 24.5%). These conclusions are driven by the fact that welfare discount factors are generally low. Recall also that the framing effect is weaker for subjects with greater welfare patience, so framing naturally affects impatient subjects more. This can, for instance, be seen in the number of subjects whose behavior would respond to a change in the frame for *all* interest rates in the relevant range. Among those who should be nudged by f_L, the share of such subjects is zero in all the six cases illustrated in Figure 4. By contrast, it varies between 0% and 8.1% among the subjects who should be nudged by f_S.

If the regulator's goal is to select a frame that is optimal for a majority of the population, our analysis gives rise to a clear recommendation: choose the frame that induces present-biased behavior over the one that induces future-biased behavior. A more complete analysis should take into account additional behavioral mechanisms and other con-

Figure 4: Optimal Nudging



sequences of savings decisions (such as externalities for the welfare state), but our results at least challenge the view that soft paternalistic interventions should generally aim at increasing savings.

9 Conclusions

For most of the paper, we have taken the usual revealed-preference perspective for a single agent. Aside from its methodological justification, this is also directly relevant for nudging, where "personalization does appear to be the wave of the future" (Sunstein, 2014, p. 100). In the digital age of big data, individual-specific data gathering and nudging is achievable, for instance by relying on cookies. However, our results also speak to the problem of nudging a population of agents. On the elicitation stage, an assumption that different agents have identical preferences, possibly after controlling for observables, or are drawn representatively from a population, would allow us to combine observations of different agents into a single data set, facilitating the preference elicitation. On the nudging stage, the necessity to determine one frame for a population of heterogeneous

agents gives rise to ordinary social choice problems, which we have mostly refrained from studying in this paper.

Our model-based approach to behavioral welfare economics should in principle be conducive to nudging. Given a conjecture about how agents with different welfare preferences act under different frames, choice data can be used to infer about welfare and to assess which framing of the decision problem helps agents avoid mistakes. It is therefore remarkable how difficult the problem still turns out to be. Welfare-based nudging is impossible for interesting classes of models, and for others it is very complex information-wise. However, our analysis also shows that seemingly small differences between behavioral models can make a big difference for nudging. For instance, a satisficing agent stops searching as soon as some aspiration level is achieved. Our results imply that it is impossible to help this agent make systematically better choices. If the agent stops searching at the end of a search engine's result page, by contrast, it is relatively easy to improve her choices by framing. This raises important questions for future research about actual decision processes.

References

Apesteguia, J. and Ballester, M. (2015). A measure of rationality and welfare. *Journal of Political Economy*, 123:1278–1310.

Benzion, U., Rapoport, A., and Yagil, J. (1989). Discount rates inferred from decisions: An experimental study. *Management Science*, 35:270–284.

Bernheim, B. (2009). Behavioral welfare economics. *Journal of the European Economic Association*, 7:267–319.

Bernheim, B., Popov, A., and Fradkin, I. (2015). The welfare economics of default options in 401(k) plans. *American Economic Review*, 105:2798–2837.

Bernheim, B. and Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124:51–104.

Camerer, C. F., Issacharoff, S., Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Regulation for conservatives: behavioral economics and the case for "asymmetric paternalism". *University of Pennsylvania Law Review*, 151:1211–1254.

Caplin, A. and Martin, D. (2012). Framing effects and optimization. Mimeo.

Cohen, J., Ericson, K., Laibson, D., and White, J. (2016). Measuring time preferences. Mimeo.

De Clippel, G. and Rozen, K. (2014). Bounded rationality and limited datasets. Mimeo.

Goldin, J. (2015). Which way to nudge? uncovering preferences in the behavioral age. *Yale Law Journal*, forthcoming.

Goldin, J. and Reck, D. (2015). Preference identification under inconsistent choice. Mimeo.

Grüne-Yanoff, T. (2012). Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare*, 38:635–645.

Kalai, G., Rubinstein, A., and Spiegler, R. (2002). Rationalizing choice functions by multiple rationales. *Econometrica*, 70:24812488.

Kőszegi, B. and Rabin, M. (2007). Mistakes in choice-based welfare analysis. *American Economic Review, Papers and Proceedings*, 97:477–481.

Kőszegi, B. and Rabin, M. (2008a). Choice, situations, and happiness. *Journal of Public Economics*, 92:1821–1832.

Kőszegi, B. and Rabin, M. (2008b). Revealed mistakes and revealed preferences. In Caplin, A. and Schotter, A., editors, *The Foundations of Positive and Normative Economics*, pages 193–209. Oxford University Press, New York.

Kőszegi, B. and Szeidl, A. (2013). A model of focusing in economic choice. *Quarterly Journal of Economics*, 128:53–104.

Loewenstein, G. (1988). Frames of mind in intertemporal choice. *Management Science*, 34:200–214.

Loewenstein, G. and Prelec, D. (1992). Anormalies in intertemporal choice: Evidence and interpretation. *Quarterly Journal of Economics*, 107:573–597.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. (2012). Revealed attention. *American Economic Review*, 102:2183–2205.

Paolacci, G., Chandler, J., and Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgement and Decision Making*, 5:411–419.

Rubinstein, A. and Salant, Y. (2006). A model of choice from lists. *Theoretical Economics*, 1:3–17.

Rubinstein, A. and Salant, Y. (2008). Some thoughts on the principle of revealed preference. In Caplin, A. and Schotter, A., editors, *Handbooks of Economic Methodologies*, pages 115–124. Oxford University Press, New York.

Rubinstein, A. and Salant, Y. (2012). Eliciting welfare preferences from behavioural data sets. *Review of Economic Studies*, 79:375–387.

Salant, Y. and Rubinstein, A. (2008). (A,f): Choice with frames. *Review of Economic Studies*, 75:1287–1296.

Shelley, M. (1993). Outcome signs, question frames and discount rates. *Management Science*, 39:806–815.

Siegel, R. and Salant, Y. (2015). Contracts with framing. Mimeo.

Spiegler, R. (2015). On the equilibrium effects of nudging. *Journal of Legal Studies*, forthcoming.

Sunstein, C. (2014). *Why Nudge? The Politics of Libertarian Paternalism*. New Haven: Yale University Press.

Thaler, R. and Sunstein, C. (2003). Libertarian paternalism. *American Economic Review, Papers and Proceedings*, 93:175–179.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press.

Weber, E., Johnson, E., Milch, K., Chang, H., Brodscholl, J., and Goldstein, D. (2007). Aymmetric discounting in intertemporal choice – a query-theory account. *Psychologcial Science*, 18:516–523.

A Proofs

A.1 Proof of Lemma 1

(i) Suppose that $B(\succeq, f) \subseteq B(\succeq, f')$ holds for each $\succeq \in P(\Lambda)$. To show that $[f]_\Lambda N(\Lambda)[f']_\Lambda$, we proceed by contradiction and assume that there exist $\succeq \in P(\Lambda)$ and $S \subseteq X$ for which $c(d(\succeq, f), S) = x$ and $c(d(\succeq, f'), S) = y$ with $x \neq y$ and $y \succeq x$. The definition of c implies $(x, y) \in d(\succeq, f)$ and $(x, y) \notin d(\succeq, f')$. Together with $(x, y) \notin \succeq$ this implies $(x, y) \in B(\succeq, f)$ but $(x, y) \notin B(\succeq, f')$, a contradiction. For the converse, suppose that there exist $\succeq \in P(\Lambda)$ and $x, y \in X$ with $(x, y) \in B(\succeq, f)$ but $(x, y) \notin B(\succeq, f')$, which requires $x \neq y$. This implies $(x, y) \in d(\succeq, f)$ and $(x, y) \notin \succeq$, hence $(x, y) \notin d(\succeq, f')$. Then $c(d(\succeq, f'), \{x, y\}) = y \succeq x = c(d(\succeq, f), \{x, y\})$, which implies that $[f]_\Lambda N(\Lambda)[f']_\Lambda$ does not hold, by Definition 1.

(ii) Reflexivity and transitivity of $N(\Lambda)$ follow from the set inclusion characterization in statement (i). To show antisymmetry, consider any $f, f' \in F$ with $[f]_\Lambda N(\Lambda)[f']_\Lambda$ and $[f']_\Lambda N(\Lambda)[f]_\Lambda$. By (i) this is equivalent to $B(\succeq, f) = B(\succeq, f')$ and thus $d(\succeq, f) = d(\succeq, f')$ for each $\succeq \in P(\Lambda)$, hence $[f]_\Lambda = [f']_\Lambda$.

A.2 Proof of Proposition 1

Suppose $\succeq$ is identifiable, which implies that $\bar{\Lambda}(\succeq)$ is not identical to $\bar{\Lambda}(\succeq')$ for any other $\succeq'$. Then $P(\bar{\Lambda}(\succeq)) = \{\succeq\}$. Consider any f with $d(\succeq, f) = \succeq$, which exists by assumption. For any $f' \in F$, we then have $B(\succeq, f) = \varnothing \subseteq B(\succeq, f')$ and hence $[f]_{\bar{\Lambda}(\succeq)} N(\bar{\Lambda}(\succeq))[f']_{\bar{\Lambda}(\succeq)}$ by Lemma 1, which implies $f \in G(\bar{\Lambda}(\succeq))$. For the converse, suppose that $\succeq$ is not identifiable, i.e., there exists $\succeq' \neq \succeq$ with $\bar{\Lambda}(\succeq') = \bar{\Lambda}(\succeq)$. Then $\{\succeq, \succeq'\} \subseteq P(\bar{\Lambda}(\succeq))$. Consider any f_1 with $d(\succeq, f_1) = \succeq$ and any f_2 with $d(\succeq', f_2) = \succeq'$, so that $B(\succeq, f_1) = \varnothing$ and $B(\succeq', f_2) = \varnothing$. Assume by contradiction that there exists $f \in G(\bar{\Lambda}(\succeq))$. Then $[f]_{\bar{\Lambda}(\succeq)} N(\bar{\Lambda}(\succeq))[f_1]_{\bar{\Lambda}(\succeq)}$ must hold, which implies $B(\succeq, f) = \varnothing$ by Lemma 1, and hence $d(\succeq, f) = \succeq$. The analogous argument for f_2 implies $d(\succeq', f) = \succeq'$, which contradicts that $\bar{\Lambda}(\succeq') = \bar{\Lambda}(\succeq)$, i.e., that $\succeq$ is not identifiable.

A.3 Proof of Proposition 2

Any behavioral model d is characterized by the collection of maximal data sets $(\bar{\Lambda}(\succeq))_{\succeq \in P}$ that it assigns to the welfare preferences. Suppose there are $m_P \geq 2$ preferences and $m_F \geq 2$ frames. Then there are $(m_P)^{m_F}$ different maximal data sets. For a given welfare preference $\succeq$, however, only

$$N(m_P, m_F) = (m_P)^{m_F} - (m_P - 1)^{m_F}$$

of them are admissible, as the others contradict the existence of a non-distorting frame for $\succeq$. The number of possible models is thus given by $N(m_P, m_F)^{m_P}$. To obtain a model with identifiable preferences, we need to assign a different maximal data set to each welfare preference. Suppose we assign one of the $N(m_P, m_F)$ admissible data sets to the first welfare preference. Then there remain at least $N(m_P, m_F) - 1$ admissible data sets for the second welfare preference (the exact number is still $N(m_P, m_F)$ if the data set assigned to the first preference was not admissible for the second preference), and so on. Observe that $N(m_P, m_F) \geq m_P$, so we can proceed iteratively and obtain the falling factorial

$$N(m_P, m_F)^{\underline{m_P}} = N(m_P, m_F) \times (N(m_P, m_F) - 1) \times \ldots \times (N(m_P, m_F) - m_P + 1)$$

as a lower bound on the number of models with identifiable preferences. Consequently,

$$S(m_P, m_F) = \frac{N(m_P, m_F)^{\underline{m_P}}}{N(m_P, m_F)^{m_P}}$$

is a lower bound on the share of models with identifiable preferences. We can rewrite

$$\begin{aligned} S(m_P, m_F) &= \frac{N(m_P, m_F)}{N(m_P, m_F)} \times \frac{N(m_P, m_F) - 1}{N(m_P, m_F)} \times \cdots \times \frac{N(m_P, m_F) - m_P + 1}{N(m_P, m_F)} \\ &= \prod_{k=1}^{m_P - 1} \left(1 - \frac{k}{N(m_P, m_F)}\right) \\ &= \exp\left(\sum_{k=1}^{m_P - 1} \log\left(1 - \frac{k}{N(m_P, m_F)}\right)\right), \end{aligned}$$

where $1 > k/N(m_P, m_F) > 0$ holds for all $k = 1, \ldots, m_P - 1$. Recall that for $x > -1$ we have $\log(1 + x) \geq x/(1 + x)$, which implies

$$\sum_{k=1}^{m_P - 1} \log\left(1 - \frac{k}{N(m_P, m_F)}\right) \geq \sum_{k=1}^{m_P - 1} -\frac{k}{N(m_P, m_F) - k}.$$

Furthermore,

$$\begin{aligned} \sum_{k=1}^{m_P - 1} -\frac{k}{N(m_P, m_F) - k} &\geq \sum_{k=1}^{m_P - 1} -\frac{k}{N(m_P, m_F) - m_P + 1} \\ &= -\frac{1}{N(m_P, m_F) - m_P + 1} \sum_{k=1}^{m_P - 1} k \\ &= -\frac{(m_P)^2 - m_P}{2(N(m_P, m_F) - m_P + 1)}. \end{aligned}$$

Altogether, we therefore have

$$S(m_P, m_F) \geq \exp\left(-\frac{(m_P)^2 - m_P}{2(N(m_P, m_F) - m_P + 1)}\right) = \tilde{S}(m_P, m_F),$$

so $\tilde{S}(m_P, m_F)$ is also a lower bound on the share of models with identifiable preferences.

We are interested in asymptotic behavior as the number of alternatives m_X and hence the number of preferences m_P grows. Holding m_F fixed and treating m_P as a real variable, it follows with l'Hôpital's rule that

$$\lim_{m_P \to \infty} -\frac{(m_P)^2 - m_P}{2(N(m_P, m_F) - m_P + 1)} = 0$$

whenever $m_F \geq 4$. We thus obtain $\lim_{m_X \to \infty} \tilde{S}(m_P(m_X), m_F) = 1$ whenever $m_F \geq 4$. Now consider the case that the number of frames $m_F(m_X)$ also depends on the number of alternatives. Observe that $\tilde{S}(m_P, m_F)$ is strictly increasing in m_F whenever $m_P \geq 2$. At the same time, $\tilde{S}(m_P, m_F) \leq 1$ always holds since $\tilde{S}(m_P, m_F)$ is a lower bound on a proportion. Hence we obtain

$$\lim_{m_X \to \infty} \tilde{S}(m_P(m_X), m_F(m_X)) = 1$$

whenever there exists $\underline{m}$ such that $m_F(m_X) \geq 4$ for all $m_X \geq \underline{m}$. Then the share of models with identifiable preferences converges to 1 as the number of alternatives grows to infinity.

A.4 Proof of Proposition 3

Consider any d with the frame-cancellation property and any data set Λ. Fix any frame $f_1 \in F$, and let $f_2 \in F$ be an arbitrary frame with $f_2 \notin [f_1]_\Lambda$. Then, by definition of $[f_1]_\Lambda$, there exists $\succeq \in P(\Lambda)$ such that $d(\succeq, f_1) = \succeq_1 \neq \succeq_2 = d(\succeq, f_2)$. By the frame-cancellation property, we have $d(\succeq_1, f) = d(d(\succeq, f_1), f) = d(\succeq, f)$ for all $f \in F$, which implies that $\succeq_1 \in P(\Lambda)$. We also obtain $d(\succeq_1, f_1) = d(\succeq, f_1) = \succeq_1$, which implies $B(\succeq_1, f_1) = \varnothing$. From $\succeq_1 \neq \succeq_2$ and the frame-cancellation property, it follows that

$$B(\succeq_1, f_2) = d(\succeq_1, f_2) \setminus \succeq_1 = d(d(\succeq, f_1), f_2) \setminus \succeq_1 = d(\succeq, f_2) \setminus \succeq_1 = \succeq_2 \setminus \succeq_1 \neq \varnothing.$$

Hence $B(\succeq_1, f_1) \subset B(\succeq_1, f_2)$, and Lemma 1 implies that $[f_2]_\Lambda N(\Lambda)[f_1]_\Lambda$ does not hold. Since f_2 was arbitrary we conclude that $f_1 \in M(\Lambda)$, and, since f_1 was arbitrary, that $M(\Lambda) = F$.

A.5 Proof of Proposition 4

We assume $k \leq m_X/2$ throughout the proof, as cases where $k > m_X/2$ can be dealt with equivalently by reversing the role of the first page f and the second page $X\backslash f$ of the search engine.

Case 1: k even. We first construct an elicitation procedure e and then show that it is optimal. Let $e(\varnothing) = f_1$ be an arbitrary subset $f_1 \subseteq X$ with $|f_1| = k$. Now fix any welfare preference $\succeq$. The procedure then generates a data set $\Lambda_1 = \{(\succeq_1, f_1)\} \in L_1$, where $\succeq_1$ agrees with $\succeq$ within the sets f_1 and $X\backslash f_1$. Let a_i denote the alternative ranked at position i within the set f_1 by $\succeq_1$, for each $i = 1, \ldots, k$. Let b_i denote the alternative ranked at position i within the set $X\backslash f_1$ by $\succeq_1$, for each $i = 1, \ldots, k, \ldots, m_X - k$. Then construct the frame $e(\Lambda_1) = f_2$ as $f_2 = \{a_1, \ldots, a_{k/2}, b_{k/2+1}, \ldots, b_k\}$. The procedure then generates a data set $\Lambda_2 = \{(\succeq_1, f_1), (\succeq_2, f_2)\} \in L_2$, where $\succeq_2$ agrees with $\succeq$ within the sets f_2 and $X\backslash f_2$. This construction is applied to all the data sets Λ_1 that are generated by the elicitation procedure for some welfare preference. The elicitation procedure can be continued arbitrarily for all other data sets.

Let $\succeq$ be an arbitrary true welfare preference. We claim that the set $T_k(\succeq)$ of top k alternatives according to $\succeq$ can be deduced from the generated Λ_2, so that the optimal nudge is identified and $n(e, \succeq) \leq 2$ follows. Observe first that none of the alternatives $b_{k+1}, \ldots, b_{m_X-k}$ (if they exist) can belong to $T_k(\succeq)$, because Λ_1 has already revealed that each $b_1, \ldots, b_k$ is preferred by $\succeq$. Now suppose that $b_k \succeq_2 a_1$ holds. We then know that $b_k \succeq a_1$ and thus $T_k(\succeq) = \{b_1, \ldots, b_k\}$. Otherwise, if $a_1 \succeq_2 b_k$ holds, we know that $a_1 \succeq b_k$ and thus $b_k \notin T_k(\succeq)$ but $a_1 \in T_k(\succeq)$. In this case we can repeat the argument for a_2 and b_{k-1}: if $b_{k-1} \succeq_2 a_2$ we know that $b_{k-1} \succeq a_2$ and thus $T_k(\succeq) = \{b_1, \ldots, b_{k-1}, a_1\}$; otherwise, if $a_2 \succeq_2 b_{k-1}$ holds, we know that $a_2 \succeq b_{k-1}$ and thus $b_{k-1} \notin T_k(\succeq)$ but $a_2 \in T_k(\succeq)$. Iteration either reveals $T_k(\succeq)$ or arrives at $a_{k/2} \succeq_2 b_{k/2+1}$, which implies $a_{k/2} \succeq b_{k/2+1}$. In this case, we know that $T_k(\succeq)$ consists of $a_1, \ldots, a_{k/2}$ and those $k/2$ alternatives that $\succeq_2$ and hence $\succeq$ ranks top within $X\backslash f_2$.

Since $\succeq$ was arbitrary, we know that $\max_{\succeq \in P} n(e, \succeq) \leq 2$. Obviously, no single observation ever suffices to deduce $T_k(\succeq)$, neither in the constructed procedure nor in any other one, hence we can conclude that $n = 2$.

Case 2: k odd and $k < m_X/2$. The construction is the same as for case 1, except that $f_2 = \{a_1, \ldots, a_{(k-1)/2}, b_{(k+1)/2+1}, \ldots, b_k, b_{k+1}\}$, where b_{k+1} exists because $k < m_X/2$. The arguments about deducing $T_k(\succeq)$ are also the same, starting with a comparison of a_1 and b_k, except that the iteration might arrive at $a_{(k-1)/2} \succeq_2 b_{(k+1)/2+1}$, in which case $T_k(\succeq)$ consists of $a_1, \ldots, a_{(k-1)/2}$ and those $(k+1)/2$ alternatives that $\succeq_2$ ranks top within $X\backslash f_2$.

Case 3: k odd and $k = m_X/2$. The construction is the same as for case 1, except that $f_2 = \{a_1, \ldots, a_{(k+1)/2}, b_{(k+1)/2+1}, \ldots, b_k\}$. The arguments about deducing $T_k(\succeq)$ are also the same, starting with a comparison of a_1 and b_k, except that the iteration might arrive at

$a_{(k-1)/2} \succeq_2 b_{(k+1)/2+1}$. In this case, we can conclude that $T_k(\succeq)$ consists of $a_1, \ldots, a_{(k-1)/2}$, plus either $a_{(k+1)/2}$ or $b_{(k+1)/2}$ but never both, and those $(k-1)/2$ alternatives that $\succeq_2$ ranks top among the remaining ones in $X \backslash f_2$. Hence there exist welfare preferences $\succeq$ for which e does not identify $T_k(\succeq)$ after two steps. Since the missing preference between $a_{(k+1)/2}$ and $b_{(k+1)/2}$ can be learned by having $e(\Lambda_2) = f_3$ satisfy $\{a_{(k+1)/2}, b_{(k+1)/2}\} \subseteq f_3$, we know that $n \leq 3$.

It remains to be shown that $n > 2$. Fix an arbitrary elicitation procedure e and denote $e(\varnothing) = f_1 = \{a_1, \ldots, a_k\}$ and $X \backslash f_1 = \{b_1, \ldots, b_k\}$, where the numbering of the alternatives is arbitrary but fixed (remember that $k = m_X/2$). Let $\succeq_1$ be the preference given (in ranking notation) by $a_1 \ldots a_k \, b_1 \ldots b_k$, and consider the data set $\Lambda_1 = \{(\succeq_1, f_1)\}$ and the subsequent frame $e(\Lambda_1) = f_2$. Since k is odd, it follows that at least one of the pairs $\{a_1, b_k\}, \{a_2, b_{k-1}\}, \ldots, \{a_k, b_1\}$ must be separated on different pages by f_2, i.e., there exists $l = 1, \ldots, k$ such that $a_l \in f_2$ and $b_{k-l+1} \in X \backslash f_2$ or vice versa. Depending on the value of l, we now construct two welfare preferences $\succeq'$ and $\succeq''$. If $l = 1$, let

$$\begin{aligned} \succeq'\text{:} \;& b_1 \ldots b_{k-1} \, b_k \, a_1 \, a_2 \ldots a_k, \\ \succeq''\text{:} \;& b_1 \ldots b_{k-1} \, a_1 \, b_k \, a_2 \ldots a_k. \end{aligned}$$

If $l = 2, \ldots, k-1$, let

$$\begin{aligned} \succeq'\text{:} \;& a_1 \ldots a_{l-1} \, b_1 \ldots b_{k-l} \, b_{k-l+1} \, a_l \, a_{l+1} \ldots a_k \, b_{k-l+2} \ldots b_k, \\ \succeq''\text{:} \;& a_1 \ldots a_{l-1} \, b_1 \ldots b_{k-l} \, a_l \, b_{k-l+1} \, a_{l+1} \ldots a_k \, b_{k-l+2} \ldots b_k. \end{aligned}$$

If $l = k$, let

$$\begin{aligned} \succeq'\text{:} \;& a_1 \ldots a_{k-1} \, b_1 \, a_k \, b_2 \ldots b_k, \\ \succeq''\text{:} \;& a_1 \ldots a_{k-1} \, a_k \, b_1 \, b_2 \ldots b_k. \end{aligned}$$

For the two constructed welfare preferences $\succeq'$ and $\succeq''$, the elicitation procedure first generates the above described data set Λ_1. Subsequently, it generates the same data set $\Lambda_2 = \{(\succeq_1, f_1), (\succeq_2, f_2)\}$, because $\succeq'$ and $\succeq''$ differ only with respect to a_l and b_{k-l+1}, which is not revealed by frame f_2. Since $T_k(\succeq') \neq T_k(\succeq'')$, it follows that $n(e, \succeq') > 2$, which implies $\max_{\succeq \in P} n(e, \succeq) > 2$. Since e was arbitrary, it follows that $n > 2$.

A.6 Proof of Proposition 5

The result follows immediately if $m_X = 2$. Hence we fix a set X with $m_X \geq 3$ throughout the proof. We denote $m = m_X!$ for convenience.

Consider an arbitrary behavioral model, given by F and d, with $m_F \geq m$ and identi-

fiable preferences. Define

$$\hat{n}(e,\succeq) = \min\{s \mid P(\Lambda_s(e,\succeq)) = \{\succeq\}\}$$

as the first step at which procedure e identifies $\succeq$, and let

$$\hat{n} = \min_{e\in E} \max_{\succeq\in P} \hat{n}(e,\succeq).$$

It follows immediately that $n \leq \hat{n}$, because $P(\Lambda_s(e,\succeq)) = \{\succeq\}$ implies $G(\Lambda_s(e,\succeq)) \neq \varnothing$. We will establish the inequality $\hat{n} < m$.

Consider any e and suppose $\hat{n}(e,\succeq) \geq m$ for some $\succeq \in P$. Since $|P| = m$, there must exist $k \in \{0,1,\ldots,m-2\}$ such that

$$P(\Lambda_k(e,\succeq)) = P(\Lambda_{k+1}(e,\succeq)).$$

Denoting $e(\Lambda_k(e,\succeq)) = \tilde{f}$ and $d(\succeq,\tilde{f}) = \tilde{\succeq}$, we thus have $\Lambda_{k+1}(e,\succeq) = \Lambda_k(e,\succeq)\cup\{(\tilde{\succeq},\tilde{f})\}$ and $d(\succeq',\tilde{f}) = \tilde{\succeq}$ for all $\succeq' \in P(\Lambda_k(e,\succeq))$. We now define elicitation procedure e' by letting $e'(\Lambda) = e(\Lambda)$, except for data sets $\Lambda \in L$ that satisfy both $\Lambda_k(e,\succeq) \subseteq \Lambda$ and $f \neq \tilde{f}$ for all $(\succeq,f) \in \Lambda$, which includes $\Lambda = \Lambda_k(e,\succeq)$. For those data sets, we define

$$e'(\Lambda) = \begin{cases} e(\Lambda\cup\{(\tilde{\succeq},\tilde{f})\}) & \text{if} \quad |\Lambda| \leq m_F - 2, \\ \tilde{f} & \text{if} \quad |\Lambda| = m_F - 1. \end{cases}$$

Note that e' is a well-defined elicitation procedure. First, $\Lambda\cup\{(\tilde{\succeq},\tilde{f})\} \in L$ holds whenever the first case applies, because $\varnothing \neq P(\Lambda) \subseteq P(\Lambda_k(e,\succeq))$ and Λ does not yet contain an observation of $\tilde{f}$. Second, the first case then applies repeatedly because $e(\Lambda\cup\{(\tilde{\succeq},\tilde{f})\}) \neq \tilde{f}$, so that e' only dictates yet unobserved frames.

Consider any $\succeq' \notin P(\Lambda_k(e,\succeq))$, so that $(\succeq_1,f) \in \Lambda_k(e,\succeq')$ and $(\succeq_2,f) \in \Lambda_k(e,\succeq)$ with $\succeq_1 \neq \succeq_2$ for some f. From $\Lambda_k(e,\succeq') \subseteq \Lambda_s(e,\succeq')$ and thus $\Lambda_k(e,\succeq) \nsubseteq \Lambda_s(e,\succeq')$ for all $s \geq k$, it follows that preference $\succeq'$ is unaffected by the modification of the procedure, i.e., $\Lambda_s(e',\succeq') = \Lambda_s(e,\succeq')$ for all $s \in \{0,1,\ldots,m_F\}$, so that $\hat{n}(e',\succeq') = \hat{n}(e,\succeq')$. Now consider any $\succeq' \in P(\Lambda_k(e,\succeq))$, including $\succeq' = \succeq$. Then $\Lambda_s(e,\succeq) = \Lambda_s(e,\succeq') = \Lambda_s(e',\succeq')$ holds for all $s \leq k$. For $k < s \leq m_F - 1$, the definition of e' implies that $\Lambda_s(e',\succeq')$ does not contain an observation of $\tilde{f}$, and that

$$\Lambda_s(e',\succeq')\cup\{(\tilde{\succeq},\tilde{f})\} = \Lambda_{s+1}(e,\succeq').$$

Thus

$$P(\Lambda_s(e',\succeq')) = P(\Lambda_s(e',\succeq')\cup\{(\tilde{\succeq},\tilde{f})\}) = P(\Lambda_{s+1}(e,\succeq')),$$

so that $\hat{n}(e', \succeq') = \hat{n}(e, \succeq') - 1$. Repeated application of this construction allows us to arrive at an elicitation procedure e^* for which $\hat{n}(e^*, \succeq) < m$ for all $\succeq \in P$, which implies that $\hat{n} < m$.

A.7 Proof of Proposition 6

We write $P = \{\succeq_1, \succeq_2, \ldots, \succeq_m\}$, where $m = m_X!$ and the numbering of the preferences is arbitrary but fixed. We number the frames such that $f_i = o(\succeq_i)$. Note that each frame f_i is non-distorting for a single preference only, the one with which it coincides. This implies $n(e, \succeq) = \hat{n}(e, \succeq)$ for all $e \in E$ and $\succeq \in P$, and thus $n = \hat{n}$, where $\hat{n}$ refers to the complexity of identifying the welfare preference as defined in the proof of Proposition 5. We will establish the equality $\hat{n} = m - 1$.

Consider an arbitrary e. Define i_1 such that $e(\varnothing) = f_{i_1}$, and i_t for $t = 2, 3, \ldots, m$ recursively such that $e(\Lambda_{t-1}) = f_{i_t}$ for the data set

$$\Lambda_{t-1} = \bigcup_{j=1}^{t-1} \{(f_{i_j}, f_{i_j})\}.$$

If $\succeq_{i_m}$ is the welfare preference, then the procedure e will generate the sequence of data sets $\Lambda_s(e, \succeq_{i_m}) = \Lambda_s$ for all $s \in \{0, 1, \ldots, m-1\}$, with $\Lambda_0 = \varnothing$. It follows from the definition of d that $P(\Lambda_s) = \{\succeq_{i_{s+1}}, \succeq_{i_{s+2}}, \ldots, \succeq_{i_m}\}$ holds for each $s \in \{0, 1, \ldots, m-1\}$. This implies $\hat{n}(e, \succeq_{i_m}) = m-1$, and hence $\max_{\succeq \in P} \hat{n}(e, \succeq) \geq m-1$. Since e was arbitrary, it follows that $\hat{n} \geq m - 1$. Together with the result $\hat{n} < m$ established in the proof of Proposition 5, this implies $\hat{n} = m - 1$.

A.8 Proof of Proposition 7

We first show that the partition $\bar{P}$ for the perfect-recall model, denoted $\bar{P}^{PR}$, is weakly finer than the one for the no-recall model, denoted $\bar{P}^{NR}$, and strictly so whenever $k < m_X - 1$. Fix any $\succeq \in P$ and consider the equivalence class $P^{PR}_{\succeq}$ in the perfect-recall model. For any $\succeq' \in P^{PR}_{\succeq}$ it holds that $T_k(\succeq') = T_k(\succeq)$, where $T_k(.)$ is the set of top k alternatives according to the respective preference. Hence $\succeq' \in P^{NR}_{\succeq}$, which implies that $P^{PR}_{\succeq} \subseteq P^{NR}_{\succeq}$ and hence that $\bar{P}^{PR}$ is weakly finer than $\bar{P}^{NR}$. If $k < m_X - 1$, we have $|X \backslash T_k(\succeq)| \geq 2$. Construct $\succeq''$ from $\succeq$ by swapping the preference between the bottom 2 alternatives. Then $T_k(\succeq'') = T_k(\succeq)$ and hence $\succeq'' \in P^{NR}_{\succeq}$, but $\succeq'' \notin P^{PR}_{\succeq}$. Thus $\bar{P}^{PR}$ is strictly finer than $\bar{P}^{NR}$ in that case.

The fact that $\bar{P}^{PR}$ is weakly finer than $\bar{P}^{NR}$ immediately implies that any nudging domain for no-recall satisficing is also a nudging domain for perfect-recall satisficing. If $k < m_X - 1$, the domain $\tilde{P} = \{\succeq, \succeq''\}$ for two preferences $\succeq$ and $\succeq''$ as constructed above

is a nudging domain for perfect-recall satisficing but not for no-recall satisficing.

A.9 Proof of Proposition 8

The proof is similar to the proofs of Propositions 5 and 6 and therefore omitted.

A.10 Proof of Proposition 9

As argued in the proof of Proposition 6, the strong priming model satisfies $n(e, \succeq) = \hat{n}(e, \succeq)$ for all $e \in E$ and $\succeq \in P$, where $\hat{n}(e, \succeq)$ denotes the first step at which procedure e identifies $\succeq$. Hence

$$\bar{n} = \min_{e \in E} \sum_{i=1}^{m} p_i \hat{n}(e, \succeq_i),$$

where we again write $m = m_X!$ for convenience. We also keep the numbering of frames such that $f_i = o(\succeq_i)$.

Consider an arbitrary e. Define i_t for $t = 1, 2, \ldots, m$ exactly as in the proof of Proposition 6, i.e., as the index of the frame prescribed by e at step t when the agent has been successfully manipulated by all previous frames. It then follows from the definition of d that $\hat{n}(e, \succeq_{i_t}) = t$ for each $t = 1, 2, \ldots, m-1$, and $\hat{n}(e, \succeq_{i_m}) = m - 1$. Hence

$$\sum_{i=1}^{m} p_i \hat{n}(e, \succeq_i) = \sum_{t=1}^{m} p_{i_t} \hat{n}(e, \succeq_{i_t}) = \sum_{t=1}^{m-1} p_{i_t} t + p_{i_m}(m-1),$$

which is a weighted average of the numbers $1, 2, \ldots, m-1, m-1$, where the weights are the prior probabilities. Since $p_1 \geq p_2 \geq \ldots \geq p_m$, this weighted average is minimized by a procedure $e \in E$ with $i_t = t$, which implies the result.

A.11 Proof of Proposition 10

Since each frame is non-distorting for exactly one preference in the strong priming model, we have $\bar{\varphi}_\Lambda = \max_{\succeq \in P(\Lambda)} \pi_\Lambda(\succeq)$. We now proceed in two steps. We first construct an elicitation procedure e, and then show that it is optimal.

Step 1. We only need to describe $e(\Lambda)$ for Λ with $|P(\Lambda)| \geq 2$, as otherwise $\bar{\varphi}_\Lambda = 1$ holds and the continuation of e is irrelevant for the generalized complexity. Given any such Λ, let j be the second-smallest index among the preferences in $P(\Lambda)$, so that $\pi_\Lambda(\succeq_j)$ is the second-highest value among the updated probabilities. Then we define $e(\Lambda) = f_j$ for this data set, where the numbering of frames is given by $f_i = o(\succeq_i)$ as before. Note that the frame f_j cannot have been observed in Λ already, since otherwise either $P(\Lambda) = \{\succeq_j\}$ or $\succeq_j \notin P(\Lambda)$ would hold. Hence the construction yields a well-defined elicitation procedure.

For instance, we obtain $e(\varnothing) = f_2$, $e(\{(f_2, f_2)\}) = f_3$, and so on. If $\succeq_1$ is the welfare preference, it follows from the definition of d that

$$P(\Lambda_k(e, \succeq_1)) = \begin{cases} \{\succeq_1, \succeq_{k+2}, \succeq_{k+3}, \ldots, \succeq_m\} & \text{if } k \leq m-2, \\ \{\succeq_1\} & \text{if } k \geq m-1, \end{cases}$$

and therefore

$$\bar{\varphi}_{\Lambda_k(e,\succeq_1)} = \begin{cases} p_1/(p_1 + \sum_{j=k+2}^m p_j) & \text{if } k \leq m-2, \\ 1 & \text{if } k \geq m-1, \end{cases} \tag{1}$$

where we once more write $m = m_X!$ for convenience. If $\succeq_i$ for $i = 2, 3, \ldots, m$ is the welfare preference, we have

$$P(\Lambda_k(e, \succeq_i)) = \begin{cases} \{\succeq_1, \succeq_{k+2}, \succeq_{k+3}, \ldots, \succeq_m\} & \text{if } k \leq i-2, \\ \{\succeq_i\} & \text{if } k \geq i-1, \end{cases}$$

and therefore

$$\bar{\varphi}_{\Lambda_k(e,\succeq_i)} = \begin{cases} p_1/(p_1 + \sum_{j=k+2}^m p_j) & \text{if } k \leq i-2, \\ 1 & \text{if } k \geq i-1. \end{cases} \tag{2}$$

Given any $k = 0, 1, \ldots, m$, the value of (1) is always weakly smaller than the value of (2). Hence $\max_{\succeq \in P} n(q, e, \succeq) = n(q, e, \succeq_1)$. The value of $n(q, e, \succeq_1)$ is given by the smallest integer $k \geq 0$ such that

$$\frac{p_1}{p_1 + \sum_{j=k+2}^m p_j} \geq q,$$

which can be rearranged to the condition in the proposition.

Step 2. Now consider an arbitrary elicitation procedure e. Define i_t for $t = 1, 2, \ldots, m$ exactly as in the proof of Proposition 6. For any $i = 1, 2, \ldots, m$ let $t(i)$ be such that $i = i_{t(i)}$, so that frame f_i is prescribed by e at step $t(i)$ when the agent has been successfully manipulated by all previous frames. We then obtain

$$P(\Lambda_k(e, \succeq_i)) = \begin{cases} \{\succeq_{i_j} |\ j = k+1, k+2, \ldots, m\} & \text{if } k \leq t(i)-1, \\ \{\succeq_i\} & \text{if } k \geq t(i), \end{cases}$$

and

$$\bar{\varphi}_{\Lambda_k(e,\succeq_i)} = \begin{cases} p_{i_{j^*(k)}}/(\sum_{j=k+1}^m p_{i_j}) & \text{if } k \leq t(i)-1, \\ 1 & \text{if } k \geq t(i), \end{cases} \tag{3}$$

where $j^*(k)$ is an index j in $\{k+1, k+2, \ldots, m\}$ for which p_{i_j} is maximal. Given any $k = 0, 1, \ldots, m$, the value of (3) is minimized when $t(i) = m$, i.e., for welfare preference $\succeq_i = \succeq_{i_m}$. Hence $\max_{\succeq \in P} n(q, e, \succeq) = n(q, e, \succeq_{i_m})$. We now claim that the value of (3) for $\succeq_{i_m}$ is weakly smaller than the value of (1), for all $k = 0, 1, \ldots, m$, from which it follows that the procedure constructed in step 1 is indeed optimal. We only need to establish the inequality

$$\frac{p_{i_{j^*(k)}}}{\sum_{j=k+1}^{m} p_{i_j}} \leq \frac{p_1}{p_1 + \sum_{j=k+2}^{m} p_j}$$

for all $k \leq m-2$. It can be rearranged to

$$p_{i_{j^*(k)}} \left(p_1 + \sum_{j \in \{k+2, \ldots, m\}} p_j \right) \leq p_1 \left(p_{i_{j^*(k)}} + \sum_{j \in \{k+1, k+2, \ldots, m\} \setminus \{j^*(k)\}} p_{i_j} \right),$$

which can further be rearranged to

$$\frac{\sum_{j \in \{k+2, \ldots, m\}} p_j}{\sum_{j \in \{k+1, k+2, \ldots, m\} \setminus \{j^*(k)\}} p_{i_j}} \leq \frac{p_1}{p_{i_{j^*(k)}}}.$$

This holds, because $p_1 \geq p_2 \geq \ldots \geq p_m$ implies that the LHS is weakly smaller than 1 while the RHS is weakly larger than 1.

A.12 Proof of Proposition 11

We only need to consider the case $\underline{q} < q \leq \overline{q}$, which presupposes the existence of a procedure e^* with which

$$\bar{\varphi}_{\varnothing} < q \leq \max_{s \in \{1, \ldots, m_F\}} \bar{\varphi}_{\Lambda_s(e^*, \succeq)}$$

for all $\succeq \in P$. We will show that, with the frame-cancellation property, for all $\succeq \in P$ it holds that $P(\Lambda_1(e^*, \succeq)) = P(\Lambda_s(e^*, \succeq))$ for all $s = 2, \ldots, m_F$, and therefore

$$\bar{\varphi}_{\Lambda_1(e^*, \succeq)} = \max_{s \in \{1, \ldots, m_F\}} \bar{\varphi}_{\Lambda_s(e^*, \succeq)}.$$

This then immediately implies $n(q) = 1$. We will in fact establish the stronger property that $P(\Lambda) = P(\Lambda')$ whenever $\varnothing \neq \Lambda \subseteq \Lambda'$.

We first show that, for any two $\succeq, \succeq' \in P$, the maximal data sets $\bar{\Lambda}(\succeq)$ and $\bar{\Lambda}(\succeq')$ are either disjoint or identical. Suppose $\bar{\Lambda}(\succeq)$ and $\bar{\Lambda}(\succeq')$ are not disjoint, so there exists

$f' \in F$ such that $d(\succeq, f') = d(\succeq', f')$. Then the frame-cancellation property implies

$$d(\succeq, f) = d(d(\succeq, f'), f) = d(d(\succeq', f'), f) = d(\succeq', f)$$

for all $f \in F$, so that $\bar{\Lambda}(\succeq) = \bar{\Lambda}(\succeq')$.

Now fix any two data sets Λ and Λ' with $\varnothing \neq \Lambda \subseteq \Lambda'$. Since $P(\Lambda') \subseteq P(\Lambda)$ always holds, we only need to show that $P(\Lambda) \subseteq P(\Lambda')$. Fix any $\succeq \in P(\Lambda)$, so that $\Lambda \subseteq \bar{\Lambda}(\succeq)$. For any $\succeq' \in P(\Lambda')$ it holds that $\Lambda \subseteq \Lambda' \subseteq \bar{\Lambda}(\succeq')$. Since $\Lambda \neq \varnothing$, this implies that $\bar{\Lambda}(\succeq)$ and $\bar{\Lambda}(\succeq')$ are not disjoint, so that $\bar{\Lambda}(\succeq) = \bar{\Lambda}(\succeq')$. Hence $\Lambda' \subseteq \bar{\Lambda}(\succeq)$ and $\succeq \in P(\Lambda')$.

A.13 Proof of Proposition 12

We first show that, for all Λ and f, $P(\Lambda, f) = d(P(\Lambda), f)$ holds under the frame-cancellation property, where $d(P(\Lambda), f)$ denotes the image of $P(\Lambda)$ under $d(., f)$. The first inclusion $P(\Lambda, f) \subseteq d(P(\Lambda), f)$ follows immediately by definition of $d(P(\Lambda), f)$. Assume then that $\succeq \in d(P(\Lambda), f)$, so there exists $\succeq' \in P(\Lambda)$ such that $d(\succeq', f) = \succeq$. The frame-cancellation property then implies $d(\succeq, f') = d(d(\succeq', f), f') = d(\succeq', f')$ for any $f' \in F$, which reveals that $\succeq \in P(\Lambda)$. Furthermore, $d(\succeq, f) = d(\succeq', f) = \succeq$. Hence $\succeq \in P(\Lambda, f)$, so that the other inclusion $d(P(\Lambda), f) \subseteq P(\Lambda, f)$ also holds. We can therefore write

$$\varphi_\Lambda(f) = \sum_{\succeq \in d(P(\Lambda), f)} \pi_\Lambda(\succeq).$$

Consider $\Lambda = \varnothing$ first. For any $f \in F$, we claim that $|d(P(\varnothing), f)| = |d(P, f)| = |\bar{P}|$. Since f already partitions P into the $|d(P, f)|$ blocks between which it distinguishes, $|d(P, f)|$ is clearly a lower bound on $|\bar{P}|$. Now suppose $|d(P, f)| < |\bar{P}|$, which implies that there exist $\succeq_1 \neq \succeq_2$ such that $d(\succeq_1, f) = d(\succeq_2, f)$ but $d(\succeq_1, f') \neq d(\succeq_2, f')$ for some $f' \in F$. Then the frame-cancellation property implies $d(\succeq_1, f') = d(d(\succeq_1, f), f') = d(d(\succeq_2, f), f') = d(\succeq_2, f')$, a contradiction. Hence

$$\varphi_\varnothing(f) = \frac{|d(P, f)|}{m_X!} = \frac{|\bar{P}|}{m_X!} = \frac{1}{\bar{s}} \tag{4}$$

for all $f \in F$, i.e., initially each frame is equally likely to be optimal.

Consider any $\Lambda \neq \varnothing$ next. It has been shown in the proof of Proposition 11 that $P(\Lambda) = P(\Lambda')$ whenever $\varnothing \neq \Lambda \subseteq \Lambda'$, which implies that $P(\Lambda) = P(\bar{\Lambda}(\succeq)) = P_\succeq$ for any $\succeq \in P(\Lambda)$. For any $f \in F$, we then obtain that $|d(P(\Lambda), f)| = |d(P_\succeq, f)| = 1$ immediately from the definition of $P_\succeq$. Hence

$$\varphi_\Lambda(f) = \frac{1/m_X!}{s_\succeq/m_X!} = \frac{1}{s_\succeq} \tag{5}$$

for all $f \in F$, where $s_{\succeq} = |P_{\succeq}|$. Again, each frame is equally likely to be optimal.

Equation (4) also implies $\underline{q} = 1/\overline{s}$. Now consider any $\succeq' \in P$ with $s_{\succeq'} \geq \overline{s}$, which must clearly exist. For any procedure e and any $s = 1, \ldots, m_F$, we then have by equation (5)

$$\bar{\varphi}_{\Lambda_s(e,\succeq')} = \frac{1}{s_{\succeq'}} \leq \frac{1}{\overline{s}} = \underline{q},$$

which implies $\overline{q} = \underline{q}$.

A.14 Proof of Proposition 13

The proof is similar to the proof of Proposition 1 and therefore omitted.

A.15 Proof of Proposition 14

The proof is similar to the proof of Proposition 1 and therefore omitted.

A.16 Proof of Proposition 15

Case 1: $\delta \leq 1/(1+\overline{r})$. Fix any $(r, y) \in C$. If either $\delta < 1/(1+\overline{r})$ or $r < \overline{r}$, or both, we have $\delta < 1/(1+r)$. Then the marginal rate of substitution of x_1 for x_2 according to the welfare utility u, which is given by $\mathrm{MRS}_u = 1/\delta$, is strictly larger than the absolute value of the slope of the budget line, which is given by $1+r$. The unique u-optimal element from any compact $S \subseteq X(r, y)$ is thus the unique alternative that maximizes x_1 in S. From $\mathrm{MRS}_{u_S} = 1/\delta_S = \gamma/\delta \geq 1/\delta$ it follows that this is also the unique u_S-optimal element in S. Hence each u_S-optimal element in S is weakly u-better than each u_L-optimal element. If instead $\delta = 1/(1+\overline{r})$ and $r = \overline{r}$ holds, we have $\mathrm{MRS}_u = 1 + r$ and all elements in any $S \subseteq X(r, y)$ are u-optimal. It again follows that each u_S-optimal element in S is weakly u-better than each u_L-optimal element. Thus f_S is a weakly successful nudge over f_L and hence an optimal nudge.

Case 2: $1/(1+\underline{r}) \leq \delta$. Analogous arguments imply that f_L is an optimal nudge in that case.

Case 3: $1/(1+\overline{r}) < \delta < 1/(1+\underline{r})$. If $\gamma = 1$, the utility functions u, u_S, and u_L all coincide, which implies that each frame is a weakly successful nudge over the other, and hence none of them is dominated. Then assume $\gamma > 1$. Choose $(r, y) \in C$ such that $\delta/\gamma < 1/(1+r) < \delta$, which exists because C is connected. Consider $S = X(r, y)$. From $\mathrm{MRS}_{u_L} < \mathrm{MRS}_u < 1+r$ it then follows that $(0, y(1+r))$ is the unique u-optimal element in S and also the unique u_L-optimal element in S. By contrast, from $\mathrm{MRS}_{u_S} > 1+r$ it follows that $(y, 0)$ is the unique u_S-optimal element in S. Hence the u_S-optimal element is not weakly u-better than the u_L-optimal element, and therefore f_S is not a weakly successful

nudge over f_L. Analogous arguments for some $(r, y) \in C$ with $\delta < 1/(1+r) < \gamma\delta$ imply that f_L is also not a weakly successful nudge over f_S. Hence none of the two frames is dominated.

B Additional Material

B.1 Complexities for the Strong Priming Model

Expected complexity, geometric distribution. Fix some $\rho \in (0, 1)$ and let

$$p_i = \rho^{i-1} \left(\frac{1-\rho}{1-\rho^m} \right)$$

for each $i = 1, 2, \ldots, m$, where $m = m_X!$ for convenience. Note that this is indeed a probability distribution, because $p_i \in (0, 1)$ and

$$\sum_{i=1}^{m} p_i = \left(\frac{1-\rho}{1-\rho^m} \right) \sum_{i=1}^{m} \rho^{i-1} = \left(\frac{1-\rho}{1-\rho^m} \right) \left(\frac{1-\rho^m}{1-\rho} \right) = 1,$$

where the second equality follows from a standard result about the geometric sequence. The expression for $\bar{n}$ in Proposition 9 can then be written as

$$\bar{n} = \left(\frac{1-\rho}{1-\rho^m} \right) \sum_{i=1}^{m-1} \rho^{i-1} i + \left(\frac{1-\rho}{1-\rho^m} \right) \rho^{m-1} (m-1).$$

Using the standard result that

$$\sum_{i=1}^{m-1} \rho^{i-1} i = \left(\frac{1-\rho^m}{(1-\rho)^2} \right) - \left(\frac{m\rho^{m-1}}{1-\rho} \right),$$

we can further simplify to

$$\bar{n} = \left(\frac{1}{1-\rho} \right) + \left(\frac{(1-\rho)(m-1)\rho^{m-1} - m\rho^{m-1}}{1-\rho^m} \right).$$

Due to $\rho \in (0, 1)$, the second term vanishes as $m \to \infty$. Hence $\lim_{m_X \to \infty} \bar{n} = 1/(1-\rho)$.

Expected complexity, uniform distribution. Let $p_i = 1/m$ for each $i = 1, 2, \ldots, m$. The expression for $\bar{n}$ in Proposition 9 can then be written as

$$\bar{n} = \frac{1}{m} \sum_{i=1}^{m-1} i + \left(\frac{m-1}{m} \right) = \left(\frac{m-1}{2} \right) + \left(\frac{m-1}{m} \right) = (m_X! - 1) \left(\frac{1}{2} + \frac{1}{m_X!} \right),$$

which is of the same order of magnitude as the previously given $n = m_X! - 1$.

Generalized complexity, geometric distribution. For the geometric distribution, the LHS of the inequality in Proposition 10 can be rewritten as

$$\sum_{j=1+k}^{m-1} p_{j+1} = \left(\frac{1-\rho}{1-\rho^m}\right) \sum_{j=1+k}^{m-1} \rho^j = \left(\frac{1-\rho}{1-\rho^m}\right)\left(\frac{\rho^{1+k}-\rho^m}{1-\rho}\right) = \frac{\rho^{k+1}-\rho^m}{1-\rho^m}.$$

Thus $n(q)$ is the smallest integer $k \geq 0$ for which

$$\frac{\rho^{k+1}-\rho^m}{1-\rho^m} \leq \left(\frac{1-\rho}{1-\rho^m}\right)\left(\frac{1-q}{q}\right),$$

or

$$\rho^k \leq \rho^{m-1} + \left(\frac{1-\rho}{\rho}\right)\left(\frac{1-q}{q}\right).$$

For $q = 1$ this implies $n(1) = m - 1$. Since the RHS of the inequality converges to $((1-\rho)/\rho)((1-q)/q)$ as $m \to \infty$, for $q < 1$ we obtain that $n(q)$ must converge to the smallest integer $k \geq 0$ for which

$$\rho^k \leq \left(\frac{1-\rho}{\rho}\right)\left(\frac{1-q}{q}\right)$$

holds. Hence

$$\lim_{m_X \to \infty} n(q) = \max\left\{\left\lceil \frac{\log\left(\frac{1-\rho}{\rho}\frac{1-q}{q}\right)}{\log \rho} \right\rceil, 0\right\}.$$

Generalized complexity, uniform distribution. For the uniform distribution, the condition in Proposition 10 becomes that $n(q)$ is the smallest integer $k \geq 0$ for which

$$k \geq (m-1) - \left(\frac{1-q}{q}\right)$$

holds. Hence we obtain

$$n(q) = \max\left\{\left\lceil (m-1) - \left(\frac{1-q}{q}\right)\right\rceil, 0\right\}.$$

B.2 Experimental Instructions

The following contains screenshots of the MTurk experiment. Figure 5 shows the description of the HIT on MTurk. Upon participation, subjects were directed to a Qualtrics page for the experiment. They first had to accept the consent form in Figure 6. We then collected some demographics (Figure 7) before subjects faced two pairs of questions about

intertemporal choice as shown in Figures 8 and 9. The value of the gift cards ($75 and $85) and the order of the questions were randomized. Finally, subjects were given an exit code with which they could demand payment on MTurk (Figure 10).

Figure 5: Screenshot of the description of the HIT on MTurk.

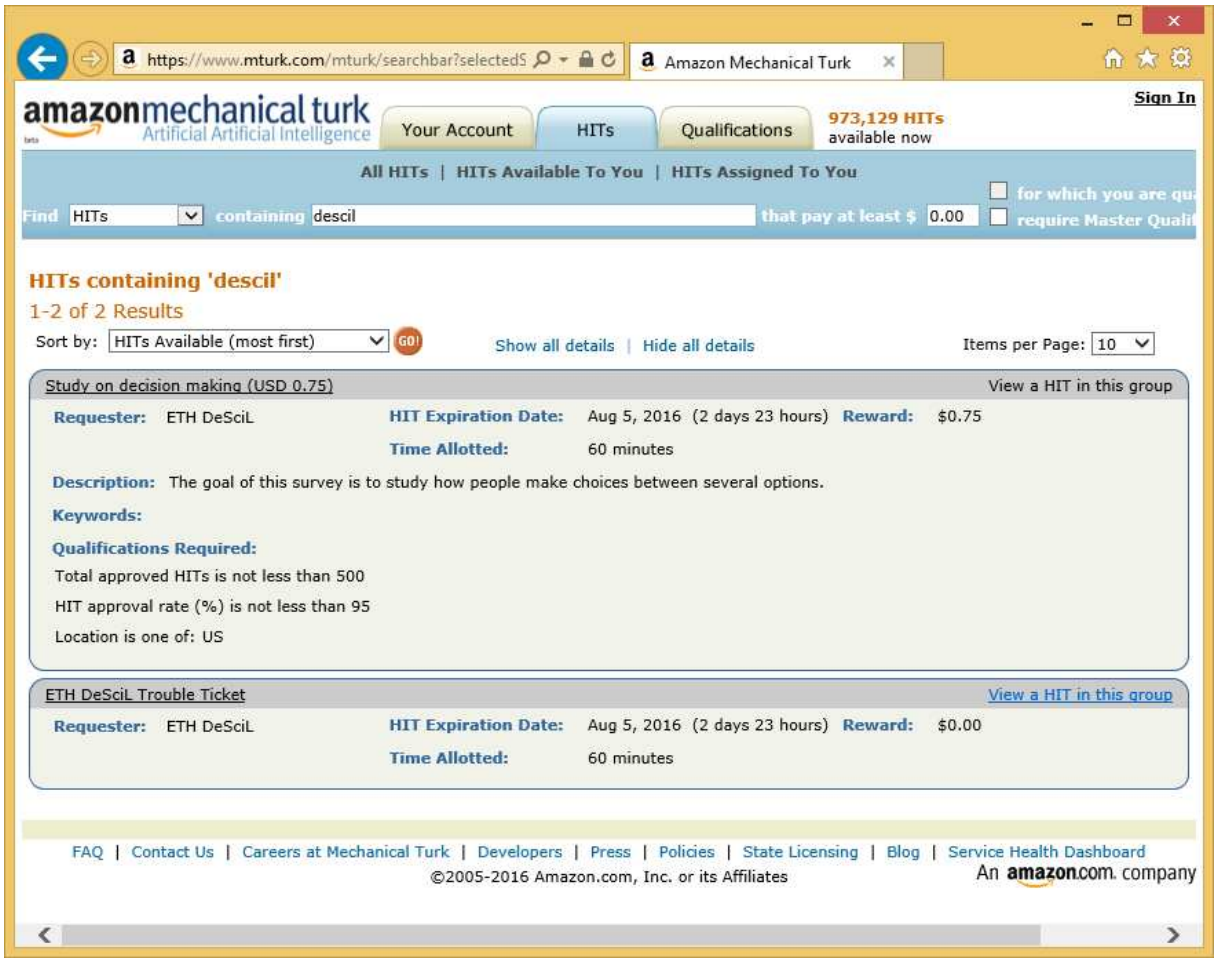


Figure 6: Screenshot of the first page of the experiment.

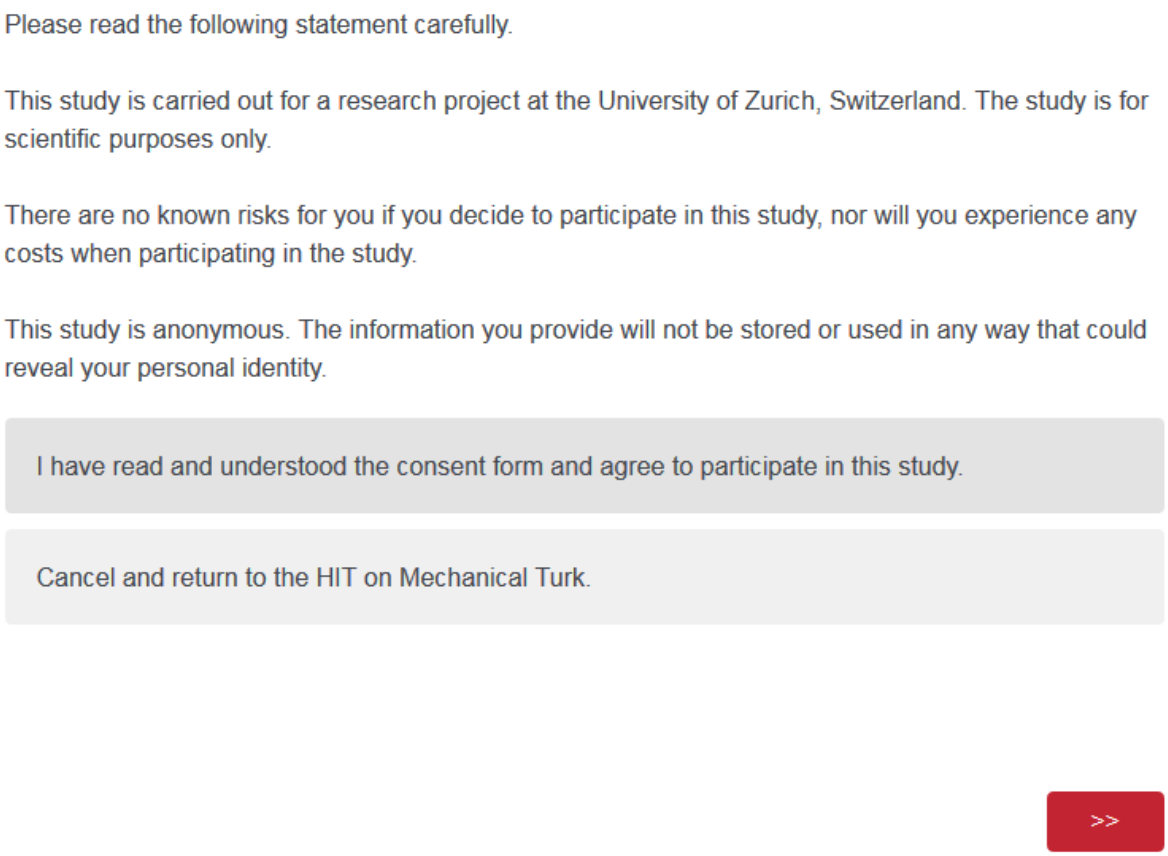


Figure 7: Screenshot of the second page of the experiment.

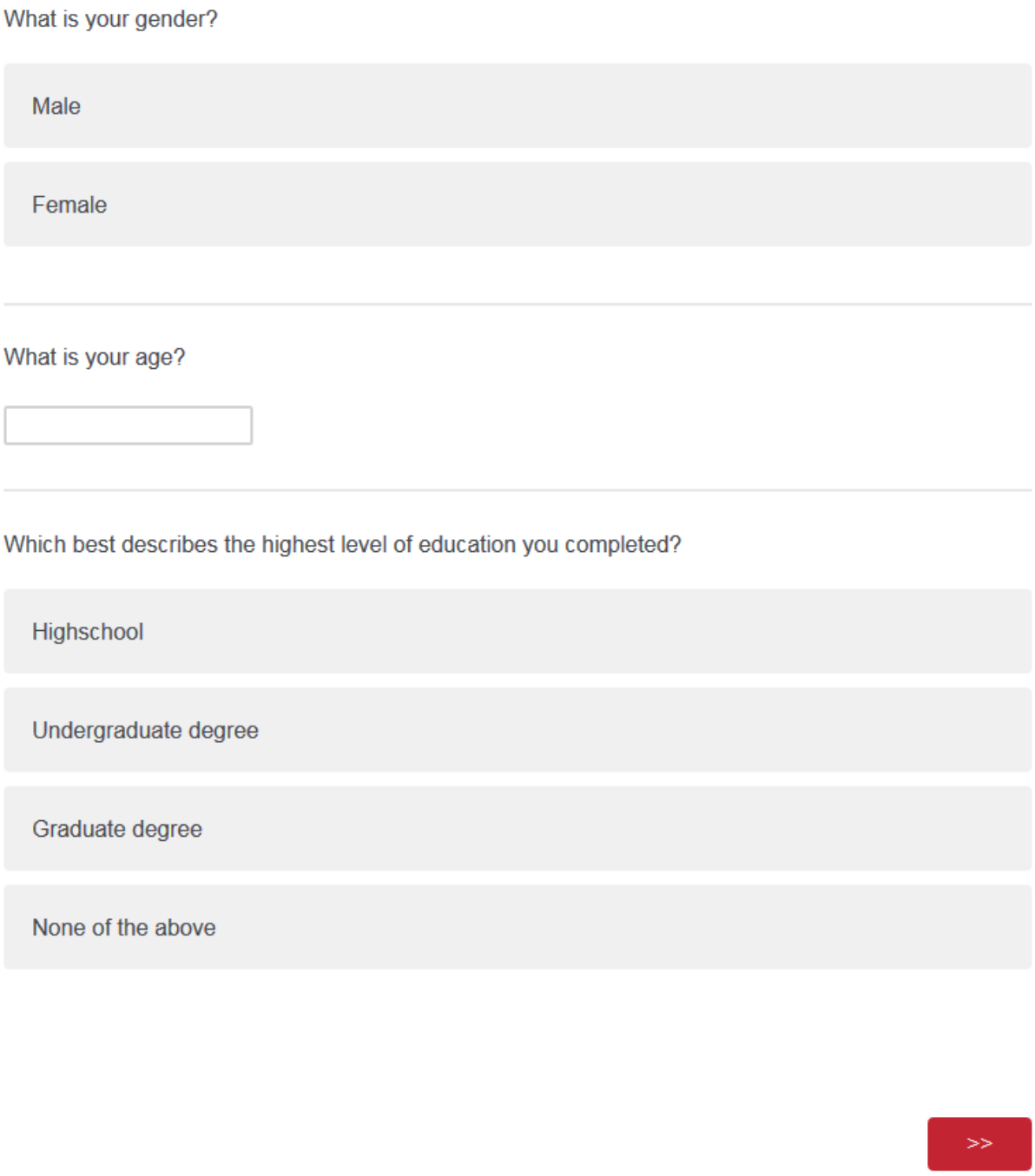


Figure 8: Screenshot of the third page of the experiment.

Suppose we offer to sell you an Amazon gift card worth **$75**. If you decide to buy, you have to pay today and you will **receive the gift card one year from today**.

How much (between $0 and $75) would you be willing to pay?

Now suppose we offer you an additional choice between the following two options:

(a) The transaction takes place as described above. You pay the amount you specified above, and you will receive the gift card one year from today.

(b) You pay an extra fee, but delivery of the gift card is sped up. You pay the amount you specified above **plus the extra fee**, and you will receive the gift card **today**.

How large (in addition to the amount you specified above) could the **extra fee** be for you to choose option (b) instead of option (a)?

>>

Figure 9: Screenshot of the fourth page of the experiment.

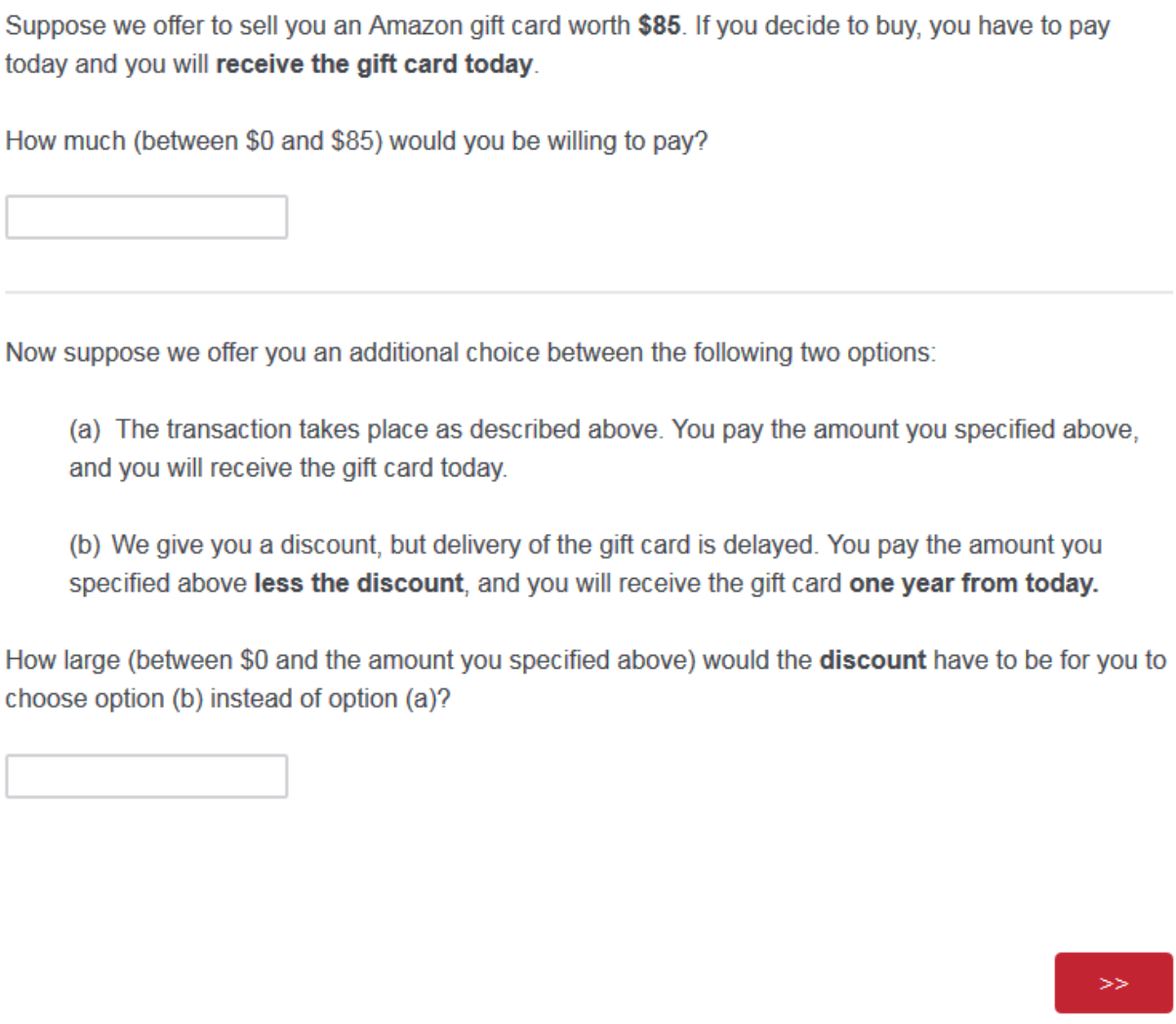


Figure 10: Screenshot of the fifth and final page of the experiment.