Preference Identification Under Inconsistent Choice*

Jacob Goldin† Daniel Reck‡

September 17, 2015

Abstract

In many settings, seemingly arbitrary features of a decision can affect what people choose. We develop an empirical framework to recover ordinal preference information from choice data when preference-irrelevant *frames* affect behavior. Plausible restrictions of varying strength permit either partial- or point-identification of preferences for the decision-makers who choose consistently across frames. Recovering population preferences requires understanding the empirical relationship between decision-makers' preferences and their consistency. We develop tools for studying this relationship and illustrate them with data on automatic enrollment into pension plans. The results suggest that 70 percent of default-sensitive employees prefer enrollment but that a default of non-enrollment may be optimal for young, low-income employees.

JEL Classification: C10, D60, I30

*The authors wish to thank Jason Abaluck, Roland Benabou, John Beshears, Sebastien Bradley, Benjamin Brooks, Charlie Brown, Raj Chetty, Henry Farber, Nikolaj Harmon, James Hines, Bo Honore, Louis Kaplow, Miles Kimball, Alvin Klevorick, Henrik Kleven, Claus Kreiner, David Lee, Jay Lu, Charles Manski, Alex Mas, Ted O'Donoghue, Søren Leth-Petersen, Wolfgang Pesendorfer, Kareen Rozen, Joel Slemrod, Jesse Shapiro, Sharon Tennyson, and seminar participants at the University of Copenhagen, Cornell, Harvard, the London School of Economics, the University of Michigan, Oxford, and Princeton for helpful discussion and comments. We are grateful to Brigitte Madrian and Aon Hewitt for providing us with data.

†Department of Economics, Princeton University and Stanford Law School; email: jgoldin@princeton.edu

‡Corresponding Author. Address: Department of Economics, University of Michigan, 611 Tappan St, Ann Arbor, MI 48109; phone: (734)763-0431; email: dreck@umich.edu

Introduction

Sometimes choices do not reveal preferences. For example, an internet company seeking to collect and use its customers' personal data might adopt an *opt-out* policy, under which it can collect a customer's data unless the customer indicates otherwise. Empirical research suggests that switching to an *opt-in* policy, under which customers must give permission before the company can collect their data, would substantially reduce the fraction of customers who allow the company to do so (Johnson, Bellman and Lohse, 2002). Suppose 40 percent of customers give permission when the policy is opt-in and 70 percent do so when the policy is opt-out. Both policies let customers control the use of their data, but the choices made under the two policies imply different conclusions about what customers prefer.

Situations such as this one pose an important challenge to behavioral economics. In many settings – from privacy controls to retirement savings (Madrian and Shea, 2001) to health insurance (Handel, 2013) to voting (Ho and Imai, 2008) – choices depend on seemingly arbitrary features of the decision-making environment, such as which option is the default, the order in which options are presented, or which features of the options are made salient. Inconsistencies in choice invalidate the traditional revealed preference approach of equating choices with preferences.[1] This difficulty contributes to widespread disagreement about when and how to use behavioral economics to inform policy.

In this paper we develop an empirical framework for analyzing choice inconsistencies and use it to derive conditions under which preference information may be recovered from choice data. Our approach consists of two steps: first, identifying preference information for the subgroup of decision-makers unaffected by the source of the inconsistency, and second, accounting for selection into this subgroup to recover population preferences. We develop new tools for carrying out both steps of the analysis and illustrate them with data on a policy that is well-known in the behavioral economics literature: automatic enrollment into employer-sponsored pension plans.

Most prior work takes one of two approaches to the problem of preference identi-

[1] By *preferences*, we mean the relative consistency of the available options with a decision-maker's objectives, whatever those may be. Preferences are not defined according to a decision-maker's observed choices; doing so would assume away the question we address by ruling out the possibility of choice reversals (see, e.g., Basu, 2003; Sen, 1973). Note that the preferences revealed under the opt-in and opt-out policies are not inconsistent if customers' preferences over their personal data happen to turn on this feature of the decision.

fication under inconsistent choice. First, researchers may utilize a positive model of behavior that fully specifies the mapping from decision-makers' preferences to their (potentially sub-optimal) behavior (e.g. Rubinstein and Salant, 2012; Benkert and Netzer, 2014; Carroll et al., 2009). Such approaches yield important insights but in many cases the resulting welfare conclusions are sensitive to the researcher's choice between competing positive models that are difficult to distinguish observationally (Bernheim, 2009; De Clippel and Rozen, 2014). An alternative approach is to restrict preference inferences to the subset of observed choice situations in which a given decision-maker chooses consistently (Bernheim and Rangel, 2009). However, in practice individual decision-makers are typically observed making only a single choice, which makes it difficult to detect which choices are consistent. Worse, this approach yield no information on the preferences of those decision-makers who exhibit systematic choice reversals – the very group whose preferences are most relevant for making optimal policy determinations regarding how choices should be designed (a claim we formalize in Appendix B). Further "refinements" can provide a path forward if the researcher can observe choices in a setting in which *all* decision-makers are known to select their most-preferred option (e.g., Chetty, Looney and Kroft, 2009), but in many applications, such as those in which behavior is sensitive to defaults or ordering, there is little reason to believe that any of the observed choice situations satisfy this condition.[2]

Our approach to this problem overcomes many of the limitations of prior work. We develop techniques to recover preference information with limited datasets – those in which each individual is observed making only one decision and observers lack ex ante knowledge about which individuals are optimizing. We do not assume that the researcher knows the exact underlying positive model that generates the choice inconsistencies, nor do we assume that all decision-makers choose optimally in any one of the observed choice situations. We provide general conditions – consistent with a broad class of behavioral models – under which the preferences of the con-

[2]Another possibility is to turn from actual to hypothetical choice data designed to elicit preference parameters (Barsky et al., 1997), or more radically, away from preference-based measures of well-being altogether (e.g., Benjamin et al., 2012; Kahneman, Wakker and Sarin, 1997). While useful, such approaches are subject to criticisms of their own: for example, any survey-based method is potentially subject to numerous framing effects (e.g., Schwarz and Clore, 1987; Deaton, 2012) and approaches divorced from individual preferences may fail to capture normatively important components of welfare (Loewenstein, 1999). A useful discussion of these and other issues related to behavioral preference recovery is provided in Beshears et al. (2008).

sistent decision-makers can be either partially- or point-identified. We also develop techniques for learning about the preferences of those decision-makers who exhibit choice inconsistencies. Together, these innovations provide a range of practical tools for better understanding the distribution of preferences in populations that exhibit inconsistent choice behavior.

We focus on binary choices in which the option chosen by some decision-makers varies according to a preference-irrelevant feature of the choice environment, which we refer to as a *frame* (Salant and Rubinstein, 2008). Examples of frames might include: (1) which option is presented as the default; (2) the order in which options are displayed; (3) whether the consequence of selecting an option is presented as a loss or a gain; (4) whether the menu of options includes an irrelevant alternative; (5) the point in time at which a decision is made; or (6) whether various consequences of the available options are made salient. Focusing on binary choices and binary frames permits us to view identification through the lens of the potential outcomes framework commonly used in the program evaluation literature (Angrist, Imbens and Rubin, 1996),[3] but the intuition we develop is useful outside of binary settings as well.

Within this framework we make two key identifying assumptions. First, when decision-makers choose consistently across frames, we assume those choices reflect their preferences – an assumption we label the *consistency principle*. This relaxation of the revealed preferences approach permits us to recover preference information in a non-paternalistic manner (Bernheim and Rangel, 2009), without taking a stance on the exact positive model that generates behavior. We also allow some decision-makers to choose inconsistently across frames, but initially we limit our analysis to settings in which the frame pulls the choices of all decision-makers in a uniform direction – an assumption we call *frame monotonicity*. Although frame monotonicity represents additional structure relative to Bernheim and Rangel (2009), it permits us to point-identify the distribution of preferences in datasets where decision-makers are observed under only one frame each. These two reduced-form assumptions are consistent with a wide array of positive models (see Appendix C).

Our first main result concerns the recovery of the preferences of consistent decision-

[3]Unlike all other applications of the potential outcomes framework, our goal is not to identify the causal effects of one variable on another, but rather to remove variation in observed choices due to framing effects, isolating the variation due to preferences.

makers – those who would select the same option under each frame.[4] To do so, we exploit the fact that under our assumptions, a decision-maker who chooses "against the frame" – for example someone who chooses the option that is not the default – is consistent and prefers the option that she chooses. This insight, along with a statistical assumption concerning the assignment of decision-makers to frames, allows us to recover preferences among the consistent decision-makers. Without frame monotonicity, the preferences of the consistent decision-makers are partially identified, and we derive the corresponding bounds.

We next develop techniques to shed light on population preferences using the preferences of consistent choosers. We begin by showing how our assumptions permit partial identification of population preferences. Full identification of population preferences requires understanding the empirical relationship between decision-makers' preferences and their consistency. Intuitively, the first step of our analysis yields preference information for a subset of the population (the consistent decision-makers). By understanding the relationship between decision-makers' likelihood of selecting into that sub-population and their likelihood of having a particular preference, we can extrapolate from the preferences of the consistent decision-makers to the full population. We develop two approaches for studying the empirical relationship between decision-makers' consistency and their preferences.

The first technique is to adjust for observable differences, such as income or education, between the consistent and inconsistent decision-makers, and then to extrapolate from the former to the latter. If consistency and preferences are uncorrelated conditional on these observable characteristics, one can recover population preferences by separately estimating the preferences of each demographic group and then re-weighting those estimates based on the distribution of observables among the inconsistent decision-makers. As in other empirical contexts, the plausibility of this matching-on-observables approach depends on what information about decision-makers can be observed.

The second technique exploits variation in the choice environment related to decision-makers' susceptibility to the frame. A *decision quality instrument* mono-

[4] An alternative interpretation of the empirical evidence concerning choice reversals is to conclude that inconsistent decision-makers simply lack normatively relevant preferences in the first place. For someone who takes that view as a starting point, the contribution of our paper is that it provides a method for isolating the preferences of the consistent decision-makers (which *are* normatively relevant) from the aggregate observed choice data.

tonically affects decision-makers' propensity to choose consistently without otherwise affecting choice. For example, a decision quality instrument could take the form of the time pressure under which a decision must be made: decision-makers faced with greater time pressure may be more likely to choose according to the frame, but time pressure is unlikely to affect which option they actually prefer. Variation in a decision-quality instrument sheds light on the empirical relationship between preferences and consistency by identifying the distribution of preferences for the set of decision-makers whose susceptibility to the frame is affected by the decision quality instrument. We then describe two extrapolation techniques for using this information to estimate population preferences.

One way to understand our techniques for recovering population preferences is by analogy to a canonical sample selection problem. For example, the matching on observables result is closely related to the commonly used technique of re-weighting of samples based on observable demographics to address sample selection. Additionally, the decision quality instrument approach is closely related to the identification of Local Average Treatment Effects (Imbens and Angrist, 1994). Although these parallels are helpful in understanding our results, there is one key difference: in our setting, whether a given individual is consistent—the analogue to selecting into the sample in the parallel—is unobservable. The techniques we develop modify existing tools to overcome this difficulty.

We illustrate our methodology using data on participation in an employer-sponsored pension plan with varying default enrollment regimes, drawn from Madrian and Shea (2001). Automatic enrollment in tax-deferred pension plans is a topic of immense policy interest, but also, due to the difficulty of policymakers' understanding preferences that motivates our paper, much controversy. Applying our approach to this setting uncovers a strong positive relationship between employees' consistency across default regimes (opt-in versus opt-out) and their preferences for enrollment in the pension plan: employees whose enrollment decisions are unaffected by the default are more likely to prefer to enroll. Our results suggest that although most of the inconsistent employees in the firm we study prefer enrollment, a sizable minority (30 percent) do not. Preferences for non-enrollment are disproportionately concentrated among younger and lower-income employees, suggesting there may be value to customizing default options based on employee characteristics.

The paper proceeds as follows: Section 1 sets up the model. Section 2 pro-

vides point- and partial-identification conditions for the preferences of the consistent choosers. Section 3 addresses the problem of recovering preferences for the full population. Section 4 illustrates our approach using data on defaults and enrollment into employer-provided pension plans. The Appendix[5] contains proofs of propositions (A); motivates our parameters of interest with a simple model of optimal frame design (B); considers the relationship between our framework and alternative structural models of default effects (C); generalizes the framework to settings with non-binary frames and non-binary menus (D); and derives standard errors for finite-sample inference (E).

1 Setup

This section introduces the notation and assumptions employed throughout the paper. We observe individual choice data from a population of density 1, with individuals denoted by i. The observed decisions are binary, $y_i \in \{0, 1\}$, and each decision-maker is observed under exactly one of two possible *frames*, denoted d_0 and d_1.[6] Let y_{1i} and y_{0i} denote what i would choose under d_1 and d_0, respectively. Population moments are given by $Y_1 \equiv E[y_{1i}|d_i = d_1]$ and $Y_0 \equiv E[y_{0i}|d_i = d_0]$. Without loss of generality, assume $Y_1 \geq Y_0$. To illustrate the notation using the privacy example from the introduction, y could indicate whether an individual allows a company to use her data, so that d_1 would indicate the opt-out regime, and d_0 would indicate the opt-in regime. We assume throughout that population moments such as these are directly observable, setting aside issues of finite-sample statistical inference.

Decision-makers have ordinal, asymmetric preferences over the available options, denoted by $y_i^* \in \{0, 1\}$. Implicit in this notation is the following assumption:

A1 (*Frame Separability*) For all individuals, y_i^* does not depend on d.

Frame separability limits which features of the decision-making environment are treated as a frame. Features of a decision that affect choice but that are relevant to

[5] Available on the authors' websites.

[6] Our definition of a frame is based on Salant and Rubinstein (2008) and Bernheim and Rangel (2009), and corresponds to what Thaler (2015) refers to as a Supposedly Irrelevant Factor. In settings where the frame is multi-dimensional, such as variation in which option is the default *and* the order in which the options are presented, we can apply this framework using the two most extreme frames – those that make decision-makers most likely and least likely to choose $y = 1$, respectively – as d_1 and d_0. See Appendix D for generalizations beyond the two-option, two-frame setting.

decision-makers' preferences over the available options are *not* frames.[7] Importantly, frame separability does not require decision-makers to be irrational; a decision-making feature that imposed a transaction cost for selecting one of the options would constitute a frame, as long as it did not also affect decision-makers' preference for *receiving* one option or the other.[8]

Decision-makers may either choose consistently or choose in a way that is sensitive to the frame. We denote consistency by $c_i \equiv 1\{y_{0i} = y_{1i}\}$. We assume throughout that the fraction of consistent decision-makers is strictly positive, $E[c_i] > 0$.

When a decision-maker chooses consistently, we assume that her choice reflects her preferences:

A2 (*The Consistency Principle*) For all individuals, $c_i = 1 \implies y_i = y_i^*$.

In the privacy settings example described above, the consistency principle implies that a customer who would choose to keep his data private under both the opt-in and opt-out frame does in fact prefer that his data be kept private. The consistency principle relaxes the instrumental rationality assumption relied on by neoclassical revealed preference analysis, in that the choices made by inconsistent decision-makers need not reveal their preferences. It fails when decision-makers suffer from biases that cause them to make the same mistake under every frame in which they are observed.

Because each decision-maker is observed under only one frame, consistency is not directly observable from the data. If consistency were directly observable, Assumptions A1 and A2 alone would permit the identification of consistent decision-makers' preferences, as in Bernheim and Rangel (2009). The following two assumptions permit us to recover this information under weaker data requirements.

A3 (*Unconfoundedness*) $(y_{0i}\,,\, y_{1i}) \perp d_i$.

Unconfoundedness is a statistical assumption about the process by which decision-makers are assigned to frames. It ensures that differences in observed choices under different frames is due to the effect of the frames rather than to differences in the

[7] For example, if a decision-maker chooses hot chocolate from {hot chocolate, ice cream} under one frame and ice cream from {hot chocolate, ice cream} under the other frame, there would be no apparent deviation from rationality if the frame indicated whether the season was winter or summer. This assumption is explicit in Salant and Rubinstein (2008) and implicit in Bernheim and Rangel (2009), who require it for determining when two potentially conflicting choice situations differ in terms of the frame or in terms of the available menu items. In this sense, frame separability is the property that distinguishes variation in frames from variation in menu items.

[8] Put differently, y_i^* indicates which option a rational individual would select in the absence of transaction costs.

decision-makers assigned to each frame. Unconfoundedness is guaranteed when frames are randomly assigned.

A4 (*Frame Monotonicity*) For all individuals, $y_{1i} \geq y_{0i}$.

Frame monotonicity requires that when a frame affects choice, it does so in the same direction for each affected decision-maker. In the privacy settings example described above, frame monotonicity fails if some customers choose to allow access to their data if and only if doing so is not the default. Much of our discussion assumes frame monotonicity, but we also derive partial identification results for settings in which the assumption fails.

2 Identifying Consistent Preferences

We initially focus on consistent decision-makers, those whose behavior is not affected by the frame. Recovering the preferences of this group would be trivial if decision-makers were observed under each frame; in that case an observer could identify which decision-makers were consistent and, using the consistency principle, which options the consistent decision-makers preferred. However, many real-world datasets do not have this property, and even when they do the order in which decision-makers are exposed to frames may itself affect behavior (LeBoeuf and Shafir, 2003). The following proposition provides conditions for the identification of consistent decision-makers' preferences when each decision-maker is observed under a single frame:

Proposition 1 *Let* $Y_c \equiv \frac{Y_0}{Y_0+1-Y_1}$.

(1.1) *Under A1 - A4,* $E[y_i^*|c_i = 1] = Y_c$.

(1.2) *Under A1 - A3,*

(i) $Y_c > \frac{1}{2} \implies E[y_i^*|c_i = 1] > Y_c$

(ii) $Y_c < \frac{1}{2} \implies E[y_i^*|c_i = 1] < Y_c$

(iii) $Y_c = \frac{1}{2} \implies E[y_i^*|c_i = 1] = Y_c$

Proof By construction, $(y_{0i}, y_{1i}) \in \{(0,0), (1,1), (1,0), (0,1)\}$. Frame monotonicity rules out $(y_{0i}, y_{1i}) = (1,0)$. Therefore we know that $y_{0i} = 1 \iff (y_{0i}, y_{1i}) = (1,1)$, and by the consistency principle, $(y_{0i}, y_{1i}) = (1,1) \implies y_i^* = 1$. Thus,

$E[y_{0i}] = p(y_i^* = 1;\ c_i = 1)$. By the same logic, frame monotonicity and the consistency principle imply that $E[1 - y_{1i}] = p(y_i^* = 0;\ c_i = 1)$. Then by definition, $p(c_i = 1) = E[y_{0i}] + E[1 - y_{1i}]$, and by the definition of conditional probability, $E[y_i^*|c_i = 1] = \frac{E[y_{0i}]}{E[y_{0i}]+E[1-y_{1i}]}$. By unconfoundedness, $Y_1 = E[y_{1i}]$ and $Y_0 = E[y_{0i}]$; substituting these into the previous expression yields 1.1.

The proofs of 1.2 and of all further results are contained in Appendix A. ■

Proposition 1.1 follows from the insight that, under frame monotonicity, only consistent decision-makers choose *against the frame* (i.e., they choose $y = 1$ under d_0 or choose $y = 0$ under d_1). Unconfoundedness guarantees that the assignment of individuals to frames is uncorrelated with preferences or consistency, which means that we can treat the set of decision-makers choosing against the frame as a representative sample of all consistent choosers. Finally, the consistency principle ensures that the observed choices of this group reveal the preferences of the corresponding decision-makers. As a result, the denominator of Y_c measures the fraction of decision-makers that are consistent and the numerator measures the subset of that group with $y_i^* = 1$.

Proposition 1.2 provides a partial identification result that is robust to failures of frame monotonicity. Borrowing terminology from Angrist, Imbens and Rubin (1996), define *frame-defiers* as the subset of inconsistent decision-makers who select $y_i = 1$ if and only if $d = d_0$. Frame-defiers would be misclassified as consistent by Proposition 1.1. Intuitively, decision-makers choosing against the frame under either frame may be either consistent choosers or frame-defiers. Because frame-defiers are assigned to the frames in equal proportions (by unconfoundedness), Proposition 1.1 will classify half of the frame-defiers as choosing $y_i = 1$ consistently and half as choosing $y_i = 0$ consistently. Misclassifying the frame-defiers as consistent therefore biases Y_c toward $\frac{1}{2}$.

Table 1 illustrates this result for hypothetical data on the online privacy controls example described in the introduction. We suppose that 70 percent of individuals allow a company to use their data ($y_i = 1$) under an opt-out policy, but that under an opt-in policy, 40 percent do so. Under frame monotonicity, we can conclude that 70 percent of individuals are consistent across default regimes and that 57 percent of those customers prefer allowing the company to use their data. Without frame monotonicity, we may only conclude that *at least* 57 percent of the consistent customers prefer allowing the company to use their data.

The preference information recovered by Proposition 1 is important for several

Table 1: Aggregate Choices by Frame

	d_1	d_0
Fraction choosing $y = 1$	$Y_1 = 0.70$	$Y_0 = \mathbf{0.40}$
Fraction choosing $y = 0$	$1 - Y_1 = \mathbf{0.30}$	$1 - Y_0 = 0.60$
Fraction consistent, $E[c_i]$, under A1-A4	$Y_0 + 1 - Y_1 = 0.4 + 0.3 = 0.70$	
Consistent preferences, $E[y_i^*\|c_i = 1]$, under A1-A4	$\frac{Y_0}{Y_0+1-Y_1} = \frac{0.4}{0.7} = 0.57$	
Bounds on $E[y_i^*\|c_i = 1]$, under A1-A3	$[0.57\,,\ 1]$	

reasons. First, if one's philosophical starting point is that inconsistent decision-makers lack normatively relevant preferences (see the discussion of this issue in Fischhoff, 1991), Proposition 1 is the end-point of the analysis; our method isolates the normatively-relevant parameter (the consistent decision-makers' preferences) from the noise induced by the frames. Second, when population preferences are known – what Bernheim and Rangel (2009) refer to as a "refinement" – Proposition 1 can be used in conjunction with that information to recover the preferences of the *inconsistent* decision-makers.[9] Such information is often valuable because optimal policy may turn on the preferences of the inconsistent decision-makers (see Appendix B), but observing aggregate population preferences under a refinement does not provide the preferences for that subgroup. Finally, the preferences of the consistent decision-makers may be used to recover the preferences of the remainder of the population by accounting for selection into the consistent sub-population, which is the task we undertake in the remaining sections.

3 Identifying Population Preferences

The remainder of the paper addresses how to use the preferences of the consistent decision-makers to gain information about the preferences of the inconsistent decision-makers or of the full population. The basic challenge to doing so is overcoming a potential selection bias: when selection into the consistent sub-population is not random, characteristics of the consistent decision-makers may be correlated with the preferences of that group. In many ways, this challenge parallels the well-known problem of

[9] Formally, when $E[y_i^*]$ is known, the law of iterated expectations allows us to recover $E[y_i^*|c_i = 0] = \frac{E[y_i^*] - E[y_i^*|c_i=1]\,E[c_i]}{1-E[c_i]}$.

selection into treatment that has been studied in the program evaluation literature. However, an important difference is that in the typical sample selection context, the researcher can identify which units have been selected into the relevant sample.[10] In contrast, whether a particular decision-maker is consistent is unobservable when each decision-maker is observed under a single frame.

To clarify the selection issue, note that we can write

$$E[y_i^*] = E[y_i^*|c_i = 1] - \frac{cov(y_i^*\,,\, c_i)}{E[c_i]}, \tag{1}$$

where the equation follows from the identity $cov(y_i^*, c_i) = E[y_i^*\, c_i] - E[y_i^*]E[c_i]$ and the fact that $E[y_i^*\, c_i] = P(y_i^* = c_i = 1) = E[y_i^*|c_i = 1]E[c_i]$. Equation (1) highlights that recovering population preferences from consistent sub-group preferences requires accounting for the correlation between preferences and consistency (the other parameters in the equation are identified under A1-A4). Moreover, the covariance is a sufficient statistic for identifying population preferences despite uncertainty about the underlying behavioral model; that is, for the purposes of identifying $E[y_i^*]$, the behavioral model only matters to the extent that it shapes $cov(y_i^*, c_i)$. Note that in the special case in which the covariance term is zero – a condition we refer to as *decision quality independence* – the preferences of the consistent decision-makers will be representative of the full population.[11]

3.1 Partial Identification

Absent information on the relationship between preferences and consistency, the distribution of preferences in the population may be partially identified in the spirit of Manski (1989), as follows:

Proposition 2

(2.1) *Under A1-A4,* $E[y_i^*] \in [Y_0,\, Y_1]$.

(2.2) *Under A1 -A3,* $\max\{Y_0 - (1 - Y_1), 0\} \leq E[y^*] \leq \min\{Y_0 + Y_1, 1\}$.

[10] For example, assessing the effect of a job training program on wages may be biased if the program induces some individuals to become employed when they would not have been employed otherwise (e.g., Lee, 2009). In that context, the researcher can observe whether a given individual has wage data and hence whether he or she has been selected into the sample of employed workers.

[11] Decision quality independence is analogous to the familiar "missing at random" assumption in the sample selection literature.

The partial identification result in Proposition 2.1 is quite intuitive: with frame monotonicity, the fraction preferring an option lies between the fraction choosing that option under the two frames. When the fraction of inconsistent decision-makers is large, the bounds will be relatively uninformative.

Without frame monotonicity, we obtain weaker, one-directional bounds for population preferences. The result follows from noting that $E[y_i^*]$ depends on three parameters: $E[y_i^*|c_i = 0]$, $E[y_i^*|c_i = 1]$, and $E[c_i]$. Although $E[y_i^*|c_i = 0]$ is unobservable, the other two parameters can be inferred from the data given information on the prevalence of frame-defiers. Knowing that $E[y_i^*|c_i = 1] \in [0, 1]$ constrains the prevalence of frame defiers, which then yields bounds on the value of $E[y_i^*]$. The further Y_0 is from $1 - Y_1$, the more informative the bounds will be.[12] Note that when frame monotonicity fails, it is possible that a majority of decision-makers choose one option under both frames even though a majority of the population in fact prefers the *other* option.

Using the hypothetical data from Table 1, we would conclude under frame monotonicity that the fraction of the population preferring that their personal data be used is between 40 and 70 percent. Without monotonicity, we can only conclude that this fraction is greater than 10 percent.

To summarize the results thus far, the degree to which $E[y_i^*]$ and $E[y_i^*|c_i = 1]$ can be identified from the data depends on the strength of the researcher's assumptions about the behavioral model. When only A1-A3 are imposed, the data permit partial identification of $E[y_i^*]$ and $E[y_i^*|c_i = 1]$, where the bounds on the former are wider than those on the latter. Adding frame monotonicity permits $E[y_i^*|c_i = 1]$ to be point identified and narrows the bounds on $E[y_i^*]$. Finally, imposing decision quality independence permits point identification of $E[y_i^*]$ as well. The remaining two sections provide alternative identification conditions for $E[y_i^*]$ that rely on frame monotonicity but not on decision quality independence.

3.2 Matching on Observables

In this section, we consider situations where the relationship between preferences and consistency depends on characteristics of decision-makers that are observable to the

[12] When $Y_0 = 1 - Y_1$ exactly, the bounds are entirely uninformative because the data do not constrain the fraction of frame-defiers and, as a result, we cannot rule out $E[c_i] = 0$. Consequently, when $Y_0 = 1 - Y_1$, any $E[y_i^*] \in [0, 1]$ is feasible.

researcher, such as income, education, age, or prior experience with the decision at hand. For example, it could be that more educated customers are less likely to prefer that companies use their personal data and are more likely to choose consistently across default regimes, but that conditional on education, preferences and consistency are independent.

Suppose that decision-makers exhibit observable characteristics $w_i \in W$. The presence of these observables permits us to relax the unconfoundedness assumption:

A3' (*Conditional Unconfoundedness*). For all observable characteristics w, $(y_{1i}, y_{0i}) \perp d_i \,|\, w_i = w$.

Using the observable characteristics to extrapolate from the preferences of consistent decision-makers requires the following assumption:

A5 (*Conditional Decision Quality Independence*) For all individuals and all observable characteristics w, $cov(y_i^*, c_i | w_i = w) = 0$.

Conditional decision quality independence requires that consistent and inconsistent decision-makers with the same observable characteristics have the same distribution of preferences. As with any matching-on-observables approach, the plausibility of this assumption will depend on the detail and quality of the observable characteristics as well as the underlying positive model of behavior. We examine this question in more detail in Appendix C and show that A5 is more likely to hold when variation in consistency is driven by heterogeneity in the cost of optimizing or in the tendency to employ a psychological heuristic (C.1.2), rather than intensity in preferences over the available options (C.2).

The identification strategy we propose in this section is: first, to estimate the preferences of consistent decision-makers with given observable characteristics; second, to extrapolate preferences from consistent to inconsistent decision-makers with the same observable characteristics; and third, to use weighted combinations based on the distribution of observable characteristics to recover preferences in the full population or the sub-population of inconsistent decision-makers.

An important barrier to employing this familiar approach in our context is that we cannot directly observe consistency. The following lemma shows that the distribution of characteristics among the consistent and inconsistent decision-makers is nonetheless identified.[13]

[13]Lemma 1 is analogous to Abadie (2003), who shows how to identify the aggregate observable characteristics of compliers with respect to an instrument when individual compliers cannot be identified. Continuing with the analogy, Proposition 3 is related to Angrist and Fernandez-Val (2013),

Lemma 1 Let $Y_j(w) = E[y_{ji}|d_i = d_j, w_i = w]$ for $j = 0, 1$, $q_w = \frac{Y_0(w)+1-Y_1(w)}{E_w[Y_0(w)+1-Y_1(w)]}$, *and* $s_w = \frac{Y_1(w)-Y_0(w)}{E_w[Y_1(w)-Y_0(w)]}$. *Under A1, A3', and A4:*

(L1.1) *For any* w, $p(w_i = w|\, c_i = 1) = q_w p(w_i = w)$

(L1.2) *For any* w, $p(w_i = w|\, c_i = 0) = s_w p(w_i = w)$.

Apart from its role as a step in the construction of the matching estimator, Lemma 1 is useful in its own right. Information on the observable correlates of consistency is important for researchers investigating the mechanisms by which frames affect decision-making and for policymakers designing interventions aimed at particular sub-groups of the population.[14] Exploiting Lemma 1 along with conditional decision quality independence, the following proposition formalizes the matching-on-observables identification strategy described above:

Proposition 3 *Let* $Y_c(w) = \frac{Y_0(w)}{Y_0(w)+1-Y_1(w)}$. *Under Assumptions A1, A2, A3', A4, and A5*:

(3.1) $E[y_i^*] = E_w[Y_c(w)]$

(3.2) $E[y_i^*|c_i = 0] = E_w[s_w\, Y_c(w)]$[15]

Table 2 illustrates the identification approach for our privacy controls example, reporting hypothetical data conditioned upon whether individuals have at least a high school education. The population moments in the third column match the moments in Table 1. The conditioning reveals that high-school-educated individuals are more likely to be consistent and the consistent choosers among them are more likely to

who exploit information on the distribution of observables to extrapolate an estimated treatment effect from one subset of a population to another.

[14]For example, Thaler and Sunstein (2008) advocate designing frames in ways that offset other decision-making biases. However, for this approach to be valid, it must be that the decision-makers subject to the bias being targeted are also the ones that are sensitive to the frame being set. Lemma 1 helps address this issue by allowing the researcher to determine which types of decision-makers are likely to be sensitive to a given frame. Lemma 1 is also valuable for assessing which types of decision-makers are "more rational" when the researcher is unable to observe repeated decisions by individual decision-makers, as required for the approach developed in Choi et al. (2014).

[15]Replacing assumption A3 with A3' in (1.1) and (1.2) implies that $E[c_i] = E_w[Y_0(w)+1-Y_1(w)]$, and$E[y^*|c_i = 1] = E_w[q_w Y_c(w)]$. The results in Proposition 3 make use of this revised estimator for $E[c_i]$. Even under random frame assignment, the revised estimator for $E[c_i]$ will be preferable for applications of Proposition 3 in finite sample, due to possible spurious correlation between observables and frame assignment. In particular, using the revised estimator ensures the weights implied by (3.1) will sum to one.

Table 2: Average Choices by Frame and High School Education

	HS = 1	HS = 0	Total
Fraction choosing $y = 1$ under d_1, $Y_1(w)$	0.66	0.76	0.70
Fraction choosing $y = 1$ under d_0, $Y_0(w)$	0.56	0.16	0.40
Fraction of population, $p(w)$	0.60	0.40	1.00
Fraction consistent, $E[c_i\|w]$	0.90	0.40	0.70
Fraction of consistent population, $p(w\|c_i = 1)$	0.77	0.23	1.00
Fraction of inconsistent population, $p(w\|c_i = 0)$	0.20	0.80	1.00
Consistent preferences, $E[{y^*}_i\|c_i = 1]$	0.62	0.40	0.57
Inconsistent preferences, $E[{y^*}_i\|c_i = 0]$	0.62	0.40	0.44
Population preferences, $E[y_i^*]$	0.62	0.40	0.53

prefer that the company use their personal data ($y_i = 1$). Under conditional decision quality independence, we conclude that 44 percent of inconsistent decision-makers, and 53 percent of the population prefer that their personal data be used. Under *unconditional* decision quality independence, both these fractions would be 57 percent and we would over-estimate the share preferring that their personal data be used, because this approach ignores the relationship between preferenes and consistency.

3.3 Decision Quality Instruments

Here we develop an approach for settings in which selection into the consistent subpopulation is driven by characteristics that are unobservable to the researcher. Specifically, we introduce the notion of a *decision quality instrument,* which exploits variation in the decision-making environment that affects decision-makers' consistency but that is orthogonal to their preferences. The key difference between this approach and canonical instrumental variables analysis (Imbens and Angrist, 1994) is that our analogue of the first-stage outcome variable, consistency, is unobservable.

Let z denote a decision quality instrument with two values, $z \in \{z_h, z_l\}$. Individual choices now depend on d and z; we denote them by by y_{ijk}, where $j \in \{0, 1\}$ indexes the frame and $k \in \{h, l\}$ indexes the instrument. Consistency is defined at each value of the instrument and denoted by $c_{ik} = 1\{y_{i1k} = y_{i0k}\}$. We denote the fraction of decision-makers observed choosing y under a given (d, z) combination by $Y_{jk} \equiv E[y_{ijk}|d = d_j, z = z_k]$. The following assumptions establish which variation constitutes a valid decision quality instrument:

A3" (*Unconfoundedness of d and z*) $(y_{i1h}, y_{i0h}, y_{i1l}, y_{i0l}) \perp (d_i, z_i)$

A6 (*Decision quality exclusivity*) For all individuals, y_i^* does not depend on z.

A7 (*Decision quality monotonicity*) For all individuals, $c_{ih} \geq c_{il}$ with $E[c_{ih} - c_{il}] > 0$.

Assumption A3" modifies the unconfoundedness assumption, which now requires that both d and z be uncorrelated with confounding factors. Assumption A6 requires that variation in the decision-making environment induced by z is irrelevant from the perspective of decision-makers' preferences; it ensures that z affects behavior by altering consistency, not by changing which option decision-makers prefer.[16] Assumption A7 requires that the effect of z on consistency is weakly monotonic for all decision-makers and strictly monotonic for some.

Variation in z might arise from natural experiments or be induced by researchers. For example, suppose that some decision-makers were randomly assigned to a treatment group aimed at manipulating their "cognitive load" – such as by memorizing a 10-digit number – prior to making the decision being studied. Such experimental designs could plausibly manipulate decision-makers' susceptibility to a frame in ways that are unrelated to their preferences. Other examples of decision quality instruments might include the time pressure for making a decision, the cost of obtaining or processing information about the available choices, the opportunity cost of cognitive resources at the time of decision-making, or the intensity of the frame (e.g., the degree to which one alternative is more salient than another).

3.3.1 Identifying Sometimes-Consistent Preferences

This section develops a reduced-form approach to recover the preferences of those decision-makers whose consistency is affected by a decision quality instrument, which sheds light on the empirical relationship between consistency and preferences.

Proposition 4. *Assume that A1, A2, and A4 hold at each fixed value of z, and assume A3", A6, and A7. Then*

$$E[y_i^* | c_{ih} > c_{il}] = \frac{Y_{0h} - Y_{0l}}{Y_{1l} - Y_{0l} - (Y_{1h} - Y_{0h})}$$

Proposition 4 is best understood by analogy to the identification of a local average

[16] Like A1, A6 does not rule out variation in z affecting welfare by altering the transaction costs associated with choosing against the frame. Indeed, exogenous variation in such costs is an excellent candidate for a decision quality instrument. See Appendix C.

treatment effect (LATE, see Imbens and Angrist, 1994). The monotonicity assumption (A7) permits us to divide the population into three groups of decision-makers: the always-consistent ($c_{ih} = c_{il} = 1$), the sometimes-consistent ($c_{ih} = 1; c_{il} = 0$), and the never-consistent ($c_{ih} = c_{il} = 0$). The denominator of the expression in Proposition 4 measures the decrease in the size of the inconsistent sub-group as we move from z_l to z_h, which identifies the size of the sometimes-consistent group, who are the analog of the compliers in the LATE framework. The expression in the numerator measures the change in the fraction choosing $y = 1$ under d_0 as z changes, which identifies the fraction of decision-makers who are sometimes-consistent *and* prefer $y = 1$.

Several other parallels to the instrumental variables literature are apparent. First, one can use Proposition 4 to motivate over-identification tests of decision quality independence along the lines of Wu (1973) and Hausman (1978). However, such a test requires $E[y^*|c_{ih} > c_{il}] = E[y^*]$, which may fail depending on the nature of selection into consistency. We explore less restrictive alternatives below. Additionally, the types of variation that will satisfy assumptions A6 and A7 depend on the underlying model of behavior that generates framing effects, reflecting a familiar interplay between structural reasoning and instrumental variables. We discuss this issue further in Appendix C. Finally, Proposition 4 may be extended beyond binary decision quality instruments, by applying Proposition 4 to each pair-wise combination of values of z or, when z is continuous, by adapting the methods of Yitzhaki (1996) (see also Heckman and Vytlacil, 2007).[17]

Table 3 illustrates this identification approach for the hypothetical data on privacy controls. We now suppose that the process for adjusting privacy settings may be either onerous (customers are required to navigate through several web pages) or streamlined (customers may adjust privacy settings with a single click). Aggregate choices under the onerous design correspond to the population moments reported in Table 1. Customers are less susceptible to default effects when the process is streamlined. We can back out the fraction of consistent choosers and the aggregate preferences of the consistent choosers when controls are onerous (z_l) or streamlined

[17]Another use for Proposition 4 is motivated by the optimal policy problem facing governments that must choose which z value to implement, for example a regulator deciding how streamlined privacy controls should be. Appendix B shows the solution to this problem trades off the cost of selecting a z that induces greater consistency against the welfare gain from doing so. The latter depends on the preferences of the decision-makers who choose consistently at one candidate z but not in another, which Proposition 4 can be used to estimate.

Table 3: Average Choices by Frame and Difficulty of Changing Privacy Settings

	Onerous (z_l)	Streamlined (z_h)
Fraction choosing $y = 1$ under d_1	0.70	0.55
Fraction choosing $y = 1$ under d_0	0.40	0.45
Fraction consistent, $E[c_i]$	0.70	0.90
Consistent preferences, $E[y_i^*\|c_{ik} = 1]$	0.57	0.50
Sometimes consistent preferences, $E[y_i^*\|c_{ih} > c_{il}]$	0.25	

(z_h) using Proposition 1, as before. Note that the fraction of consistent customers who prefer that the company use their personal data is lower under streamlined controls than onerous ones, because the variation in z affects opt-out rates (Y_1) more than opt-in rates (Y_0). Applying Proposition 4 in this example would imply that of the 20 percent of the population of customers who are sometimes-consistent, only 25 percent prefer that the company use their data, a share substantially below that of the consistent choosers at either z_h or z_l.

An interesting special case of Proposition 4 occurs when, under z_h, all decision-makers are consistent, i.e., $p(c_{ih} = 1) = 1$. For example, default effects may be eliminated by requiring all decision-makers to make an active choice (Carroll et al., 2009). In this case, $E[y_i^*|c_{ih} = 1] = E[y^*]$, so choices under z_h are a "refinement" in which the preferences of the full population is identified, as in Chetty, Looney and Kroft (2009). Furthermore, when $c_{ih} = 1$ for all individuals, $E[y_i^*|c_{ih} > c_{il}] = E[y_i^*|c_{il} = 0]$. Consequently, the statistic Y_s identifies the preferences of the inconsistent choosers at z_l. One can directly recover the preferences of the population *and of the inconsistent decision-makers* from choice data using Proposition 4 in this case.

The next two sections develop identification conditions for population and inconsistent decision-maker preferences that utilize variation in z. On its own, Proposition 4 does not identify these parameters; rather, by shedding light on the covariance between preferences and consistency, it allows us to extrapolate preference information from consistent decision-makers to other groups in the population.

3.3.2 Structural Extrapolation with Decision Quality Instruments

This section develops a latent variable model of the relationship between decision-makers' consistency and their preferences, assuming a bivariate normal distribution for the idiosyncratic terms. With this additional structure, population preference

parameters may be fully characterized using a decision quality instrument.

Suppose that consistency for individual i is determined by

$$P_i = \overline{P} + \theta z_i + \varepsilon_i \tag{2}$$

$$c_i = 1 \iff P_i > 0, \tag{3}$$

where P_i is a latent variable reflecting idiosyncratic variation ε_i and the effect of a binary decision quality instrument $z_i \in \{0, 1\}$. Note that consistency depends on i's choice under both frames, so P_i does not depend on the frame to which i is assigned. Note also that decision quality monotonicity (A7) is satisfied provided $\theta \neq 0$.

Next, suppose the distribution of preferences can also be described with a latent variable model:

$$M_i = \overline{M} + \nu_i \tag{4}$$

$$y_i^* = 1 \iff M_i > 0, \tag{5}$$

where the latent variable M_i simply reflects idiosyncratic variation in preferences, ν_i. Frame separability (A1) is satisfied because M_i does not depend on d, and decision quality exclusivity (A6) is satisfied because M_i does not depend on z_i. Unconfoundedness (A3") is satisfied provided that ε_i and ν_i are independent of z_i and d_i.

Assume that ε_i and ν_i are characterized by a bivariate standard normal distribution:

$$\begin{pmatrix} \varepsilon_i \\ \nu_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{6}$$

where $\rho \in (-1, 1)$ is the correlation between the error terms and where the normalization is without loss of generality. Note that decision quality independence is satisfied if and only if $\rho = 0$.

We close the model with the consistency principle (A2) and frame monotonicity (A4). Together, these assumptions allow us to evaluate the probability of observing a given choice for a given (d, z):

$$\forall i, \forall k = 0, 1,\ y_{i0k} = 1 \iff \varepsilon_i > -\overline{P} - \theta z_k;\ \nu_i > -\overline{M} \tag{7}$$

$$\forall i, \forall k = 0, 1,\ y_{i1k} = 0 \iff \varepsilon_i > -\overline{P} - \theta z_k;\ \nu_i < -\overline{M}. \tag{8}$$

Equations (7) and (8) can be combined with (6) to identify the parameters of the model: $\overline{P}$, $\overline{M}$, θ, and ρ. We can then recover ordinal preferences by integrating the underlying distribution: $E[y^*] = \Phi(\bar{M})$, where $\Phi()$ is the standard normal cumulative density function, and $E[y_i^* | c_{ik} = 0] = \frac{1}{1 - E[c_{ik}]} \int_{-\infty}^{\overline{P} - \theta z_k} \int_{-\overline{M}}^{\infty} \phi^{BVSN}(\varepsilon, \nu; \rho) \partial\nu \partial\varepsilon$, where

$\phi^{BVSN}(a, b; \rho)$ is the bivariate standard normal density with correlation coefficient ρ evaluated at (a, b).

The structural model described above resembles the classic bivariate normal model of selection (see e.g. Heckman, 1979). Variation in the decision quality instrument induces variation in consistency without affecting preferences; this guarantees the relationship between consistency and preferences is identified without relying on functional form alone (Puhani, 2000).[18]

Applying the model to the data from Table III yields an estimated correlation coefficient of $\rho = 0.54$; the positive estimate suggests decision-makers with a high propensity to choose consistently (so that they are consistent at z_l) are more likely to prefer $y_i = 1$ than those with a low propensity to choose consistently. The estimated parameters imply $E[y_i^*] = 0.46$. The population average is below both $E[y_i^*|c_{il} = 1]$ and $E[y_i^*|c_{ih} = 1]$ because it incorporates the preferences of the decision-makers with the very lowest propensity to choose consistently.

3.3.3 Semi-Parametric Extrapolation with Decision Quality Instruments

This section develops an extrapolation approach for recovering population preferences without relying on distributional assumptions. In particular, we model the preferences of the consistent decision-makers at a given value of the decision quality instrument as a flexible polynomial in the fraction of decision-makers who are consistent at that value of the decision quality instrument.[19]

Suppose the decision quality instrument is observed taking on $N + 1$ values, indexed $z_0, z_1, ..., z_N$, and drawn from a continuous ordered set of values, $[\underline{z}, \bar{z}] \subset \mathbb{R}$ such that $E[c_{i\underline{z}}] = 0$ and $E[c_{i\bar{z}}] = 1$. In addition, suppose that decision quality monotonicity holds with respect to any two values of z:

A7' For all individuals and all z, $z' \in [\underline{z}, \bar{z}]$ such that $z > z'$, $c_{iz} \geq c_{iz'}$ and $E[c_{iz} - c_{iz'}] > 0$.

[18] With a binary decision decision quality instrument, the model is just-identified. Additional values of z permit maximum likelihood estimation of the model's parameters.

[19] This approach shares some similarity to the literature on non-parametric identification of marginal treatment effects from local average treatment effects (Heckman and Vytlacil, 2005). An important difference is that the techniques in that literature utilize instrumental variables that drive the propensity to participate in the treatment over a range from 0 to 1. However, recall that in our context, if we were able to observe decisions made under a decision-quality state that induced everyone to choose consistently, we could simply look at the preferences revealed in that state to recover the preferences for the population.

For each individual, let $z_i^* < \bar{z}$ denote the value of z at which she begins to choose consistently, i.e., $z \geq z_i^* \implies c_{iz} = 1$. Assumption A7' implies that z_i^* is unique. Denoting the CDF of z^* by $F(.)$ and the PDF by $f(.)$, we have $E[c_{iz}] = F(z)$. In addition, note that the second part of A7' guarantees $f(z) > 0$ for all $z \in [\underline{z}, \bar{z}]$, so that $F(.)$ is strictly increasing with a well-defined inverse function over $E[c_{i\bar{z}}] \in [0, 1]$, which we denote $F^{-1}(E[c_{iz}])$.

Finally, let $g(z) = E[y_i^*|z_i^* = z]$ denote the preferences of the *marginally consistent* decision-makers at a given z.[20] To guarantee the validity of the Taylor Series approximation that underpins the following result, it will be convenient to assume that both $F(z)$ and $g(z)$ are infinitely differentiable with respect to z.

Proposition 6 *Under A1, A2, A6, and A7',* for any degree $D \in \mathbb{N}$, there are constants $a_0 ... a_D$ such that

(6.1) *For any z,*[21] $E[y_i^*|z_i^* = z] \approx a_0 + a_1 E[c_{iz}] + a_2 E[c_{iz}]^2 ... + a_D E[c_{iz}]^D$

(6.2) *For any z,* $E[y_i^*|c_{iz} = 1] \approx a_0 + \frac{a_1}{2} E[c_{iz}] + \frac{a_2}{3} E[c_{iz}]^2 + ... \frac{a_D}{D+1} E[c_{iz}]^D$

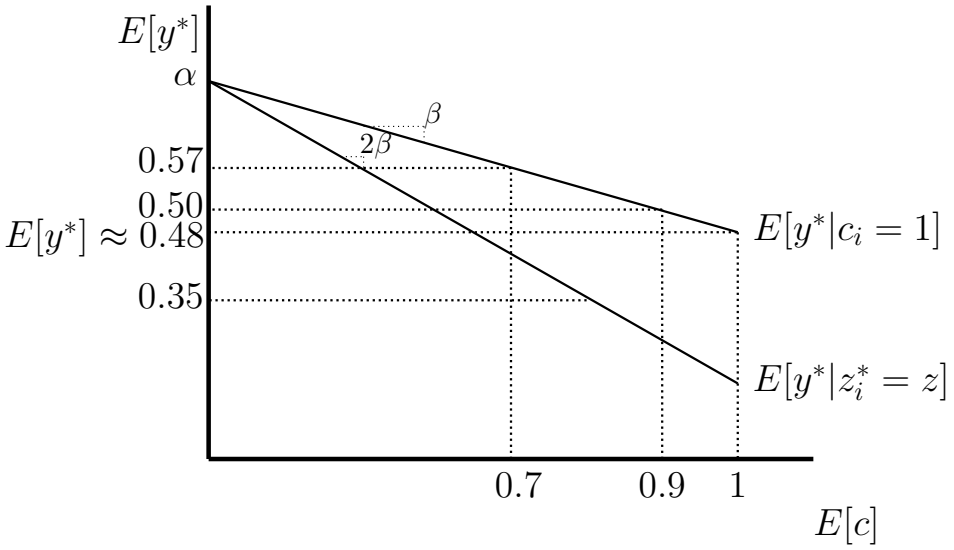
(6.3) $E[y_i^*] \approx a_0 + a_1 + ... + a_D$

Proposition 6 implies that the preferences of the consistent decision-makers at a particular value of the decision quality instrument can be approximated by a polynomial function in the fraction of decision-makers who choose consistently at that value of the instrument. Because A7' guarantees a one-to-one mapping between z and $E[c_{iz}]$, we can write the preferences of the marginally consistent decision-makers as a function of the fraction of decision-makers choosing consistently, i.e. $E[y_i^*|z_i = z] = g(F^{-1}(E[c_{iz}]))$. In addition, infinite differentiability of g and F ensure imply the composite function $h \equiv g \circ F^{-1}$ will have a well-defined Taylor Series approximation of degree D. We then obtain 6.2 by integrating the marginal preference function $h(.)$ from $E[c_{iz}] = 0$ to $E[c_{iz'}]$ and scaling by $E[c_{iz'}]$ for any arbitrary z'. Finally, 6.3 follows directly from setting $E[c_{iz}] = 1$ in 6.2. Note that when $N = D$, we will have $D + 1$ equations in $D + 1$ unknowns, so that $a_0, ..., a_D$ are just-identified. When $N > D$, we will have more equations than unknowns, and a best-fit technique such as least-squares can be used to estimate $a_0, ..., a_D$.

[20] Although by definition $z \in \mathbb{R}$, the value of z itself may be unobservable to the researcher.

[21] The approximation disregards terms of order D and higher, i.e. those of the form $E[c_{iz}]^k$ where $k \geq D$, as do the approximations in (6.2) and (6.3).

Figure 1: Extrapolation from Preferences of Consistent and Marginally Consistent Choosers



Proposition 6 allows us to quantify the relationship between consistency and preferences, which can be used to estimate population preferences, without the rigid functional form assumption underpinning the bivariate normal model in the previous section. Figure 1 illustrates the approach with data from Table 3 under a linear functional form assumption: $E[y^*|c_i(z) = 1] = \alpha + \beta E[c_i(z)]$. With two values of z, α and β are just-identified: $\alpha \approx 0.82$ and $\beta \approx -0.35$. Applying 6.2 yields $E[y_i^*] = 0.48$. As in the bivariate normal model, accounting for the relationship between consistency and preferences leads us to conclude that the fraction of the population preferring $y_i = 1$ is less than the fraction of the consistent choosers preferring $y_i = 1$ under either value of the decision-quality instrument.

4 Application to 401(k) Automatic Enrollment

In this section we illustrate our identification framework by analyzing data on enrollment decisions into employer-provided 401(k) pension plans. Depending on how the plan is designed, new employees may have to actively enroll in the plan in order to participate, or their enrollment may be automatic unless they choose to opt out. An influential body of research documents striking differences in take-up and savings behavior between such opt-in and opt-out regimes (Madrian and Shea, 2001; Choi

et al., 2006; Chetty et al., 2014). Most policy discussion of automatic enrollment takes as its starting point that employees would be better off under 401(k) plan designs that cause them to save more than they otherwise would. Our framework allows us to investigate employee preferences empirically, without imposing paternalistic assumptions of this form. Moreover, although others have studied the welfare effects of defaults in this setting (Carroll et al., 2009; Bernheim, Fradkin and Popov, 2015), an important advantage to our approach is that it does not require taking a stance on the particular positive model that generates any decision-making bias.[22] Only models that violate the consistency principle are entirely ruled out. In addition, an advantage of the reduced-form nature of our approach is that it makes transparent what role each assumption plays in identification, such as how frame monotonicity and conditional decision quality independence strengthen the conclusions that can be drawn from the data.

4.1 Data

To study the effect of automatic enrollment on pension plan participation, we use data from the large health care and insurance firm studied in Madrian and Shea (2001). This firm switched from an opt-in to an opt-out enrollment design in April 1998,[23] and Madrian and Shea (2001) find that the switch caused a large increase in the fraction of employees choosing to participate in the firm's pension plan. We use the choice data from this study to investigate employee preferences for plan participation.

We observe whether an employee enrolls in the plan (indicated by y_i) and whether the default is opt-in (d_0) or opt-out (d_1) when he or she is hired. We also observe annual compensation, age, sex, and race for each employee.[24] Table 4 describes employee

[22]Bernheim, Fradkin and Popov (2015) consider a variety of positive models of default effects and welfarist assumptions, and derive welfare inferences that are robust to a range of alternative assumptions. Importing each of these into our setting reveals that frame monotonicity and the consistency principle are satisfied for each of the possibilities they consider (see Appendix C).

[23]Under automatic enrollment, employees who were automatically enrolled faced a default contribution rate of 3 percent. Refer to Madrian and Shea (2001) for additional details regarding the data and the change in enrollment policy. Data from other studies suggests that raising the default contribution rate would increase the share opting out of enrollment under automatic enrollment (e.g. Choi et al., 2006; Bernheim, Fradkin and Popov, 2015), as would be suggested by a model of costly opt-out. In our setting, this finding implies that at a higher default contribution rate, the fraction of employees that prefer to be enrolled in a 401(k) is *lower* than our estimates for the fraction preferring enrollment under a 3 percent default.

[24]To ensure individual employees could not be identified, we were provided compensation and age

Table 4: Employee Characteristics

	Hired under opt-in	Hired under opt-out	Full sample
Compensation			
<$20K	10.5	12.7	11.8
$20K-$29K	37.7	45.6	42.2
$30K-$39K	18.6	16.5	17.4
$40K-$49K	15.2	11.2	12.9
>$50K	18.0	14.1	15.7
Age			
<30 years	30.9	37.4	34.6
30-39 years	36.0	33.3	34.5
40-64 years	33.1	29.3	30.9
Race			
White	72.4	69.5	70.8
Non-white	27.6	30.4	29.2
Gender			
Male	23.1	21.0	21.8
Female	76.9	79.0	78.1
Observations	4,185	5,702	9,887

Source: Disaggregated data from Madrian and Shea (2001) provided to the authors.
Notes: All reported tabulations are percentages of the total sample with a given characteristic. Due to data sharing agreements, a small number of observations in the original sample (1.8 percent of the original dataset) were dropped from our analysis, causing our sample characteristics to differ slightly from those reported in the original study.

characteristics. The distribution of characteristics is not substantially different across default regimes. The income, age, and racial composition of its workforce are typical of a large employer in the US, although this firm's workforce is predominately female. Employee participation rates by frame and demographic group are summarized in Table IV of Madrian and Shea (2001).

For our identification results to apply to the choices in this data, the conditions described in Section 1 must be satisfied. Frame separability seems likely to hold: it is difficult to imagine that an employee's preferences over how much to save depend on how her employer chooses to structure enrollment into its sponsored retirement plan. Unconfoundedness requires that an employee's hiring date be uncorrelated

only within a range of values. See Table 4.

with whether she chooses to participate under either plan design. This is the same assumption necessary to identify the causal effect of the change in enrollment from the change in plan design, and Madrian and Shea (2001) provide evidence this assumption is satisfied. Frame monotonicity requires that no employee chooses to enroll when enrollment is opt-in but chooses not to enroll when enrollment is opt-out, which seems plausible in this setting.

Of our assumptions, the one that is perhaps least likely to hold in this setting is the consistency principle, which requires that employees who would make the same participation decision under both opt-in and opt-out enrollment actually prefer the option that they choose. If some of the employees who choose not to participate under either plan design are present-biased, they might be better off participating in the plan – despite the fact that doing so is the opposite of their (consistently) revealed preference. Importantly, not all forms of present bias would cause the consistency principle to fail. For example, in the model of default-sensitivity studied by Carroll et al. (2009), present-bias causes individuals to procrastinate and stick with the default savings plan until they make an active choice, but when they do make an active choice the amount they choose to save is optimal. Such behavior satisfies the consistency principle because those individuals who choose consistently have selected their welfare-maximizing option. Alternatively, a government may wish to adopt the consistency principle for purposes of policy design even when it is suspect, if one of its goals is to respect individuals' choices and avoid paternalism (Bernheim and Rangel, 2009). Ultimately, the policy implications of our findings in this section depend on whether one believes that employee enrollment decisions made consistently across frames reveals normatively meaningful preference information. We return to this caveat in our discussion of the results below.

4.2 Results

To begin, we focus on the preferences of the consistent decision-makers. Table 5 columns 1 and 2 report the aggregate enrollment rates under the two policy designs. The estimated population participation rates are $\hat{Y}_1 = 0.859$ under opt-out and $\hat{Y}_0 = 0.491$. Columns 3 and 4 apply Proposition 1 to this data. Substituting the estimated population moments into the definition of Y_c in Proposition 1 yields $Y_c = 0.777$, with a standard error of 0.006. Thus, under frame monotonicity, of the 63.2 percent

of employees whose enrollment decisions are sensitive to the enrollment design, we conclude that a large majority (77.7 percent) preferred enrollment. Without assuming frame monotonicity, Proposition 1.2 implies the fraction of consistent employees that prefer enrollment is *at least* 77.7 percent.

Turning to population preferences, the bounds provided by Proposition 2 are quite coarse because of the large fraction of inconsistent decision-makers. With frame monotonicity, we can conclude that the fraction of the population preferring enrollment lies somewhere between 0.491 and 0.859. Without frame monotonicity, we can only rule out values of $E[y_i^*]$ below 0.350. Additional structure is needed to draw more precise conclusions from the data.

We next investigate heterogeneity among employees in their preferences for enrollment and their sensitivity to the enrollment regime. Although we cannot directly observe either preferences or consistency for individual employees, the results from section 3.2 allow us to investigate differences based on employees' observable characteristics. We estimate a regression of the form

$$E[y_i|d,w] = \alpha_0 + \alpha_1 1\{d = d_1\} + w_i^{'}\beta_0 + w_i^{'}\beta_1 1\{d = d_1\} \tag{9}$$

where y_i and d are defined as above and w_i is a vector of employee characteristics. Applying Proposition 1 (conditional on a given realization of employee characteristics) implies that:

$$E[c_i|w_i = w] = 1 - \alpha_1 - w^{'}\beta_1 \tag{10}$$

$$E[y_i^*|c_i = 1, w_i = w] = \frac{\alpha_0 + w^{'}\beta_0}{1 - \alpha_1 - w'\beta_1} \tag{11}$$

The results of the analysis are reported in Table 6.[25] The results suggest that both consistency and the preferences of consistent choosers vary systematically and substantially by employee characteristics. Variation in consistency is strongly related to variation in compensation, with those in the highest compensation bin (annual income over \$50K), estimated to be 40 percent more likely to choose consistently than those in the lowest bin (annual income less than \$20K). The estimated differences in consistency by income are statistically significant ($p < 0.001$).[26] When compensa-

[25]To facilitate interpretation of the results, Column 2 of Table 6 reports the average effect on $E[y_i^*|c_i = 1, w_i = w]$ of a change in each component of w (relative to a "left-out" group), holding fixed the other components of w.

[26]This finding adds to a growing literature that documents important differences in susceptibility to decision-making biases by income (e.g., Mullainathan and Shafir, 2013; Goldin and Homonoff,

Table 5: Enrollment in 401(k) Plans and Employee Preferences

	Enrollment Rates		Proposition 1: Consistent Preferences		Proposition 3: Matching on Observables	
	(1)	(2)	(3)	(4)	(5)	(6)
	Opt-out	Opt-in	Percent consistent	Percent of consistent preferring enrollment	Percent of inconsistent preferring enrollment	Percent of population preferring enrollment
Estimate	85.9	49.1	63.2	77.7	70.8	74.9
Standard Error	(0.46)	(0.77)	(0.90)	(0.63)	(0.98)	(0.70)
Observations	5702	4185	9887	9887	9887	9887

Notes: Estimates are based on calculations on data from Madrian and Shea (2001) provided to the authors. Due to data sharing agreements, a small number of observations in the original sample (1.8 percent of the original dataset) were dropped from our analysis, causing our sample characteristics to differ slightly from those in Madrian and Shea (2001). This causes a slight difference between columns (1) and (2) and the first two columns of Table IV in Madrian and Shea (2001). Estimates in columns 5 and 6 apply Proposition 3, matching on income, age, sex, and race.

tion is controlled for, differences in consistency are not significantly associated with heterogeneity in age, race, or gender.

Turning to preferences for enrollment among consistent employees, we document significant heterogeneity here as well. As with consistency, differences by income are striking. Employees in the highest compensation bin are 41 percent more likely to prefer enrollment than those in the lowest compensation bin. Unlike consistency, preferences for enrollment also vary by age, race, and gender, even after controlling for income.

The fact that both consistency and preferences for enrollment among the consistent decision-makers are associated with income casts doubt on the plausibility of decision-quality independence in this setting. To further assess the validity of the assumption, Figure 2 plots consistency and preferences for enrollment among the consistent employees in each observable cross-section of the data. The scatter plot provides additional (non-parametric) evidence against decision-quality independence: employee groups that contain a greater fraction of consistent decision-makers are also more likely to have a greater fraction of consistent decision-makers that prefer 401(k) enrollment.[27] The estimated slope of the best-fit line is 0.78, suggesting a strong positive relationship between employees' consistency and their preferences. Figure 2 also reveals that a majority of consistent employees in every group with income less than $20,000 and age less than 30 prefer non-enrollment.[28]

Because of the apparent correlation between preferences for enrollment and consistency in the population of employees, we apply the matching estimator described in Proposition 3 to estimate the distribuion of preferences for inconsistent employees and the full population. To maximize the likelihood that the conditional decision quality independence assumption is satisfied, we use a fully interacted econometric model: each combination of observables defines a unique demographic group. Consequently, for the results to be invalid there must be unobserved variation in consistency among employees of the same age, gender, race, and income that is also correlated with preferences for enrollment.[29]

2013; Choi et al., 2014). Unlike Choi et al. (2014), our approach allows us to identify patterns in consistency without observing individuals making multiple decisions.

[27]Under conditional decision-quality independence, the fraction of the consistent preferring enrollment within a cell equals the fraction of the population in that cell preferring enrollment.

[28]In fact, with one exception these were the only groups for which a majority prefer non-enrollment.

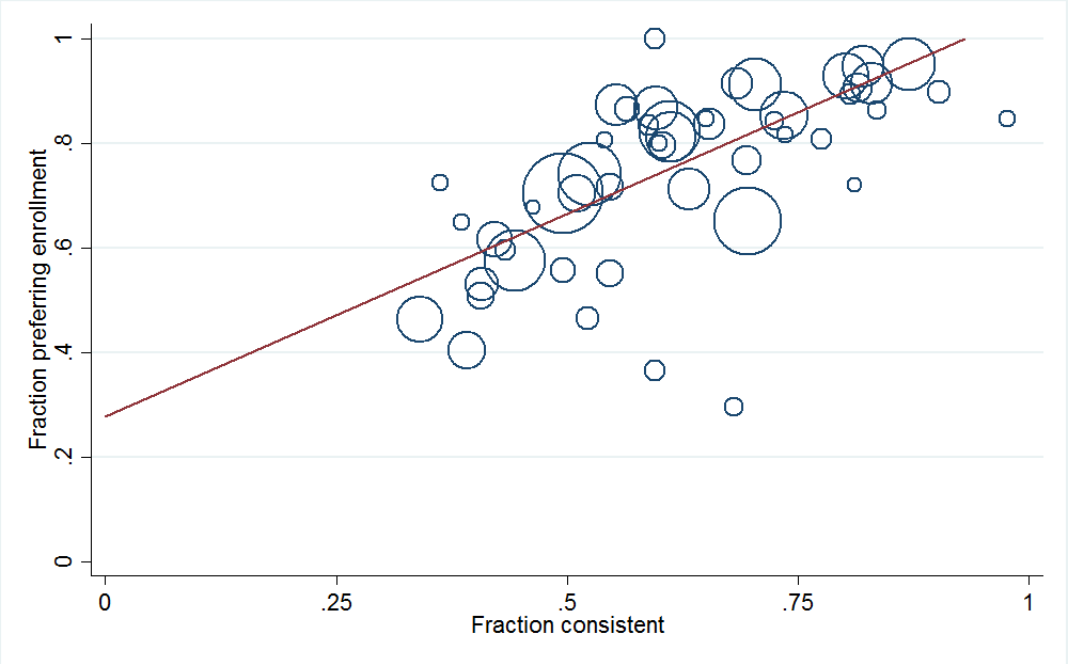
[29]For example, if cognitive ability is positively correlated with both consistency and preferences among employees of the same age, gender, race, and income, our results would yield an upwardly-

Table 6: Consistency and Preference by Observable Characteristics

	(1) Consistency	(2) Preferences of Consistent Choosers
Compensation		
$20K-$29K	0.123***	0.197***
	(0.028)	(0.032)
$30K-$39K	0.218***	0.319***
	(0.033)	(0.033)
$40K-$49K	0.267***	0.368***
	(0.035)	(0.034)
>$50K	0.398***	0.407***
	(0.034)	(0.033)
Age		
30-39 years	-0.033	0.008
	(0.022)	(0.017)
40-64 years	0.025	0.068***
	(0.023)	(0.017)
White	-0.015	0.087***
	(0.021)	(0.016)
Male	0.003	-0.059***
	(0.021)	(0.017)
Observations	9,887	9,887

Source: Disaggregated data from Madrian and Shea (2001) provided to the authors.
Notes: Estimates are based on equations (9)-(11). The left-out groups for each demographic characteristic are 1) employees with compensation less than $20K, 2) employees with age less than 30 years, 3) non-white employees, and 4) female employees. Column (1) reports the change in the probability that an employee with a given characteristic is consistent relative to the left-out group, controlling for other characteristics. Column (2) reports the average increase in the probability, relative to the left-out group, of a consistent chooser with the given characteristic preferring enrollment, holding other characteristics constant. Standard errors calculated using the delta method are reported in parentheses.
*** indicates $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Figure 2: Consistency versus Preference for Enrollment in a 401(k) Plan



Notes: Estimates are based on calculations on data from Madrian and Shea (2001) provided to the authors. Each point on the bubble scatter plot consists of all workers with given values of compensation, age, sex, and race. The fraction consistent and fraction of consistent decision-makers preferring enrollment are calculated using the take-up rates before and after automatic enrollment in each cell. The size of the cell is proportional to the area of the circle.

The last two columns of Table 5 present the results of the matching analysis. We estimate that the fraction of inconsistent employees preferring enrollment is 70.8 percent – approximately 7 percentage points lower than the corresponding fraction of consistent employees. The difference in estimated preferences between the consistent and inconsistent employees is statistically significant, allowing us to reject the hypothesis of decision quality independence ($p < 0.001$). For the full population of employees, we estimate that 74.9 percent prefer enrollment.

4.3 Discussion

Our results suggest that a sizable majority of inconsistent employees, approximately 70 percent, prefer enrollment in this employer's pension plan. This finding is consistent with the more paternalistic view that leading these individuals to save more via automatic enrollment would improve their welfare. Notably, however, we also find that a majority of the youngest and lowest-income workers do *not* prefer enrollment in the plan.[30] This could be, for example, because employees in this group perceive retirement to be far off and are therefore inattentive to retirement savings, and also prefer to save less because they anticipate higher future earnings or are currently paying off student loan debt. To the extent an employer wishes to implement a personalized default regime along the lines suggested in Sunstein (2015), our results suggest a default of non-enrollment may be welfare-maximizing for young, low-income employees.[31]

biased estimate for the preferences of the inconsistent employees. The bias in the matching estimator is given by $E[y_i^*] - E_w[Y_c(w)] = E_w\left[\frac{cov(y_i^*, c_i|w)}{E[c_i|w]}\right]$. This expression follows from the law of conditional expectations and Equation (1).

[30] This finding does not rely on frame monotonicity: since $Y_c(w) < 0.5$ for these employees, Proposition 1.2 implies $E[y_i^*|c_i = 1] < Y_c(w)$, and conditional decision quality independence then implies $E[y_i^*|c_i = 0] < 0.5$ as well.

[31] An alternative possibility is that employees in this group are simply more present-biased, and this bias leads many of them to consistently and sub-optimally opt out of the pension plan under automatic enrollment. This possibility, alluded to above, would violate the consistency principle. Additional data on the timing of decisions might help to resolve this ambiguity, but our present data cannot do so. Ultimately, our results suggest that either young, low-income employees prefer non-enrollment or that the choices of decision-makers in this group are so biased that they should be disregarded when selecting a pension enrollment regime.

5 Conclusion

Recovering preferences from choice data is a fundamental challenge in behavioral economics. Our results provide a practical framework for approaching the problem. We showed how one can recover preference information for consistent decision-makers, as long as consistent choices reveal preferences. In doing so, the problem of population preference identification is transformed into one of accounting for potentially endogenous selection into the subpopulation of consistent decision-makers. The transformed problem is both more familiar and more tractable than the original: economists have developed a wide range of tools for dealing with endogeneity challenges of this sort. The second part of the paper adapted a number of these tools to this unfamiliar setting. These techniques account for the relationship between consistency and preferences, allowing researchers to overcome the endogeneity issue under a range of conditions. Using one of these techniques to analyze enrollment in pension plans suggests that automatic enrollment benefits most workers, with the exception of younger and lower-income employees.

Like Bernheim and Rangel (2009), our approach is most appealing in settings where preference identification is important but the researcher is not confident in the precise model of behavior that generates the inconsistency. Unlike Bernheim and Rangel (2009), many of our results require assumptions beyond the requirement that consistent choices reveal preferences. However, the payoff to this additional structure is substantial: it allows us to apply our framework to weaker datasets (those in which decision-makers are observed only once and the researcher lacks ex ante knowledge over which choices are optimal) and to sheds light on the preferences of those decision-makers whose choices are sensitive to the frame. Moreover, even when our behavioral assumptions are exactly the same as those required by Bernheim and Rangel (2009), our framework provides new partial-identification results for the preferences of the consistent choosers and for the full population.

Although focusing on binary menus and binary frames simplifies the analysis, our approach is useful outside of such settings. Appendix D develops several generalizations to more complicated choice settings. Notably, a number of our results extend in a straightforward way to ordered menus with two frames and multiple options. We also develop generalizations to settings with multi-dimensional frames or multiple frames that vary in their intensity.

An important feature of our approach is its reduced-form nature. Within the wide range of models consistent with frame-monotonicity and the consistency principle, the basic identification problem – i.e., understanding the empirical correlation between decision-makers' preferences and their sensitivity to frames – is the same regardless of the details of the structural model that generates behavior. On the other hand, our approach is not a replacement for structural models of decision-making. As in other areas of empirical economics, the interpretation of the parameters identified by reduced-form approaches depends on the underlying structural model that generates behavior.[32] Finally, the reduced-form nature of our approach has the virtue of making transparent which assumptions are driving preference identification within a particular application.

The framework studied here can be thought of as a special case of a more general approach in which an observer first identifies the preferences of a reference group of decision-makers whose choices are assumed to reveal their true preferences, and subsequently extrapolates the reference group's preferences to the rest of the population. In our approach, the reference group consists of those decision-makers who choose consistently across frames. When using consistent choosers as the reference group is not feasible or credible, one might replace them with experts, experienced choosers, or those thought to be immune to the framing effect in question (e.g., Johnson and Rehavi, 2015; Bronnenberg et al., 2013; Handel and Kolstad, 2015). The identification techniques we have proposed may be utilized with these reference groups as well; for example, one might adjust the recovered preferences of experts based on observable characteristics before extrapolating their preferences to the rest of the population, or utilize exogenous variation that causes some individuals to become experts.

Our results are subject to several limitations. First, in certain applications even consistent choices may not reveal preferences. For example, decision-makers who consistently choose one retirement plan over another, regardless of the default option, may still be choosing sub-optimally based on, for example, present bias. Similarly, biases in judgment and perception – such as over-optimism or a tendency to underweight low-risk events – may manifest themselves consistently across frames. Many of these failures can be attributed within our framework to the presence of "missing" frames,

[32] As described in Appendix C, understanding the underlying structural model provides guidance about which types of control variables are needed for conditional decision quality independence to hold and about which types of variation constitute valid decision quality instruments.

which affect behavior but do not vary in the data available to the researcher. Accurately identifying preferences in such contexts requires additional data or assumptions that permit the analysis to move further away from observed choice behavior.

Second, although we have attempted to develop identification strategies that may be applied to data, the credibility of such strategies will turn on whether their assumptions are met in the application at hand. Insofar as one is skeptical that the required assumptions will be satisfied in any setting, our results highlight the difficulty in conducting even weakened forms of revealed reference analysis in the presence of framing effects. Further work, perhaps combining the current framework with data on subjective well-being, could attempt to empirically assess the validity of the underlying assumptions about welfare made here.

A third limitation is that the preference information we recover will not be sufficient to determine optimal policy in all settings. For example, when choices subject to framing effects generate externalities – such as rules for organ donations (Abadie and Gay, 2006; Johnson and Goldstein, 2003) or environmental incentives (Homonoff, 2014) – the distribution of private preferences, while still important, is not the only relevant parameter for setting policy. Similarly, when choosing against the frame causes decision-makers to incur utility costs, Appendix B shows that the optimal choice of frame depends on the intensity of preferences, not just their ordinal content, as well as the magnitude of the utility costs. As in non-behavioral settings, identifying cardinal preference information from binary choices requires additional data or richer structure than what we impose here. Developing methods to identify additional preference information and incorporate it into optimal policy prescriptions is an important task for future research.

References

Abadie, Alberto. 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics*, 113(2): 231–263.

Abadie, Alberto, and Sebastien Gay. 2006. "The Impact of Presumed Consent Legislation on Cadaveric Organ Donation: A Cross-Country Study." *Journal of Health economics*, 25(4): 599–620.

Angrist, Joshua, and Ivan Fernandez-Val. 2013. "Extrapolate-ing: External Validity and Overidentification in the LATE Framework." *Advances in Economics and Economet-*

rics, , ed. Daron Acemoglu, Manuel Arellano and Eddie Dekel. Cambridge University Press.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91(434): 444–455.

Barsky, Robert B, F Thomas Juster, Miles S Kimball, and Matthew D Shapiro. 1997. "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study." *The Quarterly Journal of Economics*, 112(2): 537–579.

Basu, Kaushik. 2003. *Prelude to Political Economy: A Study of the Social and Political Foundations of Economics.* Cambridge University Press.

Benjamin, Daniel J, Miles S Kimball, Ori Heffetz, and Alex Rees-Jones. 2012. "What Do You Think Would Make You Happier? What Do You Think You Would Choose?" *The American Economic Review*, 102(5): 2083–2110.

Benkert, Jean-Michel, and Nick Netzer. 2014. "Informational Requirements of Nudging." Working Paper.

Bernheim, B Douglas. 2009. "Behavioral Welfare Economics." *Journal of the European Economic Association*, 7(2-3): 267–319.

Bernheim, B Douglas, and Antonio Rangel. 2009. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *The Quarterly Journal of Economics*, 124(1): 51–104.

Bernheim, B Douglas, Andrey Fradkin, and Igor Popov. 2015. "The Welfare Economics of Default Options in 401(k) Plans." *American Economic Review*, 105(9): 2798–2837.

Beshears, John, James J Choi, David Laibson, and Brigitte C Madrian. 2008. "How are Preferences Revealed?" *Journal of Public Economics*, 92: 1787–1794.

Bronnenberg, Bart, Jean-Pierre Dube, Matthew Gentzkow, and Jesse Shapiro. 2013. "Do Pharmacists Buy Bayer? Sophisticated Shoppers and the Brand Premium." Working Paper.

Carroll, Gabriel D, James J Choi, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *The Quarterly Journal of Economics*, 124(4): 1639–1674.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *The American Economic Review*, 99(4): 1145–1177.

Chetty, Raj, John N Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen. 2014. "Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark." *The Quarterly Journal of Economics*, 129(3): 1141–1219.

Choi, James J, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2006. "Saving for Retirement on the Path of Least Resistance." In *Behavioral Public Finance: Toward a New Agenda.* , ed. Edward J McCaffery and Joel Slemrod. Russell Sage Foundation.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. "Who Is (More) Rational?" *The American Economic Review*, 104(6): 1518–1550.

Deaton, Angus. 2012. "The Financial Crisis and the Well-Being of Americans." *Oxford Economic Papers*, 64(1): 1–26.

De Clippel, Geoffroy, and Kareen Rozen. 2014. "Bounded Rationality and Limited Datasets." Working Paper.

Fischhoff, Baruch. 1991. "Value Elicitation: Is There Anything in There?" *American Psychologist*, 46(8): 835.

Goldin, Jacob, and Tatiana Homonoff. 2013. "Smoke Gets in Your Eyes: Cigarette Tax Salience and Regressivity." *American Economic Journal: Economic Policy.*

Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *The American Economic Review*, 103(7): 2643–2682.

Handel, Benjamin R, and Jonathan T Kolstad. 2015. "Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare." Working Paper.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica*, 46(6): 1251–1271.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*, 47(1): 153–161.

Heckman, James J, and Edward J Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments." *Handbook of Econometrics*, 6: 4875–5143.

Heckman, James J, and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Ho, Daniel E, and Kosuke Imai. 2008. "Estimating Causal Effects of Ballot Order from a Randomized Natural Experiment: The California Alphabet Lottery, 1978–2002." *Public Opinion Quarterly*, 72(2): 216–240.

Homonoff, Tatiana A. 2014. "Can Small Incentives Have Large Effects? The Impact of Taxes versus Bonuses on Disposable Bag Use." Working Paper.

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 467–475.

Johnson, Eric J, and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science*, 302(5649): 1338–1339.

Johnson, Eric J, Steven Bellman, and Gerald L Lohse. 2002. "Defaults, Framing and Privacy: Why Opting In-Opting Out." *Marketing Letters*, 13(1): 5–15.

Johnson, Erin M, and M Marit Rehavi. 2015. "Physicians Treating Physicians: Information and Incentives in Childbirth." Working Paper.

Kahneman, Daniel, Peter P Wakker, and Rakesh Sarin. 1997. "Back to Bentham? Explorations of Experienced Utility." *The Quarterly Journal of Economics*, 112(2): 375–406.

LeBoeuf, Robyn, and Eldar Shafir. 2003. "Deep Thoughts and Shallow Frames: On the Susceptibility to Framing Effects." *Journal of Behavioral Decision Making*, 16: 77–92.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies*, 76(3): 1071–1102.

Loewenstein, George. 1999. "Because It Is There: The Challenge of Mountaineering... for Utility Theory." *Kyklos*, 52(3): 315–343.

Madrian, Brigitte C, and Dennis F Shea. 2001. "The Power of Suggestion: Inertia in 401 (k) Participation and Savings Behavior." *The Quarterly Journal of Economics*, 116(4): 1149–1187.

Manski, Charles. 1989. "Anatomy of the Selection Problem." *The Journal of Human Resources*, 24(3): 343–360.

Mullainathan, Sendhil, and Eldar Shafir. 2013. *Scarcity.* Times Books.

Puhani, Patrick. 2000. "The Heckman Correction for Sample Selection and its Critique." *Journal of Economic Surveys*, 14(1): 53–68.

Rubinstein, Ariel, and Yuval Salant. 2012. "Eliciting Welfare Preferences from Behavioural Data Sets." *The Review of Economic Studies*, 79(1): 375–387.

Salant, Yuval, and Ariel Rubinstein. 2008. "(A, f): Choice with Frames." *The Review of Economic Studies*, 75(4): 1287–1296.

Schwarz, Norbert, and Gerald Clore. 1987. "Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States." *Journal of Personality and Social Psychology*, 45: 512–523.

Sen, Amartya K. 1973. "Behaviour and the Concept of Preference." *Economica*, 40(159): 241–259.

Sunstein, Cass R. 2015. *Choosing Not to Choose: Understanding the Value of Choice.* Oxford University Press.

Thaler, Richard. 2015. *Misbehaving: The Making of Behavioral Economics.* W.W. Norton.

Thaler, Richard H, and Cass R Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Yale University Press.

Wu, De-Min. 1973. "Alternative Tests of Independence Between Stochastic Regressors and Disturbances." *Econometrica*, 41(4): 733–750.

Yitzhaki, Shlomo. 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business & Economic Statistics*, 14(4): 478–486.