Revealed Preference Analysis with Framing Effects*

Jacob Goldin† Daniel Reck‡

October 1, 2018

Abstract

In many settings, decision-makers' behavior is observed to vary based on seemingly arbitrary factors. Such framing effects cast doubt on the welfare conclusions drawn from revealed preference analysis. We relax the assumptions underlying that approach to accommodate settings in which framing effects are present. 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 sensitivity to the frame. We develop tools for studying this relationship and illustrate them with data on automatic enrollment into pension plans.

JEL Classification: C10, D60, I30

*The authors wish to thank Oriana Bandiera, Doug Bernheim, Sebastien Bradley, Charlie Brown, Ryan Bubb, Raj Chetty, Henry Farber, Nikolaj Harmon, James Hines, Bo Honore, Louis Kaplow, Sarah Kotb, Miles Kimball, Alvin Klevorick, Henrik Kleven, Claus Kreiner, David Lee, Charles Manski, Alex Mas, Ted O'Donoghue, Søren Leth-Petersen, Joel Slemrod, and seminar participants at Berkeley, Carnegie Mellon, the University of Chicago, the University of Copenhagen, Cornell, the Federal Reserve Board, Harris School of Public Policy, Harvard, the London School of Economics, the University of Michigan, New York University, Oxford, Princeton, and Stanford, for helpful discussion and comments. We are especially grateful to Brigitte Madrian and Aon Hewitt for providing us with data, and to Charlie Rafkin for his excellent assistance with the data. An earlier version of the article was circulated under the title "Preference Identification Under Inconsistent Choice."

†Stanford University and NBER; email: jsgoldin@stanford.edu

‡London School of Economics; email: d.h.reck@lse.ac.uk

Introduction

In many settings, people's 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 decision are salient. Such *framing effects* cast doubt on the welfare conclusions that emerge from revealed preference analysis. For example, suppose an internet company adopts an opt-out data collection policy, under which it can use a customer's data to target advertising, unless the customer indicates otherwise. Prior research suggests that switching to an opt-in policy, under which customers must give permission before the company can use their data in this way, would reduce the number 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 65 percent do so when the policy is opt-out. Both policies let customers control the use of their data, but the choices observed under the two policies imply different conclusions about what customers prefer.

In this paper, we relax the assumptions underlying revealed preference analysis to accommodate choice data contaminated by framing effects. We focus on binary decisions in which the choices of some decision-makers vary according to a preference-irrelevant feature of the 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) the reference point from which an option is evaluated; (4) whether the menu of options includes an irrelevant alternative; (5) the point in time at which a decision is made; or (6) which features of the available options are made salient. We assume that when decision-makers choose consistently across frames, those choices reflect their preferences.[1]

Within this framework, we derive conditions for identifying preferences of various groups of decision-makers. First, we show that when a frame pulls the choices of all decision-makers in a uniform direction (frame monotonicity), one can identify the distribution of preferences for the consistent decision-makers – the subgroup unaffected by the framing effect. This is true even when each decision-maker is observed making only one decision and observers lack ex ante knowledge about which decision-makers are consistent. Under frame monotonicity, a decision-maker who chooses "against

[1]By preferences, we mean the relative degree to which the available options further a decision-maker's objectives, whatever those may be. Preferences are not defined according to observed choices; doing so would assume away the possibility of framing effects.

the frame" – for example, someone who chooses the option that is not the default – is consistent and prefers the option that she chooses. This fact, along with an exogeneity assumption concerning the assignment of decision-makers to frames, allows us to point-identify the preferences of the consistent decision-makers. Without frame monotonicity, the preferences of this group are partially identified, and we derive the corresponding bounds.

Next, we turn to the problem of identifying preferences for the full population of decision-makers. Our key insight is that this problem shares important features with the classic selection-into-treatment problem from the program evaluation literature. That is, once we have identified preferences for the subgroup of decision-makers who are consistent, we can account for selection into that subgroup to recover preferences for the overall population. Stated this way, the transformed problem is both more familiar and more tractable than the original: economists have developed a range of tools for dealing with endogeneity problems of this sort, and we adapt several to our setting.

The first approach we develop is to extrapolate the preferences of the consistent decision-makers to the inconsistent decision-makers by adjusting for observable differences between the two groups. Recovering population preferences requires that consistency and preferences be uncorrelated conditional on these observables. As in other settings where researchers rely on matching or regression, the plausibility of this assumption depends on the nature of selection and what information about decision-makers the researcher can observe.

Second, we develop decision quality instruments, which exploit variation in decision-makers' susceptibility to a frame, but not affect decision-makers' preferences. For example, decision-makers who are experimentally manipulated to have a higher shadow value of leisure (e.g., by facing greater time pressure) may be more likely to choose according to the frame, but such a manipulation is unlikely to affect which option they actually prefer. Decision quality instruments identify the distribution of preferences for those decision-makers whose susceptibility to the frame they affect. We develop techniques to extrapolate the preferences of this subgroup to the population.

Finally, we derive bounds on population preferences, based on the consistent decision-makers. The usefulness of the bounds depend on the strength of the frame – the bounds are tighter when when more decision-makers are consistent. One surprising finding from this analysis is that absent frame monotonicity, it may be that

a majority of the population prefers one option even though a majority selects the *other* option under every frame that is observed.

A growing literature confronts the problem of preference identification in settings with framing effects. One proposal is to restrict preference inferences to the subset of observed choices in which a given decision-maker chooses consistently (Bernheim and Rangel, 2009). In practice, however, individual decision-makers are typically observed choosing under only one frame, which makes it difficult to detect which choices are consistent. Worse, this approach yields no information on the inconsistent decision-makers – the very group whose behavior is shaped by the choice of frame. Further "refinements" can provide a path forward if the researcher can observe choices in a frame in which all decision-makers are known to select their most-preferred option (Chetty, Looney and Kroft, 2009; Allcott and Taubinsky, 2015), but in many applications, such as those in which behavior is sensitive to defaults or ordering effects, there is little reason to believe that any of the observed frames satisfies this condition.

A different solution is to rely on a positive model of behavior that fully specifies the mapping from decision-makers' preferences to their (potentially sub-optimal) behavior (Rubinstein and Salant, 2012). Inverting the model allows one to recover preferences from the decision-makers' observed choices. However, in many cases the resulting welfare conclusions are sensitive to the researcher's choice between competing positive models that are difficult to distinguish observationally. In some cases, even a fully specified behavioral model is insufficient to recover preferences from choice data (Benkert and Netzer, Forthcoming).

We contribute to this literature by developing a framework for preference identification that strikes a middle ground between these approaches. Relative to Bernheim and Rangel (2009), our approach requires additional structure, but the payoff to that additional structure is significant: one can apply our results to the realistic class of settings in which individual decision-makers are observed under only one frame and in which the researcher is not confident that any one of the observed frames induces all decision-makers to choose optimally. Relative to imposing a specific behavioral model, an advantage of our framework is its generality: our central behavioral assumptions – that consistent decision-makers choose optimally and frame monotonicity – hold under a wide range of models for why framing effects occur. Consequently, our approach can recover the preferences of the consistent decision-makers, as well as bounds on population preferences, while remaining reasonably agnostic about the precise under-

lying model that generates the observed framing effect. In contrast, point-identifying population preferences requires pinning down the relationship between preferences and consistency, and implicitly, restricting the behavioral model that governs which decision-makers are sensitive to the frame. We develop empirical tools to shed light on this relationship under a range of assumptions about the available data and underlying behavioral model. Finally, our framework compliments model-based approaches by making transparent the role the model's assumptions play in identification: within a broad class of models, distinguishing between behavioral or functional form assumptions matters only to the extent the assumptions offer conflicting predictions for the relationship between preferences and consistency.[2]

We illustrate our framework using data on participation in an employer-sponsored pension plan with varying default enrollment regimes. We show how the preference information that can be recovered from the data depend on the strength of the assumptions the researcher is willing to impose. Under relatively weak assumptions,we find that a sizable majority of the consistent employees prefer enrollment. We also document a strong positive relationship between employees' sensitivity to framing effects (opt-in versus opt-out enrollment) and their preferences for enrollment in the plan. We conclude that under plausible assumptions, the data imply a majority of employees prefer participation in the pension plan, but that there is significant heterogeneity. For example, employees that are likely to leave the firm within two years disproportionately prefer not to participate.

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 those settings as well, as we illustrate in an extension. We also describe how the preference information identified by our approach can be combined

[2]We also contribute to a strand of literature that uses the preferences of a reference group of decision-makers (whose choices are assumed to be optimal) as a guide to the rest of the population. In previous work, the reference group consists of experts, identified based on information about experience, occupation, or familiarity with the subject matter (Bronnenberg et al., 2015; Handel and Kolstad, 2015; Johnson and Rehavi, 2016). Our contribution is to develop a method of applying this approach to settings characterized by framing effects, in which no ex ante information is available to identify members of the reference group. Rather, inclusion in the reference group (composed of the consistent decision-makers) emerges endogenously from observable decision-making behavior.

[3]Unlike other applications of the potential outcomes framework of which we are aware, 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.

with price variation to estimate traditional measures of cardinal welfare.

The paper proceeds as follows: Section 1 presents our notation and main assumptions. Section 2 presents our results relating to the preferences of the consistent decision-makers. Section 3 presents our results relating to the full population. Section 4 illustrates our identification results using data on defaults and enrollment into employer-provided pension plans. Section 5 describes how the preference information we recover can be combined with price variation to derive conventional measures of cardinal welfare. The Online Appendix contains: proofs of propositions; derivations of standard errors; additional results relating to decision quality instrument extrapolation; supplementary material relating to the empirical application; and, generalizations relating to non-binary frames, non-binary menus, and settings in which decision-makers' assignment to frames is non-random.

1 Formal Framework

This section introduces our notation and core assumptions.

Each decision-maker i is observed to choose from a fixed menu $\mathbf{S} = \{0, 1\}$ under one of two possible frames $D_i \in \{0, 1\}$.[4] Let $Y_i(0)$ and $Y_i(1)$ denote what i would choose under frames $D_i = 0$ and $D_i = 1$, respectively. Decision-makers have strict ordinal preferences over the available options, with the most-preferred option denoted by $Y_i^* \in \{0, 1\}$. Each decision-maker is characterized by a vector of random variables $(Y_i(0),\ Y_i(1),\ D_i,\ Y_i^*)$, drawn from some underlying population distribution. For each i, the researcher observes the pair $(Y_i,\ D_i)$, where $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$. The researcher does not observe Y_i^*, $Y_i(0)$, or $Y_i(1)$.

We denote mean choices among decision-makers assigned to each frame by $\overline{Y}(1) \equiv E[Y_i|\, D_i = 1]$ and $\overline{Y}(0) \equiv E[Y_i|\, D_i = 0]$. Unless otherwise noted, the operator $E[.]$ denotes an expectation over the population distribution of the random variable inside the square brackets (which may be a function of primitive random variables). We assume for exposition that these population moments are directly observable to the researcher, deferring issues of finite-sample statistical inference to the Online Appendix. Without loss of generality, $\overline{Y}(1) \geq \overline{Y}(0)$. To illustrate the notation using the privacy example from the introduction, let Y_i indicate whether i allows a company to use her data, $D_i = 1$ indicate the opt-out regime, and $D_i = 0$ indicate the opt-in regime, so that $\overline{Y}(1) = 0.65$ and $\overline{Y}(0) = 0.40$.

[4] Because the menu is constant across decision-makers, our notation conditions on it implicitly. The Online Appendix considers generalizations to non-binary menus and non-binary frames.

Decision-makers either choose consistently or choose in a way that is sensitive to the frame. We denote consistency by $C_i = 1 - |Y_i(1) - Y_i(0)|$. Because each decision-maker is observed once, C_i is not observed. We assume that the fraction of consistent decision-makers is strictly positive, $E[C_i] > 0$.

The following four assumptions form the core of our analysis.

A1 (*Frame Separability*) For all i, Y_i^* does not depend on D.

Frame separability is an assumption about the content of decision-makers' preferences. It is implicit in the above notation because we do not index Y_i^* by D_i – otherwise, individuals would have two random variables for Y_i^* (one for each frame). A1's role is to define which features of the decision-making environment are treated as a frame.[5] Features of a decision that affect choice but that are relevant to decision-makers' preferences over the available options are *not* frames. For example, if a decision-maker chooses hot chocolate from {hot chocolate, ice cream} under $D = 0$ and ice cream from {hot chocolate, ice cream} under $D = 1$, there would be no framing effect if D indicates whether the season is winter or summer. Importantly, frame separability does not require decision-makers to be irrational; a feature of the environment that imposes a cognitive cost for selecting one of the options would constitute a frame, as long as it did not also affect decision-makers' preference for ending up with one option or the other.[6]

The remaining assumptions concern the distribution of $(Y_i(0),\ Y_i(1),\ D_i,\ Y_i^*)$ in the population.

A2 (*Frame Exogeneity*) $(Y_i(0)\,,\ Y_i(1)) \perp D_i$.

Frame exogeneity is an assumption about the data-generating process by which decision-makers are assigned to frames. The assumption ensures that differences in observed choices under different frames are due to the effect of the frames, rather than to differences in the decision-makers observed under each frame. Frame exogeneity is guaranteed when decision-makers are randomly assigned to frames.[7]

We now turn to the link between choices and preferences. The standard revealed preferences approach is to infer decision-makers' preferences directly from their

[5] 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.

[6] For a discussion of related issues, see Bernheim and Taubinsky (Forthcoming).

[7] The Online Appendix considers several generalizations in which frame exogeneity is relaxed.

choices:

RPA (*Revealed Preferences Assumption*) For all i, $Y_i^* = Y_i$.

In our setting, a *framing effect* occurs when A1 and A2 are satisfied and one observes $\overline{Y}(1) \neq \overline{Y}(0)$. By definition, framing effects violate the RPA assumption. This is because A2 implies $\overline{Y}(0) = E[Y_i(0)]$ and $\overline{Y}(1) = E[Y_i(1)]$, and $E[Y_i(0)] \neq E[Y_i(1)]$ implies $Y_i(1) \neq Y_i(0)$ for some subset of decision-makers. But for any such decision-maker, A1 and RPA imply $Y_i(1) = Y_i^* = Y_i(0)$, yielding a contradiction. The next assumption weakens RPA to accommodate choice data in which framing effects are present:

A3 (*Consistency Principle*) For all i, $C_i = 1 \implies Y_i = Y_i^*$.

Under the consistency principle, preferences are revealed by choices only for decision-makers who choose consistently across frames. Because consistency is a function of $Y_i(0)$ and $Y_i(1)$, the assumption constrains the joint distributions of $Y_i(0)$, $Y_i(1)$, and Y_i^*. It is easy to see that the consistency principle weakens RPA by the minimum necessary to accommodate an apparent framing effect. In the online privacy example described above, the assumption implies that a customer who would choose to keep her data private under both the opt-in and opt-out frames does in fact prefer that her data be kept private. Like RPA, the assumption fails when decision-makers suffer from biases that cause them to make the same mistake under every frame in which they are observed.

The following monotonicity assumption permits us to recover aggregate information about consistency even though C_i is not observable at the individual level.

A4 (*Frame Monotonicity*) For all i, $Y_i(1) \geq Y_i(0)$.

Frame monotonicity requires that when a frame affects choice, it does so in the same direction for each affected decision-maker. It thus constrains the joint distribution of $Y_i(0)$ and $Y_i(1)$. In the online privacy example, 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 will assume frame monotonicity, but we also derive partial-identification results for settings in which it fails.

Examples of Framing Effects

The following are examples of behavioral models that might generate a particular observed framing effect.

Example 1: Default Effects Our running example will concern default effects. The frame, $D \in \mathbf{S}$, encodes which option is the default. Decision-makers choose

according to their (fixed) preferences $u_i(Y)$ over a subset $\Gamma_i \subseteq \mathbf{S}$ of options that they consider (as in Masatlioglu, Nakajima and Ozbay, 2012). Decision-makers can be either active or passive. When active, decision-makers consider both options, $\Gamma_i(D) = \mathbf{S}$. When passive, they consider only the option that is the default, $\Gamma_i(D) = D$. Because u_i does not depend on the default, frame separability is satisfied. To see that the consistency principle holds, suppose some individual i chooses the same option under both defaults, $Y_i(0) = Y_i(1) = 1$. Because i chooses 1 under $D = 0$, we know that $1 \in \Gamma(0)$, and therefore that i is active under $D = 0$, with $u_i(1) > u_i(0)$. To see that frame monotonicity holds, suppose $Y_i(0) = 1$ (when $Y_i(0) = 0$, the condition holds trivially). As above, $Y_i(0) = 1$ implies i is active under $D = 0$ and prefers option 1. Because $1 \in \Gamma_i(1)$ as well, we know that $Y_i(1) = 1$. Hence, $Y_i(1) \geq Y_i(0)$.

Example 2: Time Inconsistency At date t, decision-makers choose between receiving some amount y_0 at date $t+k$ or some other amount $y_1 > y_0$ at date $t+k+1$. As in Laibson (1997), individuals choose according to the following behavioral utility function: $\tilde{u}_{it} = u(z_t) + \beta_i \sum_{k=1}^{\infty} \delta_i^t u(z_{t+k})$, where z_t is total income at time t, $\beta_i \leq 1$, and $\delta_i^t < 1 \ \forall t$. The frame denotes whether $k = 0$ ($D = 0$) or $k > 0$ ($D = 1$). Assuming the amounts in y_0 and y_1 are small relative to background income, it is straightforward to show that when $D = 0$, i chooses y_1 iff $\frac{y_0}{y_1} < \beta_i \delta_i$, and that when $D = 1$, i chooses y_1 iff $\frac{y_0}{y_1} < \delta_i$. Welfare may be evaluated either according to $\beta_i \delta_i^t$ (the short-term view) or δ_i^t (the long term view) (Bernheim, 2009). In either case, it is straightforward to verify that the consistency principle holds and that frame monotonicity holds as long as $\beta_i \leq 1$ for all i. In addition, frame separability holds as long as the welfare-relevant discount rate does not vary within an individual based on the time the decision is made (as it would under $\tilde{u}_{it}$).

Example 3: Bias Unrelated to Observed Framing Effect Consider a setting in which both default effects and present-bias are present. As above, decision-makers choose between y_0 at date $t + k$ and $y_1 > y_0$ at date $t + k + 1$, but now either y_0 or y_1 is set to be the default. Welfare is given by $u_{it} = u(z_t) + \sum_{k=1}^{\infty} \delta_i^t u(z_{t+k})$. As in Example 1.2, some decision-makers are passive, and choose whichever option is set as the default. The other decision-makers are active (but present-biased), and choose according to $\tilde{u}_{it} = u(z_t) + \beta_i \sum_{k=1}^{\infty} \delta_i^t u(z_{t+k})$. Choices are observed only at $k = 0$, with decision-makers randomly assigned across defaults. The consistency principle fails here because the decision-makers who are consistent with respect to the default are present-biased: an active decision-maker chooses y_1 if $\frac{y_0}{y_1} < \beta_i \delta_i$ but prefers y_1 if

$\frac{y_0}{y_1} < \delta_i$. Hence, some of those who consistently select y_0 would actually prefer y_1. This example highlights that our framework can recover preferences only when any mistakes are due to an *observed* framing effect. In settings where a bias is present, but no inconsistency is observed, our approach (like traditional revealed preference analysis) would incorrectly infer preferences from choices. If choices were observed under each frame and also under both $k = 0$ and $k > 0$, one could apply our approach to first eliminate the framing effect at each value of k and second to use those results to estimate the behavioral parameters β and δ.

2 Identifying Consistent Preferences

We initially focus on the consistent decision-makers – i.e., 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* $\overline{Y}_C \equiv \frac{\overline{Y}(0)}{\overline{Y}(0)+1-\overline{Y}(1)}$.

(1.1) *Under A1 - A4,* $E[Y_i^* | C_i = 1] = \overline{Y}_C$.

(1.2) *Under A1 - A3,* $\overline{Y}_C \geq \frac{1}{2} \iff E[Y_i^* | C_i = 1] \geq \overline{Y}_C$

Proposition 1.1 follows from the insight that, under frame monotonicity, only consistent decision-makers choose *against the frame* (i.e., they choose $Y_i(0) = 1$ or $Y_i(1) = 0$). Frame exogeneity 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. In turn, 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 $\overline{Y}_C$ measures the fraction of decision-makers who are consistent and the numerator measures the subset of that group with $Y_i^* = 1$. A formal proof of Proposition 1.1, and of all further results, is contained in the Online Appendix.

Proposition 1.2 provides a partial identification result that is robust to failures of frame monotonicity. Borrowing terminology from Angrist, Imbens and Rubin (1996),

Table 1: Illustration of Proposition 1

	Choose not to enroll, opt-in regime, $Y_i(0) = 0$	Choose to enroll, opt-in regime, $Y_i(0) = 1$
Choose not to enroll, opt-out regime, $Y_i(1) = 0$	0.35	0.00
Choose to enroll, opt-out regime, $Y_i(1) = 1$	0.25	0.40

Under frame monotonicity, top-right quadrant= 0; top-left quadrant= $1 - \overline{Y}(1)$; bottom-right quadrant = $\overline{Y}(0)$; bottom-left quadrant= $\overline{Y}(1) - \overline{Y}(0)$.

Fraction consistent:	$E[C_i] = 0.4 + 0.35 = 0.75$
Fraction of consistent preferring option 1:	$E[Y_i^* \mid C_i = 1] = \overline{Y}_C = \frac{0.4}{0.4+0.35} \approx 0.53$
Bounds on consistent preferences, without A4:	$0.53 \leq E[Y_i^* \mid C_i = 1] \leq 1$ $\overline{Y}_C$ is biased toward $\frac{1}{2}$, as frame defiers are equally assigned to the top-left and bottom-right quadrants by frame exogeneity (by definition they cannot be in the bottom-left quadrant).

define *frame-defiers* as the subset of inconsistent decision-makers who select $Y_i(0) = 1$ and $Y_i(1) = 0$. Frame-defiers would be misclassified as consistent by the logic underlying Proposition 1.1. To understand the intuition behind the proof, note that decision-makers choosing against the frame under either frame may be either consistent choosers or frame-defiers. Because frame-defiers are assigned to the two different frames in equal proportions (by frame exogeneity), Proposition 1.1 will classify half of the frame-defiers as choosing $Y_i = 1$ consistently and half as choosing $Y_i = 0$ consistently. Ignoring the presence of frame-defiers therefore biases $\overline{Y}_C$ toward $\frac{1}{2}$.

The reasoning behind Proposition 1 is further illustrated in Table 1, which applies the result to the online privacy example from the Introduction. The preference information recovered by Proposition 1 is important for several reasons. First, if one's philosophical starting point is that inconsistent decision-makers lack normatively relevant preferences (see Fischhoff, 1991), Proposition 1 is the end-point of the analysis; it isolates the normatively relevant parameter from the noise induced by

the frames.[8] 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] The preferences of this group can be an input into welfare calculations (see Section 5), but are not directly revealed under a refinement. 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 our focus in the next section.

3 Identifying Population Preferences

This section develops several methods for using the preferences of the consistent decision-makers to shed light on the rest of the population. To frame the problem, one can use the law of iterated expectations to write:

$$E[Y_i^*] = E[C_i]\, E[Y_i^* |\, C_i = 1] + (1 - E[C_i])\; E[Y_i^* |\, C_i = 0] \tag{1}$$

From Proposition 1.1, $E[C_i]$ and $E[Y_i^*|C_i = 1]$ are identified under A1-A4, but $E[Y_i^*|C_i = 0]$ is entirely unrestricted. Consequently, the formal challenge in recovering $E[Y_i^*]$ is the same as the standard sample selection problem that arises in the program evaluation literature (Manski, 1989).[10] Specifically, when selection into the consistent sub-population is non-random, the preferences of that group can yield a biased estimate for the preferences of the population:[11]

$$E[Y_i^*] = E[Y_i^*|C_i = 1] - \frac{1}{E[C_i]}\; \mathrm{cov}\,(Y_i^*\,,\, C_i) \tag{2}$$

However, when susceptibility to the frame is uncorrelated with preferences – a condition we refer to as *consistency independence* – Equation (2) highlights that $E[Y_i^*] = E[Y_i^*|C_i = 1]$. In that case, Proposition 1 permits identification of population pref-

[8] For example, suppose the fraction of voters supporting a referendum varies based on the question wording. It is straightforward to show that randomizing the wording evenly across two options biases the average vote share toward 0.5, which can affect the outcome if the referendum requires a super-majority to pass. In cases like this, isolating the average choices of the consistent voters may be the best option.

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

[10] One important difference is that in the typical sample selection context, the researcher can identify which units have been selected into the sample and which have not. In contrast, consistency is unobservable in our setup.

[11] Equation (2) follows from the definitions of covariance and conditional expectation.

erences without further adjustments. More generally, Equation (2) shows that under A1 - A4, the covariance between preferences and consistency is a sufficient statistic for identifying the ordinal preferences of the population; the particulars of the behavioral model only matter to the extent they shape this relationship.

The nature of the selection in (1) and (2) depends on the model determining which agents are consistent. To illustrate, consider the following two potential models for default effects, both of which are nested by the default effects example in Section 1.

Decision-Making Types Model of Default Effects Suppose that sensitivity to the default is an innate characteristic or influenced by factors such as education or prior experience that are exogenous to the specific choice being considered (as in Chetty et al., 2014). The distribution of active types depends on the following statistical model: $C_i = 1 \iff \widetilde{C}_i \geq 0$, $\widetilde{C}_i = \beta^C \theta_i^C + \eta_i^C$, where θ_i^C denotes the vector of individual characteristics that determine whether one is active, and η_i^c denotes idiosyncratic variation across individuals. Similarly, $Y_i^* = 1 \iff \widetilde{Y}_i \geq 0$, $\widetilde{Y}_i = \beta^Y \theta_i^Y + \eta_i^Y$, where θ_i^Y denotes the vector of characteristics that determine ordinal preference and η_i^Y denotes idiosyncratic variation. Assume η_i^C and η_i^Y are independent of the other random variables in the model and of one another. We then have $\text{cov}(Y_i^*, C_i) = p(\widetilde{Y}_i > 0;\ \widetilde{C}_i > 0) - p(\widetilde{Y}_i > 0)p(\widetilde{C}_i > 0)$. Consistency independence fails if θ_i^C and θ_i^Y contain common characteristics, or characteristics that are correlated in the population of decision-makers.

Bounded Rationality Model of Default Effects Suppose instead that variation in consistency is driven in part by the utility stakes of the choice at hand. Decision-makers must incur a cognitive cost, $\gamma_i \geq 0$, to choose actively and consider an option that is not the default. Decision-makers first choose whether to be active or passive, and then choose from the options they consider. Let $\Delta u_i = u_i(1) - u_i(0)$, and let $F_\Delta(.)$ denote its cumulative distribution over the population of decision-makers. Here, consistency will be determined by the net benefit to i of choosing her most-preferred option when that option is not the default, $\widetilde{C}_i = |\Delta u_i| - \gamma_i$, and we let $F_{\widetilde{C}}(.)$ describe its cumulative distribution. Decision-makers are active iff $\widetilde{C}_i > 0$, so the fraction of consistent decision-makers in the population is given by $E[C_i] = 1 - F_B(0)$. Similarly, the fraction with $Y_i^* = 1$ is given by $E[Y_i^*] = 1 - F_\Delta(0)$. One can then derive the relationship between preferences and consistency as $\text{cov}(Y_i^*, C_i) = (1 - F_\Delta(0)) \left[F_{\widetilde{C}}(0) - F_{\widetilde{C}|\Delta u>0}(0)\right]$. Setting aside the trivial case in which preferences are uniform, consistency independence requires $F_{\widetilde{C}}(0) = F_{\widetilde{C}|\Delta u>0}(0)$. This would

obtain, for instance, when the foregone utility from following the default when the default is not optimal is symmetric across agents with opposite ordinal preferences. Intuitively, the identification challenge in this model is that consistency depends on the "stakes" of the decision, and empirically, the stakes may differ among decision-makers with different ordinal preferences. When this source of variation is unimportant relative to decision-maker characteristics in explaining consistency, as would be the case when the variance of $|\Delta u_i|$ is negligible relative to the variance in γ_i, this model approximates the decision-making types model described above (see also Appendix Figure 3).

Note that in this example we assume the decision-maker knows Δu_i with certainty (see also Conlisk, 1996). One could alternatively suppose that the individual decides whether to choose actively based on whether the *expected* net benefit of doing so exceeds the cost. As it is nested by the model presented in Example 1, such a model would not violate our core assumptions. The model imposes a similar set of challenges for extrapolation to population preferences as the model with certainty. We discuss additional nuance introduced by this type of model in Online Appendix A.

These examples illustrate that consistency independence is likely to fail in many applications. The remainder of this section proposes a range of empirical methods for shedding light on the relationship between preferences and consistency.

3.1 Partial Identification

The following result clarifies the limits of what can be learned about population preferences without the imposition of additional behavioral assumptions.

Proposition 2

(2.1) *Under A1-A4,* $E[Y_i^*] \in \left[\overline{Y}(0), \overline{Y}(1)\right]$.

(2.2) *Under A1 -A3,* $\max\left\{\overline{Y}(0) - (1 - \overline{Y}(1))\,,\, 0\right\} \leq E[Y^*] \leq \min\left\{\overline{Y}(0) + \overline{Y}(1)\,,\, 1\right\}$.

The result in (2.1) follows directly from the equivalence of the standard sample selection problem and our setting, once the assumptions required for Proposition 1 are assumed (Manski, 1989). Intuitively, the fraction of the population that prefers an option lies between the fraction choosing that option under each of the two frames. As a result, the bounds will be relatively informative when the fraction of inconsistent decision-makers is small.

Without frame monotonicity, we obtain weaker, one-directional bounds for population preferences. The result in (2.2) follows from substituting the partial identification results in Proposition (1.2) into Equation (1). In particular, $E[Y_i^*|C_i = 1]$ and

$E[C_i]$ can be identified 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 $\overline{Y}(0)$ is from $1 - \overline{Y}(1)$, the more informative the bounds will be.[12] Notably, when frame monotonicity fails, (2.2) shows it is possible that a majority of decision-makers choose one option under both frames even though the *other* option is preferred by a majority of all decision-makers.

3.2 Adjusting for Observable Correlates of Consistency

A classic approach to overcoming selection problems is to condition on observables. Such an approach is useful here when the correlation between preferences and consistency is driven by characteristics of decision-makers that are observable to the researcher.

Formally, suppose that decision-makers exhibit a vector of observable characteristics, denoted by random variable $X_i \in \mathbf{X}$. Define $\overline{Y}(D, X) = E[Y_i(D)|\, D_i = D, X_i = X]$ to be the cell-specific analogs to the population means defined above. Define $q(X)$ as the ratio of consistent decision-makers with $X_i = X$ relative to all consistent decision-makers, $q(X) = \frac{\overline{Y}(0,X)+1-\overline{Y}(1,X)}{E_{X_i}\left[\overline{Y}(0,X_i)+1-\overline{Y}(1,X_i)\right]}$, where $E_{X_i}[.]$ is the expectation over the random variable X_i.We define $s(X)$ as the corresponding ratio for the inconsistent decision-makers, $s(X) = \frac{\overline{Y}(1,X)-\overline{Y}(0,X)}{E_{X_i}[\overline{Y}(1,X_i)-\overline{Y}(0,X_i)]}$. Last, we assume frame exogeneity holds conditional on each value of X:

A2' (*Conditional Frame Exogeneity*). For all $X \in \mathbf{X}$, $(Y_i(1), Y_i(0)) \perp D_i \,|\, X_i = X$.

The identification strategy we propose in such settings is to: first, estimate the preferences of consistent decision-makers with a given set of observable characteristics; second, extrapolate preferences from consistent to inconsistent decision-makers with the same observable characteristics; and third, aggregate preferences across cells based on the estimated distribution of characteristics in the full population or sub-population of inconsistent decision-makers. A barrier to employing this familiar approach in our context is that we cannot directly observe consistency. The following lemma, analogous to Abadie (2003), provides conditions under which the aggregate distribution of characteristics among the consistent and inconsistent decision-makers can nonetheless be identified.

Lemma 1 *Under A1, A2', A3, and A4:*

[12] When $\overline{Y}(0) = 1-\overline{Y}(1)$, 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 $\overline{Y}(0) = 1 - \overline{Y}(1)$, any $E[Y_i^*] \in [0, 1]$ is feasible.

(L1.1) *For any* X, $p(X_i = X | C_i = 1) = q(X)\, p(X_i = X)$

(L1.2) *For any* X, $p(X_i = X | C_i = 0) = s(X)\, p(X_i = X)$.

The next step in the identification strategy proceeds using the following assumption:

A5 (*Conditional Consistency Independence*) For all individuals i and all observable characteristics $X \in \mathbf{X}$, $cov(Y_i^*, C_i \,|\, X_i = X) = 0$.

Conditional consistency independence requires that consistent and inconsistent decision-makers with the same observable characteristics have the same distribution of preferences. The assumption is analogous to one commonly employed in the program evaluation literature, i.e., that observationally equivalent individuals do not sort on the unobserved gain to treatment (e.g., Angrist and Fernandez-Val, 2013). It is also similar to the type of assumption that has been relied on in the line of papers described in the introduction (e.g., Bronnenberg et al., 2015), in which the preferences of a reference group of experts is extrapolated to the population.

Exploiting Lemma 1 along with conditional consistency independence, the following proposition formalizes the matching-on-observables identification strategy described above:

Proposition 3 *Let* $\overline{Y}_C(X) = \frac{\bar{Y}(0,X)}{\bar{Y}(0,X)+1-\bar{Y}(1,X)}$. *Under Assumptions A1, A2', A3, A4, and A5*:

(3.1) $E\left[Y_i^*\right] = E_{X_i}\left[\overline{Y}_C(X_i)\right]$

(3.2) $E\left[Y_i^* | C_i = 0\right] = E_{X_i}\left[s(X_i)\,\overline{Y}_C(X_i)\right]$[13]

As with any matching-on-observables approach, the plausibility of this approach will depend on the detail and nature of the observable characteristics as well as the underlying positive model of behavior, as demonstrated by the following two examples.

Decision-Making Types Model of Default Effects As before, suppose that preferences and consistency in the population are (respectively) characterized by latent index models $\widetilde{Y}_i = \beta^Y \theta_i^Y + \eta_i^Y$ and $\widetilde{C}_i = \beta^C \theta_i^C + \eta_i^C$, but unlike before, we now interpret θ_i^Y and θ_i^C to denote the vector of *observable* characteristics. Accordingly, η_i^Y and η_i^C denote the unobservable determinants of preferences and consistency. Conditional consistency independence requires $\text{cov}\left(\eta_i^Y,\, \eta_i^C \mid \theta_i^Y, \theta_i^C\right) = 0$. For example, it could be that highly educated customers are less likely to prefer that companies use their personal data and are more likely to choose consistently across default regimes;

[13] Replacing assumption A2 with A2' in (1.1) implies that $E[C_i] = E_{X_i}[\bar{Y}(0, X_i) + 1 - \bar{Y}(1, X_i)]$, and $E[Y^*|C_i = 1] = E_{X_i}[q(X_i)\overline{Y}_C(X_i)]$. Even when A2 is satisfied, this revised estimator for $E[C_i]$ is preferable for applying Proposition 3 in finite samples, due to possible spurious correlation between the observables and the frame.

but that conditional on education, preferences and consistency are uncorrelated.

Bounded Rationality Model of Default Effects Let $F_{\widetilde{C}|X}(\cdot)$ denote the cumulative density of $\widetilde{C}_i$ after conditioning on $X_i = X$. It is straightforward to show that conditional consistency independence requires $F_{\widetilde{C}|X}(0) = F_{\widetilde{C}|X,\Delta u>0}(0)$ for each X. This holds if the observed characteristics absorb enough variation in $|\Delta u_i|$ and γ_i such that the remaining, unobserved variation in consistency is uncorrelated with ordinal preferences. For example, suppose the set of observables is rich enough to absorb variation in consistency associated with the utility stakes of the decision, $X_i = X_j \implies |\Delta u_i| = |\Delta u_j|$ for any two individuals i and j. Conditional consistency independence would then hold if the remaining variation in opt-out costs is uncorrelated with the remaining variation in ordinal preferences, a sufficient condition for which is that the structural parameters of the model are (conditionally) independently distributed, $cov(\gamma_i, \Delta u_i|X) = 0$. In contrast, when the (conditional) variation in consistency is driven by selection on the gains to choosing actively, both consistency and ordinal preferences are driven by the same underlying structural parameter (Δu_i); hence one would not expect their distributions to be independent, except in special cases.[14].

3.3 Decision Quality Instruments

When conditioning on observables does not yield a credible identification strategy, researchers sometimes turn to instrumental variables designs. To develop the analog to that strategy here, we introduce the notion of a *decision quality instrument,* a component of the decision-making environment that affects decision-makers' consistency but that is unrelated to their preferences.[15]

Formally, the decision quality instrument is a a new random variable, $Z_i \in \{0, 1\}$. Choices can depend on both the frame and the instrument, $Y_i(D, Z)$, so there are now four potential outcomes. Each decision-maker chooses once under a single (D, Z) combination. For each i, we observe $(Y_i,\ D_i,\ Z_i)$, where $Y_i = Y_i(D_i, Z_i)$. Consistency is defined at each value of the instrument, $C_i(Z) = 1 - |Y_i(1, Z) - Y_i(0, Z)|$. We denote

[14] A similar challenge arises in the program evaluation context, when individuals select into treatment based on the gains from doing so, as in the Roy model (Heckman et al., 1996).

[15] A standard instrumental variable might be used to identify the effect of a treatment on choice, say $E\left[Y_i(1) - Y_i(0)\right]$, when assignment to frames is confounded with individuals' potential outcomes. In contrast, we propose to use exogenous variation in whether an individual is consistent. In other words, we are "instrumenting" for C, not for D (which would be standard). To map our prior results into this notation, one can interpret $Y_i(D)$ in previous sections as $Y_i(D, Z)$ for some fixed value of Z.

the fraction of decision-makers choosing $Y_i = 1$ under a given (D, Z) combination by $\overline{Y}(D, Z) \equiv E\left[Y_i(D_i, Z_i) \mid D_i = D,\ Z_i = Z\right]$.

The following assumptions establish the type of variation that constitutes a valid decision quality instrument:

A2" (*Exogeneity of D and Z*) $(Y_i(0,0),\ Y_i(1,0),\ Y_i(1,0),\ Y_i(1,1)) \perp (D_i, Z_i)$

A6 (*Decision Quality Monotonicity*) For all i, $C_i(1) \geq C_i(0)$ and $E[C_i(1) - C_i(0)] > 0$.

A7 (*Decision Quality Exclusion Restriction*) For all i, Y_i^* does not depend on Z.

Assumption A2" modifies frame exogeneity, which now requires both D_i and Z_i to be uncorrelated with confounding factors. A6 requires the effect of Z on consistency to be weakly monotonic for all decision-makers and strictly monotonic for some. A7 requires that variation in the decision-making environment induced by Z be 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] Like frame separability, A7 does not rule out variation in Z affecting welfare by altering the real costs of choosing against the frame. Although generally untestable with the type of data we assume, A7 may be tested if one observes variation in Z affecting choices in a setting without framing effects (in which case the variation in Z should not affect behavior).

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 (e.g., Pocheptsova et al., 2009) 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, the intensity of the frame (e.g., the degree to which one alternative is more salient than another), or the complexity of the choice presented (as in Brown et al., 2017).

[16] Like frame separability, A7 is an assumption about whether Y_i^* must be indexed by Z, rather than a distributional assumption.

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

$$E\left[Y_i^* \middle| C_i(1) > C_i(0)\right] = \frac{\overline{Y}(0,1) - \overline{Y}(0,0)}{\overline{Y}(1,0) - \overline{Y}(0,0) - \left(\overline{Y}(1,1) - \overline{Y}(0,1)\right)}.$$

Proposition 4 is best understood by analogy to the identification of a local average treatment effect (LATE, see Imbens and Angrist, 1994). The monotonicity assumption (A7) permits us to divide the population into three groups: the always-consistent ($C_i(1) = C_i(0) = 1$), the sometimes-consistent $C_i\left((1) = 1;\ C_i(0) = 0\right)$, and the never-consistent ($C_i(1) = C_i(0) = 0$). The denominator of the expression in Proposition 4 measures the reduction in the size of the inconsistent sub-group as we move from $Z = 0$ to $Z = 1$, which identifies the size of the sometimes-consistent group (the analog of the compliers in the standard 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 $Y_i^* = 1$. Dividing the latter by the former yields the fraction of the sometimes-consistent with $Y_i^* = 1$. Table 2 further illustrates the intuition behind the Proposition by applying it to hypothetical data from the online privacy example first described in the Introduction. The table highlights that Proposition 4 works by essentially applying Proposition 1.1 separately by values Z, and then identifying $E\left[Y_i^* \middle| C_i(1) > C_i(0)\right]$ by comparing $E\left[Y_i^* \middle| C_i(0) = 1\right]$ to $E\left[Y_i^* \middle| C_i(1) = 1\right]$. Several other comparisons to the IV literature are worth noting. First, unlike conventional instrumental variable analyses, the variable affected by a decision quality instrument (consistency) is not directly observable to the researcher. Consequently, Proposition 4 requires frame monotonicity in addition to the standard IV monotonicity assumption (A6) in (Imbens and Angrist, 1994). Second, Proposition 4 can motivate a Wu-Hausman over-identification test of consistency independence; however, such a test requires $E[Y^*|C_i(1) > C_i(0)] = E[Y_i^*]$, which may fail depending on the nature of selection into consistency. Finally, Proposition 4 may be extended beyond binary instruments by applying the result to each pair-wise combination of Z values. Such variation allows the researcher to non-parameterically trace out the relationship between consistency and preferences, similar to identification of the marginal treatment effect in Heckman and Vytlacil (2007).

As with a standard instrument, the quantity identified by Proposition 4 corre-

Table 2: Illustration of Proposition 4

When $Z = 0$

	$Y_i(0,0) = 0$	$Y_i(0,0) = 1$
$Y_i(1,0) = 0$	0.35	0.00
$Y_i(1,0) = 1$	0.25	0.40

Fraction consistent:	$E[C_i(0)] = 0.40 + 0.35 = 0.75$
Fraction consistent *and* prefer option 1:	$E\left[C_i(0)\, Y_i^*\right] = 0.40$
Consistent preferences:	$E[Y_i^* \vert C_i(0) = 1] = \frac{0.4}{0.4+0.35} = 0.53$

When $Z = 1$

	$Y_i(0,1) = 0$	$Y_i(0,1) = 1$
$Y_i(1,1) = 0$	0.50	0.00
$Y_i(1,1) = 1$	0.05	0.45

Fraction consistent:	$E[C_i(1)] = 0.50 + 0.45 = 0.95$
Fraction consistent *and* prefer option 1:	$E\left[C_i(1)\, Y_i^*\right] = 0.45$
Consistent preferences:	$E[Y_i^* \vert C_i(1) = 1] = \frac{0.45}{0.5+0.45} = 0.47$

Fraction sometimes consistent:	$E[C_i(1)] - E[C_i(0)] = 0.95 - 0.75 = 0.20$
Fraction sometimes-consistent *and* prefer option 1:	$E\left[(C_i(1) - C_i(0))\, Y_i^*\right] = 0.45 - 0.40 = 0.05$
Sometimes consistent preferences:	$E[Y_i^* \vert C_i(1) > C_i(0)] = \frac{0.05}{0.20} = 0.25$

sponds to a specific sub-group of the population – in our setting, it is those whose sensitivity to the frame varies by Z. This quantity can be of interest for several reasons. First, consider a government deciding which value of Z to implement, for example a regulator deciding how streamlined privacy controls should be. The solution to this problem trades off the cost of selecting a value of 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.

Second, Proposition 4 can shed light on the underlying behavioral model. For a decision-quality instrument that affects present bias, for example, the "long-run" view of welfare in the quasi-hyperbolic discounting model described Section 1 predicts all of the sometimes-consistent choosers will prefer to consume the larger amount at the later date. Estimating $E[Y_i^* | C_i(1) > C_i(0)]$ can test this hypothesis.

Finally, the relationship between the preferences identified by Proposition 4 and the preferences identified by Proposition 1 can be extrapolated to shed light on the relationship between preferences and consistency for the full population.[17] Because

[17] An interesting special case occurs when, under $Z = 1$, all decision-makers are consistent, i.e., $E[C_i(1)] = 1$. In this case, $E[Y_i^* | C_i(1) = 1] = E[Y^*]$, so choices under $Z = 1$ are a "refinement" in which the preferences of the full population are identified. Furthermore, when $C_i(1) = 1$ for all individuals, $E[Y_i^* | C_i(1) > C_i(0)] = E[Y_i^* | C_i(0) = 0]$. Consequently, the empirical quantity in Proposition 4 identifies the preferences of the inconsistent choosers at $Z = 0$. Consequently, one can

this extrapolation problem depends on the positive model generating the framing effects, we shall discuss it in the context of our running examples.

Decision-Making Types Model of Default Effects In this model, a decision quality instrument exploits variation that induces a passive type to become active, or vice-versa – e.g., being warned about the bias before making the decision. Incorporating the decision quality instrument into the statistical model yields $\widetilde{C}_i = \beta^C \theta_i^C + \delta_i^C Z_i + \eta_i^C$ and $\widetilde{Y}_i = \beta^Y \theta_i^Y + \delta_i^Y Z_i + \eta_i^Y$, where we now allow η_i^C and η_i^Y to have arbitrary correlation with one another to reflect unobserved determinants of consistency and preferences. Note that A6 corresponds to $\delta_i^Y = 0\ \forall i$ and A7 corresponds to $\delta_i^C \geq 0\ \forall i$ and $\delta_i^C > 0$ for some i. Proposition 4 sheds light on the relationship between η_i^C and η_i^Y, which, depending on the functional form of their joint distribution, can be used to recover the distribution of population preferences in the spirit of Heckman (1979) (see Online Appendix Section C.1). Again, richer variation in Z would allow one to identify the joint distribution of η_i^c and η_i^y more flexibly; we describe one such approach in Online Appendix Section C.2.

Bounded Rationality Model of Default Effects In the bounded rationality model, a natural source for decision quality instruments is variation in the distribution of costs associated with choosing actively, γ_i. For example, such variation might make it easier or more difficult to select the non-default option, perhaps by simplifying the opt-out process (expanding or reducing the number of forms to fill out or the amount of red tape). Suppose that $\gamma_i = \gamma_{0i} - \delta Z_i$, where Z_i is the binary decision quality instrument and $\delta > 0$. In this case, the sometimes-consistent choosers are those with $-\delta < |\Delta u_i| - \gamma_{0i} < 0$. Reducing γ induces this group to start choosing actively. Variation in Z thus provides information on the distribution of Δu_i, F_Δ. With a binary instrument, identifying $E[Y_i^*]$ requires imposing a functional form for F_Δ. With more variation in Z, F_Δ can be estimated more flexibly. We provide additional detail in Online Appendix Section C.3.

4 Application to Automatic 401(k) Enrollment

In this section we illustrate our framework with data on enrollment decisions into employer-provided 401(k) pension plans. Such plans can either be opt-in, so that new employees must actively enroll in the plan to participate, or opt-out, so that new employees are enrolled by default. A large body of research documents strik-

recover the preferences of the population *and of the inconsistent decision-makers* without further extrapolation.

ing differences in enrollment and savings behavior between opt-in and opt-out plan designs (Madrian and Shea, 2001; Choi et al., 2006; Chetty et al., 2014). Unless employee saving preferences depend on whether enrollment is the default, such findings undermine the use of traditional revealed preference analysis in this setting. In contrast, our approach accounts for the observed framing effect to shed light on employee preferences.

4.1 Data

Our data come from the large health care and insurance firm studied in Madrian and Shea (2001). The firm switched from an opt-in to an opt-out enrollment policy in April 1998. Under the opt-out policy, passive employees were automatically enrolled at a default contribution rate of 3 percent of salary. Under both designs, the employer provided a 50% match on employee contributions of up to 6 percent, and employee contributions into the plan were capped at 15 percent.

We observe whether employee i enrolls in the plan (indicated by Y_i) and whether the default is opt-in ($D_i = 0$) or opt-out ($D_i = 1$) at the date of hire. We also observe annual compensation, age, sex, and race for each employee.[18] The income, age, and racial composition of the firm's employees are typical of a large employer in the US, although the firm's workforce is disproportionately female. Employer contributions vest in the pension after two years of employment. We refer readers to Madrian and Shea (2001) for additional details regarding the data and the change in plan design. As expected, participation is greater under opt-out than under opt-in, $\overline{y}_0 = 0.491$ and $\overline{y}_1 = 0.859$ (we use lower-case letters to denote estimated sample analogs to the population moments described in earlier sections).

4.2 Recovery of Consistent Preferences

Under A1-A4, Proposition 1 allows us to identify the preferences of the consistent employees. Frame separability (A1) requires that preferences over plan participation do not depend on whether the design is opt-in or opt-out. This seems likely to hold, as it is difficult to imagine an employee's preferences over how much to save depend on how her employer chooses to structure enrollment into its sponsored retirement plan.[19] Absent A1, the behavior observed by Madrian and Shea would not constitute

[18] For confidentiality purposes, we received binned data on compensation and age.

[19] Enrollment preferences could depend on the default if employees are uncertain over whether they should enroll in the plan and interpret the default as advice from their employer. However, employees in the firm we study who were not automatically enrolled did not shift their 401(k) contributions to the default contribution rate that applied to automatically enrolled employees, as

a framing effect and standard revealed preference analysis would suffice to recover employee preferences.

Frame exogeneity (A2) requires that an employee's hire date (within the window studied) be uncorrelated with whether she chooses to participate under either plan design. This is the same assumption required to identify the causal effect of the change in plan design on participation; Madrian and Shea present evidence that it holds by showing the observable characteristics of employees hired before and after the change in plan design are similar. Appendix Table 1 replicates that analysis for the modified sample we study; our results are quite similar to theirs.

The consistency principle (A3) requires that employees who choose to participate in the plan under both the opt-out and opt-in design actually prefer participation, and similarly, that employees who choose non-participation under both designs prefer not to participate. In contrast, the standard revealed preference assumption requires that all employees – even those whose choices depend on the participation default – prefer the option that they choose. The consistency principle will be violated if employees' choices are characterized by biases that manifest themselves across both frames; for example, if employees who consistently choose not to enroll only make that decision because of present-bias.[20] If employee participation decisions are biased for reasons unrelated to automatic enrollment (the observed framing effect), further deviations from revealed preference analysis beyond the consistency principle would be needed to accurately recover preferences.

Finally, frame monotonicity (A4) requires that no employee chooses to enroll when enrollment is opt-in, but chooses not to enroll when enrollment is opt-out. Frame monotonicity is not directly testable without observing employees making repeated choices across multiple frames, but it can be falsified if we observe a reduction in participation rates under opt-out enrollment for any subgroup of employees. Appendix Figure 2 plots employee participation by frame, for each subgroup of employees we

would be expected if they acted on the default as advice (Carroll et al., 2009). Particularly in the 401(k) context, default effects being driven primarily by trust in one's employer would be surprising given the countervailing incentives that arise when pension plans have an employer match (Bubb and Warren, 2017).

[20]Not all forms of present-bias would cause the consistency principle to fail. In the models of default-sensitivity studied by Carroll et al. (2009) and Bernheim, Fradkin and Popov (2015), for example, present-bias causes individuals to procrastinate and stick with the default until they make an active choice, at which time the amount they choose to save will be optimal. Such behavior satisfies the consistency principle because those individuals who choose consistently have selected their most-preferred option.

observe. For each group, participation is greater under opt-out enrollment, consistent with frame monotonicity.

Under A1-A4, Proposition 1.1 point-identifies the preferences of the consistent decision-makers for enrollment. Substituting the estimated population moments into the definition of $\overline{Y}_C$ in Proposition 1 yields $\overline{y}_C = 0.777$, with a standard error of 0.006.[21] Thus, under frame monotonicity, of the 63.2 percent of employees whose enrollment decisions are insensitive to the enrollment default, we conclude that 77.7 percent prefer enrollment. Without frame monotonicity, Proposition 1.2 implies the fraction of consistent employees that prefer enrollment is *at least* 77.7 percent. Thus, we conclude that a large majority of the consistent employees prefer to participate in the plan.

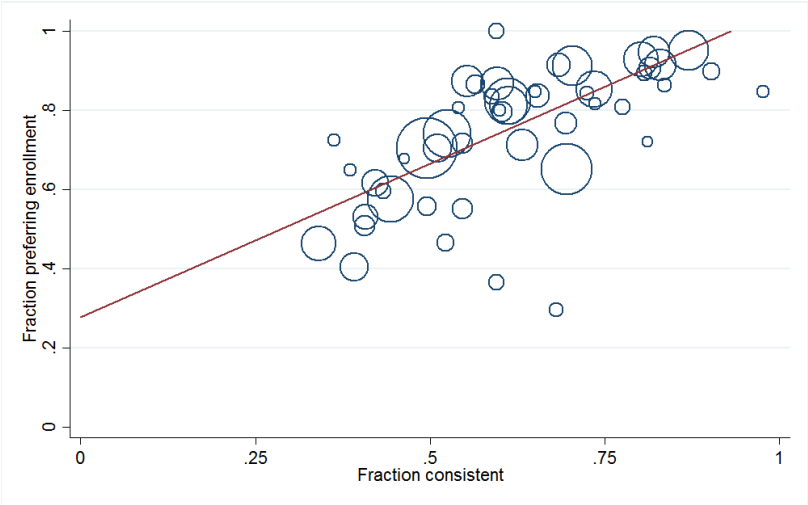
4.3 Recovery of Population Preferences

Table 3 presents the conclusions about population preferences that can be drawn from the data under assumptions of varying strength. With only A1-A3 (Column 1), the answer is not much: Proposition 2 implies one can only rule out values of $E[Y_i^*]$ below 0.350. The scope of this uncertainty is striking given that nearly 50 percent of employees choose to participate even when enrollment is opt-in. Adding frame monotonicity (Column 2) allows us to tighten these bounds significantly, yielding $0.491 \leq E[Y_i^*] \leq 0.859$. Thus, under A1-A4, an observer can conclude that at least a (near) majority of employees prefer enrollment.

Determining how large of a majority prefers enrollment requires understanding the relationship between employee preferences and consistency. As a benchmark, consistency independence (Column 3) implies $E[Y_i^*|C_i = 1] = E[Y_i^*|C_i = 0]$, so that $E[Y_i^*] = 0.777$. To shed light on the plausibility of this assumption in our data, Figure 1 plots estimates of $E[C_i]$ and $\overline{Y}_C$ for each demographic subgroup we observe. If consistency independence was satisfied, we would expect the slope of this relationship to be flat. Instead, the figure suggests a strongly positive relationship between preferences and consistency: groups with more consistent employees are also more likely to contain more employees who prefer participation. The slope of the estimated best-fit line is 0.78. Characterizing this result as a formal test of unconditional consistency independence requires assuming conditional consistency independence. Even so, for (unconditional) consistency independence to hold given this finding, it would have to

[21] Standard errors on $\overline{y}_C$ and other finite-sample statistics were obtained using the delta method. See Section B of the Online Appendix for details.

Figure 1: Consistency versus Preference for Enrollment in a 401(k) Plan



Notes: Estimates based on calculations on data from Madrian and Shea (2001) provided to the authors. Each point on the 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.

be the case that conditional consistency independence fails, and the within-subgroup correlation between preferences and consistency *exactly* offsets the observed between-subgroup correlation. Because we see little reason to expect this correlation to be positive between subgroups while negative within subgroups, we treat this possibility as remote. Hence, we interpret Figure 1 as suggestive evidence against consistency independence holding in our population. Motivated by Figure 1, an observer might feel comfortable imposing that the relationship between consistency and preferences is positive for the employees in our data, even without being confident about exactly what the relationship is. Column 4 shows the implication of this assumption for population preferences: $0.491 \leq E[Y_i^*] \leq 0.777$; roughly speaking, the assumption suggests somewhere between one-half and three-quarters of employees prefer enrollment. If the estimated slope in Figure 1 had instead been negative, the resulting bounds would be even narrower: $0.777 \leq E[Y_i^*] \leq 0.859$.

We next consider conditional consistency independence. In our setting, this assumption requires that among employees with the same income, age, gender, and race, preferences for enrollment are uncorrelated with consistency. As discussed in Section 3.2, this assumption is most likely to hold when the remaining variation in employees' sensitivity to the default is mostly driven by factors exogenous to the specific decision being considered. This can be satisfied in a types model of default

Table 3: Estimates of Population Preferences

Assumptions:	(1) A1-A3	(2) A1-A4	(3) A1-A4, Consistency Independence	(4) A1-A4, $\text{cov}(Y^*, C) \geq 0$	(5) A1-A4, Conditional Consistency Independence	(6) A1-A4, Bounded Rationality Model
Fraction of consistent	[77.7, 1]	77.7	77.7	77.7	77.7	77.7
preferring enrollment (%)	(0.6, ·)	(0.6)	(0.6)	(0.6)	(0.6)	(0.6)
Fraction of inconsistent	[0, 1]	[0, 1]	77.7	[0, 77.7]	70.9	56.1
preferring enrollment (%)	(·)	(·)	(0.6)	(·, 0.6)	(1.2)	(4.7)
Fraction of population	[35.0, 1]	[49.1, 85.9]	77.7	[49.1 , 77.7]	75.0	69.2
preferring enrollment (%)	(0.9, ·)	(0.8, 0.5)	(0.6)	(0.6, ·)	(0.8)	(3.5)

Note: This table illustrates the estimation of enrollment preferences among consistent employees, inconsistent employees, and the population, under various sets of assumptions. Column 1 applies Propositions 1.2 and 2.2 under our core assumptions excluding frame monotonicity. Column 2 adds the assumption of frame monotonicity and applies Propositions 1.1 and 2.1. Column 3 adds the assumption of consistency independence. Column 4 assumes, based on the group-level evidence in Figure 1, that the covariance between consistency and principles is positive. Column 5 assumes conditional consistency independence (Proposition 3) using the same set of employee characteristics as Figure 1. Column 6 reports results from the structural estimation of a bounded rationality model of default effects with all variation in consistency due to variation in utility stakes. All estimates are based on data from Madrian and Shea (2001) provided to the authors. Standard errors (reported in parentheses) are calculated using the delta method; see Section B of the Online Appendix for derivations.

effects, or a bounded rationality model in which the characteristics we observe capture most of the variation in preference intensity. In both cases, the remaining variation in preferences must not correlated with unobserved variation in the "stickiness" of the default.[22]

Column 5 of Table 3 presents 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 preferences for consistent employees. The difference in estimated preferences between the consistent and inconsistent employees is statistically significant ($p < 0.001$). For the full population of employees, the analysis implies that 74.9 percent prefer enrollment.

Finally, recall that under a bounded rationality model of defaults, conditional consistency independence tends to fail when most of the remaining variation in consistency (after conditioning on observables) is driven by unobserved variation in the

[22]For example, if cognitive ability was positively correlated with both consistency and preferences among employees of the same age, gender, race, and income, our results would yield an upwardly-biased estimate for the preferences of the inconsistent employees. The bias in the matching estimator is given by $E[Y_i^*] - E_X[\overline{Y}_C(X_i)] = E_X\left[\frac{\text{cov}(Y_i^*,\ C_i \mid X=X_i)}{E[C_i|X=X_i]}\right]$.

With the right data, one could attempt to recover population preferences by exploiting a decision quality instrument. For example, one might randomly assign certain employees to a streamlined process for actively choosing a plan or financial counseling services to help determine whether 401(k) participation is consistent with the employee's goals for saving and retirement.

utility "stakes" of the decision across decision-makers. Suppose that this is indeed the case, and that decision-makers behave as described by the bounded rationality model of defaults developed above. We allow the cost of active choice to vary by demographic group. Within each group, the distribution of utility gains from enrollment are normally distributed. Thus $\gamma_i = \gamma_X$ and $F_{\Delta|X}(\Delta u_i) \sim N(\mu_X, \sigma_X^2)$ for all i such that $X_i = X$, where X_i denotes i's group. Note that this setup represents the extreme case in which conditional consistency independence fails, since, by assumption, *all* of the remaining variation in consistency is driven by the stakes of the decision, $|\Delta u_i|$.[23] Within each group, observing the fraction of consistent decision-makers and the preferences of that group allow us to identify μ_X/σ_X and γ_X/σ_X, which in turn, allow us to recover the preferences of the inconsistent employees within the group. Finally, aggregating the group-specific preferences using the prevalence of each group in the population (or, using the weights in Lemma 1, the prevalence of each group among the inconsistent decision-makers) allows us to recover population preferences and preferences among the inconsistent employees.

The results of this analysis are reported in Column 6 of Table 3. Because assumptions A1-A4 are satisfied by this model of behavior, our estimate for the preferences of the consistent employees is the same as in the other columns. In contrast, this model suggests that just 56 percent of the inconsistent employees prefer enrollment. Combining the consistent and inconsistent employees, we estimate the fraction of the population preferring enrollment is approximately 69 percent. From this, we conclude that a literal application of conditional consistency independence is not necessary for obtaining the result that consistent employees prefer enrollment at a higher rate than do inconsistent employees.

To better understand why we estimate a positive relationship between enrollment preferences and consistency among the employees in our data, it is useful to investigate how differences in preferences and consistency relate to employee characteristics.

[23] In the Online Appendix, we consider the case in which γ follows a log-normal distribution. Appendix Figure 3 illustrates how different assumptions about the relative variance between γ and Δu imply different conclusions about population preferences. When the variance of γ is relatively small, estimated population preferences approaches the case in which γ is homogeneous, and all the variation in consistency is driven by Δu. In contrast, when the variance of γ is relatively large, variation in consistency is driven primarily by variation in that term. And because the model assumes that γ and Δu_i are independently distributed, estimated population preferences in this case approaches the estimates one obtains from assuming consistency independence. Consequently, one can interpret Columns 4 and 6 as two extreme forms of a bounded rationality model in which the distribution of structural parameters can vary by observable group membership.

Although we cannot directly observe either preferences or consistency for individual employees, the results in section 3.2 allow us to investigate differences based on employees' observable characteristics. We estimate a regression of the form

$$E[Y_i \,|\, D,\, X] = \alpha_0 + \alpha_1\, D + X'\,\beta_0 + X'\,\beta_1\, D \tag{3}$$

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

$$E[C_i \,|\, X_i = X] = 1 - \alpha_1 - X'\beta_1 \tag{4}$$

$$E[Y_i^* \,|\, C_i = 1,\, X_i = X] = \frac{\alpha_0 + X'\beta_0}{1 - \alpha_1 - X'\beta_1} \tag{5}$$

The results of the analysis are reported in Table 4. We find that both consistency and the preferences of consistent choosers vary systematically by employee characteristics. Variation in consistency is strongly related to variation in compensation, with those in the highest compensation bin (annual income over \$50K), 40 percent more likely to choose consistently than those in the lowest bin (annual income less than \$20K), and 41 percent more likely to prefer enrollment. After controlling for income, preferences for enrollment, but not consistency, also vary by age, race, and gender. These findings are consistent with the non-parametric evidence of a positive relationship between consistency and consistent preferences in Figure 1.

Table 4 shows that preferences for 401(k) participation are lowest among young and low-income employees (for example, in all four groups with employees below age 30 and salary below \$20,000, we estimate that a majority prefers non-participation). One explanation could be that younger and lower-income employees are more susceptible to present-bias, so that the consistency principle (A3) is more strongly violated for them than for other groups. The preferences revealed by consistent choices would then still contain some bias, as in Example 3 of Section 1. Alternatively, it could be that these employees expect their tenure at the firm to be too short for the employer matching contribution to vest, in which case even a modest preference for liquidity could lead to a preference for non-enrollment. Figure 2 investigates this hypothesis by plotting preferences for enrollment by demographic group against the fraction of the group remaining at the firm two years from the date of hire – the time at which the 50 percent employer matching contribution vests. The observed relationship is strong and positive. In addition, the best-fit line for the group-level regression has

Table 4: Consistency and Preference by Observable Characteristics

	(1) Consistency	(2) Consistent Preferences
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

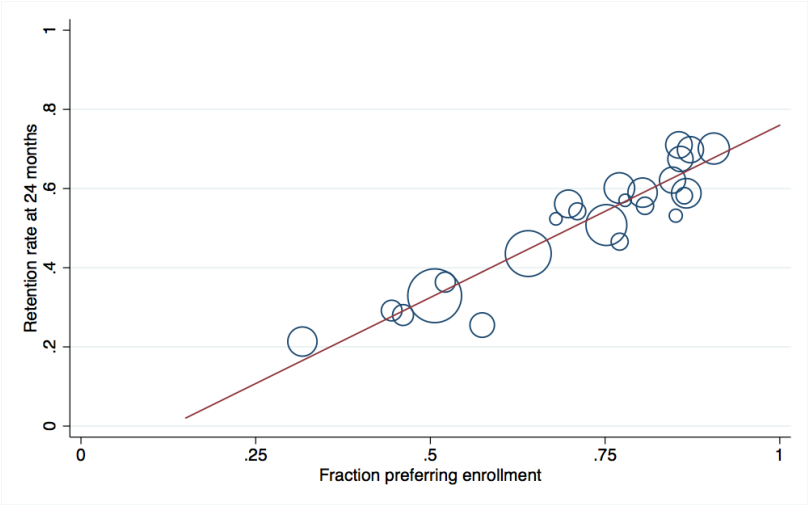
Notes: Estimates are based on equations (3)-(5), using disaggregated data from Madrian and Shea (2001) provided to the authors. The left-out groups for each demographic characteristic are 1) employees with compensation less than \$20K, 2) employees with age less than 30 years, 3) non-white employees, and 4) female employees. Column (1) examines the conditional probability that an employee is consistent . Column (2) examines the conditional probability that a consistent chooser prefers enrollment, holding other characteristics constant. Standard errors calculated using the delta method are reported in parentheses. *** indicates $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

an R^2 of over 90%, suggesting that our estimated preferences are a strong predictor of ultimate tenure at the firm for most groups. This evidence is reassuring, as it is consistent with the hypothesis that the preference information recovered by our analysis is driven by rational differences in employee expectations, rather than simply reflecting unobserved biases.

4.4 Discussion

The above results illustrate how researchers can use our proposed framework to recover more credible estimates of preference information than what would otherwise be available. Using the type of data that is already routinely collected, we show how the preference information that can be learned depends on the strength of the assumptions an observer is willing to impose. A central virtue of this approach is its transparency – researchers who rely from the start of their analysis on a specific behavioral model often implicitly impose A1-A4 in addition to some relationship between preferences and consistency. Thus, even if one ends up relying on a specific behavioral model, our approach highlights how the assumptions of the model contribute to identification – i.e., how they narrow the identification region relative to

Figure 2: Retention Rates and Preferences for Enrollment



Notes: Estimates are based on calculations on data from Madrian and Shea (2001) provided to the authors. We plot the retention rate at 24 months against the fraction preferring enrollment, for each observable group. Each point on the bubble scatter plot consists of all workers with given values of compensation, age and sex. The size of the cell is proportional to the area of the circle. Our interpretation of the figure assumes conditional consistency independence; without this assumption the x-axis can be interpreted as the fraction of consistent choosers preferring enrollment. Due to additional sample size restrictions necessary to study retention rates, we do not include race as a covariate. Earlier matching results are similar without using race as a covariate. Data on retention is calculated using captures of the relevant data at specific dates roughly 6 months apart. An employee is counted as retained in a month if he or she is still at the firm as of the most recent date of observation. Results are similar using retention at 12 or 18 months.

weaker assumptions, such as A1-A4 alone.

In the specific application we study, our results show that under relatively weak assumptions (A1-A4), the vast majority (78%) of consistent employees at the firm we study prefer 401(k) enrollment, as do between 49 and 86 percent of all employees. Adding an additional assumption about the direction of the relationship between preferences and consistency based on the non-parametric evidence in Figure 1 allows one to narrow these bounds to conclude that the fraction of all employees preferring enrollment is somewhere between 49% and 78%. Imposing a stronger, but still plausible, assumption in the form of conditional consistency independence allows one to point-estimate the fraction of employees preferring enrollment at 74%. Alternatively, one can pin down the relationship between preferences and consistency by imposing a specific behavioral model; the one we consider yields a lower estimate of population preferences (69%), but agrees with our matching estimator that the inconsistent employees tend to prefer enrollment at a lower rate than the consistent employees.

5 Cardinal Welfare Metrics and Price Variation

Thus far we have focused exclusively on the identification of ordinal preferences, the object directly identified by conventional revealed preference analysis. In this section,

we consider how the preference information recovered by our approach relates to standard money-metric welfare measures. As is the case without framing effects, constructing these welfare measures requires observing variation in the relative price of the available options.

Let $p \in \mathbb{R}$ denote the price of option 1 relative to option 0. We assume the fraction of individuals choosing option 1 is observed (or can be estimated) at any price p and under each frame $D \in \{0, 1\}$, and denote it $\overline{Y}(p, D) = E[Y_i(p, D)]$.

Holding price fixed, this model is isomorphic to the main model considered above. Our identifying assumptions have simple analogs here, supposing that they hold at any fixed price. In that case, the results we describe can be used to identify $E[Y_i^*(p)]$, for any fixed p.[24] One can then trace out this measure of *frame-free demand* for option 1 – that is, demand once framing effects have been eliminated and individuals choose according to $Y_i^*(p)$ – at each p to obtain the *frame-free demand curve*, $\overline{Y}^*(p)$. Denote the inverse of this demand curve by $p^*(\overline{Y})$.

For simplicity, assume that decision-makers have quasi-linear preferences, $U_i(Y_i,\, x_i) = u(Y_i) + x_i$ for some numeraire x_i, and budget constraint $pY_i + x_i = z_i$. Given quasi-linearity, we can express the money-metric difference in utility between the two options as $U_i(1) - U_i(0) = p_i^* - p$, where p_i^* denotes the reservation price for individual i, $p_i^* \equiv u_i(1) - u_i(0)$. Social welfare is $W(p,\, D) = \int_i U_i\left(Y_i(p,\, D),\, p\right)$.

Knowledge of $\overline{Y}^*(p)$ alone is sufficient to answer a number of questions of interest. Intuitively, the function substitutes for the standard demand curve used to estimate equivalent or compensating variation. For example, one could evaluate the welfare effect of assigning each individual with option 1 versus assigning each individual with option 0 by integrating $p^*(\overline{Y})$ over the population. As another example, one could use $\overline{Y}^*(p)$ to calculate the welfare effect on consumers of a price change that occurs under a "refinement" frame (D^*) in which all individuals choose according to $Y_i^*(p)$.

Conducting welfare comparisons when some choices are made under $D = 0$ or $D = 1$ requires additional structure because these comparisons depend on the *joint* distribution of $Y_i(0, p)$, $Y_i(1, p)$, and $Y_i^*(p)$. We sketch one strategy for dealing with this issue here:

(A8) For each individual i and frame D, there exists a unique reservation price, $p_i(D)$, such that $p < p_i(D) \iff Y_i(p,\, D) = 1$

(A9) There exists a common index i such that $Y_i(p, 0) > Y_{i'}(p, 0) \iff Y_i(p, 1) >$

[24] We set aside the assumption of exactly which approach is used to identify $\overline{Y}^*(p)$.

$Y_{i'}(p,1) \iff Y_i^*(p) > Y_{i'}^*(p)$ for any i, i'.

For example, these two assumptions would be satisfied under a model in which decision-makers are consistent whenever the intensity of their preference exceeds some threshold. Given this additional structure, we are guaranteed to have three well-defined inverse demand curves: $p(\overline{Y},1)$, $p(\overline{Y},0)$, and $p^*(\overline{Y})$, corresponding (respectively) to $\overline{Y}(p,1)$, $\overline{Y}(p,0)$, and $\overline{Y}^*(p)$. Note that frame monotonicity implies $p(\overline{Y},1) \leq p^*(\overline{Y}) \leq p(\overline{Y},0)$ for any $\overline{Y}$.

Assumptions (A8) and (A9) allow us to perform welfare comparisons for most policy changes one might be interested in. These comparisons are possible because, given the common index structure (A9), we know that at a given value of $\overline{Y}$, the values of the inverse demand curves $p(\overline{Y},1)$, $p(\overline{Y},0)$, and $p^*(\overline{Y})$ correspond to the three individual reservation prices $p_i(1)$, $p_i(0)$, and p_i^* *for the same individual.* Thus, recovering the three *aggregate* demand curves also allows us to recover the joint distribution of *individual* demand curves that is essential to the welfare calculation. The following proposition describes two such welfare comparisons.

Proposition 5 *Suppose that (A1)-(A4) hold at any price; (A8) and (A9) imply*

(5.1) *The welfare impact of an increase in price from p_0 to p_1 under the frame $D=0$ is $W(p_1,0) - W(p_0,0) = -(p_1 - p_0)\overline{Y}(p_1,0) - \int_{\overline{Y}(p_0,0)}^{\overline{Y}(p_1,0)} [p^*(\bar{Y}) - p_0] d\overline{Y}$*

(5.2) *The welfare impact of changing the frame from $D=0$ to $D=1$ is $W(p,1) - W(p,0) = \int_{\overline{Y}(p,0)}^{\overline{Y}(p,1)} [p^*(\overline{Y}) - p] d\overline{Y}$.*

Figure (3) illustrates these welfare calculations. Figure (3a) illustrates (5.1), the welfare effect of increasing the price under $D=0$. The price change reduces welfare by inducing some of those who were consuming option 1 to switch to option 0, and also by raising the price for those who continue to consume option 1. The welfare calculation for a price change under $D=1$ is analogous.

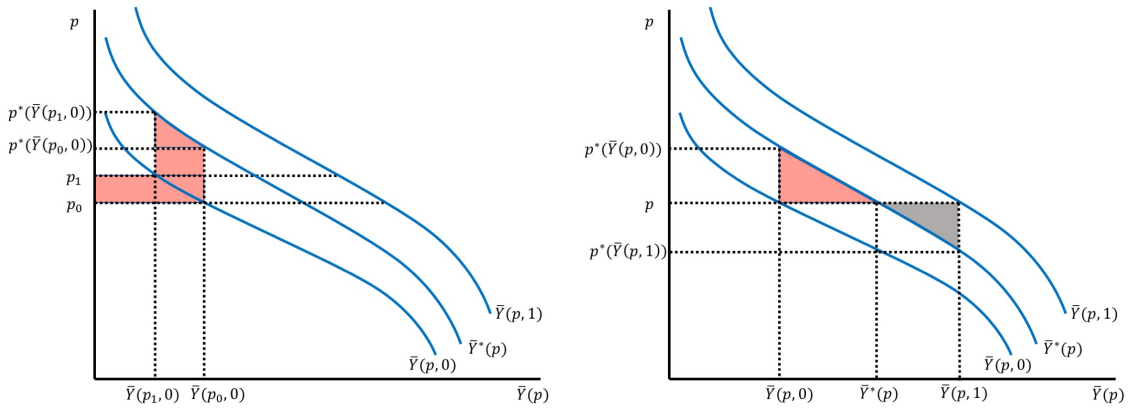
Figure 3b illustrates (5.2), the welfare effect of changing the frame from $D=0$ to $D=1$, at some fixed price p. The welfare effect of changing the frame falls entirely on inconsistent choosers in this model because these are the individuals who change their behavior when the frame changes.[25] The inconsistent choosers who prefer option 0 at price p are made worse off by a switch to $D=1$ (the red shaded region in the

[25] One could incorporate a cost of choosing against the frame into individual welfare, in which case the consistent individuals would also matter for evaluating the change in welfare (see Goldin and Reck, 2018).

Figure 3: Illustration of Welfare Calculations

(a) Change in Price in Frame $D = 0$ (b) Change frame from $D = 0$ to $D = 1$

Note:



This figure illustrates the welfare effects of a change in the price under a frame $D = 0$ (Sub-Figure a) and a change in frame from $D = 0$ to $D = 1$ at fixed price p (Sub-Figure b). Decreases in welfare are represented as red-shaded regions; increases are represented as grey-shaded regions.

graph) and the inconsistent choosers who prefer option 1 are made better off (the black shaded region). The amount by which a given individual is better or worse off is determined by $p_i^* - p$. By integrating the estimated demand functions, one can therefore determine whether welfare is higher under $D = 0$ or $D = 1$. The figure is drawn for the case in which $D = 0$ leads to higher welfare at the specified price.

6 Conclusion

Recovering preferences when framing effects are present is a fundamental challenge in behavioral economics. Our proposed approach is to maintain the revealed preference assumption unless an apparent framing effect is observed. In that case, we relax the revealed preference assumption as much as is required to accommodate the observed framing effect, but no further. We show that this transforms the original preference recovery problem into one of accounting for potentially endogenous selection into the sub-population of consistent decision-makers. Applying this approach can lead to novel insights even in well-studied settings like automatic enrollment into pension plans, as illustrated in the empirical application.

More generally, our approach offers two main advantages to empirical researchers when framing effects are present. First, it offers simple and practical tools to recover preference information without having to commit to a specific behavioral model or set of functional form assumptions. In the many settings in which our core assumptions

are plausible, Propositions 1 and 2 allow the researcher to estimate the preferences of consistent decision-makers and to bound the preferences of the population. Point-identifying population preferences raises additional challenges, and, depending on the application, the empirical tools we developed might offer a path forward. But even when our tools are unlikely to apply, understanding the preference recovery challenge as a selection problem may still provide insight. For example, a researcher's knowledge of an application may suggest that preferences and consistency are positively correlated, which, in conjunction with Equation (2), would narrow the range of values in which population preferences might fall.

The second benefit to our framework is that it makes model-based approaches to preference identification more transparent. As long as the model satisfies our core assumptions, our results shed light on which features of the model are driving the identification, namely the behavioral and distributional assumptions that pin down the relationship between preferences and consistency. Highlighting these features can help researchers choose between models and assess the credibility of their results.

At the same time, our approach is subject to at least two important limitations. First, the reduced form nature of our proposed tools might lure researchers into using them in settings in which their identifying assumptions are not met. We have tried to alleviate this concern by highlighting the types of conditions in which each tool is valid in the context of specific behavioral models.

Second, we have assumed throughout that the presence of the framing effect is the only reason that decision-makers' choices fail to reflect their preferences. When other biases cause choices to diverge from preferences, choices will not reveal preferences even once the framing effect has been removed. In such cases, applying our approach still yields the (counterfactual) choices decision-makers would make if the framing effect had not been present, but the preferences inferred from those (counter-factual) choices must be further adjusted before preferences can be recovered.

References

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

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally-Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review*, 105(8): 2501–2538.

Angrist, Joshua D., and Ivan Fernandez-Val. 2013. "ExtrapoLATE-ing: External

Validity and Overidentification in the LATE Framework." In *Advances in Economics and Econometrics: Tenth World Congress*. Vol. 3, , ed. Daron Acemoglu, Manuel Arellano and Eddie Dekel, 401–434. 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.

Benkert, Jean-Michel, and Nick Netzer. Forthcoming. "Informational Requirements of Nudging." *Journal of Political Economy*.

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, and Dmitry Taubinsky. Forthcoming. "Behavioral Public Economics." In *Handbook of Behavioral Economics*. Vol. 1, , ed. B. Douglas Bernheim, Stefano DellaVigna and David Laibson, 401–434. Elsevier.

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.

Bronnenberg, Bart, Jean-Pierre Dube, Matthew Gentzkow, and Jesse Shapiro. 2015. "Do Pharmacists Buy Bayer? Sophisticated Shoppers and the Brand Premium." *The Quarterly Journal of Economics*, 130(4): 1669–1726.

Brown, Jeffrey, Arie Kapteyn, Erzo Luttmer, and Olivia Mitchell. 2017. "Cognitive Constraints on Valuing Annuities." *Journal of the European Economic Association*, 15(2): 429–462.

Bubb, Ryan, and Patrick L. Warren. 2017. "An Equilibrium Theory of Retirement Plan Design." NYU Law and Economics Research Paper No. 16-28.

Carroll, Gabriel, James Choi, David Laibson, Brigitte 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 H. 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 McCaffery and Joel Slemrod, 304–353. Russell Sage Foundation.

Conlisk, John. 1996. "Why Bounded Rationality?" *Journal of Economic Literature*, 34(2): 669–700.

Fischhoff, Baruch. 1991. "Value Elicitation: Is There Anything in There?" *American Psychologist*, 46(8): 835–847.

Goldin, Jacob, and Daniel Reck. 2018. "Optimal Defaults with Normative Ambiguity."

Working Paper.

Handel, Benjamin R., and Jonathan Kolstad. 2015. "Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare." *American Economic Review*, 105(8): 2249–2500.

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

Heckman, James J., and Edward 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., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1996. "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method." *Proceedings of the National Academy of Sciences*, 93(23): 13416–13420.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

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, and M. Marit Rehavi. 2016. "Physicians Treating Physicians: Information and Incentives in Childbirth." *American Economic Journal: Economic Policy*, 8(1).

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *The Quarterly Journal of Economics*, 112(2): 443–477.

LeBoeuf, Robyn A., and Eldar Shafir. 2003. "Deep Thoughts and Shallow Frames: On the Susceptibility to Framing Effects." *Journal of Behavioral Decision Making*, 16(2): 77–92.

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

Masatlioglu, Yusufcan, Daisuke Nakajima, and Erkut Y. Ozbay. 2012. "Revealed Attention." *The American Economic Review*, 102(5): 2183–2205.

Pocheptsova, Anastasiya, On Amir, Ravi Dhar, and Roy F. Baumeister. 2009. "Deciding Without Resources: Resource Depletion and Choice in Context." *Journal of Marketing Research*, 46(3): 344–355.

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.