Online Appendix: Revealed Preference Analysis with Framing Effects

Jacob Goldin Daniel Reck

October 2, 2018

A Subjective Uncertainty in the Bounded Rationality Model of Default Effects

Here we relax the assumption in the bounded rationality model of default effects developed in the main text (Section 3) that the decision-maker knows $u_i(1)$ and $u_i(0)$ with certainty. We characterize behavior in this model and briefly discuss identification issues that arise.

Consider a model of framing effects in which the individual chooses either actively or passively, depending on whether the *expected* gain from choosing actively exceeds some threshold γ_i. If the individual chooses actively, the individual learns $u_i(1)$ and $u_i(0)$ with certainty and chooses the option that provides the highest utility, Y_i^*. Individuals who choose passively select whichever option is the default. The individual chooses actively when

$$E_i[u_i(Y_i^*) - u_i(D)] > \gamma_i.$$

Here, we define expectations and probabilities over each individual i's subjective beliefs about $u_i(0)$ and $u_i(1)$. Let $\Delta u_i = u_i(1) - u_i(0)$. When $D = 1$, the individual chooses actively if:

$$-p_i(Y_i^* = 0)E_i[\Delta u_i | Y_i^* = 0] > \gamma_i \tag{1}$$

When $D = 0$, the individual chooses actively if:

$$p_i(Y_i^* = 1)E_i[\Delta u_i | Y_i^* = 1] > \gamma_i \tag{2}$$

Because this model is nested by the generic model of active or passive decision-making discussed in Example 1 of Section 1, frame separability, the consistency principle, and frame monotonicity are satisfied here. This behavioral model therefore fits into the framework we develop in the main text.

How does this model differ from the model with certainty about the gains to choosing actively? Conceptually, the key is in the difference between choosing actively – i.e. paying

the cost to consider both options and learn which is most-preferred choice – and choosing consistently. In the model with certainty, decision-makers never choose actively when their most-preferred option is the default. In that model, inconsistent decision-makers choose passively in both frames, and consistent decision-makers choose actively in one frame only (when $D_i \neq Y_i^*$). The set of consistent decision-makers overlaps exactly with the set of active decision-makers.

In contrast, in the model with uncertainty, consistent decision-makers choose actively in either one or both frames, and inconsistent decision-makers choose actively in either zero or one frame. For example, a consistent decision-maker with $Y_i^* = 1$ may choose actively under $D_i = 1$ if she believes there is a sufficiently high probability that $Y_i^* = 0$. Similarly, an inconsistent decision-maker with $Y_i^* = 1$ but who believes that $Y_i^* = 0$ with high probability would choose to be active under $D_i = 1$ but not under $D_i = 0$. Consequently, unlike the model with certainty, the set of active decision-makers diverges from the set of decision-makers who are consistent. Any extrapolation from consistent choosers to the full population must therefore account not only for variation in Δu_i and γ_i across individuals, but also for variation in subjective beliefs about $u_i(1)$ and $u_i(0)$.

Finally, note that focusing solely on consistency as our measure of decision quality potentially discards some of the revealed preference information in this model: there are some inconsistent decision-makers whose active choices in one of the frames reveals their preferences. Without being able to directly observe whether a given decision-maker chooses actively, however, there is no way to identify this information separately from cases in which the decision-maker chooses according to the frame under both frames. A similar phenomenon arises elsewhere in the defaults literature, where there may be people who decide to choose actively under some default, but nevertheless end up choosing the default, so that it is impossible for an observer to distinguish them from passive decision-makers.

B Standard Errors of Proposed Estimators

The body of the paper ignores finite sample concerns in order to focus on identification. Naturally, any empirical application, including the ones we undertake in Section 4, should account for sampling variation and report standard errors for the finite-sample analogs of the population moments that identify interesting parameters. This section derives the estimates of the asymptotic variance of the finite-sample moments one would use in any empirical application to obtain standard errors. We assume the sample of decision-makers we observe is drawn randomly and independently from some underlying population. The standard errors we derive are implemented in Stata and Matlab in the empirical application in the paper. Code is available from the authors.

B.1 Standard Errors for Proposition 1

In addition to the closed-form variance estimates derived here, applied researchers can obtain the correct standard errors simply by regressing Y_i on D_i and using a pre-packaged command like Stata's "nlcom" for the proper combination of parameter estimates (see equations 3-5 in the text).

Let $\bar{y}(D)$ denote the sample averages for $D = 0$ and $D = 1$, i.e. the finite sample analogs of $\overline{Y}(D)$. We let n denote the number of individuals in the population, n_D denote the number of individuals observed in frame $D \in \{0, 1\}$, and α_D denote the fraction of individuals observed in frame D, $\alpha_D = \frac{n_D}{n}$.

Because $Y_i(0)$ and $Y_i(1)$ follow a binomial distribution, we have

$$\sqrt{n}\begin{pmatrix} \bar{y}(0) - \overline{Y}(0) \\ \overline{y}(1) - \overline{Y}(1) \end{pmatrix} \overset{a}{\sim} N\left(0, \Sigma\right),$$

where

$$\Sigma = \begin{pmatrix} \frac{1}{\alpha_0}\overline{Y}(0)(1-\overline{Y}(0)) & 0 \\ 0 & \frac{1}{\alpha_1}\overline{Y}(1)(1-\overline{Y}(1)) \end{pmatrix}.$$

The first statistic from Proposition 1 is $\overline{Y}_C$, which is defined as:

$$\overline{Y}_C = \frac{\overline{Y}(0)}{\overline{Y}(0) + 1 - \overline{Y}(1)}. \tag{3}$$

and can be estimated consistently via

$$\overline{y}_C = \frac{\bar{y}(0)}{\bar{y}(0) + 1 - \bar{y}(1)}.$$

Using the delta method, we have that $V(\overline{y}_C) \simeq \frac{1}{n}\nabla\overline{Y}'_C\,\Sigma\,\nabla\overline{Y}_C$, where $\nabla\overline{Y}_C$ is evaluated at $(\overline{Y}(0), \overline{Y}(1))$. Taking derivatives of (3) yields

$$\nabla\overline{Y}_C = \begin{pmatrix} \frac{1-\overline{Y}(1)}{\left(\overline{Y}(0)+1-\overline{Y}(1)\right)^2} \\ \frac{\overline{Y}(0)}{\left(\overline{Y}(0)+1-\overline{Y}(1)\right)^2} \end{pmatrix}.$$

Simplifying the expression for $V(\overline{y}_C)$ yields

$$V(\overline{y}_C) \simeq \frac{(1-\overline{Y}(1))^2\overline{Y}(0)(1-\overline{Y}(0))}{\alpha_0(\overline{Y}(0)+1-\overline{Y}(1))^4 n} + \frac{\overline{Y}(0)\overline{Y}(1)(1-\overline{Y}(1))}{\alpha_1(\overline{Y}(0)+1-\overline{Y}(1))^4 n},$$

which, letting $\overline{C} = \overline{Y}(0) + 1 - \overline{Y}(1)$ and using that $V(\bar{y}_D) = \frac{\overline{Y}(D)(1-\overline{Y}(D))}{n_D}$ for $D = 0, 1$, simplifies to

$$V(\overline{y}_C) = \left(\frac{1-\overline{Y}_C}{\overline{C}}\right)^2 V(\bar{y}(0)) + \left(\frac{\overline{Y}_C}{\overline{C}}\right)^2 V(\bar{y}(1)). \tag{4}$$

We can consistently estimate the variance of the asymptotic distribution of our estimator $\overline{y}_C$ by replacing all the terms in (4) with the corresponding sample means. Equation 4 shows, for example, that when the consistent subgroup is a relatively small proportion of the population, or when there is noise in the estimates of $\overline{y}(D)$ for some D due to a small sample size observed in a particular frame, the standard errors on $\overline{y}_C$ will tend to be large.

Note also that we can estimate the variance of the asymptotic distribution of our estimator for $\overline{C} = E[C_i]$. Write the estimator itself as:

$$\overline{c} = \bar{y}(0) + 1 - \bar{y}(1).$$

As this is a linear combination of other parameters, the variance of this estimator is simply:

$$V(\overline{c}) = V(\bar{y}(0)) + V(\bar{y}(1)). \tag{5}$$

B.2 Standard Errors for Proposition 3

The standard errors for the matching-on-observables estimator are somewhat more complicated, due to the presence of multiple demographic groups and the use of weights that must themselves be estimated from data. We discuss two solutions here, one based on post-estimation commands in a statistical package like Stata, and another with which one could directly calculate standard errors from primitives.

First, one can simply estimate our model via

$$E[Y_i(D,X)|D_i = D, X_i = X] = f(X,\theta) + D * g(X,\theta')$$

where $f()$ and $g()$ are specified up to vectors of parameters θ and θ', which are to be estimated from data. For instance, we could implement a linear model:

$$E[Y_i(D,X)|D_i = D, X_i = X] = \alpha + X'\beta + D(\gamma + X'\delta)$$

This equation can be estimated by a least squares linear probability model, and then the ingredients of the matching estimator are given by:

$$\overline{Y}_C(X) = \frac{\alpha + \beta}{1 - \gamma - X'\delta}$$

$$\begin{aligned}
\overline{C}(X) &\equiv E[C_i|X_i = X] = 1 - \gamma - X'\delta \\
\overline{C} &\equiv E[C_i] = E_i[1 - \gamma - X_i'\delta] \\
E[Y_i^*] &= E_i[\overline{Y}_C(X_i)] \\
E[Y_i^*|C_i = 0] &= E_i\left[\frac{1 - E[C_i|X = X_i]}{1 - E[C_i]}\,\overline{Y}_C(X_i)\right]
\end{aligned}$$

The corresponding finite sample analogs of each of these are:

$$\overline{y}_C(X) = \frac{\hat{\alpha} + \hat{\beta}}{1 - \gamma - X'\delta}$$

$$\overline{c}(X) = 1 - \hat{\gamma} - X'\hat{\delta}$$
$$\overline{c} = \frac{1}{n}\Sigma_i[\overline{c}(X_i)]$$
$$\hat{E}[Y_i^*] = \frac{1}{n}\Sigma_i[\overline{y}_C(X_i)]$$
$$\hat{E}[Y_i^*|C_i = 0] = \frac{1}{n}\Sigma_i\left[\frac{1 - \overline{c}(X_i)}{1 - \overline{c}}\,\overline{y}_C(X_i)\right]$$

This estimation strategy can be implemented via straightforward post-regression estimation in Stata, and standard errors may be estimated using the numeric delta method or a standard non-parametric bootstrap.

Alternatively, one may derive analytic standard errors via the delta method. We provide formulae for the variance and the gradients of parameters of interest, which may be straightforwardly incorporated into a matrix-based programming language, such as MATLAB, to calculate the variance of the estimators with discrete demographic groups. MATLAB code illustrating this procedure is also available upon request from the authors.

We focus on the case where the observables are discrete, $X \in \{x_1, ..., x_J\}$, as in our application. The primitive parameters of the discrete-characteristics model are, for each X,$Y(0,X) = E[Y_i(0,X)|D_i = 0, X_i = X]$, $Y(1,X) = E[Y_i(1,X)|D_i = 1, X_i = X]$, and $p(X_i = X)$.[1] For notational convenience, we will write these three primitives as $\overline{Y}_{0X}$, $\overline{Y}_{1X}$, and p_X respectively. We denote the estimators of these quantities by $\bar{y}_{0X}$, $\bar{y}_{1X}$, and $\hat{p}_X$. Now we construct the variance covariance matrix of the vector primitive parameters. Letting $\theta = (\overline{Y}_{0X_1}, \overline{Y}_{0X_2}, ..., \overline{Y}_{0X_J}, \overline{Y}_{1X_1}, ..., \overline{Y}_{1X_J}, p_{X_1}, ..., p_{X_J})'$, and $\hat{\theta} = (\bar{y}_{0X_1}, \bar{y}_{0X_2}, ..., \bar{y}_{0X_J}, \bar{y}_{0X_1}, ..., \bar{y}_{0X_J}, \hat{p}_{X_1}, ..., \hat{p}_{X_J})$, we know that

$$\sqrt{n}(\hat{\theta} - \theta) \overset{a}{\sim} N(\overrightarrow{0}, \Sigma)$$

Denoting the fraction of individuals with observable characteristic X observed in frame D by $\alpha_{DX} \equiv \frac{n_{DX}}{n}$, we can write the variance matrix as:

$$\Sigma(\theta) = \begin{pmatrix} \Sigma_0(\theta) & & \\ & \Sigma_1(\theta) & \\ & & \Sigma_p(\theta) \end{pmatrix},$$

where $\Sigma_0(\theta)$ and $\Sigma_1(\theta)$ are diagonal matrices with entries of the form $\frac{\overline{Y}_{Dx_j}(1-\overline{Y}_{Dx_j})}{\alpha_{Dx_j}}$ for

[1] When the X's are non-stochastic, such as when the researcher wishes to estimate preferences for a population with a known distribution of observable characteristics, the last of these may be excluded; the resulting modification of the variance estimation procedure below is straightforward.

$j = 1, ..., n$ and $D = 0, 1$, and

$$\Sigma_p(\theta) = \begin{pmatrix} p_{x_1}(1-p_{x_1}) & -p_{x_1}p_{x_2} & ... & -p_{x_1}p_{x_J} \\ -p_{x_1}p_{x_2} & p_{x_2}(1-p_{x_2}) & ... & -p_{x_2}p_{x_J} \\ ... & ... & ... & ... \\ -p_{x_1}p_{x_J} & -p_{x_2}p_{x_J} & ... & p_{x_J}(1-p_{x_J}) \end{pmatrix}.$$

All blank entries of the Σ matrix are zeroes.[2]

The matching approach employs many different combinations of these primitive parameters. We begin with the weights from Lemma 1 for the subset of inconsistent choosers,[3]

$$S(X) \equiv p(X_i = X | C_i = 0) = \frac{\overline{Y}_{1X} - \overline{Y}_{0X}}{\sum_{X=w}(\overline{Y}_{1w} - \overline{Y}_{0w})p_w} p_X.$$

Taking derivatives of this function and simplifying using the definition of $S(X)$, we obtain the following for any X and X':

$$\begin{aligned} \frac{\partial S(X)}{\partial \overline{Y}_{0X'}} &= -\frac{1\{X = X'\} - S(X)}{\sum_{X=w}(\overline{Y}_{1w} - \overline{Y}_{0w})p_w} p_{X'}, \\ \frac{\partial S(X)}{\partial \overline{Y}_{1X'}} &= \frac{1\{X = X'\} - S(X)}{\sum_{X=w}(\overline{Y}_{1w} - \overline{Y}_{0w})p_w} p_{X'}, \\ \frac{\partial S(X)}{\partial p_{X'}} &= \frac{1\{X = X'\}S(X) - S(X')S(X)}{p_{X'}}, \end{aligned}$$

where $1\{\}$ is an indicator function equal to 1 when the expression inside the square brackets is true and zero otherwise. These three expressions can be used to generate the entire gradient of $S(X)$.

Next we consider the weights for the subset of consistent choosers,

$$Q(X) = \frac{\overline{Y}_{0X} + 1 - \overline{Y}_{1X}}{\Sigma_{X=w}(\overline{Y}_{0w} + 1 - \overline{Y}_{1w})p_w} p_X.$$

Taking derivatives and simplifying, we obtain the following for any X and X':

$$\begin{aligned} \frac{\partial Q(X)}{\partial \overline{Y}_{0X'}} &= \frac{1\{X = X'\} - Q(X)}{\Sigma_{X=w}(\overline{Y}_{0w} + 1 - \overline{Y}_{1w})p_w} p_{X'}, \\ \frac{\partial Q(X)}{\partial \overline{Y}_{1X'}} &= -\frac{1\{X = X'\} - Q(X)}{\Sigma_{X=w}(\overline{Y}_{0w} + 1 - \overline{Y}_{1w})p_w} p_{X'}, \\ \frac{\partial Q(X)}{\partial p_{X'}} &= \frac{1\{X = X'\}Q(X) - Q(X')Q(X)}{p_{X'}}. \end{aligned}$$

[2]We know that the off-diagonal elements in $\Sigma_1(\theta)$ and $\Sigma_2(\theta)$, which govern the covariance of the various $\bar{y}_{DX}$ estimates, are zero because the estimation sample for every $\bar{y}_{DX}$ is distinct and by assumption drawn from the same underlying population, by assumption A2'. We also know that the entries of Σ governing the covariance of $\bar{y}_{DX}$and p_X are zero because of assumption A2'.

[3]Note that these are the weights implied by L1.1, rather than $s(X)$ as defined in the statement of the Lemma, i.e. $S(X) = s(X)p_X$. We do the same for inconsistent choosers below.

Next we consider the estimators for the preferences of various subgroups. First, define the preferences of the inconsistent subgroup $\overline{Y}_N \equiv E[Y_i^*|C_i = 0]$. When X is discrete, we have

$$\overline{Y}_N = \Sigma_X S(X)\overline{Y}_C(X).$$

Taking derivatives of this expression – which may be done more easily using several expressions derived above – we obtain the following for any X:

$$\frac{\partial \overline{Y}_N}{\partial \overline{Y}_{0X}} = S(X)\frac{1-\overline{Y}_C(X)}{\overline{C}(X)} + \frac{p_X}{1-\overline{C}}(\overline{Y}_N - \overline{Y}_C(X)),$$
$$\frac{\partial \overline{Y}_N}{\partial \overline{Y}_{1X}} = S(X)\frac{\overline{Y}_C(X)}{\overline{C}(X)} - \frac{p_X}{1-\overline{C}}(\overline{Y}_N - \overline{Y}_C(X)),$$
$$\frac{\partial \overline{Y}_N}{\partial p_X} = \frac{S(X)}{p(X_i = X)(1-\overline{C})}[\overline{Y}_C(X) - (1-\overline{C})\overline{Y}_N].$$

where $\overline{C}(X) \equiv \overline{Y}_{0X} + 1 - \overline{Y}_{1X}$ and $\overline{C} = \Sigma_{X=w}(\overline{Y}_{0w} + 1 - \overline{Y}_{1w})p_w$. From these three expressions we construct the gradient of Y_N.

Proceeding similarly for the full population, $\overline{Y}_{FP} = \sum_X p_X \overline{Y}_C(X)$, we obtain, for any X

$$\frac{\partial \overline{Y}_{FP}}{\partial Y_{0w}} = p_X \frac{1-\overline{Y}_C(X)}{\overline{C}(X)},$$
$$\frac{\partial \overline{Y}_{FP}}{\partial Y_{1w}} = p_X \frac{\overline{Y}_C(X)}{\overline{C}(X)},$$
$$\frac{\partial \overline{Y}_{FP}}{\partial p_X} = \overline{Y}_C(X),$$

which allows us to construct the gradient of $\overline{Y}_{FP}$.

We can also obtain the gradient of $\overline{Y}_C = \sum_X Q(X)\overline{Y}_X(X)$ in terms of the primitive parameters of this model.[4] Taking derivatives of the expression for $\overline{Y}_C$ yields

$$\frac{\partial \overline{Y}_C}{\partial \overline{Y}_{0X}} = Q(X)\frac{1-\overline{Y}_C(X)}{\overline{C}_X} + \frac{p_X}{\overline{C}}(\overline{Y}_C(X) - \overline{Y}_C),$$
$$\frac{\partial \overline{Y}_C}{\partial \overline{Y}_{1X}} = Q(X)\frac{\overline{Y}_C(X)}{\overline{C}_X} - \frac{p_X}{\overline{C}}(\overline{Y}_C(X) - \overline{Y}_C),$$
$$\frac{\partial \overline{Y}_C}{\partial p_X} = \frac{Q(X)}{p_X \overline{C}}(\overline{Y}_C(X) - \overline{CY}_C).$$

Using all of the above expressions, we can generate a gradient of each parameter of the matching-on-observables models. Putting all these expressions together, we construct a

[4]This part is not necessary to obtain a standard error on $\overline{Y}_C$, because we know how to obtain a simpler formula for the asymptotic variance of our estimator of $\overline{Y}_C$ using the result in the previous section of this Appendix. This derivation yields a numerically identical standard error estimate. The usefulness of the expressions derived here is that these expressions may be used to estimate the (asymptotic) covariance of, say, the estimators for $\overline{Y}_C$ and $\overline{Y}_N$, which is necessary for the statistical test of the null hypothesis of consistency independence against the alternative hypothesis of conditional consistency independence.

gradient matrix of the form:

$$G(\theta) = (\nabla S_{x_1}, ..., \nabla S_{x_J}, \nabla Q_{x_1}, ..., \nabla Q_{x_J}, \nabla \overline{Y}_N, \nabla \overline{Y}_{FP}, \nabla \overline{Y}_C).$$

To be clear, each of the columns of $G(\theta)$ is the gradient of a particular (nonlinear) function of the primitive parameters of the model. We can then estimate the full variance-covariance matrix of all the non-primitive parameters $(S_{X_1}, ..., S_{X_J}, Q_{X_1}, ..., Q_{X_J}, \overline{Y}_N, \overline{Y}_{FP}, \overline{Y}_C)$ by

$$\hat{V}(\hat{\theta}) = \frac{1}{n} G' \Sigma G$$

where G and Σ are evaluated at $\hat{\theta}$. The square root of the diagonals of the matrix $\hat{V}(\hat{\theta})$ will be asymptotically consistent standard errors for the parameter estimates themselves. The off-diagonal elements are the estimated covariance of different estimates, which are useful for tests of hypotheses involving more than one parameter of the model, such as tests of decision-quality independence ($H_0 : \overline{Y}_N = \overline{Y}_C$) in this setting.

B.3 Standard Errors for Proposition 4

Next we derive standard errors for the final identification result derived in the paper, in which we identify the preferences of sometimes-consistent decision-makers using decision-quality instruments. Using similar notation to before, let $\overline{Y}(D, Z) \equiv E[Y_i(D, Z)]$ for $D = 0, 1$ and $Z = 0, 1$, and denote the estimator for each population moment by $\bar{y}(D, Z)$. Similarly to before, we begin by noting that

$$\sqrt{n} \begin{pmatrix} \overline{y}(0,0) - \overline{Y}(0,0) \\ \overline{y}(0,1) - \overline{Y}(0,1) \\ \overline{y}(1,0) - \overline{Y}(1,0) \\ \overline{y}(1,1) - \overline{Y}(1,1) \end{pmatrix} \overset{a}{\sim} N(\overrightarrow{0}, \Sigma)$$

where Σ is a diagonal matrix with entries of the form $\frac{1}{\alpha_{DZ}} \overline{Y}(D, Z)(1 - \overline{Y}(D, Z))$. The new statistic in Proposition 4 is

$$\overline{Y}_S \equiv \frac{\overline{Y}(0,1) - \overline{Y}(0,0)}{\overline{Y}(1,0) - \overline{Y}(0,0) - \big(\overline{Y}(1,1) - \overline{Y}(0,1)\big)}, \tag{6}$$

which we can estimate consistently with

$$\overline{y}_S \equiv \frac{\overline{y}(0,1) - \overline{y}(0,0)}{\overline{y}(1,0) - \overline{y}(0,0) - (\overline{y}(1,1) - \overline{y}(0,1))}.$$

Using the delta method, we obtain $V(\overline{y}_S) \simeq \frac{1}{n} \nabla \overline{Y}_S' \, \Sigma \, \nabla \overline{Y}_S$, where $\nabla \overline{Y}_S$ is evaluated at $(\overline{Y}(0,0), \overline{Y}(0,1), \overline{Y}(1,0), \overline{Y}(1,1))$. Taking the gradient of (6) gives

$$\nabla \overline{Y}_S = \left(\frac{\overline{Y}(1,1) - \overline{Y}(1,0)}{\left(\Delta \overline{C}\right)^2}, \frac{\overline{Y}(1,0) - \overline{Y}(1,1)}{\left(\Delta \overline{C}\right)^2}, \frac{\overline{Y}(0,0) - \overline{Y}(0,1)}{\left(\Delta \overline{C}\right)^2}, \frac{\overline{Y}(0,1) - \overline{Y}(0,0)}{\left(\Delta \overline{C}\right)^2} \right)'$$

where $\Delta \overline{C} = \overline{Y}(1,0) - \overline{Y}(0,0) - (\overline{Y}(1,1) - \overline{Y}(0,1))$. Plugging this into the formula for $V(\overline{y}_S)$ and simplifying yields

$$V(\overline{y}_S) = \left(\frac{1 - \overline{Y}_S}{\Delta \overline{C}} \right)^2 \left[V(\overline{y}(0,1)) + V(\overline{y}(0,0)) \right] + \left(\frac{\overline{Y}_S}{\Delta \overline{C}} \right)^2 \left[V(\overline{y}(1,1)) + V(\overline{y}(1,0)) \right], \quad (7)$$

where $V(\bar{y}(D,Z)) = \frac{1}{n_{DZ}} \overline{Y}(D,Z)(1 - \overline{Y}(D,Z))$. Replacing each $\overline{Y}(D,Z)$ with the estimator $\overline{y}(D,Z)$, we obtain a consistent estimate of the asymptotic variance of $\overline{Y}_S$.

Note that when the fraction of the population who are sometimes-consistent choosers, $\Delta \overline{C}$, is relatively small, the variance of the estimator $\overline{y}_S$ can be quite large. This is analogous to a familiar practical issue with the use of conventional instrumental variables, wherein an instrument that has a small first-stage effect might yield very imprecise IV estimates. The analog of the weak instruments problem can also be shown to be an important finite-sample concern in this setting. Specifically, when $\Delta \overline{C}$ is small and there is idiosyncratic correlation between Y_i^* and Z_i (of the form that is ruled out in population moments by assumption A7), $\overline{y}_S$ may be quite biased in finite samples.

C Extrapolation from Decision Quality Instruments to Population Preferences

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

C.1 Parametric Extrapolation with Decision Quality Instruments

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

Suppose that consistency for individual i is determined by

$$P_i = \overline{P} + \theta Z_i + \varepsilon_i \tag{8}$$

$$C_i = 1 \iff P_i > 0, \tag{9}$$

where P_i is a latent variable reflecting idiosyncratic variation ε_i and the effect of a decision quality instrument $Z_i \in \{0, 1\}$. Note that consistency depends on i's choice under both frames, so P_i does not depend on the frame to which i is assigned. Note also that decision quality monotonicity (A6) is satisfied provided $\theta > 0$.

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

$$M_i = \overline{M} + \nu_i \tag{10}$$

$$Y_i^* = 1 \iff M_i > 0, \tag{11}$$

where the latent variable M_i simply reflects idiosyncratic variation in preferences, ν_i. Frame separability (A1) is satisfied because M_i does not depend on d, and the decision quality exclusion restriction (A7) is satisfied because M_i does not depend on Z_i. Exogeneity of the decision quality instrument and frame (A2") is satisfied provided that ε_i and ν_i are independent of Z_i and D_i.

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

$$\begin{pmatrix} \varepsilon_i \\ \nu_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{12}$$

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

We close the model with the consistency principle (A3) and frame monotonicity (A4). Together, these assumptions allow us to evaluate the probability of observing a given choice for a given (D, Z) combination:

$$Y_i(0,\, Z) = 1 \iff \varepsilon_i > -\overline{P} - \theta Z;\ \nu_i > -\overline{M} \tag{13}$$

$$Y_i(1, Z) = 0 \iff \varepsilon_i > -\overline{P} - \theta Z;\ \nu_i < -\overline{M}. \tag{14}$$

Equations (13) and (14) can be combined with (12) to identify the parameters of the model: $\overline{P}$, $\overline{M}$, θ, and ρ. One can then recover ordinal preferences by integrating the underlying distribution: $E[Y_i^*] = \Phi(\bar{M})$, where $\Phi(\cdot)$ is the standard normal cumulative density function, and $E[Y_i^* | \, C_i(Z) = 0] = \frac{1}{1 - E[C_i(Z)]} \int_{-\infty}^{\overline{P} - \theta Z} \int_{-\overline{M}}^{\infty} \phi^{BVSN}(\varepsilon, \nu; \rho) \partial\nu \partial\varepsilon$, where $\phi^{BVSN}(a, b; \rho)$ is the bivariate standard normal density with correlation coefficient ρ evaluated at (a, b).

The statistical model described above resembles the classic bivariate normal model of selection. Variation in the decision quality instrument induces variation in consistency

without affecting preferences; this guarantees that the relationship between consistency and preferences is identified without relying solely on functional form.

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

C.2 Semi-Parametric Extrapolation with Decision Quality Instruments

This section develops an extrapolation approach for recovering population preferences without relying on parametric distributional assumptions. In particular, we model the preferences of the consistent decision-makers at a given value of the decision quality instrument as a flexible polynomial in the fraction of decision-makers who are consistent at that value of the decision quality instrument.[5]

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

A6' For all individuals and all z, $z' \in [\underline{z}, \overline{z}]$ such that $z > z'$, $C_i(z) \geq C_i(z')$ and $E[C_i(z) - C_i(z')] > 0$.

For each individual, let $z_i^* < \overline{z}$ denote the value of z at which she begins to choose consistently, i.e., $z \geq z_i^* \implies C_i(z) = 1$. Assumption A6' implies that z_i^* is unique for each individual. Denoting the CDF of z_i^* by $F(.)$ and the PDF by $f(.)$, we have $E[C_i(z)] = F(z)$. In addition, note that the second part of A6' guarantees $f(z) > 0$ for all $z \in [\underline{z}, \overline{z}]$, so that $F(.)$ is strictly increasing with a well-defined inverse function over $E[C_i(z)] \in [0, 1]$, which we denote $F^{-1}(E[C_i(z)])$.

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

Proposition A.1 *Assume that A1, A3, and A4 hold at each fixed value of Z, and assume A2", A6', and A7.* For any degree $M \in \mathbb{N}$, there exist constants $a_0 ... a_M$ such that

(A.1.1) *For any z,* $E[Y_i^* | z_i^* = z] \approx a_0 + a_1 E[C_i(z)] + a_2 E[C_i(z)]^2 ... + a_M E[C_i(z)]^M$

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

(A.1.2) *For any z,* $E[Y_i^*|C_i(z)=1] \approx a_0 + \frac{a_1}{2}E[C_i(z)] + \frac{a_2}{3}E[C_i(z)]^2 + ... \frac{a_M}{M+1}E[C_i(z)]^M$

(A.1.3) $E[Y_i^*] \approx a_0 + a_1 + ... + a_M$

Proof Throughout the proof it will be convenient to denote $E[C_i(z)]$ by C_z. Fix any $M \in \mathbb{Z}^+$. Our technical assumptions – requiring $F(\underline{z}) = 0$, $F(\overline{z}) = 1$, and F strictly increasing – imply that F has a well-defined inverse function over the unit interval $[0,1]$. Because we have assumed $F(z)$ and $g(z)$ are continuous and infinitely differentiable, the function $h = g \circ F^{-1}$ will be continuous and infinitely differentiable as well. As a result, h has a well-defined Taylor series approximation of degree M about any point in $(0,1)$. Noting that $h(C_z) = E[Y_i^*|z_i^* = F^{-1}(C_z)] = E[Y_i^*|z_i^* = z]$ proves (A.1.1).

To prove (A.1.2), note that the preferences of the consistent choosers at some z', $E[Y_i^*|C_i(z') = 1] = E[Y_i^*|z_i^* \leq z']$, can be expressed using the definition of conditional probability as

$$E[Y_i^*|C_i(z')=1] = \frac{\int_{z=\underline{z}}^{z=z'} g(z)\, f(z)\, dz}{F(z')}$$

We employ a change of variables, letting $\overline{c} = F(z)$, $d\overline{c} = f(z)dz$. From above, $g(z) = h(C_z)$, so we obtain

$$E[Y_i^*|C_i(z')=1] = \frac{\int_{\overline{c}=0}^{\overline{c}=F(z')} h(\overline{c})d\overline{c}}{F(z')}$$

Given that the expression in (6.1) is a Taylor Series approximation of $h(\overline{c})$ by construction, we can substitute the expression in (6.1) into the numerator of the above expression, evaluate the integral in the numerator, and divide by $F(z') = C_{z'}$ to obtain (6.2).

The result in (A.1.3) follows from evaluating the expression in (A.1.2) at $\overline{c} = 1$. ■

Proposition A.1 implies that the preferences of the consistent decision-makers at a particular value of the decision quality instrument can be approximated by a polynomial function in the fraction of decision-makers who choose consistently at that value of the instrument. Because A6' guarantees a one-to-one mapping between z and $E[C_i(z)]$, we can write the preferences of the marginally consistent decision-makers as a function of the fraction of decision-makers choosing consistently, i.e. $E[Y_i^*|z_i^* = z] = g(F^{-1}(E[C_i(z)]))$. In addition, infinite differentiability of g and F ensure the composite function $h \equiv g \circ F^{-1}$ will have a well-defined Taylor series approximation of degree N. We then obtain (A.1.2) by integrating the marginal preference function $h(.)$ from $E[C_i(z)] = 0$ to $E[C_i(z')]$ and scaling by $E[C_i(z')]$ for any arbitrary z'. Finally, (A.1.3) follows from setting $E[C_i(z)] = 1$ in (A.1.2). Note that when $N = M$, we will have $N+1$ equations in $M+1$ unknowns, so that $a_0, ..., a_M$ are just-identified. When $N > M$, we will have more equations than unknowns, and a best-fit technique such as least-squares can be used to estimate $a_0, ..., a_M$.

C.3 Extrapolation in a Bounded Rationality Model

Here we describe extrapolation using a decision-quality instrument in a bounded rationality model, as alluded to in Section 3.3. This subsection also serves to flesh out some details of the costly opt-out model used in Column 6 of Table 3.

In the bounded rationality model discussed throughout Section 3, we assume a decision-maker makes an active choice when the benefits of doings so exceed some cost $\gamma_i \geq 0$. Recall that preferences and consistency in this model are given by

$$Y_i^* = 1 \iff \Delta u_i > 0$$

$$C_i = 1 \iff |\Delta u_i| > \gamma_i$$

Let $F_\Delta(.)$ denote the CDF of Δu_i, as in the main text. When the costs are homogeneous, $\gamma_i = \gamma \ \forall i$, the average (ordinal) preferences of the consistent decision-makers is given by:

$$E[Y_i^* | C_i = 1] = \frac{1 - F_\Delta(\gamma)}{1 - F_\Delta(\gamma) + F_\Delta(-\gamma)}.$$

Similarly, for the inconsistent decision-makers we have $E[Y_i^* | C_i = 0] = P\left(\Delta u_i > 0 \,|\, \Delta u_i \in (-\gamma, \gamma)\right)$, or

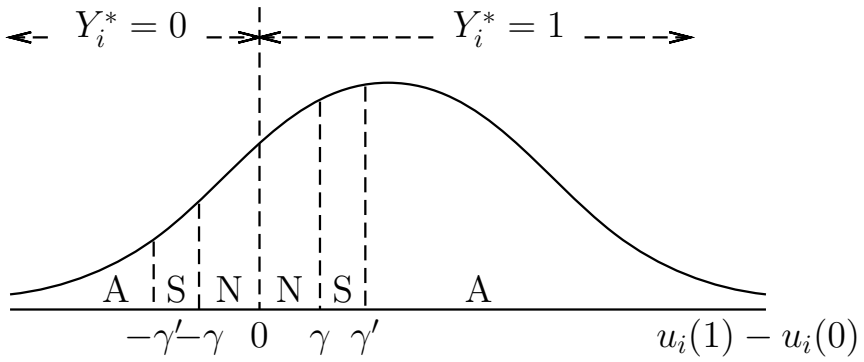
$$E[Y_i^* | C_i = 0] = \frac{F_\Delta(\gamma) - F(0)}{F(\gamma) - F(-\gamma)}.$$

In this model changes in the cost of opting out constitute valid decision-quality instruments. Reductions in these costs could be obtained, for example, by easing the administrative requirements (such as paperwork) for choosing the non-default option. Suppose that transaction costs change from γ_i to $\gamma_i' \leq \gamma_i$, with $\gamma_i' < \gamma_i$ for some i. Then variation in transactions costs will meet the criteria for being a decision-quality instrument. Letting $Z = 0$ represent choices under γ and $Z = 1$ choices under γ', we will have the following potential outcomes for $Y(D, Z)$.

$$(Y_i(0,0), Y_i(0,1), Y_i(1,1), Y_i(1,0)) = \begin{cases} (0,0,0,0) & \text{if } \Delta u_i < -\gamma_i \\ (0,0,0,1) & \text{if } \Delta u_i \in [-\gamma_i, -\gamma_i'] \\ (0,0,1,1) & \text{if } \Delta u_i \in [-\gamma_i', \gamma_i'] \\ (0,1,1,1) & \text{if } \Delta u_i \in [\gamma_i', \gamma_i] \\ (1,1,1,1) & \text{if } \Delta u_i > \gamma_i \end{cases} \tag{15}$$

The second and fourth cases correspond to the sometimes-consistent decision-makers whose ordinal preferences are captured by the statistic Y_S in Section 3.3. Figure 1 depicts the different cases in Equation (15), given two values of a decision-quality instrument. Variation in a decision-quality instrument in this model thereby allows the researcher to trace out the distribution of Δu_i, and to do so more and more flexibly as the number of values of the

Figure 1: Sometimes Consistent Choosers in the Costly Opt-out Model



instrument grows. Once the distribution of Δu_i is known, it is obviously straightforward to estimate the distribution of ordinal preferences in the population, or even the distribution of cardinal preferences, which could be useful for welfare calculations.

D Supplementary Results for Empirical Application

This section contains supplementary material from the empirical application.

E Generalizations

E.1 Non-Binary Menus

We now assume that the menu consists of a fixed set of K ordered options, $Y_i(D) \in S \equiv \{1, 2, ..., K\}$. As above, there are two frames, $D \in \{0, 1\}$. This setup might reflect an individual choosing from a menu of insurance plans, ordered from low-cost, low-benefit plans to high-cost, high-benefit plans (ignoring dominated options), where the frame makes salient either the benefits or the cost of the plan. Another example are survey questions in which the answer choices range from 1 (strongly agree) to 7 (strongly disagree), and the frame reflects variation in whether the question is worded positively or negatively.

The most-preferred option of individual i is represented by $Y_i^* \in \{1, 2, ...K\}$.[6] As before, frame separability requires that Y_i^* does not depend on D. Frame monotonicity is the same as well: $Y_i(1) \geq Y_i(0)$ for all i.

We next define *partition consistency at* $k \in S$, $C_i(k)$, as follows: $C_i(k) = 1$ if $\{Y_i(1) \leq k$ and $Y_i(0) \leq k\}$ or $\{Y_i(1) > k$ and $Y_i(0) > k\}$, and $C_i(k) = 0$ otherwise. Intuitively, partition consistency

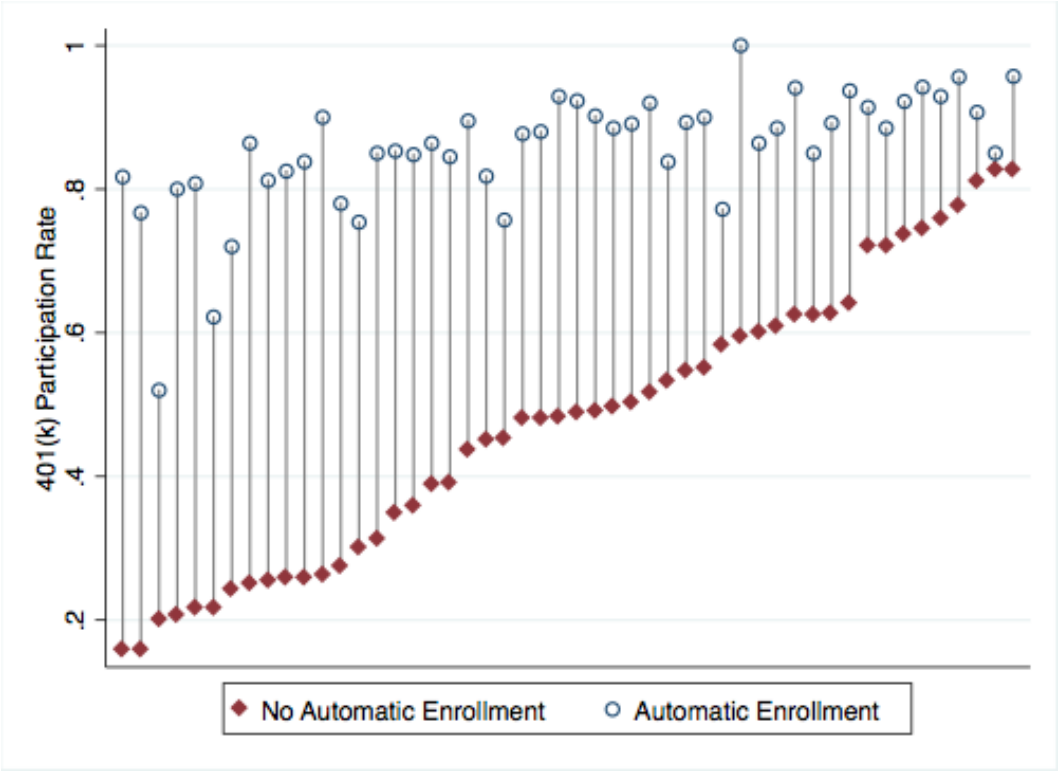
[6] We focus on identifying the distribution of Y_i^* holding the menu fixed; with variations in the menu one might naturally seek to recover richer information about individuals' preference relations.

Table 1: Worker Characteristics by Automatic Enrollment Cohort

	Not Automatically Enrolled (WINDOW Cohort)	Automatically Enrolled (NEW Cohort)
Income		
<$20k (%)	10.5	12.7
$20k-$29k (%)	37.7	45.6
$30k-$39k (%)	18.7	16.5
$40k-$49k (%)	15.2	11.2
>$50k (%)	18.0	14.1
Age		
<30 (%)	30.9	37.4
30-39 (%)	36.0	33.3
40-64 (%)	33.1	29.3
Gender		
Female (%)	76.9	79.0
Male (%)	23.1	21.0
Ethnicity		
White (%)	72.4	69.5
Non-White (%)	27.5	30.5
Number of Observations	4,185	5,702

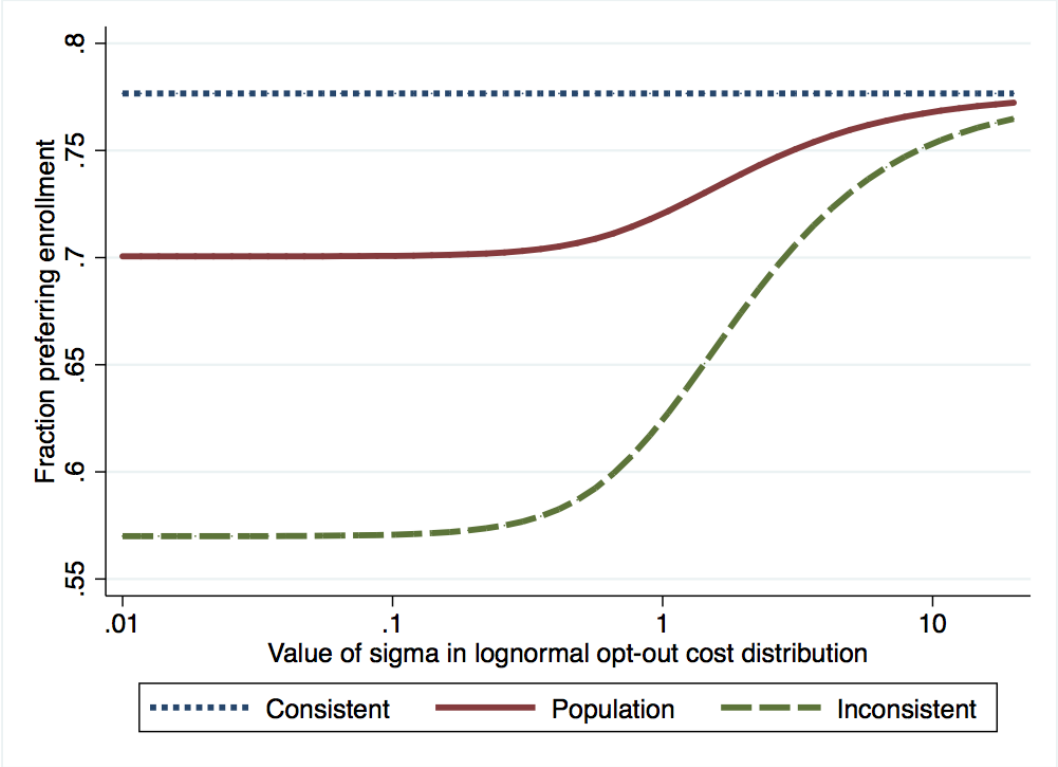
Note: This table shows that there is a similar distribution of observable characteristics in the automatically enrolled cohort (those hired within one year of the switch to automatic enrollment) and in the cohort who were not automatically enrolled (those hired within one year before the switch). These are called the NEW and WINDOW cohorts, respectively, in Madrian and Shea (2001). A similar set of summary statistics appears in Table II of Madrian and Shea (2001). The numbers here are very similar to the numbers in Madrian and Shea. The slight differences between the two sets of tabulations are driven by the fact that a small number of observations (110 in the NEW cohort and 72 in the OLD cohort) are dropped due to cell size restrictions in the aggregated data provided to the authors from Madrian and Shea.

Figure 2: Test of Monotonicity Assumption Using Subgroups



Note: This figure shows that participation rates are higher under automatic enrollment than without automatic enrollment for every demographic group (aligned on the x axis) in the cell-level data provided to us from Madrian and Shea (2001). Demographic cells are constructed on the basis of income group, age group, gender, and ethnicity. Table 1 lists the specific divisions of these characteristics used to construct the cells. The fact that participation rates are higher under automatic enrollment in every group, for finely defined groups, is consistent with the assumption of frame monotonicity (A4).

Figure 3: Variance of γ_i and Implied Population Moments in the Bounded Rationality Model



Note: This figure plots consistent preferences ($E[Y_i^*|C_i = 1]$), inconsistent preferences ($E[Y_i^*|C_i = 0]$), and population preferences ($E[Y_i^*]$) in a bounded rationality model of default effects in which Δu_i follows a normal distribution with mean $\mu_{\Delta u}$ and variance normalized to 1 and γ_i follows a log-normal distribution with mean μ_γ and variance σ_γ, and in which γ_i is assumed to be independent of Δu_i. As it contains three unknown parameters, this model is not identified with the two population moments, $\overline{Y}_1$ and $\overline{Y}_0$. The illustration here: (1) varies the assumed value of σ_γ on the x-axis; (2) estimates $\mu_{\Delta u}$ and μ_γ from the two observed population moments; and then (3) calculates the fraction of consistent, inconsistent, and all individuals preferring enrollment. We observe that at low values of σ_γ, the model is well-approximated by the assumption of a homogeneous γ. At high values of σ_γ, however, inconsistent and population preferences are well-approximated by assuming consistency independence ($cov(Y_i^*, C_i) \approx 0$). In a sense, therefore, preference estimates based on a model with homogeneous γ and estimates based on consistency independence represent two extreme implementations of the same behavioral model of default effects.

at k captures whether the decision-maker chooses consistently with respect to the binary choice between the two partitions of S, $\{1, 2, ...k-1\}$ and $\{k, k+1, ...K\}$.

Finally, in place of the consistency principle, we assume the following:

A10 (Partition Consistency Principle) For all individuals i and options $k \in \{1, 2, ...K\}$, $C_i(k) = 1$ implies

$$\begin{aligned} Y_i \geq k &\implies Y_i^* \geq k \\ Y_i \leq k &\implies Y_i^* \leq k. \end{aligned} \tag{16}$$

The partition consistency principle states that if a decision-maker selects an option greater than k under each frame, his or her most-preferred option is also greater than k (and similarly if the decision-maker chooses less than k under each frame). Note that the partition consistency principle implies the consistency principle used in previous sections: if $Y_i(0) = Y_i(1)$, then (16) implies $Y_i = Y_i^*$. Combining frame monotonicity and the partition consistency principle, we have $Y_i(1) \geq Y_i^* \geq Y_i(0)$.

With this structure, a number of our earlier results extend naturally to this new setting:

Proposition A.2 *Let $G_D(k) \equiv P(Y_i(D) \leq k | D_i = D)$ for $k = 1, ..., N$, $D = 0, 1$ and let $G_D(0) \equiv 0$. Let $Y_C(k) \equiv \frac{G_0(k)}{G_0(k)+1-G_1(k)}$ for $k = 0, ..., K$. Frame separability (A1), frame exogeneity (A2), frame monotonicity (A4), and the partition consistency principle (A8), imply that for $k = 1, ..., K$,*

(A.2.1) *The fraction of partition-consistent individuals at k with $Y_i^* \leq k$ is given by $P(Y_i^* \leq k | C_i(k) = 1) = Y_C(k)$.*

(A.2.2) *The fraction of partition-consistent individuals at k is given by $E[C_i(k)] = G_0(k) + 1 - G_1(k)$.*

(A.2.3) *The fraction of the population who prefer option k is bounded as follows: $p(Y_i^* = k) \in [G_0(k) - G_1(k-1),\ G_1(k) - G_0(k-1)]$.*

Proof Throughout the proof, we denote the fraction of individuals preferring some option k by $\bar{\phi}_k \equiv p(Y_i^* = k)$.

Proof of (A.2.1) and (A.2.2): Fix some $k \in \{1, ..., K-1\}$. Let $X' = \{x_1, ...x_{k-1}\}$ and $X'' = \{x_k, .., x_K\}$ Note that we can write the many-choices problem into a binary menu choice problem between X' and X''. Similarly, note that (A1), (A2), (A4), and (A10) imply the binary analogs to these assumptions (A1)-(A4). As such, these results follow directly from the application of Proposition 1 to this problem.

Proof of (A.2.3): First suppose that $k = 1$. Applying Proposition 2 to the binary menu choice problem with $X' = \{1\}$ and $X'' = \{2, ..., K\}$ implies that

$$\bar{\phi}_1 \in [G_0(1), G_1(1)] \tag{17}$$

Note that this confirms the desired result for $k = 1$ since $G_1(0) = G_l(0) = 0$ by definition. Next, applying the same proposition for $k = 2$, we have $\overline{\phi}_1 + \overline{\phi}_2 \in [G_l(2), G_h(2)]$. Combined with (17), this implies

$$\overline{\phi}_2 \in [G_0(2) - G_1(1),\, G_1(2) - G_0(1)]. \tag{18}$$

Similarly with $k = 3$, we have that $\overline{\phi}_1 + \overline{\phi}_2 + \overline{\phi}_3 \in [G_0(3),\, G_1(3)]$, and applying (17) and (18) implies that $\overline{\phi}_3 \in [G_0(3) - G_1(2),\, G_1(3) - G_0(2)]$. Proceeding recursively, suppose that for some k, we know that for any $k' < k$,

$$\overline{\phi}_{k'} \in [G_0(k') - G_1(k'-1),\, G_1(k') - G_0(k'-1)] \tag{19}$$

Then application of Proposition 2 to the binary menu choice problem with $X' = \{x_1, ..., x_k\}$ yields $\overline{\phi}_1 + \overline{\phi}_2 + ... + \overline{\phi}_k \in [G_0(k),\, G_1(k)]$, so $\overline{\phi}_k \in [G_0(k) - (\overline{\phi}_1 + \overline{\phi}_2 + ... + \bar{\phi}_{k-1}),\, G_1(k) - (\overline{\phi}_1 + \overline{\phi}_2 + ... + \bar{\phi}_{k-1})]$. Applying the lower and upper bounds from (19) and simplifying yields the desired result. ■

Discussion of Proposition A.2. The results in (A.2.1) and (A.2.2) provide information about consistent decision-makers in non-binary settings, using the notion of partition consistency. With this information, the techniques developed in Section 3 can be employed to extrapolate from consistent choosers for any partition to the full population. By identifying $P(Y_i^* \leq k)$ for the population at each k, one can trace out the full population cumulative density function of Y^*. Proposition (A.2.3) generalizes the population bounds derived in Section 2 of the main text. In this case, the bounds have a new and interesting property. Even if individuals are highly susceptible to framing effects when they prefer some option far away from k, our estimate for the fraction of people preferring option k can still be precise because the partition consistency principle permits us to ignore individuals who consistently choose options above or below k.

E.2 Non-Binary Frames

We next extend the model beyond a simple binary frame in two ways. First, we consider the situation in which the intensity of the frame varies. We show how this reduces to a binary, two-frame problem under intuitive assumptions. With additional structure, the intensity of framing is a valid decision-quality instrument. Second, we show that one can analyze data with multiple dimensions of framing in a very similar fashion to how we analyze frame intensity, using a modified set of assumptions.

E.2.1 Varying Framing Intensity

We now suppose that i chooses from a binary menu $Y_i \in \{0,1\}$ under one of multiple frames that vary in the intensity with which they pull the decision-maker toward option 1, $D \in \{0,1,...,J\}$. The following modification to our frame monotonicity assumption captures this idea:

(A11) (Pairwise Frame Monotonicity) $\forall i\,,\, D \geq D' \implies Y_i(D) \geq y_i(D')$

Pairwise frame monotonicity implies that the frames can be ordered according to their intensity; if a decision-maker chooses an option under one frame, he or she will also choose that option under any frame that pushes towards it more intensely. For example, $D = 0$ might indicate a frame in which option 0 is the default and the choice environment makes the cognitive cost of opting out very high, and $D = J$ might indicate a frame where option 1 is the default and the cognitive cost of opting out is very high; the intermediate frames would correspond to situations where the opt-out costs were lower. Alternatively, the decision could be one in which the decision-maker must choose whether to purchase a good for a given price, and the frame describes the reference point with which the decision-maker has been anchored.

We also modify the consistency principle, weakening it so that choices are only assumed to reveal preferences when the decision-maker is consistent across all observed frames:

(A12) (Global Consistency Principle) $Y_i(D) = Y_i(D')\ \forall D,\, D' \implies Y_i(D) = Y_i^*$

When (A11) and (A12) hold, the multi-frame setting can be reduced to the binary one studied in the rest of the paper, with respect to the most intense frames observed in each direction. Specifically, consider the binary frame $\widetilde{D}$, defined as $\widetilde{D} = 1$ when $D = J$ and $\widetilde{D} = 0$ when $D = 0$. For other values of D, $\widetilde{D}$ is undefined. Note that with respect to $\widetilde{D}$, (A11) and (A12) imply that frame monotonicity and the consistency principle are each satisfied. The earlier propositions in the paper can then be directly applied.

Analog to Decision-Quality Instruments Additional structure beyond the global consistency principle allows one to recover even more information on preferences. To illustrate this, we note that this problem has an interesting relationship to the model of decision-quality instruments presented in Section 3.3. We alluded in Section 3.3to the idea that valid decision-quality instruments can include those varying the intensity of a given framing effect. Adopting the notation of Section 3.3, let the four possible frames $\tilde{D} \in \{0,1,2,3\}$ be given by $\tilde{D} = 0$ when $(D,Z) = (0,0)$, $\tilde{D} = 1$ when $(D,Z) = (0,1)$, $\tilde{D} = 2$ when $(D,Z) = (1,1)$ and $\tilde{D} = 3$ when $(D,Z) = (1,0)$. The consistency principle from Section 3.3, at $Z = 0$ is equivalent to the global consistency principle (A9). The idea in the previous paragraph, using the two most extreme frames, is exactly analogous to recovering the preferences of the

consistent group at $Z = 0$: $\overline{Y}_C(0) = E[Y_i^*|C_i(0) = 1]$. The consistency principle at $Z = 1$ in Section 3.3also implies a second condition for consistency across $D = 1$ and $D = 2$, which will imply that $\overline{Y}_C(1) = E[Y_i^*|C_i(1) = 1]$. Frame separability in this model implies decision quality exclusion, and frame monotonicity implies decision quality monotonicity, so that all changes in behavior between $D = 0$ and $D = 1$, and between $D = 2$ and $D = 3$ tell us about the preferences of individuals consistent across $(0, 3)$ but not across $(1, 2)$, which allows us to recover the preferences of decision-makers whose choices depend on the intensity of framing, $Y_S = E[Y_i^*|C_i(1) > C_i(0)]$.

E.2.2 Multi-Dimensional Frames

We next consider choice settings in which frames differ along multiple dimensions, so that frames cannot be ordered by intensity. Individual i chooses from a binary menu $Y_i \in \{0, 1\}$ under a *frame vector* $D = (D_1, ..., D_K)$, so that each component of D encodes some feature of the choice environment. We assume each frame component D_k of D is discrete with two possible realizations: $D_k \in \{0, 1\}$.[7] For example, a decision-maker's choice between two options might be affected both by which option is presented first and by which option is framed as the default. In this example, D_1 could describe the order of the options and D_2 could describe which option is the default.

As before, denote choices under frame D by $Y_i(D)$. We will assume component-wise frame monotonicity:

(A11') $\forall i\,, \forall k, \forall D_{\neg k},\ Y_i(1, D_{\neg k}) \geq Y_i(0, D_{\neg k})$, where $D_{\neg k}$ is a vector consisting of all frame components other than k and we have re-written $Y_i(D)$ as $Y_i(D_k, D_{\neg k})$.

Assumption (A11') implies that frame monotonicity holds for each component of the frame vector when all other components are held fixed. It also requires that the direction of the effect of any one decision characteristic on choice be independent of other decision characteristics. For example, it must not be the case that making option 0 the default induces more decision-makers to select $Y_i = 0$ when 0 is listed first but that making 0 the default induces more decision-makers to select $Y_i = 1$ when 0 is listed second.

As before, we assume the global consistency principle (A12). Whenever the individual would choose the same option in *every* frame, she prefers the option that she chooses.

When (A11') and (A12) hold, we can proceed similarly to the previous subsection using the two most extreme frames as a binary frame $\tilde{D}$, where $\tilde{D} = 0$ when $D = (0, 0, ..., 0)$ and $\tilde{D} = 1$ when $D = (1, 1, ..., 1)$. The results from the main part of the paper are then obtained with this binary frame.

Further possibilities are generated by considering what additional structure might allow us to learn about preferences from the behavior of individuals who are consistent with respect

[7] It is straightforward to combine this approach with the one in which frames vary by intensity.

to all frame components but one, all frame components but two, and so on. For instance, in the case where there are two frame components $D = (D_1, D_2)$, we might be willing to impose that for all individuals, D_1 always exerts a stronger pull on behavior than D_2, i.e. $Y_i(1,1) \geq Y_i(1,0) \geq Y_i^* \geq Y_i(0,1) \geq Y_i(0,0)$ for all individuals. For example, ordering might exert a weaker influence on behavior than the default. This naturally generates an ordering of the entire frame vector and the problem becomes isomorphic to the analysis of framing intensity, including the relationship to decision-quality instruments just described. In this instance, the analog of the decision-quality instrument would indicate whether ordering influenced decisions in the same direction or the opposite direction as the default. When ordering reinforces the default, the inconsistency is relatively large, and when ordering mitigates it, the inconsistency is small.

E.3 Relaxing Frame Exogeneity

Throughout the body of the text, the frame exogeneity assumption allowed us to set aside the challenge of identifying the effect of the frame on behavior to focus on the preference identification problem. Specifically, from observing $\overline{Y}(0) \equiv E[Y_i(0)|D = 0]$ and $\overline{Y}(1) \equiv E[Y_i(1)|D = 1]$, frame exogeneity allowed us to recover $E[Y_i(0)]$ and $E[Y_i(1)]$. The principle was the same when these moments were observed for specific subgroups of decision-makers or at specific values of a decision quality instrument.

In practice, frame exogeneity is most plausible when decision-makers are randomly assigned to frames or when substantively identical cohorts are observed under different frames, as in our empirical example. In this section, we show that our proposed methods extend naturally to settings in which the effect of the frame on behavior must be estimated using other conventional tools of causal inference. These methods allow us to recover moments analogous to $E[Y_i(0)]$ and $E[Y_i(1)]$, which can then be incorporated as the primitives in our main identification results.

E.3.1 Matching on Observables

Suppose that frame exogeneity fails unconditionally, but holds once the researcher conditions on decision-makers' observable characteristics. In this case, the researcher can use a standard regression or matching-on-observables type approach to estimates of $E[Y_i(0)]$ and $E[Y_i(1)]$. In particular, the conditional frame exogeneity assumption (A2') introduced in Section 3.2 is exactly the assumption necessary to identify these two primitive parameters using this technique. With that assumption, one can recover $E[Y_i(0)]$ and $E[Y_i(1)]$ by using that $E[Y_i(1)] = E_X[E[Y_i(1)|D_i = 1, X]]$ and $E[Y_i(0)] = E_X[E[Y_i(0)|D_i = 0, X]]$.

E.3.2 Difference in Differences

Decision-makers are assigned to one of two groups, the treatment or the control, denoted respectively by $A_i \in \{0,1\}$, and are observed choosing from $\mathbf{X}$ in two time periods, $T \in \{0,1\}$. All individuals make their choice under $D_i = 0$ when $T = 0$, as do individuals in the control when $T = 1$. Individuals in the treatment group choose under $D_i = 1$ when $T = 1$. Choices may potentially depend on both D and T: $Y_i = Y_i(D,T)$. The researcher thus observes four population moments: $E[Y_i(0,0)|A_i = 0]$, $E[Y_i(0,1)|A_i = 0]$, $E[Y_i(0,0)|A_i = 1]$, and $E[Y_i(1,1)|A_i = 1]$.

In this setup, the standard parallel trends assumption corresponds to

$$E[Y_i(0,1)|A_i = 0] - E[Y_i(0,0)|A_i = 0] \;=\; E[Y_i(0,1)|A_i = 1] - E[Y_i(0,0)|A_i = 1]$$

Under this assumption, it is straightforward to show that the researcher can recover the average effect of the frame on the behavior of the treated group in time period 1, $E[Y_i(1,1) - Y_i(0,1)|A_i = 1]$, and/or the counterfactual behavior of the same group in the same time period, $E[Y_i(0,1)|A_i = 1]$. Hence, replacing frame exogeneity with the parallel trends assumption allows us to identify $E[Y_i(0,1)|A_i = 1]$ and $E[Y_i(1,1)|A_i = 1]$, the analogs of $\overline{Y}(0)$ and $\overline{Y}(1)$ for decision-makers assigned to the treatment group. From these, we can construct our statistics of interest like $\overline{Y}_C$ to obtain preference information for consistent decision-makers *in the treatment group*, $E[Y_i^*|C_i = 1, A_i = 1]$, or preferences for the entire sub-population of decision-makers assigned to the treatment group, $E[Y_i^*|A_i = 1]$. To recover preferences for the full population of consistent decision-makers (or the full population) requires extrapolating the Average Treatment Effect on the Treated (ATET) to the Average Treatment Effect (ATE) – a problem which has been well-studied in the program evaluation literature.

E.3.3 Instrumental Variable Analysis

Suppose the researcher uses an instrumental variable to estimate the causal effect of the frame on behavior. (To be clear, the type of instrument that would be used here is entirely distinct from the concept of a decision quality instrument developed in Section 3.3.) Each decision-maker is assigned to one of two values of the instrument, $W_i \in \{0,1\}$, and suppose the frame to which one is assigned is a function of the instrument, $D_i(W_i)$, with $D_i(1) \geq D_i(0)$ and $E[D_i(1) - D_i(0)] > 0$. Choice may still depend on D, but does not otherwise depend on Z, so we can continue to write $Y_i(D)$. An individual is thus represented by the following vector of random variables: $(Y_i(0), Y_i(1), Y_i^*, D_i(0), D_i(1), W_i)$. Finally, we assume assignment of the instrument is exogenous with respect to behavior, $W_i \perp (Y_i(0), Y_i(1))$.

With this setup, it is straightforward to recover $E[Y_i(1)|D_i(1) > D_i(0)]$ and $E[Y_i(0)|D_i(1) >$

$D_i(0)]$ (Imbens and Rubin, 1997), which can be used in place of $\overline{Y}(0)$ and $\overline{Y}(1)$ in our proposed estimators. For example, Proposition 1 would recover the preferences of the consistent compliers, $E[Y_i^*|C_i = 1, D_i(1) > D_i(0)]$, and the subsequent propositions would recover the preferences of the overall population of compliers, $E[Y_i^*|D_i(1) > D_i(0)]$. To the extent that the compliers have different preferences than the rest of the population, one could implement methods for extrapolating the LATE to the ATE in this setting (e.g., Angrist and Fernandez-Val, 2013).

Finally, note that the same logic would apply if a regression discontinuity design is used to estimate a causal framing effect. In this case, one could identify information about the preferences of the subgroup for whom the discontinuity induces treatment, i.e. the compliers at the threshold of the running variable (or all individuals with a running variable at the cutoff in the case of a sharp regression discontinuity design).

References

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.

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

Imbens, Guido W., and Donald B. Rubin. 1997. "Estimating Outcome Distributions for Compliers in Instrumental Variables Models." *Review of Economic Studies*, 64(4): 555–574.