Essays in Behavioral Public Economics

by

Daniel Reck

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in the University of Michigan
2016

Doctoral Committee:

Professor Joel B. Slemrod, Chair
Professor James R. Hines Jr.
Professor Miles S. Kimball
Professor Stefan Nagel

For my parents.
Mom and Dad, I didn’t do this just to make you proud.

Acknowledgements

I am forever indebted to my advisor Joel Slemrod, with whom I had a great many conversations that made the research contained in this dissertation possible. I have also been extremely fortunate to receive invaluable advice and feedback from my dissertation committee: Jim Hines, Miles Kimball, and Stefan Nagel.

To my friends in the University of Michigan and Ann Arbor communities, I owe gratitude for countless conversations about all kinds of ideas, and for the open and supportive community in which we lived. In economics, these include Eric Chyn, Morris Hamilton, Enda Hargaden, Aristos Hudson, Gaurav Khanna, Ben Meiselman, Daniela Morar, and Danny Schaffa. In the broader community, Kenzie Allen, Emily Clader, Julia Goldberg, Sam Heidepriem, Jake Levinson, Amy Wilson, Sunny Yao, Fangting Yu, the many volunteers and staff of 826michigan, and many others I surely accidentally left out.

My academic work has been blessed by the opportunity to collaborate with incredible coauthors and participate in discussions in several research communities. I especially benefited from my collaboration with Jacob Goldin, with whom I wrote some of the material contained here. I have been fortunate to work and collaborate within fantastic scholarly communities at Michigan, the University of Copenhagen, the University of Oklahoma, and the Internal Revenue Service. Working in behavioral public economics has given me the good fortune to be a part of an exciting community of mutually supportive young scholars, with whom I have enjoyed many conversations. These include Sebastien Bradley, Naomi Feldman, Tatiana Homonoff, Ben Lockwood, Alex Rees-Jones, and Dmitry Taubinsky. I owe a special debt to John Beshears and Brigitte Madrian, who graciously shared the data used in the first chapter of this dissertation, and who provided valuable discussion of the research itself.

Finally, I thank my partner, Linea, for her unconditional support, and for reminding me to enjoy life from time to time.

Table of Contents

List of Tables

List of Figures

List of Appendices

Introduction

Behavioral economics is a blessing and a curse. A large body of evidence from behavioral economics suggests that individuals often behave in ways that contradict the predictions of rigid rational choice models. These discoveries have made economics more scientific, and they have lead to fruitful efforts to build more realistic theories of decision-making. In the policy sphere, discoveries from the field behavioral economics have introduced powerful new tools for influencing behavior. All of these promising developments come with a catch, though. Economists' conception of optimal policymaking always relies in one way or another on the principle that preferences are revealed by choices. When we lose rational choice, we also lose revealed preference. As a result, there is intense controversy over how to use the powerful new policy tools introduced by behavioral economics, and also a lingering concern that our old reasoning about conventional policies might be fundamentally flawed.

How do we understand welfare when some choices do not reveal preferences? What should we do with all these new policy tools? For that matter, what do we really know about our old policy tools? In many ways, these questions have been a long time coming. When critics of economics dismiss the the field and its research agenda because "people are clearly not rational," they mean, not that choices cannot be explained by a complete and transitive relation on the menu of available objects in the positive sense, but rather that choices often do not reveal normative preferences. The controversy in behavioral welfare economics is really an older controversy about rationality. The evolution of behavioral economics has merely brought that debate inside the discipline of economics.

In the chapters that comprise my dissertation, I will attempt to shed some new light on this issue. I do so not by laying out a set of normative assumptions I believe that everyone should always make, but rather describing several lines of argument that *could* be applied to choice data and describing the conditions under which each of them lead to valid inference about preferences and optimal policy. After all, a long tradition in public economics defers ultimate normative judgements, such as the value of economic equality, to policymakers, but nevertheless to describe how policymakers with particular beliefs about welfare can design optimal policies.

The first chapter contained herein examines the use of revealed preference analysis in situations where choices depend on seemingly arbitrary features of the decision-making environment, called *frames.* Frames could be, for example, default options, the order in which options are displayed, or the degree to which options or features of the options are made salient. Many of the new policy tools introduced by behavioral economics can be thought of as a deliberate manipulation of frames by policymakers. How to recover preferences is an important first step towards understanding the welfare consequences of these policies. The key insights of this work are that 1) information on the preferences of persons immune to frames can be recovered even with limited data, and, given this finding, that 2) the preferences of persons who *are* influenced by frames can be regarded as a familiar sample selection problem. Though they often prove difficult to solve, economists have developed a deep understanding of sample selection problems, such as those arising in the presence of missing data, self-selected samples, and selection into treatment. Viewing the behavioral revealed preference problem through this lens leads one to several new tools for addressing the problem, described in this chapter. Some of these tools are then applied to perhaps the most well-known policy tool introduced by behavioral economists: automatic enrollment into retirement savings plans. The resulting analysis suggests that while most individuals prefer participation in employer-sponsored retirement savings plans, and thus would be made better-off by automatic enrollment, the majority of younger, low-income individuals do not. This finding raises questions about whether policymakers should automatically enroll such individuals in retirement savings plans.

The second chapter applies similar methods to the question of how to analyze survey data characterized by framing effects. In the survey context, frames could encode the order in which answers or questions are provided, or the wording of questions. While a large literature, summarized below, documents the existence of framing effects, no formal work has considered how to extract useful information from surveys despite framing effects, perhaps for the same reason that the optimal policy problem in behavioral economics has proven difficult to address. One ubiquitous technique for attempting to accommodate framing effects, used by researchers across the social sciences and many prominent survey and political polling organizations, is to randomize respondents across frames and then to pool the data in estimation. This chapter shows that this technique does not actually solve the problem in any meaningful way. The results from randomizing and pooling will instead be a confused mixture of the distribution of the quantity the survey was designed to measure and the distribution of the effects of framing. As in the first chapter, however, careful statistical reasoning can provide preferable techniques for extracting useful information from surveys. The second chapter lays out several of these techniques and discusses their underlying assumptions and usefulness in the survey context.

The third chapter concerns the implications of behavioral economics findings for a more conventional topic in public economics: optimal tax policy. A large and growing literature has shown that individuals are often inattentive to tax incentives when making decisions, or they mis-perceive tax schedules. One might be tempted to conclude that taxes that individuals ignore are actually desirable policies. For example, policymakers often worry that taxing people will cause them to work less, but if people do not realize they are being taxed, then they may not work less, which makes the tax more like the canonical ideal of a "lump-sum" tax. This paper generalizes this reasoning to any type of misperception, points out two important caveats, and argues that these caveats deserve greater attention in the empirical examination of these topics. The first caveat concerns what happens when individuals are selectively attentive to higher-stakes incentives, i.e. they may be more attentive when the hidden tax is larger than when it is smaller. In this case, attempting to exploit mistakes for social gain, by relying more heavily on under-perceived taxes, can backfire. The social consequences of this backfiring, called the "curse of debiasing," can reverse the conclusion of whether policymakers should exploit inattention to taxes, especially when attention is costly for individuals to allocate. The second caveat concerns budget constraints: individuals who misperceive relative prices will also misperceive their budget constraint. How individuals nonetheless satisfy their budget constraint has important implications for welfare, which may not be fully captured by empirical analysis based on field experiments. Both the phenomenon of debiasing and of budget adjustment are in principle empirically examinable, and such empirical work would substantially improve the debate over optimal taxation in the presence of mis-perception.

This research does not fully solve the problem of how to conduct optimal analysis in a behavioral economic world, but it takes a few steps in the right direction. While the research does contain assumptions about welfare that may, at least in some contexts, be subject to debate, it does so in a completely transparent fashion, which will allow future research to consider alternative assumptions using similar tools. In the course of studying this problem, I have become more optimistic about the possibility of building a consensus on how economists can approach this problem without relying exclusively on paternalistic judgements. Whereas the difficulty with optimal policy in behavioral economics is often cast as a philosophical impasse, leading some to become outright nihilistic in their views about welfare, this work shows that a combination of economic and econometric theory can make the problem parsimonious, tractable, and even a little familiar.

Chapter 1

Preference Identification Under Inconsistent Choice[1]

by Jacob Goldin (Stanford Law School) and Daniel Reck (University of Michigan)

Abstract

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

1 Introduction

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

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

is opt-out. Both policies let customers control the use of their data, but the choices made under the two policies imply different conclusions about what customers prefer.

Situations such as this one pose an important challenge to behavioral economics. In many settings – from privacy controls to retirement savings (Madrian and Shea, 2001) to health insurance (Handel, 2013) to voting (Ho and Imai, 2008) – choices depend on seemingly arbitrary features of the decision-making environment, such as which option is the default, the order in which options are presented, or which features of the options are made salient. Inconsistencies in choice invalidate the traditional revealed preference approach of equating choices with preferences.[2] This difficulty contributes to widespread disagreement about when and how to use behavioral economics to inform policy.

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

Most prior work takes one of two approaches to the problem of preference identification under inconsistent choice. First, researchers may utilize a positive model of behavior that fully specifies the mapping from decision-makers' preferences to their (potentially suboptimal) behavior (e.g. Rubinstein and Salant, 2012; Benkert and Netzer, 2014; Carroll et al., 2009). Such approaches yield important insights but in many cases the resulting welfare conclusions are sensitive to the researcher's choice between competing positive models that are difficult to distinguish observationally (Bernheim, 2009; De Clippel and Rozen, 2014). An alternative approach is to restrict preference inferences to the subset of observed choice situations in which a given decision-maker chooses consistently (Bernheim and Rangel, 2009). However, in practice individual decision-makers are typically observed making only a single choice, which makes it difficult to detect which choices are consistent. Worse, this approach yield no information on the preferences of those decision-makers who exhibit systematic choice reversals – the very group whose preferences are most relevant for making optimal policy determinations regarding how choices should be designed (a claim we formalize in Ap-

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

pendix B). Further "refinements" can provide a path forward if the researcher can observe choices in a setting in which *all* decision-makers are known to select their most-preferred option (e.g., Chetty, Looney and Kroft, 2009), but in many applications, such as those in which behavior is sensitive to defaults or ordering, there is little reason to believe that any of the observed choice situations satisfy this condition.[3]

Our approach to this problem overcomes many of the limitations of prior work. We develop techniques to recover preference information with limited datasets – those in which each individual is observed making only one decision and observers lack ex ante knowledge about which individuals are optimizing. We do not assume that the researcher knows the exact underlying positive model that generates the choice inconsistencies, nor do we assume that all decision-makers choose optimally in any one of the observed choice situations. We provide general conditions – consistent with a broad class of behavioral models – under which the preferences of the consistent decision-makers can be either partially- or point-identified. We also develop techniques for learning about the preferences of those decision-makers who exhibit choice inconsistencies. Together, these innovations provide a range of practical tools for better understanding the distribution of preferences in populations that exhibit inconsistent choice behavior.

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

Within this framework we make two key identifying assumptions. First, when decision-

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

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

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

Our first main result concerns the recovery of the preferences of consistent decision-makers – those who would select the same option under each frame.[5] To do so, we exploit the fact that under our assumptions, a decision-maker who chooses "against the frame" – for example someone who chooses the option that is not the default – is consistent and prefers the option that she chooses. This insight, along with a statistical assumption concerning the assignment of decision-makers to frames, allows us to recover preferences among the consistent decision-makers. Without frame monotonicity, the preferences of the consistent decision-makers are partially identified, and we derive the corresponding bounds.

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

The first technique is to adjust for observable differences, such as income or education, between the consistent and inconsistent decision-makers, and then to extrapolate from the former to the latter. If consistency and preferences are uncorrelated conditional on these

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

observable characteristics, one can recover population preferences by separately estimating the preferences of each demographic group and then re-weighting those estimates based on the distribution of observables among the inconsistent decision-makers. As in other empirical contexts, the plausibility of this matching-on-observables approach depends on what information about decision-makers can be observed.

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

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

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

The paper proceeds as follows: Section 2 sets up the model. Section 3 provides point- and partial-identification conditions for the preferences of the consistent choosers. Section 4 addresses the problem of recovering preferences for the full population. Section 5 illustrates our approach using data on defaults and enrollment into employer-provided pension plans. The Appendix[6] contains proofs of propositions (A); motivates our parameters of interest with a simple model of optimal frame design (B); considers the relationship between our framework and alternative structural models of default effects (C); generalizes the framework to settings with non-binary frames and non-binary menus (D); and derives standard errors for finite-sample inference (E).

2 Setup

This section introduces the notation and assumptions employed throughout the paper. We observe individual choice data from a population of density 1, with individuals denoted by i. The observed decisions are binary, $y_i \in \{0, 1\}$, and each decision-maker is observed under exactly one of two possible *frames*, denoted d_0 and d_1.[7] Let y_{1i} and y_{0i} denote what i would choose under d_1 and d_0, respectively. Population moments are given by $Y_1 \equiv E[y_{1i}|d_i = d_1]$ and $Y_0 \equiv E[y_{0i}|d_i = d_0]$. Without loss of generality, assume $Y_1 \geq Y_0$. To illustrate the notation using the privacy example from the introduction, y could indicate whether an individual allows a company to use her data, so that d_1 would indicate the opt-out regime, and d_0 would indicate the opt-in regime. We assume throughout that population moments such as these are directly observable, setting aside issues of finite-sample statistical inference.

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

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

Frame separability limits which features of the decision-making environment are treated as a frame. Features of a decision that affect choice but that are relevant to decision-makers' preferences over the available options are *not* frames.[8] Importantly, frame separability does

[6] Available on the authors' websites.

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

[8] For example, if a decision-maker chooses hot chocolate from {hot chocolate, ice cream} under one frame and ice cream from {hot chocolate, ice cream} under the other frame, there would be no apparent deviation from rationality if the frame indicated whether the season was winter or summer. This assumption is explicit

not require decision-makers to be irrational; a decision-making feature that imposed a transaction cost for selecting one of the options would constitute a frame, as long as it did not also affect decision-makers' preference for *receiving* one option or the other.[9]

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

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

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

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

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

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

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

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

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

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.

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

also derive partial identification results for settings in which the assumption fails.

3 Identifying Consistent Preferences

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

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

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

(1.2) *Under A1 - A3,*

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

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

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

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

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

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

Table 1: Aggregate Choices by Frame

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

preferences of the corresponding decision-makers. As a result, the denominator of Y_c measures the fraction of decision-makers that are consistent and the numerator measures the subset of that group with $y_i^* = 1$.

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

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

The preference information recovered by Proposition 1 is important for several reasons. First, if one's philosophical starting point is that inconsistent decision-makers lack normatively relevant preferences (see the discussion of this issue in Fischhoff, 1991), Proposition 1 is the end-point of the analysis; our method isolates the normatively-relevant parameter (the consistent decision-makers' preferences) from the noise induced by the frames. Second, when population preferences are known – what Bernheim and Rangel (2009) refer to as a "refinement" – Proposition 1 can be used in conjunction with that information to recover the preferences of the *inconsistent* decision-makers.[10] Such information is often valuable because optimal policy may turn on the preferences of the inconsistent decision-makers (see Appendix B), but observing aggregate population preferences under a refinement does not provide the preferences for that subgroup. Finally, the preferences of the consistent decision-makers may be used to recover the preferences of the remainder of the population by accounting for selection into the consistent sub-population, which is the task we undertake in the remaining sections.

4 Identifying Population Preferences

The remainder of the paper addresses how to use the preferences of the consistent decision-makers to gain information about the preferences of the inconsistent decision-makers or of the full population. The basic challenge to doing so is overcoming a potential selection bias: when selection into the consistent sub-population is not random, characteristics of the consistent decision-makers may be correlated with the preferences of that group. In many ways, this challenge parallels the well-known problem of selection into treatment that has been studied in the program evaluation literature. However, an important difference is that in the typical sample selection context, the researcher can identify which units have been selected into the relevant sample.[11] In contrast, whether a particular decision-maker is consistent is unobservable when each decision-maker is observed under a single frame.

To clarify the selection issue, note that we can write

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

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

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

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

a condition we refer to as *decision quality independence* – the preferences of the consistent decision-makers will be representative of the full population.[12]

4.1 Partial Identification

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

Proposition 2

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

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

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

Without frame monotonicity, we obtain weaker, one-directional bounds for population preferences. The result follows from noting that $E[y_i^*]$ depends on three parameters: $E[y_i^*|c_i = 0]$, $E[y_i^*|c_i = 1]$, and $E[c_i]$. Although $E[y_i^*|c_i = 0]$ is unobservable, the other two parameters can be inferred from the data given information on the prevalence of frame-defiers. Knowing that $E[y_i^*|c_i = 1] \in [0, 1]$ constrains the prevalence of frame defiers, which then yields bounds on the value of $E[y_i^*]$. The further Y_0 is from $1 - Y_1$, the more informative the bounds will be.[13] Note that when frame monotonicity fails, it is possible that a majority of decision-makers choose one option under both frames even though a majority of the population in fact prefers the *other* option.

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

To summarize the results thus far, the degree to which $E[y_i^*]$ and $E[y_i^*|c_i = 1]$ can be identified from the data depends on the strength of the researcher's assumptions about the

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

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

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

4.2 Matching on Observables

In this section, we consider situations where the relationship between preferences and consistency depends on characteristics of decision-makers that are observable to the researcher, such as income, education, age, or prior experience with the decision at hand. For example, it could be that more educated customers are less likely to prefer that companies use their personal data and are more likely to choose consistently across default regimes, but that conditional on education, preferences and consistency are independent.

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

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

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

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

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

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

An important barrier to employing this familiar approach in our context is that we cannot directly observe consistency. The following lemma shows that the distribution of characteristics among the consistent and inconsistent decision-makers is nonetheless identified.[14]

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

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

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

Apart from its role as a step in the construction of the matching estimator, Lemma 1 is useful in its own right. Information on the observable correlates of consistency is important for researchers investigating the mechanisms by which frames affect decision-making and for policymakers designing interventions aimed at particular sub-groups of the population.[15] Exploiting Lemma 1 along with conditional decision quality independence, the following proposition formalizes the matching-on-observables identification strategy described above:

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

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

(3.2) $E[y_i^*|c_i = 0] = E_w[s_w\, Y_c(w)]$[16]

Table 2 illustrates the identification approach for our privacy controls example, reporting hypothetical data conditioned upon whether individuals have at least a high school education. The population moments in the third column match the moments in Table 1. The

[14] Lemma 1 is analogous to Abadie (2003), who shows how to identify the aggregate observable characteristics of compliers with respect to an instrument when individual compliers cannot be identified. Continuing with the analogy, Proposition 3 is related to Angrist and Fernandez-Val (2013), who exploit information on the distribution of observables to extrapolate an estimated treatment effect from one subset of a population to another.

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

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

Table 2: Average Choices by Frame and High School Education

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

conditioning reveals that high-school-educated individuals are more likely to be consistent and the consistent choosers among them are more likely to prefer that the company use their personal data ($y_i = 1$). Under conditional decision quality independence, we conclude that 44 percent of inconsistent decision-makers, and 53 percent of the population prefer that their personal data be used. Under *unconditional* decision quality independence, both these fractions would be 57 percent and we would over-estimate the share preferring that their personal data be used, because this approach ignores the relationship between preferenes and consistency.

4.3 Decision Quality Instruments

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

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

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

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

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

Assumption A3" modifies the unconfoundedness assumption, which now requires that both d and z be uncorrelated with confounding factors. Assumption A6 requires that variation in the decision-making environment induced by z is irrelevant from the perspective of decision-makers' preferences; it ensures that z affects behavior by altering consistency, not by changing which option decision-makers prefer.[17] Assumption A7 requires that the effect of z on consistency is weakly monotonic for all decision-makers and strictly monotonic for some.

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

4.3.1 Identifying Sometimes-Consistent Preferences

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

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

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

Proposition 4 is best understood by analogy to the identification of a local average treatment effect (LATE, see Imbens and Angrist, 1994). The monotonicity assumption (A7) permits us to divide the population into three groups of decision-makers: the always-consistent ($c_{ih} = c_{il} = 1$), the sometimes-consistent ($c_{ih} = 1; c_{il} = 0$), and the never-consistent ($c_{ih} = c_{il} = 0$). The denominator of the expression in Proposition 4 measures the decrease in the size of the inconsistent sub-group as we move from z_l to z_h, which identifies

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

the size of the sometimes-consistent group, who are the analog of the compliers in the LATE framework. The expression in the numerator measures the change in the fraction choosing $y = 1$ under d_0 as z changes, which identifies the fraction of decision-makers who are sometimes-consistent *and* prefer $y = 1$.

Several other parallels to the instrumental variables literature are apparent. First, one can use Proposition 4 to motivate over-identification tests of decision quality independence along the lines of Wu (1973) and Hausman (1978). However, such a test requires $E[y^*|c_{ih} > c_{il}] = E[y^*]$, which may fail depending on the nature of selection into consistency. We explore less restrictive alternatives below. Additionally, the types of variation that will satisfy assumptions A6 and A7 depend on the underlying model of behavior that generates framing effects, reflecting a familiar interplay between structural reasoning and instrumental variables. We discuss this issue further in Appendix C. Finally, Proposition 4 may be extended beyond binary decision quality instruments, by applying Proposition 4 to each pair-wise combination of values of z or, when z is continuous, by adapting the methods of Yitzhaki (1996) (see also Heckman and Vytlacil, 2007).[18]

Table 3 illustrates this identification approach for the hypothetical data on privacy controls. We now suppose that the process for adjusting privacy settings may be either onerous (customers are required to navigate through several web pages) or streamlined (customers may adjust privacy settings with a single click). Aggregate choices under the onerous design correspond to the population moments reported in Table 1. Customers are less susceptible to default effects when the process is streamlined. We can back out the fraction of consistent choosers and the aggregate preferences of the consistent choosers when controls are onerous (z_l) or streamlined (z_h) using Proposition 1, as before. Note that the fraction of consistent customers who prefer that the company use their personal data is lower under streamlined controls than onerous ones, because the variation in z affects opt-out rates (Y_1) more than opt-in rates (Y_0). Applying Proposition 4 in this example would imply that of the 20 percent of the population of customers who are sometimes-consistent, only 25 percent prefer that the company use their data, a share substantially below that of the consistent choosers at either z_h or z_l.

An interesting special case of Proposition 4 occurs when, under z_h, all decision-makers are consistent, i.e., $p(c_{ih} = 1) = 1$. For example, default effects may be eliminated by

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

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

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

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

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

4.3.2 Structural Extrapolation with Decision Quality Instruments

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

Suppose that consistency for individual i is determined by

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The structural model described above resembles the classic bivariate normal model of selection (see e.g. Heckman, 1979). Variation in the decision quality instrument induces variation in consistency without affecting preferences; this guarantees the relationship between consistency and preferences is identified without relying on functional form alone (Puhani, 2000).[19]

Applying the model to the data from Table 3 yields an estimated correlation coefficient of $\rho = 0.54$; the positive estimate suggests decision-makers with a high propensity to choose consistently (so that they are consistent at z_l) are more likely to prefer $y_i = 1$ than those with a low propensity to choose consistently. The estimated parameters imply $E[y_i^*] = 0.46$.

[19] 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.

The population average is below both $E[y_i^*|c_{il}=1]$ and $E[y_i^*|c_{ih}=1]$ because it incorporates the preferences of the decision-makers with the very lowest propensity to choose consistently.

4.3.3 Semi-Parametric Extrapolation with Decision Quality Instruments

This section develops an extrapolation approach for recovering population preferences without relying on distributional assumptions. In particular, we model the preferences of the consistent decision-makers at a given value of the decision quality instrument as a flexible polynomial in the fraction of decision-makers who are consistent at that value of the decision quality instrument.[20]

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

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

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

Finally, let $g(z) = E[y_i^*|z_i^* = z]$ denote the preferences of the *marginally consistent* decision-makers at a given z.[21] To guarantee the validity of the Taylor Series approximation that underpins the following result, it will be convenient to assume that both $F(z)$ and $g(z)$ are infinitely differentiable with respect to z.

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

(6.1) *For any z,*[22] $E[y_i^*|z_i^* = z] \approx a_0 + a_1 E[c_{iz}] + a_2 E[c_{iz}]^2 ... + a_D E[c_{iz}]^D$

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

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

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

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

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

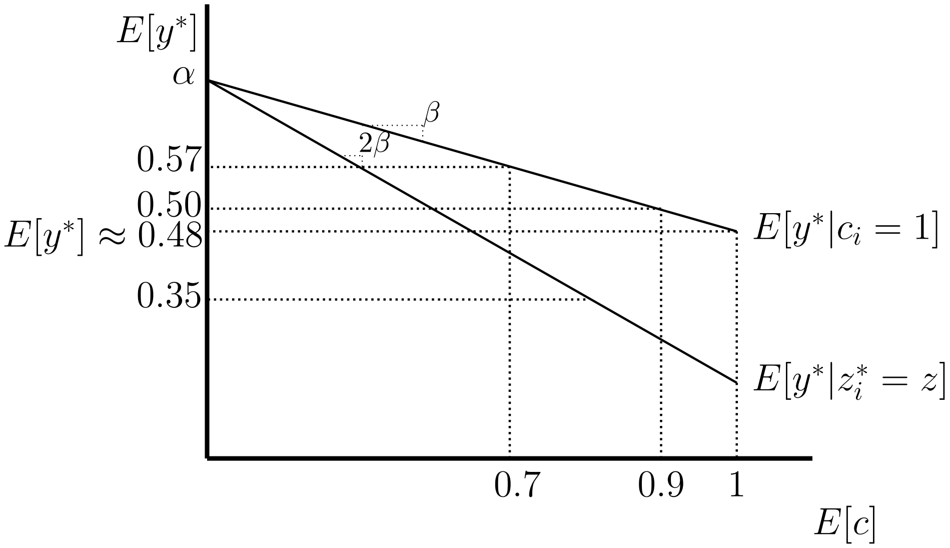
Proposition 6 implies that the preferences of the consistent decision-makers at a particular value of the decision quality instrument can be approximated by a polynomial function in the fraction of decision-makers who choose consistently at that value of the instrument. Because A7' guarantees a one-to-one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

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

5 Application to 401(k) Automatic Enrollment

In this section we illustrate our identification framework by analyzing data on enrollment decisions into employer-provided 401(k) pension plans. Depending on how the plan is designed, new employees may have to actively enroll in the plan in order to participate, or their enrollment may be automatic unless they choose to opt out. An influential body of research documents striking differences in take-up and savings behavior between such opt-in and opt-out regimes (Madrian and Shea, 2001; Choi et al., 2006; Chetty et al., 2014). Most policy discussion of automatic enrollment takes as its starting point that employees would be better off under 401(k) plan designs that cause them to save more than they otherwise

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



would. Our framework allows us to investigate employee preferences empirically, without imposing paternalistic assumptions of this form. Moreover, although others have studied the welfare effects of defaults in this setting (Carroll et al., 2009; Bernheim, Fradkin and Popov, 2015), an important advantage to our approach is that it does not require taking a stance on the particular positive model that generates any decision-making bias.[23] Only models that violate the consistency principle are entirely ruled out. In addition, an advantage of the reduced-form nature of our approach is that it makes transparent what role each assumption plays in identification, such as how frame monotonicity and conditional decision quality independence strengthen the conclusions that can be drawn from the data.

5.1 Data

To study the effect of automatic enrollment on pension plan participation, we use data from the large health care and insurance firm studied in Madrian and Shea (2001). This firm switched from an opt-in to an opt-out enrollment design in April 1998,[24] and Madrian

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

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

Table 4: Employee Characteristics

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

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

and Shea (2001) find that the switch caused a large increase in the fraction of employees choosing to participate in the firm's pension plan. We use the choice data from this study to investigate employee preferences for plan participation.

We observe whether an employee enrolls in the plan (indicated by y_i) and whether the default is opt-in (d_0) or opt-out (d_1) when he or she is hired. We also observe annual compensation, age, sex, and race for each employee.[25] Table 4 describes employee characteristics. The distribution of characteristics is not substantially different across default regimes. The income, age, and racial composition of its workforce are typical of a large employer in the US, although this firm's workforce is predominately female. Employee participation rates by frame and demographic group are summarized in Table IV of Madrian and Shea (2001).

401(k) is *lower* than our estimates for the fraction preferring enrollment under a 3 percent default.

[25] To ensure individual employees could not be identified, we were provided compensation and age only within a range of values. See Table 4.

For our identification results to apply to the choices in this data, the conditions described in Section 2 must be satisfied. Frame separability seems likely to hold: it is difficult to imagine that an employee's preferences over how much to save depend on how her employer chooses to structure enrollment into its sponsored retirement plan. Unconfoundedness requires that an employee's hiring date be uncorrelated with whether she chooses to participate under either plan design. This is the same assumption necessary to identify the causal effect of the change in enrollment from the change in plan design, and Madrian and Shea (2001) provide evidence this assumption is satisfied. Frame monotonicity requires that no employee chooses to enroll when enrollment is opt-in but chooses not to enroll when enrollment is opt-out, which seems plausible in this setting.

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

5.2 Results

To begin, we focus on the preferences of the consistent decision-makers. Table 5 columns 1 and 2 report the aggregate enrollment rates under the two policy designs. The estimated population participation rates are $\hat{Y}_1 = 0.859$ under opt-out and $\hat{Y}_0 = 0.491$. Columns 3 and 4 apply Proposition 1 to this data. Substituting the estimated population moments into the definition of Y_c in Proposition 1 yields $Y_c = 0.777$, with a standard error of 0.006. Thus, under frame monotonicity, of the 63.2 percent of employees whose enrollment decisions are sensitive to the enrollment design, we conclude that a large majority (77.7 percent) preferred enrollment. Without assuming frame monotonicity, Proposition 1.2 implies the fraction of consistent employees that prefer enrollment is *at least* 77.7 percent.

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

We next investigate heterogeneity among employees in their preferences for enrollment and their sensitivity to the enrollment regime. Although we cannot directly observe either preferences or consistency for individual employees, the results from section 4.2 allow us to investigate differences based on employees' observable characteristics. We estimate a regression of the form

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

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

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

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

The results of the analysis are reported in Table 6.[26] The results suggest that both consistency and the preferences of consistent choosers vary systematically and substantially by employee characteristics. Variation in consistency is strongly related to variation in compensation, with those in the highest compensation bin (annual income over \$50K), estimated to be 40 percent more likely to choose consistently than those in the lowest bin (annual income less than \$20K). The estimated differences in consistency by income are statistically significant ($p < 0.001$).[27] When compensation is controlled for, differences in consistency are not significantly associated with heterogeneity in age, race, or gender.

Turning to preferences for enrollment among consistent employees, we document signif-

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

[27] This finding adds to a growing literature that documents important differences in susceptibility to decision-making biases by income (e.g., Mullainathan and Shafir, 2013; Goldin and Homonoff, 2013; Choi et al., 2014). Unlike Choi et al. (2014), our approach allows us to identify patterns in consistency without observing individuals making multiple decisions.

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

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

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

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

The fact that both consistency and preferences for enrollment among the consistent decision-makers are associated with income casts doubt on the plausibility of decision-quality independence in this setting. To further assess the validity of the assumption, Figure 2 plots consistency and preferences for enrollment among the consistent employees in each observable cross-section of the data. The scatter plot provides additional (non-parametric) evidence against decision-quality independence: employee groups that contain a greater fraction of consistent decision-makers are also more likely to have a greater fraction of consistent decision-makers that prefer 401(k) enrollment.[28] The estimated slope of the best-fit line is 0.78, suggesting a strong positive relationship between employees' consistency and their preferences. Figure 2 also reveals that a majority of consistent employees in every group with income less than \$20,000 and age less than 30 prefer non-enrollment.[29]

Because of the apparent correlation between preferences for enrollment and consistency in the population of employees, we apply the matching estimator described in Proposition 3 to estimate the distribuion of preferences for inconsistent employees and the full population. To maximize the likelihood that the conditional decision quality independence assumption is satisfied, we use a fully interacted econometric model: each combination of observables defines a unique demographic group. Consequently, for the results to be invalid there must be unobserved variation in consistency among employees of the same age, gender, race, and income that is also correlated with preferences for enrollment.[30]

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

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

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

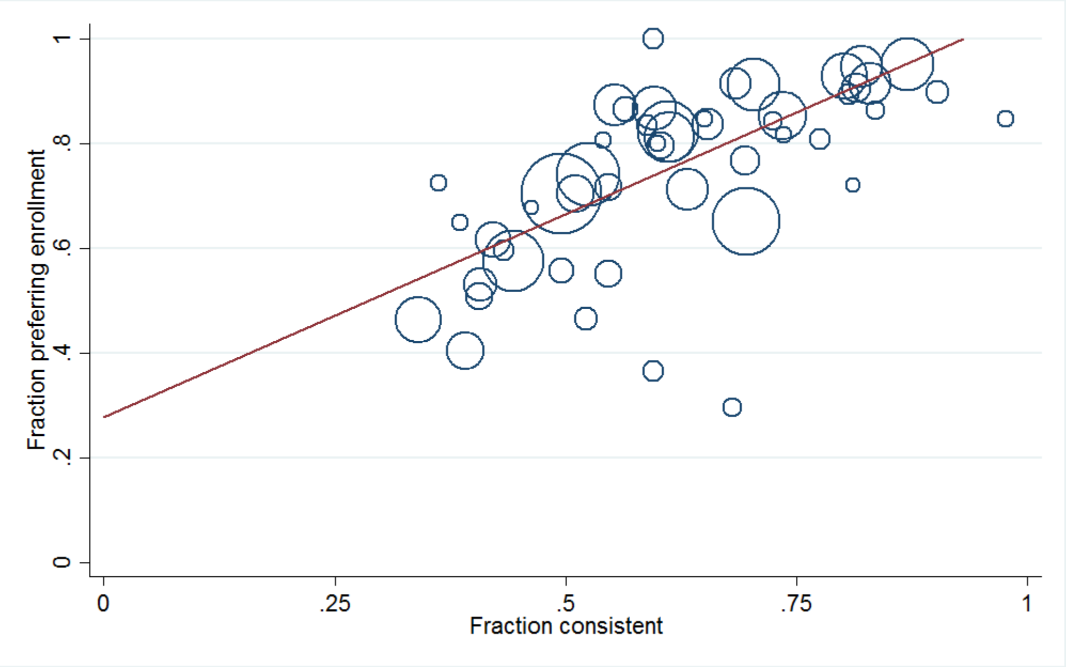
[30] For example, if cognitive ability is positively correlated with both consistency and preferences among employees of the same age, gender, race, and income, our results would yield an upwardly-biased estimate for the preferences of the inconsistent employees. The bias in the matching estimator is given by $E[y_i^*] - E_w[Y_c(w)] = E_w\left[\frac{cov(y_i^*, c_i|w)}{E[c_i|w]}\right]$. This expression follows from the law of conditional expectations and Equation (1).

Table 6: Consistency and Preference by Observable Characteristics

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

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

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



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

5.3 Discussion

Our results suggest that a sizable majority of inconsistent employees, approximately 70 percent, prefer enrollment in this employer's pension plan. This finding is consistent with the more paternalistic view that leading these individuals to save more via automatic enrollment would improve their welfare. Notably, however, we also find that a majority of the youngest and lowest-income workers do *not* prefer enrollment in the plan.[31] This could be, for example, because employees in this group perceive retirement to be far off and are therefore inattentive to retirement savings, and also prefer to save less because they anticipate higher future earnings or are currently paying off student loan debt. To the extent an employer wishes to implement a personalized default regime along the lines suggested in Sunstein (2015), our results suggest a default of non-enrollment may be welfare-maximizing for young, low-income employees.[32]

6 Conclusion

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

Like Bernheim and Rangel (2009), our approach is most appealing in settings where

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

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

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

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

An important feature of our approach is its reduced-form nature. Within the wide range of models consistent with frame-monotonicity and the consistency principle, the basic identification problem – i.e., understanding the empirical correlation between decision-makers' preferences and their sensitivity to frames – is the same regardless of the details of the structural model that generates behavior. On the other hand, our approach is not a replacement for structural models of decision-making. As in other areas of empirical economics, the interpretation of the parameters identified by reduced-form approaches depends on the underlying structural model that generates behavior.[33] Finally, the reduced-form nature of our approach has the virtue of making transparent which assumptions are driving preference identification within a particular application.

The framework studied here can be thought of as a special case of a more general approach in which an observer first identifies the preferences of a reference group of decision-makers whose choices are assumed to reveal their true preferences, and subsequently extrapolates the reference group's preferences to the rest of the population. In our approach, the reference group consists of those decision-makers who choose consistently across frames. When using consistent choosers as the reference group is not feasible or credible, one might replace them with experts, experienced choosers, or those thought to be immune to the framing effect in question (e.g., Johnson and Rehavi, 2015; Bronnenberg et al., 2013; Handel and Kolstad,

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

2015). The identification techniques we have proposed may be utilized with these reference groups as well; for example, one might adjust the recovered preferences of experts based on observable characteristics before extrapolating their preferences to the rest of the population, or utilize exogenous variation that causes some individuals to become experts.

Our results are subject to several limitations. First, in certain applications even consistent choices may not reveal preferences. For example, decision-makers who consistently choose one retirement plan over another, regardless of the default option, may still be choosing sub-optimally based on, for example, present bias. Similarly, biases in judgment and perception – such as over-optimism or a tendency to underweight low-risk events – may manifest themselves consistently across frames. Many of these failures can be attributed within our framework to the presence of "missing" frames, which affect behavior but do not vary in the data available to the researcher. Accurately identifying preferences in such contexts requires additional data or assumptions that permit the analysis to move further away from observed choice behavior.

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

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

References

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

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

Angrist, Joshua, and Ivan Fernandez-Val. 2013. "Extrapolate-ing: External Validity and Overidentification in the LATE Framework." *Advances in Economics and Econometrics*, , ed. Daron Acemoglu, Manuel Arellano and Eddie Dekel. Cambridge University Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 2:

The Analysis of Survey Data with Framing Effects[1]

by Jacob Goldin (Stanford Law School) and Daniel Reck (University of Michigan)

Abstract

A well-known difficulty in survey research is that respondents' answers to questions can depend on arbitrary features of a survey's design, such as the wording of questions or the ordering of answers. We develop an empirical framework for analyzing survey data characterized by such *framing effects*. We show that the conventional approach to analyzing data with framing effects – randomizing across frames and pooling the data for estimation – does not generally identify a useful parameter. We consider a range of alternatives, and describe the conditions under which each yields a parameter of interest. One approach attempts to isolate the answers of respondents that are unaffected by framing. A second approach attempts to recover information on the distribution of "true answers" in the population – when true answers exist – using a technique analogous to post-survey reweighting for non-response bias.

1 Introduction

A well-known difficulty in survey research is that how survey-takers answer a question may depend on seemingly arbitrary details about how the question is asked. For example, respondents may answer differently depending on the order of questions (Moore, 2002; Deaton, 2012), the order in which answers are listed (Holbrook et al., 2007), the grouping of responses

[1] For valuable comments, we thank Angus Deaton, Edward Freeland, Michael Gideon, Allyson Holbrook, Bo Honore, David Lee, Yair Listokin, Charles Manski, Alex Rees-Jones, Joel Slemrod, Justin Wolfers, and participants in the Princeton Industrial Relations Lunch and the Michigan Summer Seminar. All errors are our own.

into categories (Schwarz, 1990), or any number of minor variations in the manner in which questions or responses are worded (Schuman and Presser, 1981; Krosnick and Fabrigar, Forthcoming; Chong and Druckman, 2007). Such *framing effects* arise in many contexts; large literatures in psychology, political science, communications, and marketing are devoted to documenting and explaining their presence.

Despite the attention paid to framing effects in recent decades, the range of practical solutions available to survey researchers remains limited. The conventional wisdom is that when framing effects are unavoidable, researchers should *balance* them by randomly assigning an equal number of respondents to each version of the survey questionnaire and then analyze the pooled data.[2] Some researchers acknowledge problems with the conventional balancing approach but note the lack of better alternatives.[3]

In this paper we develop a simple theoretical framework for the analysis of survey data characterized by framing effects.[4] The framework highlights important shortcomings with the conventional balancing approach. In its place, we develop a range of practical alternatives, which we argue are more likely to shed light on substantive questions of interest.

Our formal analysis focuses on binary-response questions in which an arbitrary feature of the survey – the *frame* – affects the responses of a subset of survey-takers. We assume that each individual respondent is observed answering a given survey question only once, under one of two possible frames. We label respondents as *consistent* if they would select the same answer under both frames and *inconsistent* if their answer would vary depending on the frame to which they are assigned.[5] The primitive parameters of interest for discrete

[2] The following statements are typical of the literature: "Randomization ... does not reduce the impact of context at the level of individual respondents. It simply ensures that these influences result in random noise rather than systematic bias in the sample as a whole." (Sudman, Bradburn and Schwarz, 1995). "Our findings suggest that survey organizations should routinely rotate the order of response choices to guard against creating bias in results." (Holbrook et al., 2007). "Acquiescence bias can be reduced by balancing scales so that the affirming response half the time is in the direction of the construct and half the time is in the opposite direction (e.g. six agree/disagree items on national pride, with the patriotic response matching three agree and three disagree responses)." (Presser et al., 2004).

[3] For example, Schwarz and Oyserman (2001) write, "[R]esearchers ... may reverse the order in which the items are listed for half of the respondents. Although this ensures that the researcher becomes aware of possible response order effects, it remains unclear what to do with the results aside from the less than satisfying solution of averaging over both sets of answers."

[4] In other work, we apply a similar framework to study the identification of decision-makers' preferences in settings characterized by inconsistent choice data (Goldin and Reck, 2015). The goal of the current project is to apply this framework to framing effects in survey research.

[5] Our approach utilizes a potential outcomes framework of the type that has been widely applied in causal inference analysis (e.g., Holland, 1986; Imbens and Angrist, 1994). In those studies, researchers utilize experimental or non-experimental data to estimate the causal effect of some treatment on an observed outcome variable. To our knowledge, we are the first to apply this approach to a setting in which the goal is not to estimate the *effect* of the treatment (i.e., the frame) on a variable but rather to recover the distribution of a variable after removing the treatment's effect.

choice survey questions are typically *response shares*, the fraction of respondents selecting a particular answer, perhaps within a given subgroup of respondents.

Using this model, we show that balancing respondents evenly across frames does not eliminate the bias induced by framing effects. Rather, the balancing approach yields a response share that is a weighted sum of (1) the response share of the consistent respondents, and (2) the fraction of respondents assigned to each frame, typically 0.5. The weights in this weighted sum correspond to the fraction of respondents who are consistent. To understand why, consider two extreme cases. First, suppose that all respondents are completely immune to framing effects, so that each answers consistently across frames. In this case, the pooled response share yields the share of respondents (consistently) selecting each answer. Second, suppose instead that *all* respondents are influenced by framing effects, so that each respondent's answer is determined by the frame to which he or she is assigned. In this case, randomly assigning respondents evenly across frames and pooling the responses yields a pooled response share equal to 0.5. Finally, when some, but not all, respondents are influenced by the frame, the pooled response share is a weighted sum of the response share for the consistent respondents and 0.5 for the inconsistent respondents. Outside of a narrow set of applications, this weighted sum is unlikely to be the parameter of interest to the researcher. In its place, we propose several alternatives and describe the conditions under which they are valid.

First, we consider identification of the consistent responses – the responses of the subset of respondents who are unaffected by the frame. Consistent responses are most likely to of primary interest to the researcher in settings where frame-varying responses indicate incoherent or unformed opinions or beliefs.[6] Under the assumption that the frame affects respondents in a uniform direction, an assumption we label *frame monotonicity*, we show that the consistent responses are point-identified. Intuitively, only consistent respondents ever select an answer that is "against the frame," for example choosing the second option when the answer order favors the first option. As such, examining the fraction of respondents who answer against the frame sheds light on the fraction of consistent respondents selecting each response.

Whether the assumption of frame monotonicity is plausible in a particular context is frequently debatable. In settings where researchers are unwilling to assume frame monotonicity, we show that the consistent preferences are only partially identified. Importantly,

[6] For example, suppose a survey asks, "In forming its laws, should a society prioritize democracy or economic growth?" and some people choose a different answer depending on the order that the answer choices are listed. A researcher might conclude that the inconsistent respondents simply lack a meaningful answer to the question being asked, and instead focus on recovering the responses of those whose answers do not depend on the question order.

however, we also show that when frame monotonicity is assumed erroneously, the technique we propose for recovering the consistent responses will be biased, but it will be less biased than the conventional balancing approach.

In other applications, even the respondents who answer inconsistently will have well-defined answers to the question being asked that the researcher hopes to recover. That is, although inconsistent respondents do not directly reveal an answer to the survey question through their responses, a well-defined answer nonetheless exists for each respondent being asked. Plainly, in some cases, such as questions about past behavior (Schwarz, 1990), true answers do exist for even respondents whose answers depend on the frame.[7]

We take no stance on when researchers should assume that a true answer exists, but we provide a set of additional results for applications in which researchers do wish to make that assumption – i.e., where their goal is to recover the distribution of true answers among the entire population of respondents. In particular, we derive a new re-weighting technique that exploits respondents' observable characteristics to recover population responses. One can understand this technique by analogy to the use of post-survey weights to adjust for survey non-response bias. In the analogy, the true answers of inconsistent respondents are viewed as missing data, which must be imputed using the responses of consistent respondents with similar characteristics. An important difference from existing methods of post-survey re-weighting is that in our framework the researcher cannot directly observe the identity of – and hence the characteristics associated with – the inconsistent respondents. We overcome this difficulty by providing techniques to recover the distribution of characteristics among the consistent and inconsistent respondents, despite the fact that members of these groups cannot be individually identified.[8] In addition to the weighting estimator, we derive worst-case bounds for the distribution of true answers in the population. We also discuss the interpretation of other intuitive modes of analysis, such as simply reporting response shares separately under various frames.

As we describe the various techniques for analyzing survey data and characterize their

[7] For example, suppose a survey asks respondents how much television they watch per week, and the wording of the answer choices affects the amount of television respondents report watching (Schwarz, 1990). If the goal of the study is to learn about the television habits of the population, the researcher will be interested in the behavior of both consistent and inconsistent respondents.

[8] Researchers may wish to understand the characteristics associated with consistent and inconsistent respondents as end in itself. For example, suppose an advocacy organization surveys people about their support for a policy and the organization discovers that the framing of the question affects the fraction of respondents expressing support for the policy. The campaign may wish to understand which types of people are most sensitive to the framing effects, which may indicate that the respondent does not yet have a fully-formed opinion of the options. Our results allow researchers to estimate the distribution of observable characteristics for the consistent and inconsistent respondents, which can then be used to predict whether or not a particular respondent is likely to be sensitive to a frame.

interpretation, we illustrate our proposed techniques with data from a mix of classic and recent studies characterized by survey framing effects. Doing so highlights that our proposed techniques can provide novel insights, using only the type of data that is routinely collected.

2 Model and Notation

We presume to have data on a sample of individuals, denoted by i. Each individual answers a binary question, with their response denoted by $y_i \in \{0, 1\}$. Many survey questions have this binary form. For example, y_i might capture whether an individual agrees with a particular statement or supports a particular candidate for public office.

Each individual answers the question once, under one of the two possible *frames*. As described in the introduction, the frame denotes the manner in which the question is asked, such as whether it is worded positively or negatively, or the order in which the answer choices are listed. Let d_i denote the frame to which individual i is assigned, $d_i \in \{d_0, d_1\}$. Let Y_0 denote the mean response among individuals observed under frame d_0, $Y_0 = E[y_i|d_i = d_0]$, and let $Y_1 = E[y_i|d_i = d_1]$. We assume these response shares are directly observable to the researcher, setting aside issues of finite sample size. We assume that framing effects are observed, so $Y_0 \neq Y_1$, and without loss of generality, we assume $Y_1 > Y_0$.

Let $y_i(d)$ denote how i would answer if asked the survey question under frame d. That is, $y_i(d) = y_i$ when $d_i = d$. Of particular interest for our analysis will be respondents who are and are not affected by survey framing. Let $c_i \in \{0, 1\}$ indicate whether an individual's response would be consistent across frames: $c_i \equiv 1\{y_i(d_0) = y_i(d_1)\}$. The fraction of respondents affected by the frame is given by $p(c_i = 0) = 1 - E[c_i]$. We will at times refer to the *consistent response share*, i.e., the response share among those individuals who are unaffected by the frame: $\rho_c = E[y_i|c_i = 1]$. Note that by the definition of c_i, ρ_c does not depend the frame assigned to any individual. Also note that if individual respondents were observed under multiple frames, the researcher could directly observe c_i for each individual and easily compute ρ_c.

Finally, we will assume throughout this paper that the frame under which an individual answers the question is unrelated to how the individual would respond under either frame.

Assumption 1: Unconfoundedness. Frame assignment d_i is independent of the vector $(y_i(d_0),\, y_i(d_1))$.

This assumption is satisfied, for example, if respondents are randomly assigned across frames.

Examples

The above framework accommodates a number of types of survey questions and framing effects. We list several here, and briefly discuss their differences. Some of these differences become material for the proposed methods for analyzing survey data described below.

- Answer order. A large literature suggests that polls about candidate preferences can be influenced by the order in which candidates are listed. The data in this example come from a Gallup telephone survey conducted on October 15 and October 16, 2012, and relate to the 2012 presidential election (Gallup Organization, 2012). For example, a survey from October 15-16, 2012 on the Presidential election asked respondents:[9]

  Suppose that the presidential election were being held today, and it included Barack Obama and Joe Biden as the Democratic candidates and Mitt Romney and Paul Ryan as the Republican candidates. Would you vote for Mitt Romney and Paul Ryan, the Republicans [Barack Obama. and Joe Biden, the Democrats] or Barack Obama and Joe Biden [Mitt Romney and Paul Ryan, the Republicans]?

  When the Republican candidates were listed second, the average level of support for Romney was 55%. When the Democrats were listed second, Romney's support fell to 48%.[10]

- Question wording (acquiescence bias). Schuman and Presser (1981) document the presence of numerous survey framing effects, including one that they refer to as "acquiescence bias" – the tendency of respondents to agree with the question being asked, regardless of the content. For example, in one study individuals were randomly assigned one of two versions of the same question. The fraction agreeing with the stated proposition, the estimates of Y_0 and Y_1 in our notation, is provided in brackets.

  Individuals are more to blame than social conditions for crime and lawlessness in this country [60%]

  Social conditions are more to blame than individuals for crime and lawlessness in this country [57%]

[9] Individuals who indicated they were unsure were asked "As of today, do you lean more toward...", followed by the candidates, presented in the same order as in the first question. Respondents indicating a preference in the follow-up question are coded according to their stated preference in this question. The numbers we report discard observations where the individual indicated a preference for a third candidate, or did not know or refused the follow-up question.

[10] The data thus indicate a strong *recency* effect. Recency effects are commonly observed on oral surveys, see Holbrook et al. (2007).

- Question wording (gain/loss framing). Tversky and Kahneman (1981) asked experimental participants about their willingness to accept risky policies that have the potential to save large numbers of lives. Respondents were asked two versions of a question after being randomly divided into the gain frame and the loss frame:

  Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:

  If Program A is adopted, 200 people will be saved. If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. Which of the two programs would you favor? [Program A: 72%, Program B: 28%]

  If Program C is adopted 400 people will die. If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.Which of the two programs would you favor? [Program C: 22%, Program D: 78%]

  Note that Programs A and C are identical, as are Programs B and D.

- Question order. Moore (2002) documents question-order effects in a 1997 Gallup survey. Respondents were asked the following:

  Do you generally think [Bill Clinton / Al Gore] is honest and trustworthy?"

  Respondents' answers varied depending on which politician they were asked about first. When the Clinton question was asked first, 50% of respondents reported thinking that Clinton was trustworthy, whereas 57% reported thinking so when they were first asked about Al Gore. Conversely, 68% reported believing Al Gore to be trustworthy when the Gore question was first, but only 60% did so when the Gore question was second.

These examples illustrate that the potential for framing effects is largely unavoidable. The logical content of the question is the same under different frames.[11] Questions and answers

[11] The acquiescence bias example could be an exception to this, as individuals believe that the two causes are equally to blame for crime and lawlessness might respond in the negative to both versions of the question. However, this would affect answers in the opposite direction from acquiesence bias. Similarly, in the Tversky and Kahneman (1981) gain/loss question, perhaps listing both the number of deaths and the number of lives saved in the same question would elicit information not subject to framing effects. Even then, though, one would need to decide which number to list first, lives saved or deaths.

must be provided to respondents in some order, and they must be worded in some way. Thus, while using questions that elicit useful information while being unaffected by framing may be the ideal scenario for survey researchers, determining a method of dealing with framing effects will oftentimes be inescapable.

3 The Conventional Pooling Approach

As described in the introduction, the conventional approach to analyzing survey data that exhibit framing effects is to randomize respondents across frames and then to proceed using the pooled data. For example, political polling surveys, recognizing the potential for response-order effects, typically rotate which major candidate is listed first and which is listed second; however, after this initial randomization, candidate order plays no further role in the analysis. This section shows that, contrary to the conventional wisdom, simply randomizing respondents across frames does not eliminate the difficulty associated with framing effects.

Suppose that respondents are randomly assigned to each frame, with equal probability of being selected for each. This is the conventional approach, and we refer to it as the pooled mean because it entails pooling the data collected under each frame. The pooled response share, Y_p, is defined as:[12]

$$Y_p \equiv \frac{Y_1 + Y_0}{2}$$

The following proposition characterizes the pooled response share.

Proposition 1: The Pooled Response Share.

The pooled response share identifies the following weighted sum: $Y_p = E[c_i]\rho_c + (1 - E[c_i])\left(\frac{1}{2}\right)$.

Proof By the law of iterated expectations,

$$Y_0 = E[y_i(d_0)|c_i = 1, d = d_0]\, p(c_i = 1|d = d_0) + E[y_i(d_0)|c_i = 0, d = d_0]\, p(c_i = 0|d = d_0).$$

[12] It is straightforward to relax the assumption of equal probabilities of assignment to each frame, instead allowing some generic fraction of respondents to be assigned to each frame, i.e. $Y_p = \omega Y_0 + (1 - \omega)Y_1$. We know of no survey implementing unequal randomization in practice, so we shall proceed with the simpler case here, setting $\omega = 0.5$.

Applying Unconfoundedness yields

$$Y_0 = E[y_i(d_0)|c_i = 1]\, p(c_i = 1) + E[y_i(d_0)|c_i = 0]\, p(c_i = 0).$$

Because $y_i(d_0) = y_i(d_1)$ when $c_i = 1$, we use the definition of ρ_c to write

$$Y_0 = \rho_c\, p(c_i = 1) + E[y_i(d_0)|c_i = 0]\, p(c_i = 0). \tag{1}$$

Similarly, one can show that

$$Y_1 = \rho_c\, p(c_i = 1) + E[y_i(d_1)|c_i = 0]\, p(c_i = 0). \tag{2}$$

Finally, note that when $c_i = 0$, $(y_i(d_0), y_i(d_1)) \in \{(0,1),(1,0)\}$, and thus $y_i(d_0) + y_i(d_1) = 1$. Therefore

$$E[y_i(d_0)|c_i = 0] + E[y_i(d_1)|c_i = 0] = 1. \tag{3}$$

Substituting (1) and (2) into the definition of Y_p and applying (3) yields the desired result.■

Discussion When survey respondents are evenly divided between the two frames, Proposition 1 shows that the pooled response share is a weighted average of the consistent response share and 0.5. To understand why intuitively, consider two extreme examples. First suppose that all respondents are consistent, so $E[c_i] = 1$. Then the pooled response share would estimate the share of the population that (consistently) selects $y_i = 1$, $Y_p = E[y_i|c_i = 1] = E[y_i]$. Next, consider the opposite extreme, where all respondents are sensitive to the frame, so $E[c_i] = 0$. In this case, assigning half of the respondents to d_0 and half to d_1 would result in half of the respondents selecting each answer, so we would obtain $Y_p = 0.5$. Finally, when there are both consistent and inconsistent respondents in the population, the pooled response share will simply be the weighted average of these two extremes, where the weights depend on the fraction of respondents who are consistent.

We can think of few practical settings in which a researcher would want to identify the weighted average associated with the pooled response share. In particular, the pooled response share will depend on factors related to the design of the survey, in particular the fraction of respondents assigned to each frame and the size of the framing effect. For example, a pooled response share close to 0.5 could indicate either that many of the respondents are influenced by the frame or that an equal number of respondents would consistently select each of the two available answers.[13] In the next sections, we propose an alternative approach for

[13] One narow set of cases in which the pooled response share does identify a useful parameter is when the goal of a survey is to predict a future choice that is characterized by the same framing effect as the

dealing with framing effects in survey analysis that attempts to shed light on the consistent response share, ρ_c.

4 Identifying Consistent Responses

This section considers how researchers might attempt to recover information about the consistent response share, ρ_c. In many cases, ρ_c is likely to capture information of interest to the researcher. In particular, for subjective survey questions designed to measure the respondent's attitudes or beliefs, inconsistent respondents may lack a well-formed opinion on the question being asked. In such cases, researchers may wish to isolate the consistent response share from those whose responses depend on the frame. Another advantage of ρ_c is that it provides information about the quantity the survey was designed to measure in a way that is not mechanically related to the frame in which respondents are observed. For example, when a survey of likely voters yields different response shares depending on which candidate is listed first, the consistent response share is equal to the fraction planning to vote for each candidate from the sub-population of respondents who select the same candidate regardless of the order in which the options are listed.

Identifying ρ_c would be trivial if the researcher could observe individual respondents under both frames; the inconsistent respondents could be individually identified and their responses discarded.[14] In contrast, when each respondent is only observed under one frame, the degree to which ρ_c can be identified depends on whether the frame in question affects all respondents in a uniform direction. In particular, consider the following assumption:

Assumption 2: Frame monotonicity. For every individual i, $y_i(d_1) \geq y_i(d_0)$.

For example, if one observes that a higher fraction of respondents answers "Yes" in response to version A of a question as opposed to version B, frame-monotonicity implies that there are no respondents who answer "No" to version A but "Yes" to version B. Clearly this assumption will be more plausible in some settings, such as default effects (where it is difficult to imagine that a non-trivial number of respondents always select whichever response is not marked as the default), and less likely to hold in other settings, such as response-order effects (where one might imagine that some respondents always select the first option while

survey itself. For example, suppose one finds that election polling data exhibits candidate order effects. If actual voting behavior will exhibit the same candidate order effects as the survey, and if the order of candidate names on the actual ballot will itself be randomized, then the pooled response share will estimate the weighted average associated with the actual election results.

[14] Even if obtaining this information were possible, in many contexts it may be that surveying a respondent under one frame would bias their subsequent answers under alternate frames (LeBoeuf and Shafir, 2003).

others always select the most recent option). The following proposition describes the role of frame monotonicity in identifying ρ_c.

Proposition 2: The Consistent Response Share

Let $Y_c = \frac{Y_0}{Y_0+1-Y_1}$ and let $\rho_c = E[y_i|c_i = 1]$.

2.1 Under frame monotonicity, $E[c_i] = Y_0 + 1 - Y_1$ and $Y_c = \rho_c$.

2.2 Without frame monotonicity, $|1 - Y_1 - Y_0| \leq E[c_i] \leq Y_0 + 1 - Y_1$, and $Y_c \geq \frac{1}{2} \implies \rho_c \in [Y_c, 1]$ while $Y_c \leq \frac{1}{2} \implies \rho_c \in [0, Y_c]$.

Proof of 2.1 For all individuals, $(y_i(d_0), y_i(d_1)) \in \{(1,1), (0,0), (1,0), (0,1)\}$. Frame monotonicity rules out the last possibility, that $(y_i(d_0), y_i(d_1)) = (1,0)$. Thus when $y_i(d_0) = 1$, we know that $(y_i(d_0), y_i(d_1)) = (1,1)$ and when $y_i(d_1) = 0$, $(y_i(d_0), y_i(d_1)) = (0,0)$. Thus a respondent choosing *against the frame* will be choosing consistently, and the share of respondents choosing the against-the-frame answer will be the fraction of respondents who are consistent *and* select that answer. It follows that $E[c_i] = Y_1 + 1 - Y_0$, and, using the definition of conditional probability, that $Y_c = \rho_c$. ■

Proof of 2.2 Let $\alpha = p[(y_i(d_0), y_i(d_1)) = (1,0)]$. Frame monotonicity imposes $\alpha = 0$, which leads to the result in the previous proof. Otherwise, it follows from the four possibilities for $(y_i(d_0), y_i(d_1))$ that

$$Y_0 = \rho_c E[c_i] + \alpha \tag{4}$$

and also that

$$1 - Y_1 = (1 - \rho_c)E[c_i] + \alpha. \tag{5}$$

To obtain bounds for $E[c_i]$, note that by (4) and (5),

$$Y_0 + 1 - Y_1 = E[c_i] + 2\alpha \tag{6}$$

The uppwer bound for $E[c_i]$ is immediately implied by (6) and the fact that $\alpha \geq 0$. The lower bounds for $E[c_i]$ are implied by the fact that $\alpha \leq Y_0$, and $\alpha \leq 1 - Y_1$ from (4) and (5). Combining these with (6) we obtain $E[c_i] \geq 1 - Y_1 - Y_0$ and $E[c_i] \geq Y_0 + Y_1 - 1$. Combining these yields desired the given lower bound.[15]

[15] Note that which of these two constraints binds depends on whether $Y_0 \gtrless 1 - Y_1$, i.e. whether $Y_c \gtrless \frac{1}{2}$, so the two cases match the two cases of the second part of the result.

To obtain bounds for ρ_c, we combine 4 and 5 in the definition of Y_c:

$$Y_c = \frac{\rho_c E[c_i] + \alpha}{E[c_i] + 2\alpha}. \tag{7}$$

Subtract $\frac{1}{2}$ from both sides of equation (7) and simplify to obtain

$$Y_c - \frac{1}{2} = \left(\rho_c - \frac{1}{2}\right) \frac{E[c_i]}{E[c_i] + 2\alpha}. \tag{8}$$

Noting that $\frac{E[c_i]}{E[c_i]+2\alpha} \in [0, 1]$, equation 8 implies the desired bounds for ρ_c. ■

Proposition 2.1 establishes that the consistent response share is point identified under frame monotonicity. Intuitively, frame monotonicity implies that only consistent individuals choose against the frame so that the only respondents selecting $y_i = 1$ under d_0 are those who would also select $y_i = 1$ under d_1. Consequently, we know that Y_0 respondents were consistent *and* chose option 1. Similar logic implies that $1 - Y_1$ of respondents were consistent *and* chose option 0. Scaling the former by the share of respondents that are consistent yields the share of consistent respondents who select $y_i = 1$, i.e., ρ_c.

Proposition 2.2 establishes that without frame monotonicity, the consistent response share is partially identified. Borrowing terminology from Imbens and Angrist (1994), we can divide the set of inconsistent respondents into two groups, those who are effected by the frame in the same direction as the majority of inconsistent respondents (the *frame-compliers*) and those who are affected by the frame in the opposite direction (the *frame-defiers*). That is, a frame-defier selects $y_i = 1$ if and only if the frame is d_0. Intuitively, the presence of frame-defiers means that some respondents who are inconsistent will be misclassified as consistent in the computation of Y_c; the misclassified group will contain the frame-defiers, plus an offsetting number of frame-compliers. The frame-defiers will respond with $y_i = 1$ if and only if they are assigned to d_0; the frame-compliers if and only if they are assigned to d_1. Since there are an equal number of frame-defiers and frame-compliers in the group of misclassified respondents, the group, on average, answers $y_i = 1$ half of the time under both frames. Because the misclassified group's behavior – in the aggregate – is the same under each frame, the group as a whole appears to be consistent (even though, in reality, each individual member of the group is actually inconsistent). And because the behavior of this group is attributed to the consistent respondents, the misclassification will bias Y_c upwards from ρ_c when ρ_c is in fact below 0.5 and downwards when the opposite is true. Thus the failure of frame-monotonicity implies that Y_c is biased away from ρ_c towards 0.5.

The following corollary highlights an important practical implication of this result for survey researchers. Recall that Y_p denotes the pooled response share obtained from randomly assigning half of the respondents to each frame, $Y_p = \frac{1}{2}(Y_0 + Y_1)$.

Corollary to Proposition 2.2: Bias in Y_c versus Y_p

Without frame monotonicity, either 1) $Y_p < Y_c \leq \rho_c$, 2) $Y_p > Y_c \geq \rho_c$, or 3) $Y_p = Y_c = \rho_c = \frac{1}{2}$..

Proof Using the expression for Y_c in (7) and the expression for Y_p in Proposition 1, it follows that

$$Y_c - Y_p = \frac{(\rho_c - \frac{1}{2})p(c_i = 1)[1 - p(c_i = 1) - 2\alpha]}{p(c_i = 1) + 2\alpha} \tag{9}$$

One can derive that $\alpha < \frac{1}{2}(1 - E[c_i])$ from the definitions of Y_0 and Y_1,[16] so $1 - p(c_i = 1) - 2\alpha > 0$. When $Y_c \geq \frac{1}{2}$, we know that $\rho_c \geq Y_c$ from 2.2 and $Y_c \geq Y_p$ from equation (9). Once we know that $\rho_c \geq Y_c \geq Y_p$, the desired result is established for this case. A similar line of reasoning establishes the result for the case when $Y_c \leq \frac{1}{2}$. In the knife-edge case where $Y_c = \frac{1}{2}$, we know from 2.2 that $\rho_c = \frac{1}{2}$, and then Proposition 1 implies that $Y_p = \frac{1}{2}$. ■

Although neither Y_c nor Y_p will generally be equal to the consistent response share (ρ_c) in the absence of frame monotonicity, the corollary states that the bias in the former will be smaller than the bias in the latter. Intuitively, the misclassified group causing the bias in Y_c is necessarily smaller than the group of inconsistent respondents causing the bias in Y_p. Consequently, a researcher who finds $Y_c < Y_p$ may conclude $\rho_c \leq Y_c < Y_p$; or similarly, if $Y_p < Y_c$, the researcher may conclude $Y_p < Y_c \leq \rho_c$. Finally, note that when individuals are observed under both frames, ρ_c is easily identified and the bias from using Y_c as an estimator for ρ_c, which generally depends on the nuisance parameter α, can be directly identified.

5 Which Respondents are Consistent?

Thus far our focus has been on eliminating the bias caused by frame-sensitive respondents in order to learn about the distribution of the survey variable. In some applications, the consistency (or lack thereof) of the respondents will itself be an issue of primary interest to the researcher. For example, a political pollster working for an electoral campaign may be quite interested in likely voters whose stated preferences between two candidates depend on the order in which the candidates are listed, or upon which features of the candidates

[16] Specifically, we assumed $Y_1 > Y_0$. Plugging the formulas for Y_1 and Y_0 in the proof of 2.1 into this inequality gives the result. An interesting special case occurs in the limiting case where $\alpha = \frac{1}{2}(1 - E[c_i])$, which also implies $Y_0 = Y_1$. In this case there are just as many frame defiers as frame compliers, and the two cancel each other out in the observations of Y_1 and Y_0. Then, although $Y_1 = Y_0$ and the aggregtae data show no inconsistency, ρ_c and $E[c_i]$ can take *any value* from zero to one. We rule this out when we assume that framing effects are observed, i.e. $Y_1 \neq Y_0$, but it is important to note the implication of this reasoning: framing effects can be present even when they are unobserved.

are made salient in the survey. Understanding who such voters are could be quite useful for better targeting political messages. Similarly, both advertisers and advocacy groups may wish to know which types of consumers are most responsive to particular types of messaging. Understanding which respondents are consistent is also the first step in the method for recovering the distribution of true answers we describe in the next section.

Without exposing a single respondent to multiple frames, it is impossible to identify precisely which individuals are consistent and which are not. Under the assumption of frame monotonicity, however, researchers may estimate the aggregate distribution of characteristics of the consistent and inconsistent decision-makers.[17] Formally, we suppose that individuals are endowed with observable characteristics $g \in G$; these characteristics could be discrete, continuous, or even some vector of discrete and continuous characteristics. Let $f(g)$ denote the probability distribution function over G; this distribution is directly identified up to finite sample concerns, which we ignore. We will denote the response shares under the two frames, conditional on observable characteristics, by $Y_0(g) \equiv E[y_i|d_i = d_0; g_i = g]$ and $Y_1(g) \equiv E[y_i|d_i = d_1; g_i = g]$.

Proposition 3: Relating Consistency and Observable Characteristics[18] Under frame monotonicity,

3.1 The distribution of g among the consistent respondents is given by $p(g_i = g \,|\, c_i = 1) = \frac{Y_0(g)+1-Y_1(g)}{Y_0+1-Y_1} f(g)$.

3.2 The distribution of g among the inconsistent respondents is given by $p(g_i = g \,|\, c_i = 1) = \frac{Y_1(g)-Y_0(g)}{Y_1-Y_0} f(g)$.

Proof First note that by the same logic as in Proposition 2.1, unconfoundedness and frame monotonicity imply that

$$p(c_i = 1|g_i = g) = Y_0(g) + 1 - Y_1(g). \tag{10}$$

Bayes rule implies that

$$p(g_i = g|c_i = 1) = \frac{p(c_i = 1|g_i = g)f(g)}{p(c_i = 1)}. \tag{11}$$

[17] The result in this section is analogous to that of Abadie (2003), who, in the context of instrumental variable estimation, shows how to identify the aggregate characteristics of the "compliers," despite the fact that individual members of that population cannot be identified.

[18] Throughout the paper we have been assuming unconfoundedness, that $d_i \perp (y_i(d_0), y_i(d_1))$. Technically, Propositions 4 and 5 can be proven under the weaker assumption of conditional unconfoundedness, only requiring that for any g, $d_i \perp (y_i(d_0), y_i(d_1))|g_i = g$. However, we do not emphasize this fact, as our focus is on settings where respondents are randomized across frames by design, so unconfoundedness is guaranteed.

Substituting (10) and the expression for $p(c_i = 1)$ obtained in Proposition 2.1 yields 3.1. The proof of 3.2 is analogous. ■

Intuitively, the distribution of g in the consistent population will be the distribution of g in the population adjusted for the relative propensity of individuals with characteristic g to be consistent relative to the rest of the population (see equation 11). Under monotonicity, researchers can recover the fraction of individuals with a given characteristic that are consistent, and the fraction of the population that are consistent using Proposition 2. As a result, we can directly recover the distribution of g in the consistent, and inconsistent populations.

6 The Distribution of True Answers

In some cases, researchers will be interested in the distribution of true answers in the full population. For example, framing effects have been documented in surveys that solicit self-reported behavioral frequency data, such as the frequency with which respondents watch television or engage in risky health behaviors (see Schwarz and Oyserman (2001) for a number of examples). Another example is exit polling data for an election, where answer order may bias which candidate a voter recalls voting for (especially in more obscure contests such as judicial elections), but pollsters nevertheless want to know how the population voted. In such cases, there exists a well-defined "true" answer to the survey variable even for those respondents whose reported answers vary by frame; their susceptibility to framing effects simply prevents that true value from being revealed to the researcher.[19] More generally, questions involving past behavior will typically have true answers, so long as they are logically unambiguous. In other settings, such as quesitons about personal opinions, future behavior, or hypothetical decisions, whether a true answer exists for a survey question is debatable. We take no stand on whether researchers should assume that true answers exist in such situations. Often, in the authors' experience, economists are willing to assume that true answers exist, while other social scientists, especially psychologists, are not.[20] We leave to applied researchers the question of whether they should assume that true answers exist in

[19] Understood this way, the problem parallels the one addressed in the "Contaminated Outcomes" literature, in that the goal is to learn about a parameter of interest in settings where the observed data contain some fraction of erroneous observations (Joel L. Horowitz and Charles F. Manski, n.d.) (here, the erroneous observation occurs when a survey-taker responds according to the frame).

[20] For an interesting discussion of this issue, see Fischhoff (1991).

any particular context, and consider, in this section, what researchers who do make the assumption can learn about the distribution of true answers.

Assumption 3: Existence of True Answers. For every individual i, there exists a true answer, y_i^*.

Researchers wishing to identify the distribution of true answers will need to make an assumption about how respondents choices relate to their true answers. We will focus primarily on the case where consistent answers reveal true responses:

Assumption 4: Consistent Answers Are True Answers For every individual i, $c_i = 1$ implies $y_i = y_i^*$.

In words, Assumption 4 requires that those individuals providing the same answer across frames are reporting their true answers. The assumption represents a weakening of the typical baseline assumption made in survey research, that survey-takers responses represent the true answer to the question being asked. [21]

We begin by deriving worst-case bounds for the distribution of true answers (as in Manski, 1989).

Proposition 4: Partially Identifying the Distribution of True Answers Under Assumptions 1, 3, and 4,

4.1 Frame monotonicity implies $Y_0 \leq E[y_i^*] \leq Y_1$.

4.2 Without frame monotonicity, $Y_c \geq \frac{1}{2} \implies E[y_i^*] \in [Y_0 - (1 - Y_1), 1]$ and $Y_c \leq \frac{1}{2} \implies E[y_i^*] \in [0, Y_0 + Y_1]$.

Proof of 4.1 The law of iterated expectations implies

$$E[y_i^*] = E[y_i^*|c_i = 1]p(c_i = 1) + E[y_i^*|c_i = 0]p(c_i = 0) \tag{12}$$

[21] This assumption is easily relaxed to accomodate multiple framing effects for a single question, such as those in which both answer order and question wording cause framing effects. It will fail when there are unobserved framing effects; intuitively, when unobserved factors are biasing respondents' answers away from their true answers, recovering the distribution of true answers will be difficult. Relatedly, we will primarily focus on situations where researchers assume that consistent respondents report true answers, but one can imagine some alternatives. One alternative would be to suppose that under some ideal frame d^*, all respondents provide true answers, i.e. $y_i = y_i^*$ when $d_i = d^*$. Identifying the distribution of true answers in this case requires simply observing a representative sample under the frame d^*. Another alternative would be to relax assumption 4 and assume that framing effects unavoidably cause deviations in reported answers from true answers, but that the deviation occurs in a known direction, e.g. for some frame d, $d_i = d$ implies $y_i^* \leq y_i$, or that conssitent answers are subject to an unobserved bias so $c_i = 1$ implies $y_i^* \leq y_i$. One can naturally imagine a partial identification strategy for the learning about the distribution of true answers in such settings.

Under Assumption 4, $\rho_c = E[y_i^*|c_i = 1]$. Applying what we know about ρ_c and $p(c_i = 1)$ from Proposition 2.1 and simplifying yields $E[y_i^*] = Y_0 + E[y_i^*|c_i = 0](Y_1 - Y_0)$. The bounds then follow from the fact that $0 \leq E[y_i^*|c_i = 0] \leq 1$. ■

Proof of 4.2 We first consider the lower bound. Because $E[y_i^*|c_i = 0] \geq 0$, equation (12) implies that$E[y_i^*] \geq \rho_c p(c_i = 1)$. This is equivalent to $E[y_i^*] \geq Y_0 - \alpha$ by equation (4). Finding a suitable maximum value for α will yield a lower bound. Re-arranging equation (5) yields $1 - \rho_c = \frac{1-Y_1-\alpha}{E_c}$, and $1 - \rho_c \leq 1$, so $\alpha \leq 1 - Y_1$. Substituting our maximum value for α gives the desired lower bound. Note that this lower bound is non-trivial only when $Y_c \geq \frac{1}{2}$; otherwise $Y_0 - (1 - Y_1) \leq 0$.

We next derive an upper bound using similar reasoning. Because $E[y_i^*|c_i = 0] \leq 1$, equation (12) implies that$E[y_i^*] \leq \rho_c p(c_i = 1) + 1 - p(c_i = 1)$. This is equivalent to $E[y_i^*] \leq Y_1 + \alpha$ by equation(5). Once again we require a suitable maximum value for α. In this case the non-trivial bound comes from equation (4), which implies $\rho_c = \frac{Y_0-\alpha}{E[c_i]} \geq 0$, so $\alpha \leq Y_0$. Note that this upper bound is non-trivial only when $Y_c \leq \frac{1}{2}$; otherwise $Y_0 + Y_1 \geq 1$. ■

Proposition 4 reveals an interesting fact about the recovery of population preferences from survey data. A conservative researcher wishing to analyze data in a robust fashion despite framing effects might simply report the distribution of answers under alternative frames. In our notation, this means reporting Y_0 and Y_1 separately. It might be tempting to claim that the distribution of true answers, $E[y_i^*]$, lies somewhere between Y_0 and Y_1. Proposition 4 reveals, perhaps surprisingly, that such an interpretation, conservative as it might seem, is only valid when the researcher assumes not only that consistent answers are true answers, *but also assumes frame monotonicity.* Intuitively, when there are a large number of individuals choosing against the frame, the framing effect can appear much smaller than the true size of the survey response inconsistency. All hope is not lost, however, even in such cases. Proposition 4.2 shows that without frame monotonicity, the distribution of answers under the two frames still provides some information about the true distribution. These one-directional bounds will be more informative when Y_0 and $1 - Y_1$ are very different, which tends to occur when the observed framing effect $(Y_1 - Y_0)$ is small.

Some researchers may wish to go further, making additional assumptions to point-identify the distribution of true answers in the population. Under the assumption that consistent answers are true answers, Proposition 2.1 point identifies the distribution of true answers among consistent choosers. The problem of recovering the full population distribution then parallels the well-studied question of inference under sample selection.[22]

[22] See Manski (2003). An important difference from the standard setting is that researchers can typically observe whether individual respondents are included in the sample or not (e.g., whether they respond to the survey); in contrast, the researcher cannot directly observe whether an individual respondent is consistent.

We will develop an approach to this problem using techniques analogous to the stratification weights commonly used in survey analysis in the presence of missing data. The key assumption is that conditional on observable characteristics, respondents whose answers are sensitive to framing and those reporting their true answers across frames have the same distribution of true answers:[23]

Assumption 5: Conditional Consistency Independence

$$\forall g \in G,\ cov(c_i, y_i^*|g_i = g) = 0$$

where g are observable characteristics as in the previous section.

Proposition 5: The Distribution of True Answers Under Conditional Consistency Independence Let $Y_c(g) = \frac{Y_0(g)}{Y_0(g)+1-Y_1(g)}$. Under Assumptions 1-4 and Conditional Consistency Independence, $E[y_i^*] = E_g[Y_c(g)]$.

Proof By the law of iterated expectations, $E[y_i^*] = E_g[E[y_i^*|g_i = g]]$. Proposition 2.1 implies that for any g, $Y_c(g) = E[y_i^*|g_i = g, c_i = 1]$. Conditional consistency independence then implies that $Y_c(g) = E[y_i^*|g_i = g]$.[24] Substituting this into the first expression yields the result. ■

To implement the approach suggested by Proposition 5 in practice, one would first calculate $Y_c(g)$ for each observable characteristic (or as a continuous function of observable characteristics), and then take a weighted sum of $Y_c(g)$, using weights equal to the distribution of g in the full population (denoted $f(g)$ in the previous section). The technique is analogous to post-stratification weights frequently employed in survey analysis (Holt and Smith, 1979), which correct for the fact that some respondents are more likely to select into the sample than others. In our setting, consistency weights correct for the fact that some respondents are more likely to select into the consistent subgroup of the population – the subgroup whose true responses to the survey can be inferred.

Carrying the analogy further, for conventional post-survey non-response weights to eliminate selection bias, it must be the case that respondents' propensity to participate in the survey is uncorrelated with unobservable correlates of the variable being investigated. Our

As described below, this difference complicates the task of adapting the standard sample selection techniques to this setting.

[23] An alternative approach, analogous to instrumental variable methods, for when this assumption is not credible is described in Goldin and Reck (2015).

[24] Recall that two-valued variables are independent if they have zero covariance. One can also show formally that $E[y_i^*] = Y_c - \frac{cov(c_i, y_i^*)}{p(c_i=1)}$. Conditioning this on g and taking its expectation also leads to the desired result.

conditional consistency independence assumption guarantees exactly this; it will fail when respondents' consistency is related to the distribution of y in unobservable ways. As such, the more individual characteristics the researcher can observe that are potentially correlated with a respondent's consistency, the more confident the researcher can be that using consistency weights will recover the full population average.

7 Conclusion

In this paper we have developed a straightforward framework to study a ubiquitous problem in survey research: the sensitivity of responses to seemingly-arbitrary features of the survey's design. We showed how the conventional approach to dealing with this problem does not actually solve it, and proposed several practical alternatives. Many of these techniques focus on the distribution of answers of consistent respondents and the characteristics of consistent respondents; others attempt to go further and recover the distribution of "true answers" in the population.

Once one adopts the formal framework we propose for analyzing survey response inconsistencies, several parallels emerge to classic ideas about identification and inference. As ever, the degree to which interesting parameters can be identified from the data depend on the assumptions the researcher is willing to impose. With very little structure, one can recover information on the responses of consistent respondents – those unaffected by framing – and the characteristics of consistent individuals. Recovering true answers in the population requires assuming that true answers exist in the first place, and assuming away unobserved framing effects. In both of these cases, the sharpness of the identification depends on whether framing effects are assumed to be monotonic. Finally, assuming that consistency across frames and true answers are conditionally independent permits point identification of the distribution of true answers in the full population.

Two limitations of our work deserve consideration in future research and applications. First, we have focused on the relatively simple setting of binary survey variables with two frames; applying the approaches in settings with additional frames or answer choices requires further assumptions. It is difficult to imagine that the difficulty with randomizing across frames and pooling the data will be eliminated in a more complicated setting, but credible alternatives may be somewhat more difficult to obtain. Second, our approach is aimed at eliminating the bias induced by framing effects, but other sources of bias could still be a problem. Generalizing the approach proposed here to non-binary survey questions and to settings characterized by other types of bias – such as random choice, forgetfulness, or selection effects – are important directions for future research.

References

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

Chong, Dennis, and James Druckman. 2007. “Framing Theory.” *Annual Review of Political Science*, 10: 103–106.

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

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

Gallup Organization. 2012. “Gallup News Service Poll: Monthly Indicators Update.” Roper Center Public Opinion Archives Survey Dataset.

Goldin, Jacob, and Daniel Reck. 2015. “Preference Identification Under Inconsistent Choice.”

Holbrook, Allyson, Jon Krosnick, David Moore, and Roger Tourangeau. 2007. “Response Order Effects in Dichotomous Categorical Questions Presented Orally.” *Public Opinion Quarterly*, 71(3): 325–348.

Holland, Paul. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association*, 81(396): 945–960.

Holt, D. Tim, and T.M. Fred Smith. 1979. “Post Stratification.” *Journal of the Royal Statistical Society*, 142(1): 33–46.

Horowitz, Joel L., and Charles F. Manski. n.d.. “Identification and Robustness with Contaminated and Corrupted data.” *Econometrica*, 63.

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

Krosnick, John, and Lee R. Fabrigar. Forthcoming. *The Handbook of Questionnaire Design.* Oxford University Press.

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

Manski, Charles. 1989. “Anatomy of the Selection Problem.” *The Journal of Human Resources*, 24(3).

Manski, Charles. 2003. *Partial Identification of Probability Distributions (Springer Series in Statistics.* Springer.

Moore, David W. 2002. “Measuring New Types of Question-Order Effects: Additive and Subtractive.” *Public Opinion Quarterly*, 66(1): 80–91.

Presser, Stanley, Mick Couper, Judith Lessler, and Elizabeth Martin. 2004. *Methods for Testing and Evaluating Survey Questions.* Wiley.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context.* SAGE.

Schwarz, Norbert. 1990. “Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Construction.” *Review of Personality and Social Psychology*, 11: 98–119.

Schwarz, Norbert, and Daphna Oyserman. 2001. “Asking Questions About Behavior: Cognition, Communication, and Questionnaire Construction.” *American Journal of Evaluation*, 22(2): 127–160.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1995. *Thinking about Answers : The application of Cognitive Processes to Survey Methodology.* Jossey-Bass.

Tversky, Amos, and Daniel Kahneman. 1981. “The Framing of Decisions and the Psychology of Choice.” *Science*, 211(4481): 453–458.

Chapter 3:

Taxes and Mistakes: What's in a Sufficient Statistic? [1]

by Daniel Reck (University of Michigan)

Abstract

What determines the efficiency cost of taxation in the presence of optimization errors? Employing recent results quantifying efficiency cost when consumers are subject to biases, this paper shows how budget adjustment rules, debiasing, and the nature of tax perception affect efficiency cost. Budget adjustment rules govern how taxpayers meet their budget constraint in spite of misperceptions. Complete consideration of budget adjustment rules shows why simply detecting misperception of taxes is insufficient for welfare. Furthermore, if consumers "debias" at sufficiently high stakes, policymakers' attempts to exploit biases to reduce inefficiency—like switching from high- to low-salience taxes—can actually *increase* inefficiency. Any cognitive costs of debiasing exacerbate this "curse of debiasing." I demonstrate that the model can be applied to even complicated misperceptions using the example of "ironing," which leads to a clarification of prior welfare analysis of ironing.

1 Introduction

A vitally important idea from the normative theory of taxation is that taxes distort relative prices, causing inefficiency (Harberger, 1964). Yet recent evidence suggests that individuals make mistakes by ignoring or failing to understand tax incentives. For example, some individuals ignore less salient taxes,[2] react differently to taxes that are framed

[1] The author thanks David Albouy, Johannes Becker, Sebastien Bradley, Michael Gideon, Jacob Goldin, Morris Hamilton, James Hines, Miles Kimball, Yusufcan Masatlioglu, Nathan Seegert, Joel Slemrod, Justin Wolfers, seminar participants at the Michigan Public Finance Lunch Seminar, the International Institute of Public Finance Annual Conference, and the National Tax Association Annual Conference for helpful discussion and comments. Any errors are my own.

[2] Here, salience refers to the visibility of a tax at the time when a taxpayer makes a decision. For evidence of salience effects, see Chetty, Looney and Kroft (2009); Finkelstein (2009); Cabral and Hoxby (2012); Feldman and Ruffle (2015); Goldin and Homonoff (2013). Chetty et al. (2014) provide evidence of inattention to savings incentives. Many more recent studies finding surprisingly small responses to plausibly misperceived policies offer salience as an explanation, see for example Saez (2010).

differently,[3] confuse average and marginal tax rates,[4] or fail to understand the rules for tax avoidance and evasion.[5] Understanding the normative implications of mis-perception and mistakes has proven difficult, because the traditional approach to welfare analysis in economics relies on the principle that preferences are revealed by choices. When people make mistakes, some choices do not reveal preferences. So how can we understand welfare?

A promising answer to that question has been proposed by Chetty, Looney and Kroft (2009) (henceforth CLK).[6] These authors use a sufficient-statistics approach to tax salience to understand how mistakes can affect the excess burden of taxation. The critical assumption is that individuals always respond optimally to price changes in absence of taxation, but that the change in demand from a tax rate change may be suboptimal. Two sufficient statistics emerge: responses to price changes summarize preferences, while responses to tax changes summarize behavior. Goldin (2012) explores how these statistics may be used to decide how much tax the government should raise via low-salience taxes. A central issue in this work is the finding that low salience taxes can cause less inefficiency than salient ones. This finding has led to serious contemplation of the policy implications of the salience bias (Schenk, 2011; Galle, 2009; Gamage and Shanske, 2011).

This paper explores several under-studied questions on the relationship between behavioral bias and efficiency cost, questions which naturally arise when we consider applying the existing work to the real world. First, how does the individual's needs to meet the budget constraint, even under misperception of prices, matter for inefficiency? Second, what happens if individuals correct their mistakes when policy makes those mistakes more costly? Third, how do we apply the theory to biases other than the salience bias?

To answer these questions, I show how different assumptions about budget adjustment rules, debiasing, cognitive costs, and the nature of mis-perceptions influence behavior and the efficiency cost of taxation, and by implication the optimal tax policy.[7] Primary results of interest are 1) a complete description of how budget adjustment rules affect efficiency cost, 2) expressions for the efficiency cost of taxation that directly examine for any type of mis-perception, and 3) the finding that attempting to exploit optimization errors to reduce inefficiency can give rise to a "curse of debiasing," which actually increases inefficiency.

2 Informal Discussion

In this paper, I develop a static, two-good model of tax mistakes, generalizing Chetty, Looney and Kroft (2007) and Liebman and Zeckhauser (2004), and deriving the statistics of interest for welfare analysis using the CLK (2009) framework. Three components are added to the standard model: misperception, budget adjustment, and debiasing. First, individuals mis-perceive the tax rate, and try to choose optimally given their perception. For example, individuals may ignore a sales tax not posted in the price of a good, they may have false beliefs about the value of a tax rate, or they may believe their average tax rate is their marginal tax rate, known as "ironing" (Liebman and Zeckhauser, 2004).

[3] Gallagher and Muehlegger (2011); Sahm, Shapiro and Slemrod (2010); McCaffery and Baron (2006); Homonoff (2015).

[4] Sheffrin (1994); de Bartolome (1995); Fujii and Hawley (1988); de Bartolome (1995). For other work on apparent misperception of marginal tax rates, see Feldman, Katuščák and Kawano (2016).

[5] Andreoni, Erard and Feinstein (1998). Another example is individuals' apparent belief that estate tax rates are significantly higher than they truly are (Slemrod, 2006).

[6] See also Chetty (2009).

[7] Optimal tax implications only obtain insofar as the objective of the optimal tax program is to minimize inefficiency (Ramsey, 1927). The relevant distributional effects of misperceptions are not discussed here; for some exploration of them see Goldin and Homonoff (2013) and Taubinsky and Rees-Jones (2016). Neither do I address administrative and compliance issues.

When an individual does not choose based on the true marginal price, she will tend not to satisfy her budget constraint. The second addition to the model is the budget adjustment rule, which I parameterize as the share of the necessary reduction in expenditure that falls on the taxed good. For example, an individual may simply reduce her consumption of an untaxed good to meet her budget constraint, in which case none of the reduction in expenditure falls on the taxed good.[8]

2.1 Budget Adjustment Rules

Given these two additions to the standard model, I derive expressions for how budget adjustment rules affect the efficiency cost of taxation. Recall that the efficiency cost of taxation can be defined as the loss in utility relative to a revenue-equivalent lump sum tax. I derive the lump-sum-equivalent budget adjustment rule, where the change in demand due to budget adjustment exactly equals the effect of a lump-sum tax in the amount of the budget shortfall. Intuitively, the individual's choice under this budget adjustment rule fully accounts for the effect of taxes on income, despite misperceiving marginal prices. With this budget adjustment rule, a tax that individuals ignore becomes identical to a lump-sum tax. The excess burden of a tax individuals ignore then increases quadratically as the share of budget adjustment borne by the taxed good moves away, hypothetically, from its lump-sum-equivalent value. When budget adjustment causes a sufficient decline in expenditure on the taxed good, low-salience taxes will not be socially desirable.

2.2 Debiasing

The third addition to the standard model reflects the idea of bounded rationality:[9] the individual will make an optimal choice when the gain to doing so is sufficiently high. For example, a consumer may pay attention to a six percent sales tax on a car but not on a cup of coffee. I call the switch to full optimization "debiasing," following the literature in psychology and behavioral economics (Fischoff, 1981; Kahneman, Slovic and Tversky, 1982).

There are two ways of introducing this intuitive idea into the model. The first is to assume, axiomatically, that at a sufficiently high tax rate, individuals always optimize. When we consider a shift from perfectly salient taxes to low-salience taxes under this assumption, we find that any reductions in the efficiency cost of taxation arising from mis-perception will be erased at some high level of reliance on low-salience taxes. On the margin, this erasure of the social benefits of mis-perception leads low-salience taxes to cause *more* excess burden than salient taxes.[10] This

[8] In a model with more than two goods, the budget adjustment rule would be a vector of such shares, where the sum of the components is unity. Each component of the vector would specify one good's share of the reduction in expenditure to meet the budget constraint.

[9] The literature defines bounded rationality differently depending on the context. Here I define bounded rationality as a tendency to correct behavioral biases when they are sufficiently costly, but otherwise tolerate them due to the limited nature of cognitive resources. A second strand of the literature defines bounded rationality according to partitions of the information set, where the individual can only know whether the state of the world is in a particular subset of the information set. In the latter, individuals are typically assumed to be expected utility maximizers who "know what they don't know." This evidence is difficult to square with evidence on tax perceptions, where individuals "don't know what they don't know." As such, I assume individuals act as though they have full information but their perceptions are biased. If individuals are boundedly rational according to the definition I use here, they should correct this bias when they stand to gain sufficient utility from so doing.

[10] One natural way governments might try to avoid the consequences of debiasing is to implement many distinct, small, non-salient taxes. I will not examine such a policy in detail. Whether it really would avoid the curse of debiasing is ambiguous: do consumers debias for all tax at once or for individual taxes? Does the pressure to debias as mistakes become more costly build across goods and types of taxes or does it build independently for particular goods and

result is the tax parallel of what Gabaix and Laibson (2006) call the "curse of debiasing," wherein debiased consumers exploit policies which attempt to exploit biased consumers.[11]

I also consider a fully specified positive model of bounded rationality to further understand how debiasing drives the result, and to examine cognitive costs. To do this I assume, as in Chetty, Looney and Kroft (2007), that the individual makes the same mistakes until she debiases, and a parameter governs how much must be at stake to induce complete debiasing. One can view this parameter as a cognitive cost of debiasing, perhaps related to the opportunity cost of attention or the mental anguish (or joy) of thinking about taxes, or one can assume that there are no cognitive costs and the decision to debias only affects utility through the change in consumption.

When there are no cognitive costs, the curse of debiasing bites when sufficiently many individuals choose to debias in response to a marginal increase in the tax rate. The curse of debiasing only affects marginal excess burden with no cognitive costs. Total excess burden will, in this model, always be weakly lower for the low-salience tax than for a rate-equivalent high-salience tax. Real cognitive costs of debiasing, however, exacerbate the curse of debiasing for marginal excess burden, and *reverse the direction of the total excess burden comparison* when sufficient debiasing occurs.

To understand the policy relevance of the curse of debiasing, consider a simple example. Suppose the government introduces a small sales tax, which is typically excluded from the posted price of a good. Upon finding that inattention makes the tax less inefficient than other tax instruments on the margin,[12] the government increases the price-exclusive tax rate. Then consumers debias, because ignoring the tax becomes too costly. Once everyone has debiased, the excess burden of this sales tax will be the excess burden of an equivalent price-*inclusive* tax, *plus* the cost of individuals' having to think hard about the exclusive tax. Previously desirable, non-salient taxes become undesirable if individuals are boundedly rational and the government relies too heavily on them.[13]

It is worth noting that the curse of debiasing result is not generic: the tax rate at which enough debiasing occurs to make low-salience taxes undesirable might be very high, so high that governments can raise all the revenue they need without encountering this problem. What the theory really demonstrates formally is the importance of assumptions about debiasing and cognitive costs for the argument that the government ought to exploit biases for social gain.

2.3 Ironing

We can apply this same general model outside the context of tax salience by specifying how misperceptions work in alternative situations. I demonstrate this using the example of what Liebman and Zeckhauser (2004) call "ironing." Ironing occurs when individuals act as if their average tax rate (ATR) is their marginal tax rate (MTR). In this case, the excess burden of taxation depends on the ratio of average and marginal tax rates. Applying the CLK result for excess burden points yields two equivalent formulas for excess burden under ironing, using two different (compensated) elasticities. One formula uses the observed elasticity of demand with respect to one minus the MTR—the conventional *behavioral* elasticity of taxable income—and the other uses the elasticity of demand with respect to one minus the ATR—what I will call the *preferences* elasticity. Liebman and Zeckhauser use the preference elasticity formula in

taxes? These important questions have not yet been addressed in the literature.

[11] Gabaix and Laibson (2006) examined the curse of debiasing in the context of competing firms attempting to exploit consumers' biases with hidden add-on prices.

[12] Figuring this out would require careful consideration of budget adjustment rules or robust estimation of the sufficient statistics of the CLK model, but for the example, assume that policymakers properly considered this question.

[13] This kind of thinking suggests that cognitive burden could be considered to be part of the marginal efficiency cost of funds (Slemrod and Yitzhaki, 2002) Future work could integrate the cognitive burden in taxation into the optimal tax systems framework, by adding cognitive burden to traditional concepts like compliance costs, avoidance, and evasion.

their calculations of the welfare impact of ironing, but these authors evaluate the expression using an estimated value of the behavioral elasticity. Liebman and Zeckhauser (2004) note that it may be improper to use this elasticity because it may not reveal preferences; the analysis here formalizes that concern and provides a simple means of making the desired correction. The analysis also demonstrates that examining welfare effects requires careful consideration of how preferences are revealed when individuals make mistakes.

The rest of the paper proceeds as follows: Section 3 sets up the two-good model, Section 4 examines budget adjustment rules and their welfare implications, Section 5 adds debiasing to the model and proves the curse of debiasing results, and Section 6 examines the example of ironing. Section 7 concludes with a discussion of optimal policy arguments involving tax misperceptions, and describes empirical questions raised by the analysis, answers to which questions should improve our ability to design optimal policies.

3 Model

Assume that a consumer spends a fixed endowment, Z, on two goods, x and y.[14] The individual has additively separable utility $U(x, y) = u(x) + v(y)$. Assume that utility is increasing at a decreasing rate in x and y: $u'(x) > 0$, $u''(x) \leq 0$, $v'(y) > 0$, and $v''(y) \leq 0$.[15] The pre-tax price of x is p and the price of y is normalized to 1. The government taxes x at (per unit) rate t and y is not taxed. When multiple types of tax are in play we divide the tax into different types of tax indexed by j, writing $t = \sum_j t_j$. For example, we could have $t = t_I + t_E$ where t_E is the price-exclusive tax rate and t_I is the tax included in the posted price. For simplicity, assume there are constant returns to scale on the production side of the economy, so that the producer price p is exogenous.[16] The individual's choices must satisfy the budget constraint $(p + t)x + y = Z$. Note that if we normalize the price to 1, our per-unit tax will be the same as an *ad valorem* tax. When I use such a normalization so I will denote the tax rate by τ for clarity. This notation is more conventional for the income tax applications considered in Section 6.[17]

To introduce mis-perception, assume that the individual perceives the tax to be $\hat{t}(t_1, ..., t_J)$ when it is in fact t. For inattention to taxes not included in the posted price of the good, for example, we would have $\hat{t} = t_I$: the consumer only notices the price-inclusive tax. Assume that the consumer will *try* to maximize utility based on the *perceived* marginal tax rate $\hat{t}$. Define the *planned consumption bundle* $(\hat{x}, \hat{y})$ as the solution to the perceived decision problem:[18]

$$u'(\hat{x}) = (p + \hat{t})v'(\hat{y}) \tag{1}$$

$$(p + \hat{t})\hat{x} + \hat{y} = Z. \tag{2}$$

Whenever $\hat{t} \neq t$, the bundle $(\hat{x}, \hat{y})$ does not satisfy the budget constraint. We specify the budget adjustment rule, ρ as x's share of the change in consumption necessary to meet the budget constraint. Actual consumption of x and y

[14] The traditional interpretation of a two-good model would be that y is either that it is a "numeraire", or "background" good, or it is leisure, which is not taxable. In section 4, I discuss how the model, especially as far as budget adjustment rules are concerned, might change depending how one interprets good y.

[15] Note also that $U_{xy} = 0$ due to additive separability.

[16] This assumption may be relaxed without affecting any important conclusions. My focus is on consumer behavior and optimization errors, so I make simplifying assumptions about the production side of the economy. See Chetty (2009) for work on endogenous producer prices in this context.

[17] None of the important conclusions of the model depend on whether one assumes *ad valorem* or per-unit taxes. Using one or the other is a matter of convenience and convention.

[18] The budget constraint will be satisfied with equality due to the assumption of monotonicity.

is given by the following:

$$(p+t)x = (p+t)\hat{x} - \rho(t-\hat{t})\hat{x} \tag{3}$$

$$y = \hat{y} - (1-\rho)(t-\hat{t})\hat{x} \tag{4}$$

which identifies a unique point on the true budget line.[19]

4 The Budget Adjustment Rule

This section expands the model to incorporate misperception and budget adjustment rules (BARs), and derives expressions for excess burden that can be used for any misperception of marginal prices and any BAR. I derive an approximation for the lump-sum-equivalent BAR, compute efficiency cost as a function of the BAR, and show how the desirability of low-salience taxes depends on the BAR. Together, these results imply that budget adjustment rules have important welfare implications in the presence of mis-perceived marginal prices. While it is true that the sufficient statistics of the CLK model incorporate the budget adjustment rule, estimating these statistics requires observing behavior in a setting where the individuals use the same BAR that they would given a permanent, real-world change in a government's reliance on misperceived taxes.

4.1 Discussion and Examples

The budget adjustment rule could be the result of a conscious process, the order of decision-making, or other aspects of the setting in which a choice is made. Chetty, Looney and Kroft (2007) propose three intuitive budget adjustment rules, each corresponding to a value of ρ.

If the individual chooses x based on the perceived marginal price and budget constraint, and adjusts y to meet the true budget constraint, then $\rho_1 = 0$. We can think of this as choice rule in which x is chosen "actively," based on conscious thought and perception of prices, and y is purchased with whatever income is left over. If the individual instead chose y actively and purchases x with the left over resources, then $\rho_2 = 1$: all of the change in consumption necessary to meet the budget constraint will fall on good x.

A third possibility is that the individual will choose x and y jointly based on the perceived MRS condition in equation (1), but with her income implicitly reduced[20] (or increased) by the excess tax she owes (or does not owe) due to misperception. In this case we will have:

$$\rho_3 \simeq (p+t)\left.\frac{\partial x}{\partial Z}\right|_{t=\hat{t}} \frac{x}{\hat{x}} = \omega_x|_{t=t}\, \eta_{x,z}|_{t=\hat{t}} \tag{5}$$

[19] One possibility not pursued here is that the budget adjustment process works differently for different tax misperceptions. In this case ρ would need to depend on the type of tax j. None of the results in this paper compare two different misperceived taxes, so this possibility is not material for the results.

[20] Behavior with this budget rule can be characterized according to the MRS condition under mis-perception

$$u'(x) = (p+\hat{t})v'(y)$$

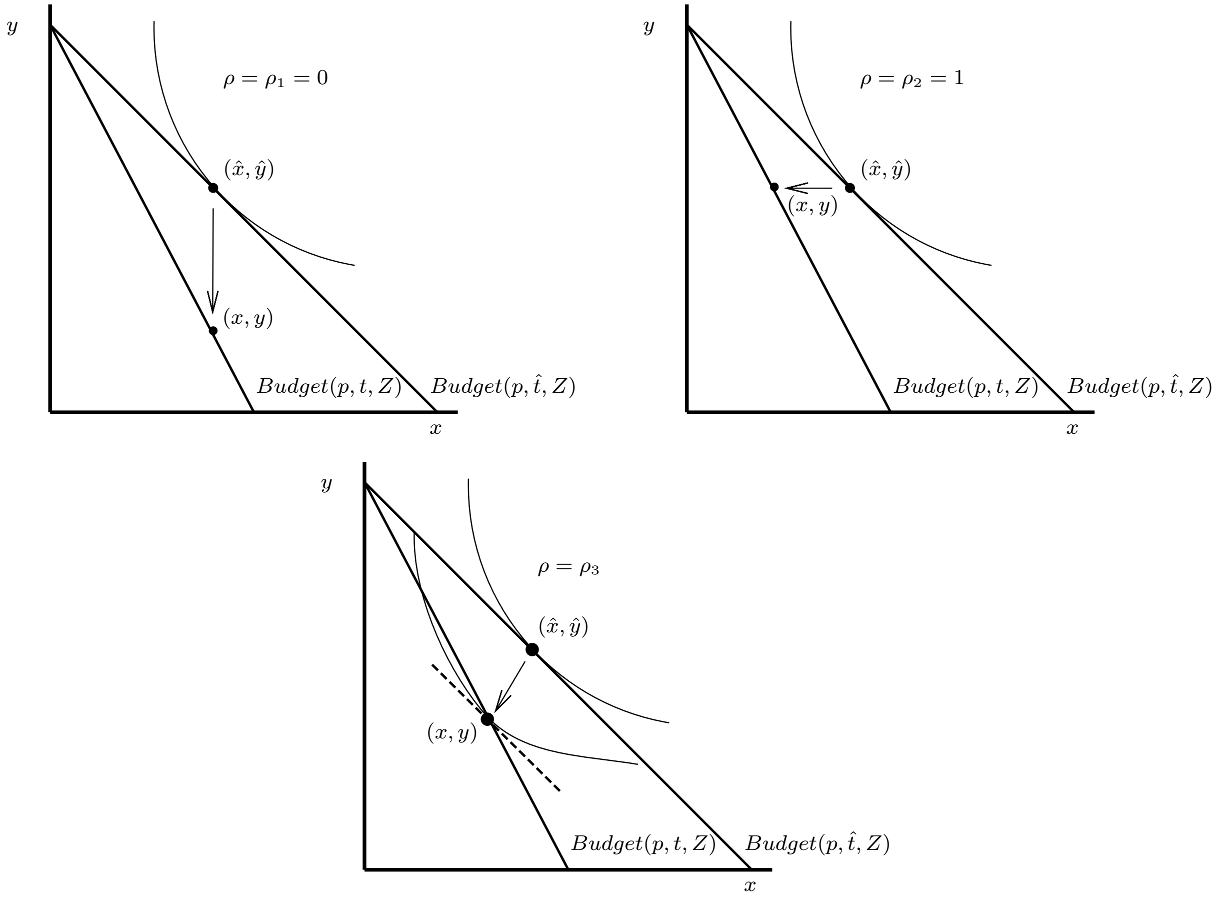
and the true budget constraint

$$(p+t)x + y = Z$$

$$\Leftrightarrow (p+\hat{t})x + y = Z - (t-\hat{t})x.$$

Re-writing the equation the second way highlights that income is "implicitly reduced" by $(t-\hat{t})x$.

Figure 3: Budget Adjustment Rules



Notes: Each panel depicts the change in the choice bundle as the individual moves from the preferred choice under mis-perception $(\hat{x}, \hat{y})$ to the adjusted bundle that satisfies the budget constraint (x, y). The first two panels correspond to $\rho = 0$ and $\rho = 1$.The third panel corresponds $\rho = \omega_x \eta_{x,Z}$; the dotted line in this panel is the budget constraint if income were reduced by $L = (t - \hat{t})x$. It should be apparent that change in behavior from the third budget adjustment rule is exactly the same as the change in behavior from introducing a lump-sum tax of L. When $\hat{t} \neq 0$, this means adding a lump sum tax onto an existing per unit tax of $\hat{t}$.

where $\eta_{x,Z}$ is the income elasticity of demand for x, $\eta_{x,z} = \frac{\partial x}{\partial Z}\frac{Z}{x}$, evaluated at the perceived tax rate, and $\omega_x = \frac{(p+t)x}{Z}$ is x's share of the budget evaluated at the true tax rate. The derivation of this expression is in Appendix A.

Intuitively, the individual acts like her income is reduced in lump-sum fashion during budget adjustment. When the income elasticity of demand for x is higher, or when x constitutes a larger share of the budget, x then bears a larger portion of the change in expenditure from budget adjustment. Useful for welfare calculations will be the fact that when the tax is completely perceived by the consumer at first—so $\hat{t} = t$ and $\hat{x} = x$—we evaluate both ω_x and $\eta_{x,z}$ at the true tax rate and we have simply $\rho_3 = \omega_x \eta_{x,z}$.

These three budget adjustment rules are depicted in Figure 3. For different values of ρ, the bundle (x, y) ends up at very different locations on the true budget line. This foreshadows the finding that different budget adjustment rules have substantially different welfare implications for the efficiency cost of taxation.

If we wish to understand the consequences of a particular kind of mistake, all we need to do is specify $\hat{t}$ and ρ. For example, suppose an individual ignores the sales tax on a cup of coffee whenever that tax is not included in the posted price, because of a salience bias, and there are no other taxes on coffee. Suppose x is the amount of coffee consumed, and y is cash on hand. Because our individual totally ignores the sales tax, we will have $\hat{t} = 0$. In this setting, it is intuitive to assume that the individual has quasilinear preferences and she simply reduces her consumption of a numeraire good y to meet the budget constraint. So we would have $\rho = 0$. A second example is ironing (Liebman and Zeckhauser, 2004). After slightly modifying the model to consider an income tax application, we should assume $\hat{\tau}$ is the individual's average tax rate, $\hat{\tau} = \int_0^x \tau(s)ds/x$, where $\tau(s)$ is the marginal tax rate when $x = s$. In their paper, Liebman and Zeckhauser assume behavior is equivalent to that under budget adjustment rule ρ_3.[21] For these specifications of ρ and $\hat{\tau}$, this model becomes exactly like the ironing model in their paper. I consider this example more thoroughly in Section 6.

4.2 Excess Burden

Propositions from CLK and Chetty (2009) allow us to characterize excess burden in this setting by calculating two demand elasticities. Note that we have not yet introduced debaising, which is discussed in Section 5.

The results in CLK and Chetty (2009) we use relies on two assumptions. Let $x(p, \mathbf{t}, Z) = x(p, t_1, ..., t_j, Z)$ be demand for x and let $y(p, \mathbf{t}, Z)$ be demand for y:

A1. Taxes affect utility only through their effect on observable choices. Thus, indirect utility given some tax rate t is

$$V(p, \mathbf{t}, Z) = u(x(p, \mathbf{t}, Z)) + v(y(p, \mathbf{t}, Z))$$

A2. In the absence of taxation, the individual behaves fully optimally:

$$\forall p,\ x(p, \mathbf{0}, Z) = \arg\max_x u(x) + v(Z - px)$$

where $x(p, \mathbf{0}, Z)$ is demand for x in the no-tax equilibrium. Intuitively, assumption **A1** rules out any unobservable cognitive costs caused by optimization errors. Assumption **A2** requires that taxes alone are the cause of sub-optimal decision-making.[22] Both axioms hold in the model used here, until we add cognitive costs in section 5.2.

Before arriving at the results, we need to introduce some notation. Define the following two terms:

$$\frac{\partial x^c}{\partial p} \equiv \frac{\partial x}{\partial p} + x\frac{\partial x}{\partial Z}$$

[21] This choice is motivated by the assumption that the individual believes that she faces a linear tax equal to her average tax rate, with no exemption, and that the individual must correctly perceive her average tax rate. This behavior is the same as using the budget adjustment rule ρ_3 because the individual accounts for the effect of taxes on her income and on her *virtual* income—the extra income she gets because the marginal tax rate is lower at lower levels of income, unlike a linear tax with no exemptions—but not on relative prices. A depiction of this kind of budget adjustment rule for a two-bracket tax is provided in Figure 8 in the appendix.

[22] This assumption may be relaxed depending on the setting to which the model is applied; if we wish to study the welfare consequences of confusing price schedules instead of confusing taxes, for example, then **A2** would no longer be appropriate. Nevertheless, we can replace **A2** with any "refinement" stating under which conditions individuals will reveal their true preferences, and then use behavior under this condition to understand welfare (Bernheim and Rangel, 2009).

Mathematically, the summary statistic of interest is

$$\theta_j^c = \frac{\partial x^c}{\partial t_j} \bigg/ \frac{\partial x^{*c}}{\partial t_j}.$$

A2 implies that $\frac{\partial x^{*c}}{\partial t_j} = \frac{\partial x^c}{\partial p}$ when $t = \hat{t}$, but there could be other ways of obtaining $\frac{\partial x^{*c}}{\partial t_j}$ using choice data.

$$\frac{\partial x^c}{\partial t_j} \equiv \frac{\partial x}{\partial t_j} + x\frac{\partial x}{\partial Z}.$$

The first of these is a conventional compensated price effect, derived from the Slutsky equation. Whenever the individual behaves optimally—when $\hat{t} = t$—we will have $\frac{\partial x^c}{\partial p} < 0$. The second is a definition similar to a conventional compensated tax effect, but which need not have the properties of conventional compensated price effects because it does not result from full optimization. We denote elasticity of demand for x with respect to a change in the after tax price coming from a price change as $\varepsilon_{x,q|p} = \frac{\partial x}{\partial p}\frac{p+t}{x}$. Similarly, let the elasticity of demand with respect to the after-tax price coming from a change in tax j be $\varepsilon_{x,q|t_j} = \frac{\partial x}{\partial t_j}\frac{p+t}{x}$, and denote compensated versions of these elasticities by $\varepsilon^c_{x,q|p}$ and $\varepsilon^c_{x,q|t_j}$, respectively.

CLK and Chetty (2009) introduce the following statistics, which summarize the deviation from full optimization for a tax t_j. I will call these the *degree of uncompensated error,* θ_j, and the *degree of compensated error,* θ^c_j, defined as follows:

$$\theta_j \equiv \frac{\partial x}{\partial t_j}\bigg/\frac{\partial x}{\partial p}$$

$$\theta^c_j \equiv \frac{\partial x^c}{\partial t_j}\bigg/\frac{\partial x^c}{\partial p}$$

Let $x_0 = x(p, \mathbf{t_0}, Z)$ denote demand at some vector of initial tax rates $\mathbf{t_0}$, and denote the true initial tax rate—the sum over the elements of $\mathbf{t_0}$—by t_0. Following Auerbach (1985) and Mohring (1971), CLK define the excess burden of introducing a tax using equivalent variation:

$$EB(\mathbf{t}) = Z - e(p, \mathbf{0}, V(p, \mathbf{t}, Z)) - R(0, \mathbf{t}, Z).$$

where $e(.)$ is the expenditure function and $R(.)$ is tax revenue at a tax rate t.[23]

Proposition 1 (Chetty (2009))**.** *Assume there is only one type of misperceived tax, t_j.*[24] *Under assumptions **A1** and **A2**, the excess burden of a small tax increase Δt_j starting from a small initial tax rate t_0 is approximately given by:*

$$EB(\Delta t_j|t_0) \simeq -\frac{1}{2}(\Delta t_j)^2\theta^c_j\frac{\partial x^c}{\partial t_j} - \Delta t_j\frac{\partial x^c}{\partial t_j}[\hat{t}_0 + \theta^c_j(t_0 - \hat{t}_0)] \quad (6)$$

$$= -\frac{1}{2}(\Delta t_j)^2\theta^c_j x_0\frac{\varepsilon^c_{x,q|t_j}}{p+t_0} + \Delta t_j x_0\frac{\varepsilon^c_{x,q|t_j}}{p+t_0}[\hat{t}_0 + \theta^c_j(t_0 - \hat{t}_0)] \quad (7)$$

where all derivatives are evaluated at $t = \hat{t}_0 \equiv \hat{t}(\mathbf{t_0})$.

Whether misperception increases or decreases excess burden depends on whether $\theta^c_j \gtrless 1$. Panel I of Figure 4 gives a graphical depiction of the result for the case where there are no income effects on the demand for x and $t_0 = 0$, so that we observe total excess burden. Panel II depicts marginal excess burden, when $t_0 > 0$. To obtain excess burden

[23] Traditionally, one thinks of excess burden as the amount the individual would willingly pay in order to replace the current tax regime with efficient lump-sum taxation. That intuition does not carry through to the world of optimization errors, because the individual does not rationally value all transactions. One can think of excess burden instead as the amount that an individual who was fully aware of her biases, despite making mistakes, would be willing to pay to replace the current tax regime with lump-sum taxation. Alternatively one can use different intuition and think of excess burden as "the amount of additional tax revenue that could be collected from the consumer while keeping his utility constant if the distortionary tax were replaced with a lump-sum tax", or "the total value of the purchases that fail to occur because of the tax." (Chetty, 2009).

[24] Proposition 1 may be modified to accommodate multiple mis-perceived taxes but such an extension is not useful for the results in this paper.

when there are no income effects, we can integrate the difference between marginal utility and price over the change in x due to the tax rate change. To calculate excess burden as a function of misperceptions and the BAR, we must calculate the compensated tax effect and relate it to the compensated effect of a price change or, equivalently, a change in the perceived tax rate $\hat{t}$.

Lemma 1 (The Compensated Tax Effect with No Debiasing)**.** *When individuals behave in the manner described by equations* (1) *through* (4) *and there is no debiasing, then the compensated effect of a change in the tax rate starting from any point where $\hat{t} = t$ is given by*

$$\frac{\partial x^c}{\partial t_j} = \frac{\partial x^c}{\partial p} * \frac{\partial \hat{t}}{\partial t_j} + \left[x\frac{\partial x}{\partial Z} - \rho\frac{x}{p+t}\right]\left(1 - \frac{\partial \hat{t}}{\partial t_j}\right) \tag{8}$$

$$\varepsilon^c_{x,q|t_j} = \varepsilon^c_{x,q|p}\frac{\partial \hat{t}}{\partial t_j} + [\omega_x \eta_{x,Z} - \rho]\left(1 - \frac{\partial \hat{t}}{\partial t_j}\right) \tag{9}$$

where all expressions are evaluated at the current tax rate, $t = \hat{t}$.

The proof of this lemma is given in the appendix. Equation (8) illustrates that a tax rate increase has two effects on demand. First, planned consumption $(\hat{x}, \hat{y})$ may change, depending on how the rate change is perceived. The first term of equation (8) corresponds to the change in planned choices due to a change in perceptions. When a change in tax rates goes completely unnoticed by the consumer, $\partial\hat{t}/\partial t_j = 0$ and this term disappears. Notably, $\partial\hat{t}/\partial t_j$ can depend on how the tax change is implemented. With price-exclusive and price-inclusive taxes, if we assume $\hat{t} = t_I$, we will have $\partial\hat{t}/\partial t_I = 1$ and $\partial\hat{t}/\partial t_E = 0$.

The second effect of a tax change is that new budget adjustment is required to meet the budget constraint. The second term in equation (8) shows how the budget adjustment process affects the compensated tax effect. When the tax change is fully perceived, $\partial\hat{t}/\partial t_j = 1$ and the second term is zero because no new budget adjustment is necessary. When the individual accounts for the effect of misperception on income but not relative prices, $\rho = \rho_3 = \omega_x \eta_{x,Z}$ and this term disappears. For any other value of ρ budget adjustment causes the tax to have a distortionary income effect, because the individual fails to properly account for the effect of taxes on her purchasing power.

Proposition 2 (How Does Excess Burden Depend on the Budget Adjustment Rule?)**.** *Assume that the government introduces a small misperceived tax t_j into an untaxed market. Behavior is characterized by the function $\hat{t}(t_j)$, ρ, and equations* (1) *through* (4)*. The excess burden of this tax is approximately*

$$EB(t_j) \simeq -\frac{1}{2}t_j^2\left(\frac{\varepsilon^c_{x,q|p}\frac{\partial \hat{t}}{\partial t_j} + [\omega_x\eta_{x,Z} - \rho]\left(1 - \frac{\partial \hat{t}}{\partial t_j}\right)}{\varepsilon^c_{x,q|p}}\right)^2 x\frac{\varepsilon^c_{x,q|p}}{p} \tag{10}$$

$$= -\frac{1}{2}t_j^2\left(\frac{\partial \hat{t}}{\partial t_j}\right)^2 x_0\frac{\varepsilon^c_{x,q|p}}{p} - \xi\frac{t_j^2 x}{p}\left(\frac{\partial \hat{t}}{\partial t_j} + \frac{1}{2}\frac{\xi}{\varepsilon_{x,q|p}}\right) \tag{11}$$

where $\xi = (\omega_x\eta_{x,Z} - \rho)\left(1 - \frac{\partial \hat{t}}{\partial t_j}\right)$ and all expressions are evaluated at $t = 0$.

The proof is given in the appendix. Like equation (8), the excess burden of a misperceived tax has a relative price component and a budget adjustment component. The first term of equation (11) shows how the misperception of relative prices mutes ($\partial\hat{t}/\partial t_j < 1$) or amplifies ($\partial\hat{t}/\partial t_j > 1$) traditional excess burden directly. The second term of this equation shows that the additional excess burden is a quadratic function of the budget adjustment rule and the distance between the true change in the tax rate and the perceived change in the tax rate, $\left(1 - \frac{\partial \hat{t}}{\partial t_j}\right)$. When the tax is fully ignored, so $\partial\hat{t}/\partial t_j = 0$, we will find that the further ρ is from its lump-sum equivalent value, the larger is the efficiency cost from budget adjustment. Finally, note that whenever $\rho = \rho_3$ the second term is eliminated and the excess burden of this tax is the same as a tax of rate $\hat{t}$, since $t^2\left(\frac{\partial\hat{t}}{\partial t_E}\right)^2 \simeq \hat{t}^2$.[25]

Proposition 3 (How does the desirability of low-salience taxes depend on the budget adjustment rule?)**.** *Suppose that the government has recourse to two types of taxes, t_I and t_E, and that*

1. *Consumers only notice t_I: $\hat{t} = t_I$.*
2. *No individuals debias (see section 5.2).*

Then the introduction of a small tax that individuals ignore, t_E, causes less excess burden than the introduction of a rate-equivalent tax that individuals perceive fully, t_I, if and only if $\rho < -\varepsilon_{x,q|p}$, where $\varepsilon_{x,q|p}$ is evaluated at $t_I = t_E = 0$.

See the appendix for the proof. When consumption of the taxed good x is reduced sufficiently in the course of budget adjustment in the presence of low-salience taxes, low-salience taxes are no longer desirable on efficiency grounds. To understand the necessary condition for this to occur, consider a one percent increase in the after-tax price. If this tax is perfectly salient, it results in an $-\varepsilon_{x,q|p}$ percent reduction in demand for x. If the tax is perfectly non-salient, it results in a ρ percent decrease in consumption of x. Whichever tax generates the larger reduction in demand for the taxed good causes more inefficiency.

While it is true that the budget adjustment rule is "in θ," meaning that a consistent estimate of the two elasticities in θ would be sufficient for policy recommendations, the influence of budget adjustment rules on the compensated tax effect (the numerator of θ) implies that small-scale, experimental evidence on behavioral biases may not allow researchers to properly estimate the correct compensated demand elasticity. An experiment typically makes a small tax more or less salient for a short amount of time. In the experimental setting, we are likely thinking of good y as a numeraire or background good, so it makes sense to think that $\rho = 0$, and y bears the full burden of budget adjustment. When we think about optimal tax, though, we think of good y as an untaxable good like leisure. To apply the experimental estimates of θ to this case, we would need to assume that the budget adjustment rule used in the small-scale experiment is the same as that used in a broader context. When an individual ignores a large tax on consumption, does she then work more to offset the fact that her consumption expenditure is higher than she planned? We do not yet know, so budget adjustment rules deserve more attention.

Another possibility, not addressed here, is that the budget adjustment rule depends on the tax rate, or on income. Assuming the individual only has recourse to the three budget adjustment rules described in Figure 3, one can show that as tax rates change, which of the three budget adjustment rules leads to the highest utility can change. Another possibility is that some budget adjustment rules are more costly to employ than others. Furthermore, a sufficiently sophisticated budget adjustment rule could even lead to a fully optimal choice ex ante, even in the presence of misper-

[25] More specifically, take a Taylor Series expansion of $\hat{t}(t)$ to write

$$\hat{t}(t) = \hat{t}(0) + \frac{\partial\hat{t}}{\partial t}t + \frac{1}{2}\frac{\partial^2\hat{t}}{\partial t^2}t^2 + ...$$

Note that $\hat{t}(0) = 0$. Square both sides and ignore third order and higher terms—those of the form λt^n where λ is a constant and $n \geq 3$—to find that $t_j^2\left(\frac{\partial\hat{t}}{\partial t_j}\right)^2 \simeq \hat{t}^2$. This approximation will be accurate up to the inclusion of third order and higher terms, just like all excess burden approximations in this paper.

ception.[26] Such logic brings the model closer to the general notion of bounded rationality covered in section 5.1, but more empirical evidence on the determinants of budget adjustment rules would be necessary to fully understand this issue.

5 Debiasing

Do individuals rationally allocate scarce decision-making tools, so that they will make better decisions when more utility is at stake? Theories of optimal attention and bounded rationality suggest that the answer is yes, while psychological theories of hard-wired bias in perception and judgment imply that the answer is no.[27] Behavioral economics has not come to a consensus on this question, and it is possible that the answer differs by the kind of bias one considers. There is evidence that sometimes framing and salience effects do exist for large taxes.[28] A related complication concerns imperfect knowledge rather than imperfect perception: individuals should make mistakes indefinitely if their information set is flawed and they never realize it—i.e. if they "don't know what they don't know" about some tax. It is plausible that people eventually find out about sufficiently costly mistakes, perhaps through exogenous shocks to the information set from conversations with financial professionals, acquaintances, or reading the news. Direct evidence on information search suggests that some individuals do search for information at opportune times for decision-making (Hoopes, Reck and Slemrod, 2015).

This section provides some insights about how the efficiency cost of taxation and optimal tax policy depend on assumptions about bounded rationality and debiasing. We explore how the curse of debiasing can cause naive applications of the theory of taxes and mistakes to backfire, and how cognitive costs of debiasing affect the desirability of misperceived taxes.

To address this issue, we begin with a technical extension of proposition 1 to allow for non-linear dependence between demand for the taxed good x and the tax rate. We will maintain the local linearity assumptions on prices and incomes for simplicity. To do so, we distinguish between the average degree of error, which is the average value of the relative responsiveness over the range of tax not perceived, and the marginal degree of error, which is the relative responsiveness to an incremental change in the tax rate.

For simplicity we will limit ourselves to one type of misperceived tax t_j. Define the *marginal degree of compensated error* at some tax rate t_0 to be

$$\theta^c_{j,t_0} \equiv \frac{\left.\frac{\partial x^c}{\partial t_j}\right|_{t=t_0}}{\left.\frac{\partial x^c}{\partial p}\right|_{t=\hat{t}_0}}.$$

Second, define the *average degree of compensated error* at some tax rate t_0 as

$$\begin{aligned}\Theta^c &\equiv \frac{\frac{1}{t_0-\hat{t}_0}\int_{\hat{t}_0}^{t_0}\left.\frac{\partial x^c}{\partial t_j}\right|_{t=t'}dt'}{\left.\frac{\partial x^c}{\partial p}\right|_{t=\hat{t}_0}} \\ &= \frac{1}{t_0-\hat{t}_0}\int_{\hat{t}_0}^{t}\theta^c_{j,t'}dt'\end{aligned}$$

[26] Specifically, budget adjustment leads to a fully optimal choice when the individual pre-determines how to adjust her budget, i.e. the value of ρ, with full knowledge of the tax rate, despite mis-perceiving taxes in choosing $\hat{x}$ and $\hat{y}$. Formally, individuals would maximize over ρ their realized utility given their behavior, as determined by equations (1) through (4). This kind of choice behavior might explain the phenomenon of "mental accounting" (Thaler, 1980).

[27] See chapter 2 of Congdon, Kling and Mullainathan (2011) for a more thorough overview of these questions.

[28] See for example Gallagher and Muehlegger (2011) Cabral and Hoxby (2012), Chetty et al. (2014).

Note that were we to revert to the assumption that demand is linear in tax rates, so $\partial x^c/\partial t_j$ is constant, we would have $\theta^c_{j,t_0} = \Theta^c = \theta^c_j$, and proposition 1 is all that is necessary to characterize excess burden. Let Θ and θ_{j,t_0} denote the uncompensated versions of Θ^c and θ^c_{j,t_0}, respectively.

Proposition 4 (Extension to Non-Linearities in Tax Rates)**.** *Assume there is only one type of misperceived tax, t_j. Under assumptions **A1** and **A2**, and the assumption that demand is locally linear in $(p + \hat{t})$ and Z but not in tax rates, the excess burden at a tax rate t_j is*

$$\begin{aligned} EB(t_j) &= -\frac{1}{2}t_j^2\Theta^c\frac{1}{t-\hat{t}}\int_{\hat{t}}^{t}\left.\frac{\partial x^c}{\partial t_j}\right|_{t=t'}dt' && (12)\\ &= -\frac{1}{2}t_j^2(\Theta^c)^2\frac{\partial x}{\partial p} && (13)\\ &= -\frac{1}{2}t_j^2(\Theta^c)^2 x_0\frac{\varepsilon_{x,q|p}}{p} && (14)\end{aligned}$$

where $x_0 = x(p, \mathbf{0}, Z)$ and $\frac{\partial x}{\partial p}$ is evaluated at $t = 0$. The excess burden of a small tax increase Δt_j starting from an initial tax rate t_0 is approximately:

$$\begin{aligned} EB(\Delta t_j|t_0) &\simeq -\frac{1}{2}(\Delta t_j)^2\theta^c_{j,t_0}\left.\frac{\partial x^c}{\partial t_j}\right|_{t=t_0} - \Delta t_j\left.\frac{\partial x^c}{\partial t_j}\right|_{t=t_0}[\hat{t}_0 + \Theta^c(t_0-\hat{t}_0)] && (15)\\ &= \frac{1}{2}(\Delta t_j)^2\theta^c_{j,t_0}x_0\frac{\varepsilon^c_{x,q|t_j}}{p+t_0} + \Delta t_j x_0\frac{\varepsilon^c_{x,q|t_j}}{p+t_0}[\hat{t}_0 + \Theta^c(t_0-\hat{t}_0)] && (16)\end{aligned}$$

where $\frac{\partial x^c}{\partial t_j}$ is evaluated at $t = t_0$, and $\frac{\partial x^c}{\partial p}$ at $t = \hat{t}_0$.

The proof of this theorem is in the appendix. The average degree of error Θ^c matters for total excess burden. Marginal excess burden depends on the marginal degree of error θ^c because this parameter governs the difference between demand at the initial tax rate and demand at the new tax rate. Marginal excess burden also depends on the average degree of error because it determines demand at the initial tax rate. Panels III and IV of Figure 4 depict excess burden and marginal excess burden in this setting, when there are no income effects.

5.1 An Axiomatic Approach to Debiasing

We can derive some results for debiasing without a positive model, using an axiomatic approach. In addition to assumptions **A1** and **A2**, we now also assume the following:

A3. At a sufficiently high tax rate, all individuals optimize fully.[29]

$$\forall j\ \exists\ \bar{t}_j \text{ s.t.} \forall t_j \geq \bar{t}_j, x(p, t_0, ...t_j, ...t_J, Z) = \arg\max_x u(x) + v(Z - (p+t)x).$$

This assumption states that, whatever optimization errors occur at low tax rates, they are eliminated at sufficiently high tax rates.[30] Referring to Panels III and IV of Figure 4, **A3** requires that the tax demand curve become the same as the price demand curve at tax rates above some high tax rate $\bar{t}$. This assumption incorporates a variety of ways through

[29] Note that in the applications explored here, a high tax rate corresponds to a high utility cost of ignoring the tax. A more general formulation of **A3** would require that sufficiently costly mistakes cannot occur.

[30] In the fully specified model in the next section, this assumption will be true if there are no cognitive costs and the support of the distribution $f(\delta)$ is bounded above, e.g. if there are finitely many individuals with a fixed cost of debiasing.

which debiasing may occur, including the deliberate decision to optimize in high-stakes decisions and endogeneity of the budget adjustment rule.

Proposition 5 (The Curse of Debiasing)**.** *Suppose that the government has recourse to two taxes, t_I and t_E, and that*

1. *Consumers respond perfectly to changes in t_I, so $\theta^c_{t_I} = 1$.*
2. *Consumers initially under-respond to changes in t_E, so $\theta^c_{t_E} < 1$ when $t_E = 0$.*
3. *Assumptions **A1**, **A2**, and **A3** are true.*
4. *Optimal demand x^* is strictly positive at all prices and tax rates.*

Then there exists a tax rate t'_E at which an increase in t_E causes higher *marginal excess burden than an increase in the noticed tax t_I.*

Panels III and IV of Figure 4 shows how the curse of debiasing works. Panel III depicts the first two assumptions of the proposition: the misperceived tax initially causes lower excess burden than fully perceived tax. Initially, debiasing does not matter, so excess burden in panels I and III are similar. Assumption **A3** forces there to be a tax rate at which a change in t_E causes higher marginal excess burden than a change in t_I. The excess burden of a tax change where the curse of debiasing bites is depicted in Panel IV of Figure 4. Marginal excess burden is much higher in panel IV, where we assume debiasing occurs, than in Panel II, where it does not. A more extreme example, in which debiasing occurs discretely at the tax rate $\overline{t_E}$, is depicted in Figure 5. In either case, what we call the curse of debiasing erases the social benefits of misperceived taxes when the government relies too heavily on them. Note also that when Θ is always less than 1—which can be justified by a model like the one in the next section—the total excess burden of a misperceived tax is still weakly less than the total excess burden of a fully perceived tax, despite marginal excess burden being subject to the curse of debiasing.

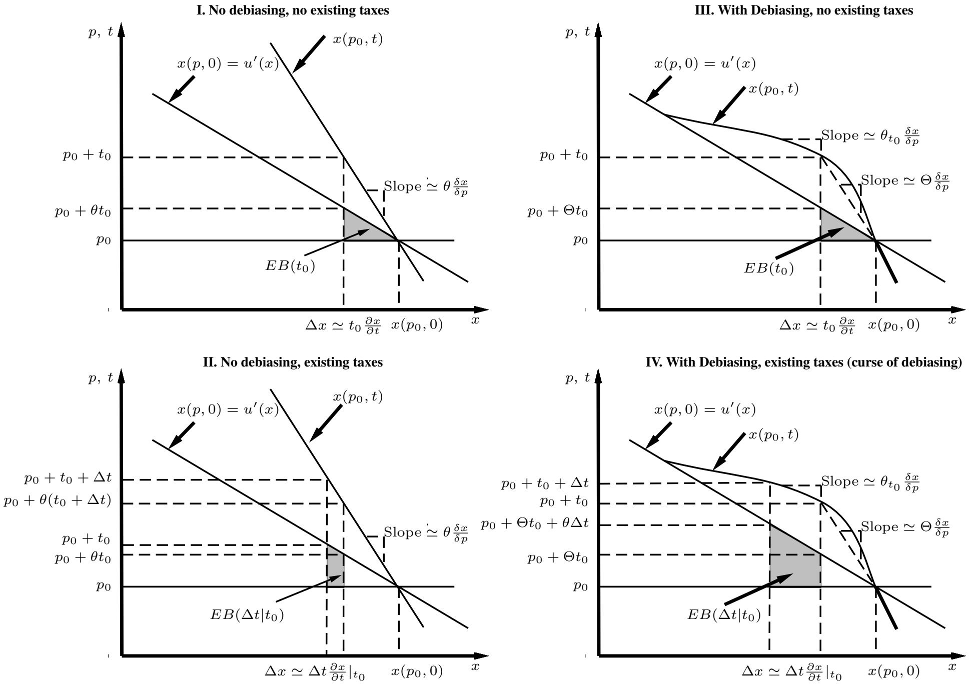
That with too much reliance on a particular tax instrument, that instrument becomes undesirable relative to other tax instruments is a nearly universal feature of optimal tax theory. Another way to think of this result is that it implies that the same holds true for choices along the lines of whether the government should use taxes that are not salient (or are otherwise misperceived) or taxes that individuals respond to perfectly, provided that individuals debias.

Most of the motivation for debiasing in this section comes from the idea of bounded rationality, where the cost of perfectly optimizing is high enough that the consumer tolerates some error, but **A3** need not be limited to this idea. Upon realizing that the government is attempting to exploit their biases (even benevolently), consumers could adversely react by deliberately debiasing, like individuals who shelter income from the taxation because they dislike government policies. Nothing substantive about the model changes in this scenario, though upcoming results about cognitive costs could not be applied to debiasing out of distaste for policy.

5.2 A Positive Approach to Debiasing

Using the fully specified positive model, this section examines when the curse of debiasing will occur and how adding cognitive costs affects efficiency cost and the attractiveness of misperceived taxes. We will start with the model in section 3. Assume further that there are two types of taxes t_E and t_I and that the individual only pays attention to the latter, so $\hat{t} = t_I$. To further simplify the model, assume $\rho = \rho_3$, so that the individual takes account of the impact of her mis-perceptions on income but not on relative prices, and that there are no income effects on demand, so $\rho_3 = 0$. The qualitative results in this section will not be changed so long as the conditions in proposition 5 still obtain.

Figure 4: Excess Burden and Marginal Excess Burden, with and without Debiasing (No Income Effects)


I. No debiasing, no existing taxes
$p,\ t$
$x(p,0)=u'(x)$
$x(p_0,t)$
p_0+t_0
$p_0+\theta t_0$
p_0
Slope $\simeq \theta \frac{\delta x}{\delta p}$
$EB(t_0)$
$\Delta x \simeq t_0 \frac{\partial x}{\partial t}$
$x(p_0,0)$
x
III. With Debiasing, no existing taxes
$p,\ t$
$x(p,0)=u'(x)$
$x(p_0,t)$
Slope $\simeq \theta_{t_0} \frac{\delta x}{\delta p}$
Slope $\simeq \Theta \frac{\delta x}{\delta p}$
p_0+t_0
$p_0+\Theta t_0$
p_0
$EB(t_0)$
$\Delta x \simeq t_0 \frac{\partial x}{\partial t}$
$x(p_0,0)$
x
II. No debiasing, existing taxes
$p,\ t$
$x(p,0)=u'(x)$
$x(p_0,t)$
$p_0+t_0+\Delta t$
$p_0+\theta(t_0+\Delta t)$
p_0+t_0
$p_0+\theta t_0$
p_0
Slope $\simeq \theta \frac{\delta x}{\delta p}$
$EB(\Delta t|t_0)$
$\Delta x \simeq \Delta t \frac{\partial x}{\partial t}|_{t_0}$
$x(p_0,0)$
x
IV. With Debiasing, existing taxes (curse of debiasing)
$p,\ t$
$x(p,0)=u'(x)$
$x(p_0,t)$
$p_0+t_0+\Delta t$
p_0+t_0
$p_0+\Theta t_0+\theta\Delta t$
$p_0+\Theta t_0$
p_0
Slope $\simeq \theta_{t_0} \frac{\delta x}{\delta p}$
Slope $\simeq \Theta \frac{\delta x}{\delta p}$
$EB(\Delta t|t_0)$
$\Delta x \simeq \Delta t \frac{\partial x}{\partial t}|_{t_0}$
$x(p_0,0)$
x

Notes: Panels I and II are based on CLK (2009) and Chetty (2009). Panels III and IV add debiasing. In all panels, the tax demand curve $x(p_0,t)$ governs behavior and the price demand curve $x(p,0)=u'(x)$ determines welfare. The curves differ due to the presence of misperception, whose consequences are determined by the marginal and average degree of error θ and Θ. In panels I and III, we assume that there are no existing taxes and a misperceived tax at rate t_0 is introduced. We integrate the marginal utility curve over the change in x to obtain excess burden, represented by the shaded gray area. In panels II and IV, we add a tax of Δt onto a tax of t_0. The debiasing assumption **A3** causes the tax demand curve to bend toward and eventually unite with the price demand curve. The curse of debiasing (panel IV) causes the excess burden to be much larger than it would be if consumers did not debias (panel II). The figure does not depict the excess burden of perfectly perceived taxes for clarity, but these may be examined using the price demand curve only, in the usual way.

Now we add debiasing. Denote choices under optimization errors by (x, y) and denote choices from full optimization, after debiasing, by (x^*, y^*). Let $G(t_E)$ denote the gain to an individual of debiasing fully, in money-metric utility:[31]

$$G(t_E) = \frac{[u(x^*) + v(y^*)] - [u(x) + v(y)]}{\frac{\partial V}{\partial Z}},$$

where $\frac{\partial V}{\partial Z}$ is evaluated at (x^*, y^*). The numerator is the gain to debiasing in utility units, and we divide by $\frac{\partial V}{\partial Z}$, the marginal value of wealth at the optimal consumption level, to obtain money-metric utility.

Assume that individuals know the value of $G(t_E)$.[32] Then they will consume (x^*, y^*) whenever $G(t_E)$ is greater than some parameter δ, and they will otherwise consume (x, y).

Intuitively, we can think of δ as reflecting the opportunity cost of attention, the mental anguish of thinking about some taxes, or simply a component of an arbitrary cognitive system determining which taxes an individual notices. As such, we will consider both the case where δ does not enter the utility function, and the case where welfare is reduced by δ whenever an individual chooses (x^*, y^*) instead of (x, y).

Next, we assume that δ is heterogeneous in the population, following Chetty, Looney and Kroft (2007). Assume that δ_i is distributed identically and independently across a measure one continuum of individuals i with a probability density function $f(\delta)$ and a cumulative density function $F(\delta)$. The fraction of debiased agents at a tax rate t_E—those for whom $\delta_i < G(t_E)$—can be written as $F(G(t_E))$. Differences between individuals could be related to a heterogeneous opportunity cost of attention or heterogeneous cognitive abilities. Another way to think of the heterogeneity would be to think of a single representative individual who makes transactions involving the purchase of x repeatedly, and has a varying cognitive cost each time due to her mental state, whether she is in a hurry, and so on.[33]

Lemma 2 (The Compensated Tax Effect with Debiasing[34])**.** *Suppose individuals behave in the manner described in Section 3 and the government has recourse to two taxes t_E and t_I. Suppose also that*

1. *Consumers only notice t_I, $\hat{t} = t_I$.*

2. *The budget adjustment rule is $\rho = \rho_3$.*

3. *There are no income effects on the demand for x under full optimization.*

4. *A fraction $F(G)$ of individuals debias.*

Then the compensated effect of an increase in t_E of Δt_E for an initial tax rate t_{E0} is given by

$$\left.\frac{\partial X^c}{\partial t}\right|_{t_{E0}} = f(G)\frac{dG}{dt_E}(x^* - x) + F(G)\frac{\partial x^{*c}}{\partial p} \tag{17}$$

where X is aggregate demand for good x, $G = G(t_{E0})$.

[31] The approach used by Chetty (2009) to approximate the function G implies that G is indeed a function of t_E. More generally we could write G as a function of $(t - \hat{t})$. Other variables affecting the gain to optimizing G, such as demand elasticities and x's share of the budget, affect the function $G(t_E)$ and $\partial G/\partial(t - \hat{t})$ in intuitive ways but are suppressed here for clarity.

[32] We must assume the value of G is known to avoid the problem of infinite regress, which plagues models of bounded rationality. The difficulty with this assumption is part of the reason I take a more axiomatic approach in Section 5.1.

[33] This characterization also has the advantage that if we observe multiple purchases at a given tax rate, we should still expect individual demand not to exhibit large discrete jumps when the tax rate increases only slightly.

[34] Taubinsky and Rees-Jones (2016) also derive this result and generalize it to a heterogeneous agents framework. The result appearing here was derived independently and pre-dates these authors' work.

The proof is in the appendix to this paper.[35] We can use lemma 2 and the definition of Θ_t^c and θ_t to derive that the average degree of compensated error is the fraction of optimizing agents

$$\Theta_{t_{E0}}^c = F(G(t_E)),$$

and the marginal degree of error is the fraction of optimizing agents *plus* a term approximating the change due to debiasing at a tax rate t_{E0}:

$$\theta_{t_{E0}}^c \simeq F(G) + f(G)\frac{\partial G}{\partial t_E} t_{E0}.$$

The derivations of both of these facts are also given in the appendix.

Assume first that there are no cognitive costs. Then by direct applications of proposition 4 and lemma 2, we can understand total and marginal excess burden in this model. The average degree of error is the fraction of debiased individuals, which is always weakly less than 1. As such, the total excess burden of an ignored tax t_E will always be less than the total excess burden of a rate-equivalent tax that individuals notice t_I. The same will be true for any specification of $\hat{t}$ and ρ where $\partial x^c/\partial t_j < \partial x^c/\partial p$ when $t_j = 0$. The marginal degree of error, which affects only marginal excess burden, is not always less than one, because it also depends on the number of individuals who are indifferent between debiasing and not, $f(G)$, and the change in the gain to optimizing from an increase in the tax rate, $\partial G/\partial t_E$. From the perspective of marginal excess burden, low-salience taxes are undesirable if these quantities are sufficiently large, so that $\theta_{t_{E0}} > 1$. Proposition 5 implies that this *must* happen at some tax rate, as long as everyone eventually debiases (so the support of $f(\delta)$ is bounded above). With estimates of the distribution f and the function G—which can be shown to depend on familiar demand elasticities—the tax rate where enough individuals debias—where $f(G)\frac{\partial G}{\partial t_E} t_{E0}$ is large enough—to cause $\theta_{t_{E0}} > 1$ can be fully characterized.

Now assume instead that debiasing incurs a cost equal to δ_i measured in money metric utility. Just as the tax rate is high enough that individual's with cost δ_i debias, excess burden is increased by $f(\delta_i)\delta_i$. In this case, the efficiency cost of taxation will be the excess burden as measured above, plus cognitive costs incurred by any individuals who debias. Denoting the former by $EB_0(t_E)$, we will have

$$\begin{aligned} EB(t_E) &= EB_0(t_E) + \int_0^{G(t_E)} \delta f(\delta) d\delta && (18)\\ &= -\frac{1}{2} t_E^2 (\Theta^c)^2 \frac{\partial x^c}{\partial p} + \Theta \mathbb{E}[\delta | G(t_E) > \delta] && (19)\\ &= \Theta^2 EB(t_I) + \Theta \mathbb{E}[\delta | G(t - \hat{t}) > \delta], && (20) \end{aligned}$$

where $EB(t_I)$ is the excess burden of a rate-equivalent perfectly perceived tax t_I. For small values of t_E, we will have that $EB(t_E) < EB(t_I)$. When t_E is high enough that $\Theta = 1$, we will have that $EB(t_E) > EB(t_I)$. Finally, equation (20) is always increasing in t_E. These three statements are sufficient to prove the following:

Proposition 6 (The Curse of Debiasing with Cognitive Costs)**.** *Assume the government has recourse to two taxes t_E and t_I and that*

1. *Consumers only notice t_I, $\hat{t} = t_I$*
2. *The budget adjustment rule is $\rho = \rho_3$.*

[35] Note that under the simplifying assumptions on ρ and $\hat{t}$ employed here, $\frac{\partial x^c}{\partial t_E} = 0$ by equation (8) (see also the discussion of equation (34) in the appendix). In a less restrictive setting, we would add $(1 - F(G))\frac{\partial x^c}{\partial t_E}$ to the right-hand-side of equation (17).

3. *There are no income effects on demand for x under full optimization.*

4. *Individual i debiases whenever $G(t_E) > \delta_i$.*

5. *Debiasing incurse a cognitive cost of δ_i.*

6. *The support of the distribution of cognitive costs $f(\delta)$ is bounded above.*

Then there will be a unique tax rate t'_E such that $t_E > t'_E$ if and only if the total excess burden of taxation is higher under an ignored tax t_E than under a rate-equivalent perceived tax t_I.

Figure 5 shows that in the case where everyone debiases at the same tax rate, moving the tax rate from just below the debiasing threshold for just above it causes total excess burden to jump discretely from zero to twice the excess burden of a rate-equivalent perfectly perceived tax. When low-salience taxes cause cognitive costs, the curse of debiasing affects not just marginal excess burden, but also average excess burden, because causing individuals to debias also causes them to inefficiently incur cognitive costs.[36]

6 The Example of Ironing

To illustrate that the model outlined above yields useful insights outside the relatively simple context of tax salience, this section describes in detail an application to what Liebman and Zeckhauser (2004) call ironing. Doing so provides new insights on the welfare implications of ironing and clarifying the analysis of Liebman and Zeckhauser. A more general lesson to be taken from this section is that one can always examine welfare effects so long as one has a mechanism for recovering preferences, but doing so requires requires careful attention to how preferences are revealed by choices, even in the context of a fully specified positive model.

Motivated by survey evidence suggesting that some individuals think their average tax rate is their marginal tax rate Liebman and Zeckhauser (2004) propose a model of "ironing." To address this issue, we first need to import the model to an income tax setting, such as Saez (2001). Denote taxable income by x, and assume utility is decreasing in x.[37] As in Saez (2001), y is now disposable income. The individual attempts to maximize utility $v(y) - u(x)$ subject to the budget constraint $y = (1 - \hat{\tau})x + Z$ where Z is now virtual income, income gained or lost from having different marginal tax rates at lower values of x than the chosen x. The price p is normalized to 1, reflecting the units of earnings. This process generates planned choices $\hat{x}$ and $\hat{y}$, like before. The budget adjustment rule then determines how x and y differ from $\hat{x}$ and $\hat{y}$.

Imagine we begin in the no-tax equilibrium, and then add a non-linear income tax. The individual reveals her preferences according to the change in compensated demand from a change in the perceived tax rate $\hat{\tau}$, i.e. $\partial x^c / \partial \hat{\tau}$,[38] which determines the denominator of the statistic θ^c.

Liebman and Zeckhauser assume that individuals use the third budget adjustment rule described in section 3, ρ_3. The individual accounts for the effect of taxes on (virtual) income but not on marginal prices.[39] For welfare calculations, Liebman and Zeckhauser also assume that income effects on x are negligible, which further simplifies

[36] It may be possible to generalize this finding under a more general rationality concept, assuming that whenever the individual chooses to change her choice rule, she incurs a cost equal to the utility gain from doing so.

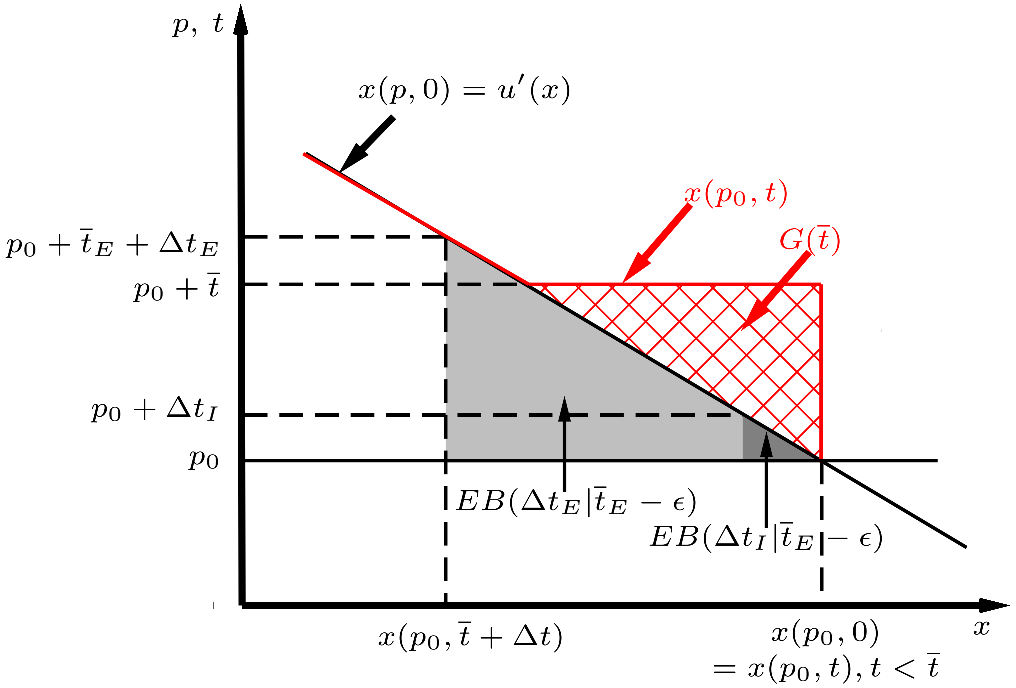
[37] In recent models examining the income tax, earnings are denoted by z rather than x. In this paper I always call the taxed behavior x for clarity.

[38] An identical alternative would be to continue to let the price vary and evaluate derivatives at $p = 1$.

[39] The main result in this section could also be derived from the equations characterizing ironing in Liebman and Zeckhauser (2004):

$$u'(x) = -(1 - ATR)v'(y)$$

Figure 5: Debiasing and Marginal Excess Burden



Notes: The figure is drawn assuming $t_I = 0$ initially. The tax demand curve is drawn in red, and is vertical until the individual debiases because we assume $\rho = 0$ and $\hat{t} = t_I$, so demand will not change at all as t_E increases, until debiasing occurs. Then, the individual moves discretely to the fully optimal choice, so the tax and price demand curves become the same curve. The marginal excess burden of increasing t_E just over the debiasing cutoff $\overline{t}_E$ is the area of the dark and light shaded areas, while the marginal excess burden of increasing t_I given the same existing tax rate for t_E is given by just the darker shaded triangle. The red triangle with diagonal hatching is the gain to optimizing at a tax rate of $\overline{t}$, which should be added onto excess burden calculations whenever $t_E > \overline{t}$ and debiasing incurs a cognitive cost commensurate with the utility gain from so doing.

the budget adjustment rule to $\rho = \rho_3 = 0$. I shall ignore debiasing for expository clarity–Liebman and Zeckhauser also do not include it)–but one could easily add it.

Proposition 7. *Assume that*

1. *Individuals engage in ironing, so $\hat{\tau} = ATR \equiv \int_0^x \tau(s)ds/x$.*
2. *The budget adjustment rule is $\rho = 0$.*
3. *There are no income effects on demand for x.*
4. *No individuals debias (see section 5.2).*

Then the excess burden of a non-linear tax $\tau(x)$ will be given by

$$\begin{aligned} EB(\tau(x)) &\simeq -\frac{1}{2}\tau^2 \frac{ATR}{\tau}\frac{\partial x}{\partial \tau} & (21)\\ &= \frac{1}{2}\tau ATR x_0 \frac{\varepsilon_{x,1-\tau|\tau}}{1-ATR} & (22)\\ &= \frac{1}{2}ATR^2 x_0 \frac{\varepsilon_{x,1-\tau|ATR}}{1-ATR} & (23)\end{aligned}$$

where $\varepsilon_{x,1-\tau|\tau} = \frac{\partial x}{\partial(1-\tau)}\frac{1-\tau}{x}$ is the elasticity of taxable income with respect to the net-of-tax rate, and $\varepsilon_{x,1-\tau|ATR} = \frac{\partial x}{\partial(1-ATR)}\frac{1-ATR}{x}$ is the elasticity of taxable income under full optimization, i.e. the elasticity of demand with respect to the net of tax rate following the introduction of a linear tax equal to ATR (which also equals the elasticity of x with respect to the price). Derivatives and elasticities are evaluated at the no-tax equilibrium $\tau = ATR = 0$.

Equation (21) highlights that in this context, $\theta = \frac{ATR}{\tau}$. When Liebman and Zeckhauser approximate the deadweight loss under ironing, they use equation (23), which resembles the Harberger-Browning approximation but with the "preferences" elasticity, which will not equal the "behavioral" elasticity for a non-linear tax. Proposition 7 indicates that using an estimate of the behavioral elasticity in this context, as Liebman and Zeckhauser do, is incorrect, because it uses the average tax rate where it should use the marginal tax rate. Such an approximation integrates over the "tax demand curve" instead of the "price demand curve" (see Panel I of Figure 4). This approximation does not account for the fact that the elasticity that reveals preferences (i.e. the elasticity with respect to the true marginal tax rate) is *different* from the elasticity that determines behavior (i.e. the elasticity with respect to perceived marginal tax rate). Liebman and Zeckhauser note the possibility of an error in approximation in a footnote in their paper. Proposition 7 tells us how to make the desired correction, which is simply to multiply their calculated excess burden by τ/ATR.

Using an elasticity of taxable income with respect to the marginal tax rate of 0.4, from Gruber and Saez (2002),[40] Liebman and Zeckhauser (2004) approximate the efficiency cost of US income taxation to be \$56.7 billion under

$$y = ATRx.$$

Note that the second equation is equivalent to $y = (1-\tau)x + Z$ where Z is virtual income. For a more detailed illustration of why these two equations are behaviorally equivalent to the model in Section 3 with $\rho = \rho_3$, even in the presence of a non-linear tax and the requirement that the individual correctly perceive her ATR, refer to the discussions of budget adjustment rules and ironing in the appendix.

[40] It is still not clear whether this estimate is appropriate. If individuals really iron, then the estimate of the elasticity of taxable income from Gruber and Saez (2002) is incorrect, because $\frac{d\log(x)}{d\log(1-\tau)}$ from some small change in the shape of the tax schedule is no longer a proper estimate of the elasticity of taxable income over the entire tax schedule. How an individual changes behavior from some change in the shape of the tax schedule is governed by how much the average tax rate changes, not the marginal tax rate. Furthermore, the change in the average tax rate will be different for individuals at different income levels. For this reason, calculations using estimates of the preferences elasticity are more practical in this situation.

ironing and \$109.0 if individuals behaved fully optimally. Applying the correction factor τ/ATR using the aggregate statistics from 2003 in Liebman and Zeckhauser[41] we obtain that the excess burden under ironing should is approximately \$77.4 billion. This does not affect the qualitative result that ironing reduces excess burden relative to full optimization—though that result does rely on assumptions about the budget adjustment rule and the size of income effects, which can be examined by a straightforward application of proposition 2. However, the finding here illustrates that the efficiency implications of misperception are sensitive to distinguishing between preferences elasticities and behavioral elasticities when quantifying welfare effects using data on suboptimal choices.

7 Conclusion

How economists should characterize optimal policy in the presence of suboptimal behavior is a central, unanswered question in behavioral economics. The analysis presented in this paper shows that one reason policy implications are slow to emerge is that detecting misperceptions alone provides little to no about optimal policy. The theory suggests that, in order to know how to design efficient tax policy in the presence of optimization errors, we also need to know the answers to the following empirical questions:

1. whether and how individuals debias when policy changes make mistakes more costly,
2. if so, whether debiasing incurs cognitive costs,
3. what budget adjustment rules individuals use, and
4. which mis-perceptions the most for behavior.

The first two points reflect the results in this paper that when behavioral biases initially make a tax socially desirable relative to fully perceived taxes, increasing our reliance on these types of taxes in an attempt to minimize inefficiency can backfire due to the curse of debiasing. I have shown how debiasing increases the efficiency cost of the tax, reduces the desirability of exploiting biases for social gain, and can also inefficiently impose cognitive costs on taxpayers. Whether individuals debias should be examined empirically, perhaps by examining how estimated biases change as tax rates change or when tax rates interact with other policies that make mistakes costly. Cognitive costs could be inferred by how much individuals are willing to pay to have their mis-perceptions corrected by an outside party such as a financial adviser.

Since the first draft of this paper appeared, two teams of authors have examined debiasing in an experimental setting. Feldman, Goldin and Homonoff (2015) found a null result; Taubinsky and Rees-Jones (2016) implemented a larger-scale experimental approach with considerably more statistical power and did find significant evidence of debiasing. Whether these phenomena can be found with real-world tax policies is still an open question, and, perhaps naturally, neither of these studies addresses the question of whether debiasing incurs cognitive costs.

[41] In this calculation, I use aggregate statistics on taxable income, revenue, and deadweight loss from Liebman and Zeckhauser, with a marginal tax rate of $\tau = 0.314$ and an average tax rate of $ATR = 0.230$ in the correction factor. These values match their approximations of total excess burden. The first is obtained by finding the marginal tax rate that generates Liebman and Zeckhauser's results for the deadweight loss under rational behavior, i.e. solving

$$EB(t) = \frac{1}{2}\tau^2 x \frac{\epsilon_{x,1-\tau}}{1 - ATR}$$

for τ using their aggregate statistics and deadweight loss approximation. The second is obtained by dividing aggregate taxable income \$4233.3 billion in 2003, by total income tax revenue \$974.7 billion. A more precise analysis would simulate the deadweight loss for a sample of taxpayers with varying incomes.

The third item above reflects that mis-perception alone does not fully determine behavior or utility without a budget adjustment rule. Given any mis-perception, it is possible to specify a budget adjustment rule for which the efficiency cost of taxation is either improved or worsened by the presence of mis-perceptions, relative to a perfectly perceived tax. For example, whether ironing really reduces the excess burden of a tax whenever average tax rates are less than marginal tax rates will depend on the budget adjustment rule. Furthermore, the budget adjustment rule used by biased individuals in a small scale experiment may be different from the budget adjustment rule used for large, permanent taxes. Budget adjustment rules could be examined empirically by observing the difference between the planned choice bundle based on misperception, as measured by surveys or financial planning, and the actual consumption bundle. One could also recover budget adjustment rules by comparing choices in a situation where some people notice a tax and some do not. Directly examining budget adjustment rules and applying the results in this paper would significantly expand our ability to obtain policy implications from a wide range of findings in behavioral economics.

The fourth item in the list above refers to the fact that for complex taxes like the US income tax, there are several candidates for behavioral biases with implications for welfare. These include the diminished salience of tax collected by withholding, mis-perception of the non-linearity of the tax schedule (e.g. ironing), ignorance of various tax expenditures or tax preferences, ignorance of the penalties for evasion, and reactions to the uncertainty created by frequent changes to tax rates. Improving policy to be mindful of behavioral biases would require understanding which mis-perceptions matter most, an important question for future research. Using a model like the one in this paper allows us to derive different empirical predictions for each of these biases, which could be tested.

Another way to think of the results is that it allows us to examine efficiency cost and optimal policy in situations where we are willing to make assumptions, perhaps with some empirical support, about misperceptions, budget adjustment and perhaps even debiasing, but where we lack the ideal dataset with which to estimate the compensated elasticities that are the sufficient statsitics of the Chetty, Looney and Kroft (2009) model. Given the frictionless uncompensated elasticity of demand (e.g. the price elasticity), the income elasticity, a specified misperception, a budget adjustment rule, and an assumption about debiasing, one can use the theory in this paper to recover efficiency cost without ever having to estimate the compensated elasticity of demand with respect to the tax rate. Estimating this compensated elasticity can be difficult because of issue with budget adjustment rules in field experiments or short-run analysis described above, and becuse the variation in tax rates necessary to do so may be unavailable. The theory above describes a way around these difficulties.

Finally, the insights from this analysis are not unique to the theory of taxation. Hidden incentives are a ubiquitous feature of the real world. How individuals meet their budget constraints when incentives are hidden is of substantive interest for characterizing economic behavior in general. The basic model of misperception used in this paper can be easily generalized to any misperception of prices. The parameterization of the budget adjustment rule employed here has useful features, namely that it is flexible and it interacts with well-understood demand elasticities in intuitive ways. Future work could use this kind of model of consumer behavior to examine thoroughly the normative consequences of misperception outside the world of taxation. In addition, a second instance of Gabaix and Laibson (2006) hints that the curse of debiasing is probably a very general concept, occurring whenever one agent attempts to exploit the biases of another, boundedly rational one.

References

Andreoni, James, Brian Erard, and Jonathan Feinstein. 1998. “Tax Compliance.” *Journal of Economic Literature*, 36(2): 818–860.

Auerbach, Alan. 1985. “The Theory of Excess Burden and Optimal Taxation.” In *Handbook of Public Economics*. Vol. 1, , ed. Alan Auerbach and Martin Feldstein. Elsevier.

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.

Cabral, Marika, and Caroline Hoxby. 2012. “The Hated Property Tax: Salience, Tax Rates, and Tax Revolts.” National Bureau of Economic Research Working Paper 18514.

Chetty, Raj. 2009. “The Simple Economics of Salience and Taxation.” National Bureau of Economic Research Working Paper 15246.

Chetty, Raj, Adam Looney, and Kory Kroft. 2007. “Salience and Taxation: Theory and Evidence.” National Bureau of Economic Research Working Paper 13330.

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

Chetty, Raj, John N Friedman, Soren Leth-Petersen, Torben Nielsen, and Tore Olsen. 2014. “Active vs. Passive Decisions and Crowdout in Retirement Savings Accounts: Evidence from Denmark.” *Quarterly Journal of Economics*, 123(3): 1141–1219.

Congdon, William, Jeffrey Kling, and Sendhil Mullainathan. 2011. *Policy and Choice: Public Finance through the Lens of Behavioral Economics.* Brookings Institution Press.

de Bartolome, Charles. 1995. “Which Tax Rate Do People Use: Average or Marginal?” *Journal of public Economics*, 56(1): 79–96.

Feldman, Naomi E, and Bradley J Ruffle. 2015. “The Impact of Including, Adding, and Subtracting a Tax on Demand.” *American Economic Journal: Economic Policy*, 7(1): 95–118.

Feldman, Naomi, Jacob Goldin, and Tatiana Homonoff. 2015. “Raising the Stakes: Experimental Evidence on the Endogeneity of Taxpayer Mistakes.” Unpublished Working Paper.

Feldman, Naomi, Peter Katuščák, and Laura Kawano. 2016. “Taxpayer Confusion over Predictable Tax Liability Changes: Evidence from the Child Tax Credit.” *American Economic Review*, 106(3).

Finkelstein, Amy. 2009. “E-Z Tax: Tax salience and tax rates.” *The Quarterly Journal of Economics*, 124(3): 969–1010.

Fischoff, Baruch. 1981. “Debiasing.” DTIC Document Unpublished working paper.

Fujii, Edwin, and Clifford Hawley. 1988. “On the Accuracy of Tax Perceptions.” *The Review of Economics and Statistics*, 344–347.

Gabaix, Xavier, and David Laibson. 2006. “Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets.” *The Quarterly Journal of Economics*, 121(2): 505–540.

Gallagher, Kelly, and Erich Muehlegger. 2011. “Giving Green to Get Green? Incentives and Consumer Adoption of Hybrid Vehicle Technology.” *Journal of Environmental Economics and Management*, 61(1): 1–15.

Galle, Brian D. 2009. “Hidden Taxes.” *Washington University Law Review*, 87: 59–114.

Gamage, David, and Darien Shanske. 2011. “Three Essays on Tax Salience: Market Salience and Political Salience.” *Tax L. Rev.*, 65: 19.

Goldin, Jacob, and Tatiana Homonoff. 2013. “Smoke Gets in Your Eyes: Cigarette Tax Salience and Regressivity.” *American Economic Journal: Economic Policy*, 5(1): 302–336.

Gruber, Jon, and Emmanuel Saez. 2002. “The Elasticity of Taxable Income: Evidence and Implications.” *Journal of Public Economics*, 84(1): 1–32.

Harberger, Arnold. 1964. “The Measurement of Waste.” *The American Economic Review*, 54(3): 58–76.

Homonoff, Tatiana. 2015. “Can Small Incentives Have Large Effects? The Impact of Taxes versus Bonuses on Disposable Bag Use.” Princeton University Unpublished Working Paper.

Hoopes, Jeffrey L, Daniel H Reck, and Joel Slemrod. 2015. “Taxpayer search for information: Implications for rational attention.” *American Economic Journal: Economic Policy*, 7(3): 177–208.

Kahneman, Daniel, Paul Slovic, and Amos Tversky. 1982. *Judgment Under Uncertainty.* Cambridge University Press.

Liebman, Jeffrey, and Richard Zeckhauser. 2004. “Schmeduling.” Harvard University Unpublished Working Paper.

McCaffery, Edward, and Jonathan Baron. 2006. “Thinking About Tax.” *Psychology, Public Policy & Law*, 12: 106–108.

Mohring, Herbert. 1971. “Alternative Welfare Gain and Loss Measures.” *Economic Inquiry*, 9(4): 349–368.

Ramsey, Frank P. 1927. “A Contribution to the Theory of Taxation.” *The Economic Journal*, 37(145): 47–61.

Saez, Emmanuel. 2001. “Using Elasticities to Derive Optimal Income Tax Rates.” *The Review of Economic Studies*, 68(1): 205–229.

Saez, Emmanuel. 2010. “Do Taxpayers Bunch at Kink Points?” *American Economic Journal: Economic Policy*, 2(3): 180–212.

Sahm, Claudia, Matthew Shapiro, and Joel Slemrod. 2010. “Household Response to the 2008 Tax Rebate: Survey Evidence and Aggregate Implications.” In *Tax Policy and the Economy*. Vol. 24, 69–110. The University of Chicago Press.

Schenk, Deborah. 2011. “Exploiting the Salience Bias in Designing Taxes.” *Yale Journal on Regulation*, 28: 253.

Sheffrin, Steven. 1994. “Perceptions of Fairness in the Crucible of Tax Policy.” *Tax Progressivity and Income Inequality. Cambridge*, 309–334.

Slemrod, Joel. 2006. “The Role of Misconceptions in Support for Regressive Tax Reform.” *National Tax Journal*, 59(1): 57.

Slemrod, Joel, and Shlomo Yitzhaki. 2002. “Tax Avoidance, Evasion, and Administration.” In *Handbook of Public Economics*. Vol. 3, , ed. Alan Auerbach and Martin Feldstein, 1423–1470. Elsevier.

Taubinsky, Dmitry, and Alex Rees-Jones. 2016. “Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment.” Unpublished Working Paper.

Thaler, Richard. 1980. “Toward a Positive Theory of Consumer Choice.” *Journal of Economic Behavior & Organization*, 1(1): 39–60.

Appendix to Chapter 1 (Preference Identification Under Inconsistent Choice)

A Proofs

Proof of Proposition 1

The proof of 1.1 is provided in the body of the paper.

To prove 1.2, let $\alpha = p(y_{i0} = 1;\ y_{i1} = 0)$ denote the fraction of frame defiers and note that now $p(c_i = 0) = p(y_{i1} = 1;\ y_{i0} = 0) + \alpha$. It is straightforward to show that

$$E[y_{i0}] = p(c_i = 1)E[y_i^*|c_i = 1] + \alpha \tag{1}$$

$$E[y_{i1}] = p(c_i = 1)E[y_i^*|c_i = 1] + p(c_i = 0) - \alpha \tag{2}$$

Substituting these into the definition of Y_c, we have

$$Y_c = \frac{p(c_i = 1)E[y^*|c_i = 1] + \alpha}{p(c_i = 1) + 2\alpha} \tag{3}$$

Subtract $\frac{1}{2}$ from both sides of (3) to obtain

$$Y_c - \frac{1}{2} = \frac{p(c_i = 1)(E[y^*|c_i = 1] - \frac{1}{2})}{p(c_i = 1) + 2\alpha} \tag{4}$$

In addition, subtracting $E[y^*|c_i = 1]$ from both sides of (3) yields

$$Y_c - E[y_i^*|c_i = 1] = \frac{\alpha(1 - 2E[y^*|c_i = 1]))}{2\alpha + p(c_i = 1)} \tag{5}$$

This expression gives the bias in Y_c when frame monotonicity fails. To complete the proof, note that by (4), $Y_c > \frac{1}{2} \implies E[y^*|c_i = 1] > \frac{1}{2}$, and then by (5), $E[y^*|c_i = 1] > \frac{1}{2} \implies E[y_i^*|c_i = 1] > Y_c$. This proves 1.2.i. Repeating these two steps with the inequalities reversed proves 1.2.ii. Finally, 1.2.iii follows directly from 4. ■

Proof of Proposition 2

By the law of iterated expectations, we can write:

$$E[y_i^*] = E[y_i^*|c_i = 1]p(c_i = 1) + E[y_i^*|c_i = 0]p(c_i = 0) \tag{6}$$

First we assume frame monotonicity to prove 2.1. In the proof of proposition 1.1 we showed that $E[y_i^*|c_i = 1]p(c_i = 1) = p(y_i^* = 1;\ c_i = 1) = Y_0$, and $p(c_i = 0) = Y_1 - Y_0 > 0$. Substituting these into (6) yields

$$E[y_i^*] = Y_0 + E[y_i^*|c_i = 0](Y_1 - Y_0) \tag{7}$$

Proposition 2.1 follows from the fact that (7) is strictly monotonic in $E[y_i^*|c_i = 0]$ and $E[y_i^*|c_i = 0] \in [0, 1]$.

To prove 2.2, note that by (1) and (2),

$$p(c_i = 1) = Y_0 + 1 - Y_1 - 2\alpha \tag{8}$$

$$E[y_i^*|c_i = 1] = \frac{Y_0 - \alpha}{Y_0 + 1 - Y_1 - 2\alpha} \tag{9}$$

Note that (9) implies that consistent preferences are point-identified when the prevalence of frame defiers, α, is known. Substituting (8) and (9) into (6) yields

$$E[y_i^*] = Y_0 - \alpha + E[y_i^*|c_i = 0](Y_1 - Y_0 + 2\alpha) \tag{10}$$

Because $p(c_i = 0) = Y_1 - Y_0 + 2\alpha \geq 0$, this expression is increasing in $E[y_i^*|c_i = 0]$.

For the lower bound, set $E[y_i^*|c_i = 0] = 0$: $E[y_i^*] \geq Y_0 - \alpha$. This expression is decreasing in α, so we obtain a lower bound with the maximum possible α. By (9), $E[y_i^*|c_i = 1] \leq 1$ implies $\alpha \leq 1 - Y_1$. It follows that $E[y_i^*] \geq Y_0 - (1 - Y_1)$, which is only binding when $Y_0 - (1 - Y_1) \geq 0$.

For the upper bound, set $E[y_i^*|c_i = 0] = 1$ in (10): $E[y_i^*] \geq Y_1 + \alpha$. This expression is increasing in α, so we obtain an upper bound by setting the maximum possible α. By (9), $E[y_i^*|c_i = 1] \geq 0$ implies $\alpha \leq Y_0$.[1] It follows that $E[y_i^*] \leq Y_0 + Y_1$, and this upper bound is binding whenever $Y_0 + Y_1 \leq 1$. ■

Proof of Lemma 1

Throughout the proofs of Lemma 1 and Proposition 3 we suppress the notation for conditioning on $w_i = w$, so that, for example, $E[y_i^*|w] \equiv E[y_i^*|w_i = w]$.

[1] Combining insights from these two cases, it follows that $\alpha \leq \min\{Y_0, 1 - Y_1\}$. Which of these two constraints is binding determines whether we obtain an upper or a lower bound for $E[y_i^*]$.

To prove L1.1, note that Bayes Rule implies

$$p(w_i = w|c_i = 1) = \frac{p(c_i = 1|w)}{p(c_i = 1)} p(w_i = w) \tag{11}$$

In the proof of Proposition 1.1, we showed that $p(c_i = 1) = Y_0 + 1 - Y_1$ under unconditional unconfoundedness (A3). Repeating the proof of Proposition 1 while conditioning on w, under conditional unconfoundedness (A3'), yields $p(c_i = 1|w) = Y_0(w) + 1 - Y_1(w)$. By the law of total probability, $p(c_i = 1) = E_w[Y_0(w) + 1 - Y_1(w)]$. Substituting these two expressions into (11) yields $p(w_i = w|c_i = 1) = q_w p_w$.

The proof that $p(w_i = w|c_i = 0) = s_w p_w$ is analogous. ■

Proof of Proposition 3

To prove 3.1, note that by the law of iterated expectations

$$E[y_i^*] = E_w[E[y_i^*|w]] \tag{12}$$

Repeating the proof of Proposition 1.1 while conditioning on w yields $E[y_i^*|c_i = 1, w] = Y_c(w)$. Conditional decision quality independence (A5) implies $E[y_i^*|w] = Y_c(w)$. Substituting this into (12) yields the desired result.

To prove 3.2, we begin by observing that

$$E[y_i^*|c_i = 0] = \frac{p(y_i^* = 1;\ c_i = 0)}{p(c_i = 0)}$$

Applying the law of iterated expectations to the numerator yields

$$E[y_i^*|c_i = 0] = \frac{E_w[p(y_i^* = 1;\ c_i = 0|w)]}{p(c_i = 0)} \tag{13}$$

By the definition of conditional probability, we know that for any w

$$p(y_i^* = 1;\ c_i = 0|w) = E[y_i^*|c_i = 0, w]\, p(c_i = 0|w)$$

As above, conditional decision quality independence (A5) implies $E[y_i^*|c_i = 0, w] = E[y_i^*|w] = Y_c(w)$. Equation (13) then implies[2] that

$$[y_i^*|c_i = 0] = E_w\left[\frac{p(c_i = 0|w)}{p(c_i = 0)} Y_c(w)\right]$$

Substituting the result from L1.2 that $\frac{p(c_i=0|w)}{p(c_i=0)} = p(w_i = w|c_i = 0) = s_w$ yields the desired

[2] The fact that $p(c_i = 0)$ is constant with respect to w allows us to move it inside the expectations operator in this expression.

result. ■

Proof of Proposition 4

By decision quality monotonicity (A7), we can divide the population into three groups based on (c_{ih}, c_{il}): the always consistent (A) with $c_{ih} = c_{il} = 1$, the sometimes consistent (S) with $c_{ih} = 1;\ c_{il} = 0$, and the never consistent (N) with $c_{ih} = c_{il} = 0$. Let π_j denote the share of the population in each group for $j = A, S, N$, and let $E[y_i^*|j]$ denote the fraction of each group preferring y^*.

For each fixed $z = z_k$, the conditions are identical to those in Proposition 1.1. Following the same logic as in the proof of Proposition 1.1, we have $Y_{0k} = p(y^* = 1; c_{ik} = 1) = E[y^*|c_{ik} = 1]\, p(c_{ik} = 1)$.

At each given z we know which groups are consistent, so we know that

$$Y_{0l} = E[y_i^*|A]\pi_A$$

$$Y_{0h} = E[y_i^*|A]\pi_A + E[y_i^*|S]\pi_S$$

$$Y_{0h} - Y_{0l} = E[y_i^*|S]\pi_S \tag{14}$$

We also showed in the proof of Proposition 1.1 that $p(c_{ik} = 1) = Y_{0l} + 1 - Y_{0h}$. It follows that

$$p(c_{il} = 1) = \pi_A = Y_{0l} + 1 - Y_{1l}$$

$$p(c_{ih} = 1) = \pi_A + \pi_S = Y_{0h} + 1 - Y_{1h}$$

$$\pi_S = Y_{1l} - Y_{0l} - (Y_{1h} - Y_{0h}) \tag{15}$$

Dividing (14) by (15) yields the desired result. ■

Proof of Proposition 5

Throughout the proof we let $\bar{c}(z) = E[c_{iz}]$. Fix any $D \in \mathbb{Z}^+$. Our technical assumptions – requiring $F(\underline{z}) = 0$, $F(\bar{z}) = 1$, 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, it has a well-defined Taylor Series approximation of degree D about any point in $(0, 1)$. Noting that $h(\bar{c}(z)) = E[y_i^*|z_i^* = F^{-1}(\bar{c}(z))] = E[y_i^*|z_i^* = z]$ proves 5.1.

Now note that the preferences of the consistent choosers at some z', $E[y_i^*|c_{iz'} = 1] =$

$E[y_i^*|z_i^* \leq z']$, can be expressed using the definition of conditional probability as

$$E[y_i^*|c_{iz'} = 1] = \frac{\int_{z=\underline{Z}}^{z=z'} g(z)f(z)dz}{F(z')}$$

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

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

Substituting our approximation for $h(\overline{c})$ into 5.1, evaluating the integral in the numerator, and dividing by $F(z') = \overline{c}(z')$ yields the desired result in 5.2.

The result in 5.3 follows from evaluating the expression in 6.2 at $\overline{c} = 1$. ■

B The Optimal Choice of Frame

This section motivates the parameters we focus on in the body of the paper by highlighting their relevance for setting the optimal policy. In this Section we derive formulas for the optimal frame as well as for the optimal decision-quality state (when the decision-quality state is a choice variable for the policymaker). The model we consider is simple but appealing in that the welfare conclusions are robust to a range of alternative positive models (in the spirit of Chetty (2009)).

We derive three results. First, when decision-makers' welfare depends solely on the option they end up selecting, the optimal frame depends solely on the preferences of the inconsistent decision-makers. Intuitively, the choice of frame does not affect the choices made by the consistent decision-makers and consequently the planner should ignore the preferences of that group when determining the optimal frame.

Second, when decision-makers experience normatively-relevant *opt-out costs* from choosing against the frame – for example by selecting an option other than the default – the preferences of the consistent decision-makers become relevant as well. In particular, when the planner's goal is to maximize the fraction of decision-makers that select their most-preferred option (e.g., when preference intensity is homogenous) and minimize the fraction incurring an opt-out cost, the optimal frame depends on the weighted average of consistent and inconsistent decision-makers' ordinal preferences, where the weights depend on the magnitude of the opt-out costs and the fraction of consistent choosers in the population.

Third, we consider the problem faced by a planner who must decide whether to adopt a (potentially more expensive) decision-quality state, i.e., one in which more decision-makers will choose consistently. We show that the social welfare benefits achieved by improving individuals' decision-making in this way depend on the difference in preferences between the decision-makers who are inconsistent at the high decision-quality state and the decision-makers who would be induced to choose consistently by the policy change. Intuitively, when this difference is large, a greater fraction of the sometimes-consistent decision-makers benefit from the increase in the decision-quality state and the social planner may provide the never-consistent with a better-tailored frame.

B.1 Setup

Assume a continuum population of measure 1 chooses from a fixed menu $X = \{0, 1\}$. A benevolent planner selects which option is favored by the frame, $d \in \{0, 1\}$. The decision quality state is given by $z \in \mathbb{Z}$, which could be fixed (if $\mathbb{Z}$ is a singleton set) or set by policy, with some associated implementation cost $\kappa(z)$.[3] We assume that the planner's objective is to maximize the probability that individuals choose their preferred option, and possibly also to minimize the probability that they choose the option not favored by the frame (due to the presence of opt-out costs, for example). For simplicity, the relative weight the planner attaches to these two objectives is described by a single parameter, γ. Formally, the planner's problem is

$$\max_{d\in\{0,1\},z\in\mathbb{Z}} \int_i \left[1\{y_{id} = y_i^*\} - \gamma 1\{y_{id} \neq d\} \right] di - \kappa(z)$$

where $1\{\}$ is an indicator function equal to 1 when the expression inside the brackets is true and zero otherwise. This simple objective function corresponds to the case in which the planner seeks to maximize social welfare with 1) homogeneous preference intensity, 2) homogeneous opt out costs, and 3) equal pareto weights across individuals. In the more general setting, heterogeneity in these variables is reflected in the solution to the optimal policy problem, but much of the intuition behind the characteristics of the optimal policy will still hold. These restrictions require that an agent ending up with her preferred option contributes equally to social welfare whether her preferred option is 0 or 1, and that opting out is equally costly for all decision-makers who opt out, regardless of their preference or the default. [4]

[3] The units of $\kappa(z)$ is the number of individuals the planner would need to give their preferred option to justify incurring a cost of $\kappa(z)$.

[4] With a different social welfare function, such as one that incorporates the intensity of each individual's preference for the available options, the parameters we identify will be insufficient to fully characterize optimal policy without additional restrictions. In general, identifying the intensity of individuals' preferences requires richer data and/or stronger positive assumptions than the type we employ here. In particular, implementing the solution to a utilitarian planner's problem would require estimating the joint distribution of (1) decision-makers' cardinal preferences and (2) the welfare cost to a decision-maker of choosing against the frame.

B.2 Results

We prove two simple propositions characterizing the solution to the planner's problem. The first considers the optimal frame when the decision-quality state is fixed. The second considers the joint choice of the optimal frame and the optimal decision-quality state, assuming for simplicity that there are no opt-out costs.

Proposition A.1

Suppose $\mathbb{Z}$ is singleton, $\mathbb{Z} = \{z\}$. Assume the consistency principle (A2) and frame monotonicity (A4).[5] *Let $\bar{c} = E[c_{iz}]$. Let $\rho = \frac{\gamma\bar{c}}{\gamma\bar{c}+1-\bar{c}}$. The optimal frame is d_1 if and only if*

$$(1-\rho)E[y_i^*|c_i = 0] + \rho E[y_i^*|c_i = 1] > \frac{1}{2} \tag{16}$$

Proof: Note that the planner's problem above is equivalent to the following

$$\max_{d\in\{1,0\},z\in\mathbb{Z}} p(y_{id} = y_i^*) - \gamma p(y_{id} \neq d) - \kappa(z)$$

Since z is fixed by assumption, the solution to the planner's problem simplifies to the comparison of the objective function evaluated at $d = 1$ and $d = 0$. We will therefore have that $d = 1$ is superior if and only if

$$\underbrace{p(y_{i1} = y^*)}_{1} - \underbrace{\gamma p(y_{i1} = 0)}_{2} > \underbrace{p(y_{i0} = y^*)}_{3} - \underbrace{\gamma p(y_{i0} = 1)}_{4} \tag{17}$$

We next derive each of these probabilities.

By frame monotonicity, the consistency principle, and the law of iterated expectations, terms one and three simplify to:

$$p(y_{i1} = y_i^*) = p(c_i = 1) + E[y_i^*|c_i = 0]p(c_i = 0)$$

$$p(y_{i0} = y^*) = p(c_i = 1) + E[1 - y_i^*|c_i = 0]p(c_i = 0)$$

By frame monotonicity, terms two and four simplify to:

$$\gamma p(y_{i1} = 0) = \gamma p(y_{i1} = y_{i0} = 0) = \gamma E[1 - y_i^*|c_i = 1]p(c_i = 1)$$

$$\gamma p(y_{i0} = 1) = \gamma p(y_{i1} = y_{i0} = 1) = \gamma E[y^*|c_i = 1]p(c_i = 1)$$

[5] The assumption of frame separability is embedded in the planner's problem. Unconfoundedness is necessary for the statistical estimation of the relevant preference parameters from data, but not for understanding the relationship between these parameters and optimal policy, our focus here.

Combining terms and simplifying, we have that $d = 1$ is optimal if and only if

$$E[y_i^*|c_i = 0]\frac{p(c_i = 0)}{\gamma p(c_i = 1) + p(c_i = 0)} + E[y_i^*|c_i = 1]\frac{\gamma p(c_i = 1)}{\gamma p(c_i = 1) + p(c_i = 0)} > \frac{1}{2}$$

Substituting the definitions of ρ and $\bar{c}$ yields the desired result. ■

Discussion of Proposition A.1 The optimal frame will tend toward $d = 1$ when (1) the fraction of inconsistent choosers who prefer $y = 1$ is large, and (2) the fraction of consistent choosers who prefer $y = 1$ is large. The first group is helped by the frame being $d = 1$, because the default directly influences their choices. The second group is helped by the frame being $d = 1$, because they will not incur an opt-out cost to receive their preferred option. The left-hand side of (16) shows that the relative importance of the consistent group is determined by a weight ρ, which is increasing in opt-out costs γ and the size of the consistent subgroup ($\bar{c}$). When choosing against the frame is assumed to have no normatively relevant cost, the optimal policy depends only on the preferences of inconsistent choosers. To see this, note that when $\gamma = 0$, $\rho = 0$, so the condition for optimality of $d = 1$ simplifies to $E[y_i^*|c_i = 0] > \frac{1}{2}$.

Finally, note that the optimal policy does not depend on γ when the majority of consistent and inconsistent decision-makers prefer the same option. The magnitude of the opt-out costs only matters when consistent and inconsistent choosers are better off under different defaults. Otherwise the default that minimizes opt-outs, as recommended by Thaler and Sunstein (2003), will tend to be socially optimal. In contrast, when inconsistent and consistent choosers have different preferences (which requires that decision-quality independence fails) and the inconsistent choosers preferences dominate (i.e. ρ is small), minimizing opt-outs will lead to a sub-optimal policy.

Proposition A.2

Suppose the planner can choose between two decision quality states, $\mathbb{Z} = \{z_h, z_l\}$. Suppose $\kappa(z^h) > \kappa(z^l)$, and let $\Delta\kappa = \kappa(z^h) - \kappa(z^l)$ be the change in per-person cost of increasing the decision-quality state from z_l to z_h. Assume the consistency principle (A2), frame monotonicity (A4), decision-quality exclusivity (A6), decision quality monotonicity (A7), and that $\gamma = 0$. Then the solution to the planner's problem is given case-wise by

1. $(1, z_l)$ *if*

   (a) $Y_N^* > \frac{1}{2}$ *and* $\frac{Y_N^* \pi_N + Y_S^* \pi_S}{\pi_N + \pi_S} > \frac{1}{2}$, *and* $\Delta\kappa > (1 - Y_S^*)\pi_S$, *OR*

   (b) $Y_N^* > \frac{1}{2}$ *and* $\frac{Y_N^* \pi_N + Y_S^* \pi_S}{\pi_N + \pi_S} < \frac{1}{2}$ *and* $\Delta\kappa > Y_S^* \pi_S + (2Y_N^* - 1)\pi_N$

2. $(0, z_l)$ *if*

   (a) $Y_N^* < \frac{1}{2}$ *and* $\frac{Y_N^*\pi_N + Y_S^*\pi_S}{\pi_N+\pi_S} < \frac{1}{2}$ *and* $\Delta\kappa > Y_S^*\pi_S$*, OR*

   (b) $Y_N^* < \frac{1}{2}$ *and* $\frac{Y_N^*\pi_N + Y_S^*\pi_S}{\pi_N+\pi_S} > \frac{1}{2}$ *and* $\Delta\kappa > \pi_S(1 - Y_S^*) + (1 - 2Y_N^*)\pi_N$,

3. $(0, z_h)$ *if*

   (a) $Y_N^* < \frac{1}{2}$ *and* $\frac{Y_N^*\pi_N + Y_S^*\pi_S}{\pi_N+\pi_S} < \frac{1}{2}$ *and* $\Delta\kappa < Y_S^*\pi_S$*, OR*

   (b) $Y_N^* > \frac{1}{2}$ *and* $\frac{Y_N^*\pi_N + Y_S^*\pi_S}{\pi_N+\pi_S} < \frac{1}{2}$ *and* $\Delta\kappa < Y_S^*\pi_S + (2Y_N^* - 1)\pi_N$

4. $(1, z^h)$ *if*

   (a) $Y_N^* > \frac{1}{2}$ *and* $\frac{Y_N^*\pi_N + Y_S^*\pi_S}{\pi_N+\pi_S} > \frac{1}{2}$ *and* $\Delta\kappa < (1 - Y_S^*)\pi_S$*, OR*

   (b) $Y_N^* < \frac{1}{2}$ *and* $\frac{Y_N^*\pi_N + Y_S^*\pi_S}{\pi_N+\pi_S} > \frac{1}{2}$ *and* $\Delta\kappa < (1 - Y_S^*)\pi_S + (1 - 2Y_N^*)\pi_N$

where $Y_N^* \equiv E[y_i^*|c_{ih} = c_{il} = 0]$, $Y_S^* \equiv E[\phi_i|c_{ih} > c_{il}]$, $\pi_N = 1 - E[c_{ih}]$, *and* $\pi_S = E[c_{ih}] - E[c_{il}]$.

Proof: When $\gamma = 0$ the planner's objective evaluated at each of the four possible d by z_k combinations is

$$p(y_{idk} = y_i^*) - \kappa(z_k)$$

In the proof of Proposition A.1, we showed that these four expressions can be re-written for fixed z_k as

$$d = 1: \quad p(c_{ik} = 1) + E[y_i^*|c_{ik} = 0]p(c_{ik} = 0) - \kappa(z_k) \tag{18}$$

$$d = 0: \quad p(c_{ik} = 1) + E[1 - y_i^*|c_{ik} = 0]p(c_{ik} = 0) - \kappa(z_k) \tag{19}$$

By decision-quality monotonicity and the existence of consistent choosers, we can divide the population into always-consistent (A), never-consistent (N) and sometimes-consistent (C), exactly as in Propositions 4 and A.1. The fraction preferring $y = 1$ in each population are given by Y_A^*, Y_N^*, and Y_S^*, respectively, and the size of each population is given by π_A, π_N, and π_S, respectively.

Using this notation, $p(c_{il} = 1) = \pi_A$, $p(c_{ih} = 0) = \pi_N$, $p(c_{il} = 0) = 1 - \pi_A = \pi_N + \pi_S$. By the law of iterated expectations and the definition of conditional probability,

$$E[y_i^*|c_{il} = 0] = \frac{Y_S^*\pi_S + Y_N^*\pi_N}{\pi_S + \pi_N}$$

$$E[1 - y_i^*|c_{il} = 0] = \frac{(1 - Y_S^*)\pi_S + (1 - Y_N^*)\pi_N}{\pi_S + \pi_N}$$

Using these expressions to simplify equations (18) and (19) for both values of z_k yields:

$$d = 1, z = z_l : \quad \pi_A + Y_S^* \pi_S + Y_N^* \pi_N - \kappa(z^l) \tag{20}$$

$$d = 0, z = z_l : \quad \pi_A + (1 - Y_S^*)\pi_S + (1 - Y_N^*)\pi_N - \kappa(z^l) \tag{21}$$

$$d = 1, z = z_h : \quad \pi_A + \pi_S + Y_N^* \pi_N - \kappa(z^h) \tag{22}$$

$$d = 0, z = z^h : \quad \pi_A + \pi_S + (1 - Y_N^*)\pi_N - \kappa(z^h) \tag{23}$$

Note that the first three terms in each of these will be the total number of individuals who receive their preferred option when the planner chooses that (d, z) combination.

First, consider situations where the planner chooses $d = 1$ regardless of z. This requires (20)>(21) and (22)>(23), which simplify to the first two conditions in (1a) and (4a). The planner will set z_h if (22)>(20), which simplifies to $\Delta\kappa < (1 - Y_S^*)\pi_S$, which yields (4a). With the inequality reversed, we get (1a).

Second, consider situations where the planner chooses $d = 0$ regardless of z. This requires (20)<(21) and (22)<(23), which simplify to the first two conditions in (2a) and (3a). Then the planner chooses z_h if (23)>(21), which simplifies to $\Delta\kappa < Y_S^* \pi_S$. This yields the final condition in (2a) and (3a).

Third, consider the situation where the planner would want to choose $d = 1$ under z_h and $d = 0$ under z_l. This requires (20)<(21) and (22)>(23), which provides the first two conditions in (1b) and (3b). In this situation, the planner chooses z^h if (22)>(21) and z^l otherwise. Performing this comparison, we have that the planner chooses z^h if $\Delta\kappa < Y_S^* \pi_S + (2Y_N^* - 1)\pi_N$, which is the final condition in (3b). When the inequality is reversed, we obtain the final condition in (1b).

Finally, consider the situation where the planner would want to choose $d = 0$ under z_h and $d = 1$ under z_l. This requires (20)>(21) and (22)<(23), which provide the first two conditions in (2b) and (4b). In this situation, the planner chooses z_h if (23)>(20). Comparing these, we see that the planner chooses z^h if $\Delta\kappa < (1-Y_S^*)\pi_S + (1-2Y_N^*)\pi_N$,which is the final condition in (4b). When the inequality is reversed, we obtain the final condition in (2b). ■

The planner should switch to z_h from z_l if the number of individuals who receive their preferred option increases by enough to justify the increase in implementation cost $\Delta\kappa$. We divide the cases first based on two possibilities for the solution to the problem: the optimal choice of frame either depends on the choice of decision-quality environment or it does not. When the optimal choice of frame does not depend on z, switching to z_h from z_l helps only those individuals who are consistent at z_h but not z_l (group S), and who prefer the option not associated with the frame. In other words, when the never-consistent choosers prefer the same frame as the sometimes- and never-consistent choosers together, the planner should only take into account how many sometimes-consistent choosers will receive their preferred option under the improved decision-quality state. Parts (1a), (2a), (3a), and (4a) correspond to these possibilities. In the second set of possibilities, the optimal frame changes as the planner increases from z_l to z_h. This case obtains if the individuals who become consistent at z_h (group S) have different preferences from the group who never choose consistently (group N). Here, moving from z_h to z_l not only gives individuals in group S who prefer the non-framed option their preferred option, but it also allows the planner to set a better default for group N. For example, suppose the planner would want to set $d = 0$ in z_l but $d = 1$ in z_h. This corresponds to Parts (1b) and (3b) of the proposition. Then we must have that $Y_S^* < \frac{1}{2}$, $Y_N^* > \frac{1}{2}$, and the preferences of the S group dominate when determining optimal policy under z_l, which occurs if there are more of them or their preferences are more homogenous. In this case the benefit of switching to z_h includes not only the benefit of giving those in group S who prefer the non-framed option their preferred option, but also the benefit of setting a frame in accordance with the preferences of the remaining group who are never consistent, group N. How large this benefit is depends on how far Y_N^* is from $\frac{1}{2}$ (i.e. how bad the z_l frame was for this group) as well as the size of group N.

C Alternative Positive Models of Default Sensitivity

In this Appendix we consider the relationship between our framework and positive models of framing effects. In many cases, a range of alternative models of framing effects will be observationally equivalent given the available data, meaning that any one of them might explain decision-makers' observed sensitivity to the frame in question. As emphasized by Bernheim (2009), such model uncertainty poses a challenge for welfare analysis, since the preferences implied by a decision-maker's observed choice behavior depend on the positive model of behavior that maps the decision-maker's preferences into his or her choices. An important advantage of our approach is that it can shed light on decision-makers' preferences without specifying the exact positive model of behavior that generates the observed framing effect, at least within the class of positive models consistent with the assumptions set out in the body of the paper (such as frame monotonicity and the consistency principle).

To illustrate, we consider the application of our framework to alternative behavioral models that have been proposed to explain why decision-makers exhibit sensitivity to default effects. An important lesson of this exercise is that our reduced-form framework is not a replacement for structural modeling. That is, although the framework is sufficiently general to accommodate many of the models we consider, the specific positive model shapes the interpretation of the reduced-form parameters that our proposed approaches recover and has important implications for which of the proposed approaches are likely to succeed in a given setting.

In each model, we assume that every decision-maker (DM) i chooses from a fixed menu $X = \{0, 1\}$. DM's valuations of the two options are given by $u_i(0)$ and $u_i(1)$, with the difference denoted by $\bar{u}_i = u_i(1) - u_i(0)$. We assume that the distribution of $\overline{u}_i$ is described by cumulative density function $F(.)$.[6] We denote the default by $d \in \{d_0, d_1\}$, where the subscript indicates which option is the default. We continue to denote $y_i^* = 1\{u_i(y) > u_i(x)\}$ and $c_i = 1\{y_{i1} = y_{i0}\}$.

C.1 Inattention

One reason why people might be drawn to the default option is that they do not pay attention to the other menu items. Here, we develop a model of default sensitivity based on inattention, using the approach of Masatlioglu, Nakajima and Ozbay (2012).

C.1.1 Setup

In the model of Masatlioglu, Nakajima and Ozbay (2012), one assumes that an individual pays attention only to some subset of the menu X, but that she maximizes her preferences over the alternatives that she considers.[7] In order to incorporate framing effects, we must specify an *attention filter* Γ which depends on d. The attention filter Γ is a mapping from (X, d) to a subset of X, $\Gamma(X, d) \subseteq X$.[8] Given a utility function representation of individual i's preferences, $u_i(.)$, we can write the consumer's choice as the solution to the utility-maximization problem restricted to $\Gamma_i(X, d_i)$.

$$y_i(d_i) = \arg \max_{y \in \Gamma_i(X, d_i)} u_i(y) \tag{24}$$

Claim: When X is binary, frame monotonicity and the consistency principle will be satisfied if the individual always pays attention to the default option. Formally, the sufficient condition

[6] Our focus in the body of the paper is only with decision-makers' ordinal preferences, but in some of the models considered here, differences in preference intensity will explain some of the variation in decision-makers' consistency.

[7] Instead of reflecting inattention to a subset of the menu, an alternative possibility for why some decision-makers tend to select the default option is that they follow the default as a decision-making heuristic, or shortcut. The model is formally equivalent to the limited attention model for binary menus, but the two models depart with more than two options. That is, the heuristic model is a special case of the limited attention model where the attention filter consists solely of a single option (whichever option happens to be the default).

[8] Masatlioglu et al.'s key assumption is that $\forall X$, $\Gamma_i(X, d) = \Gamma_i(X \backslash x, d)$ whenever $x \notin \Gamma_i\{X, d\}$. This assumption is not directly relevant to our setting because our focus is on binary choices. However, in the non-binary case it will place additional restrictions on when preferences are revealed by choices.

is

$$\forall i, \forall j \in \{0,1\} \; j \in \Gamma_i(X, d_j) \tag{25}$$

Proof: Suppose condition (25) is satisfied. Recall that frame monotonicity is only violated when $(y_{i0}, y_{i1}) = (1,0)$. Suppose that $y_{i0} = 1$. Then $1 \in \Gamma_i\{x, d_0\}$ by (24). By (25), $0 \in \Gamma_i\{x, d_0\}$. By (24) again then, $u_i(1) > u_i(0)$. Finally, $1 \in \Gamma_i\{x, d_1\}$ by (25), so we know that $y_{i1} = 1$. This $y_{i0} = 1 \implies y_{i1} = 1,\ c_i = 1$, and $y_i^* = 1$. Using a similar set of steps, we know that $y_{i1} = 0 \implies y_{i0} = 1,\ c_i = 1,\ y_i^* = 0$. ■

Intuitively, when DMs choose $y = 1$ under $d = 0$, they are "revealing" that they pay attention to $y = 1$ under default d_0, since a DM cannot choose an alternative not in the attention set $\Gamma(.)$. Given the assumption that all DMs also pay attention to the default option, by choosing $y = 1$ a DM reveals that she prefers y to x. Thus, when inattention drives default effects and individuals always pay attention to the default option, the assumptions underlying our reduced-form approach will obtain.[9] In related work on stochastic choice, Manzini and Mariotti (2014) make an assumption analogous to Equation (25) to model DMs who (with some probability) may or may not consider options besides the default.

In this model, variation in whether decision-makers are consistent is governed by variation in Γ_i. The relationship between preferences and consistency is thus governed by the relationship between u_i and Γ_i. In the next two sub-sections, we consider two alternative possibilities for what might drive variation in Γ_i and what each implies for the relationship between consistency and preferences.

C.1.2 Heterogeneous Attention Costs

Here we consider a model in which paying attention to the non-default option requires the DM to incur some utility cost $k_i \geq 0$. We assume decision-makers fail to consider the non-default option whenever the cost of paying attention to the non-default option, k_i exceeds a threshold value $\overline{k}$. For example, k_i may reflect the decision-maker's cognitive ability, other demands on his or her attention, or prior experience with the choice being made. A decision-maker is consistent, so $c_i = 1$, if and only if $k_i < \overline{k}$.

Assuming that k_i is distributed in the population with a cumulative distribution function $G(.)$, we will have that $E[c_i] = G(\overline{k})$. Whether decision-quality independence holds in this model depends on the empirical correlation between the determinants of decision-makers'

[9] One could imagine extending this approach to incorporate additional data. Note that, like frame monotonicity, property (25) could be tested if one is able to observe an individual's choice across multiple frames, or if one could observe attention directly (such as by interviewing decision-makers after their choice or by employing an eye scanner). In addition, note that when we move beyond the binary case, assumption (25) would justify the assumption that active choices reveal a preference for the chosen option over the default option, but not the stronger assumption that active choices reveal preferences over the entire menu.

attention costs, k_i, and their preferences, represented by $\overline{u}_i$. In particular, we will have $cov(c_i, y_i^*) = 0 \iff p(k_i < \overline{k}; \bar{u}_i > 0) = p(k_i < \overline{k})p(\bar{u}_i > 0)$. Thus a sufficient condition for decision-quality independence is if k_i is distributed independently of $\overline{u}_i$.

When k_i and $\bar{u}_i$ are correlated, this correlation should be taken into account to estimate the distribution of preferences in the population. There are two empirical strategies one might employ. The first is to collect data on variables likely to be highly correlated with k_i and $\bar{u}_i$ and implement a matching-on-observables approach. Our reasoning in the application in Section 4 was similar: younger, lower-income workers may pay less attention to savings decisions because retirement is further off, and they may prefer to save less because they will make more later or are currently paying down debt. When these are the dominant factors determining attention and preferences, matching on these observable characteristics yields a credible identificaiton strategy.

A second strategy would be to use decision-quality instruments to exogenously increase k_i or decrease $\bar{k}$. For example, providing decision-makers with practice making similar decisions or manipulating the time required to attend to other available options are natural candidates for decision-quality instruments.

With variation in a decision-quality instrument, this model begins to resemble the one in Section 3.3.2 of the paper. With a joint normal distribution for $\bar{u}_i$ and $log(k_i)$ and a homogeneous effect of a change in the decision-quality instrument z on $log(k_i)$, this becomes identical to the latent variable model outlined in that section, so we can trace out the propensity to optimize as a function of the fraction of optimizers at a given level of z, and extrapolate to recover the full distribution of $\bar{u}_i$ and k_i. One can naturally imagine similar models relying on weaker functional form assumptions.

C.1.3 Stakes-Based Attentiveness

In this model, the DM decides whether to pay attention by comparing the cost of doing so against the benefit they stand to gain. In making this decision, the agent knows the absolute value of the difference in utility between the available options, but not which option has the higher utility. That is, agents know the utility amount at stake, but not which option they prefer. For example, consider an employee selecting a retirement savings plan. The employee may know how much selecting the right retirement savings plan matters to her and how costly it is to learn about the menu of plans, but she may not actually know which plan is best for her without incurring utility costs from making the comparison.

Assume the DM knows the utility "stakes" of the meta-decision of whether to consider the non-default option, $|\bar{u}_i|$, and must decide whether to incur the cost of attention, k_i.[10] As

[10] If agents could not account for the amount at stake, this model would reduce to the previous one in

above, we assume that individuals who pay the cost select their most-preferred option under both defaults whereas individuals who do not simply select the default.

Suppose that decision-makers believe (ex-ante) that $y = 0$ is best with probability ω_i. Consider the DM's problem when the default is d_0. If she pays attention, with probability ω she will end up staying with $y = 0$ and with probability $1 - \omega$ she will discover that she prefers $y = 1$ (i.e., that $\bar{u} > 0$) and pick $y = 1$. Whichever option she prefers, she will incur the attention cost C. If she doesn't pay attention, she will end up with $y = 0$ with certainty. Thus the net utility gain to paying attention when the default is d_0 is given by $(1 - \omega) |\bar{u}| - C$. The decision-maker will choose to pay attention to both options under the following condition:

$$\Gamma_i(X, d_0) = X \iff (1 - \omega_i) |\bar{u}_i| - k_i > 0$$

Similarly, when the default is d_1, the DM will choose to consider the non-default option when

$$\Gamma_i(X, d_1) = X \iff \omega_i |\bar{u}_i| - k_i > 0$$

Thus for an agent to choose to consider both options under both frames, it must be the case that

$$\Gamma_i(X, d_0) = \Gamma_i(X, d_1) = X \implies \frac{k_i}{|\bar{u}_i|} < \omega < 1 - \frac{k_i}{|\bar{u}_i|}$$

Note that the condition is guaranteed to fail when $k_i \geq |\bar{u}_i|$ (since $\omega \in [0, 1]$), i.e. when the cost of attentiveness exceeds the potential benefits. Note that we can also write the above condition for attentiveness in both frames as $k_i < \min\{(1 - \omega_i) |\bar{u}_i|, \omega_i |\bar{u}_i|\}$, which highlights that a DM with sufficiently low k_i will always consider both options.

Because this model is a special case of the general inattention model described above, we know that frame monotonicity and the consistency principle will be satisfied. Note though that although DMs who are fully attentive under both defaults will be consistent, the converse is not true. For example, a DM who prefers $y = 1$ and has $\omega < \min\left\{\frac{|\bar{u}|}{C}, 1 - \frac{|\bar{u}|}{C}\right\}$ will choose $y = 1$ consistently, even though she follows the default under d_1 and only considers both options under d_0.

In general, decision-quality independence will not hold in this setting. Even when attention costs k_i are orthogonal to preferences, $|\bar{u}|_i$ may nonetheless be correlated with both. That is, individuals with high $|\bar{u}_i|$ will be more likely to pay attention to both options (and thus to choose consistently) and also may be more likely to prefer one option to the other (i.e., preference intensity may be correlated with ordinal preferences). However, conditional decision-quality independence may still hold when one can observe sufficient characteristics

which heterogeneity in attentiveness depended solely on individual characteristics. If agents knew the precise utility gain that they would achieve by considering the non-default option, the model would become formally identical to the costly opt-out model considered below.

to control for both k_i and $|\bar{u}_i|$, in which case matching on observables will allow us to recover population preferences. That is, under stakes-based attention models, the observer should control for variation among decision-makers associated with the costs of attention and the perceived utility stakes in the underlying decision. In the 401(k) application in Section 4, we know that income is highly correlated with the value of deferred tax on pension contributions, so conditioning on income should help control for variation in the stakes of the decision at hand. One could also solicit and control for 1) the individual's knowledge of the definitions of various aspects of retirement plans, and 2) for the self-reported importance of the savings decision. Similarly, valid decision-quality instruments will consist of variation in the choice environment that affects the cost of attention *or* the perceived stakes of the decision monotonically for all individuals. For example, a researcher might emphasize the importance of the decision to some participants in an experimental intervention.

C.2 Costly Opt-Outs

In this model, we assume a DM can incur a perceived utility cost $\gamma_i \geq 0$ in order to choose an option that is not the default. Here we assume that the utility cost is neoclassical, such as an administrative fee for opting into an alternative retirement plan. Although we initially focus on the case in which γ_i reflects a real cost (i.e., the perceived cost equals the true cost), this formal model also captures "as if" transaction costs that nonetheless shape decision-makers' behavior, as explored in Appendix C.3 below. In independent work, Bernheim, Fradkin and Popov (2015) develop a model like this one, with a little more structure on the distribution of u_i and γ_i, to estimate the distribution of privately optimal contribution rates to a 401(k) plan and opt-out costs. Because we are primarily focused on the relationship between positive models and the identification strategies described in the body of our paper, however, our discussion of the implications of these models is largely orthogonal to their work,

A defining feature of this class of model is that the decision-maker knows the potential utility gain from choosing the non-default option and selects it if and only if the benefit from doing so exceeds the perceived opt-out cost. DM's choice is thus given by

$$y_i(d) = \arg\max_{y \in X} u_i(y) - \gamma_i 1\{c_i \neq d\}$$

When $d_i = d_0$, the solution to this problem is given by $y_{i0} = 1 \iff \bar{u}_i > \gamma_i$. When $d = d_1$, the solution is given by $y_{i1} = 1 \iff -\bar{u}_i < \gamma_i$. We can summarize the three distinct possibilities for the choices of individual i as follows:

$$(y_{i0}, y_{i1}) = \begin{cases} (0,0) & \text{if } -\bar{u}_i > \gamma_i \\ (0,1) & \text{if -}\bar{u}_i < \gamma_i,\ \bar{u}_i < \gamma_i \\ (1,0) & \text{if } \bar{u}_i > \gamma_i \end{cases} \tag{26}$$

From (26), it is straightforward to verify that the consistency principle and frame monotonicity will hold. The two statistics studied in our paper will be given in this model by

$$y_i^* = 1 \iff \bar{u}_i > 0$$

$$c_i = 1 \iff |\bar{u}_i| > \gamma_i$$

When transaction costs are homogenous, $\gamma_i = \gamma\ \forall i$, the average (ordinal) preferences of the consistent decision-makers is given by: $E[y_i^*|c_i = 1] = P\left(\overline{u}_i > 0 \,|\, \overline{u}_i \in (-\infty, -\gamma_i] \cup [\gamma_i, \infty)\right)$, or

$$E[y_i^*|c_i = 1] = \frac{1 - F(\gamma)}{1 - F(\gamma) + F(-\gamma)}$$

Similarly, for the inconsistent decision-makers we have $E[y_i^*|c_i = 0] = P\left(\overline{u}_i > 0 \,|\, \overline{u}_i \in (-\gamma, \gamma)\right)$, or

$$E[y_i^*|c_i = 0] = \frac{F(\gamma) - F(0)}{F(\gamma) - F(-\gamma)}$$

Note that heterogeneity in decision-makers' consistency in this model is driven by heterogeneity in the intensity of their preferences as well as the size of their transaction costs. Consequently, decision-quality independence will not generally be satisfied:[11]

$$cov(y_i^*, c_i) = p(\bar{u}_i > \gamma_i) - p(\bar{u} > 0)p(\bar{u}_i < -\gamma_i \text{ or } \bar{u}_i > \gamma_i)$$

which will equal zero if and only if the distribution of preferences happens to satisfy $p(\overline{u}_i > \gamma_i|\overline{u}_i > 0) = p(\overline{u}_i < -\gamma_i|\overline{u}_i < 0)$. That decision-quality independence usually fails here is not surprising: whether an individual is consistent in this model depends strongly on her preferences.

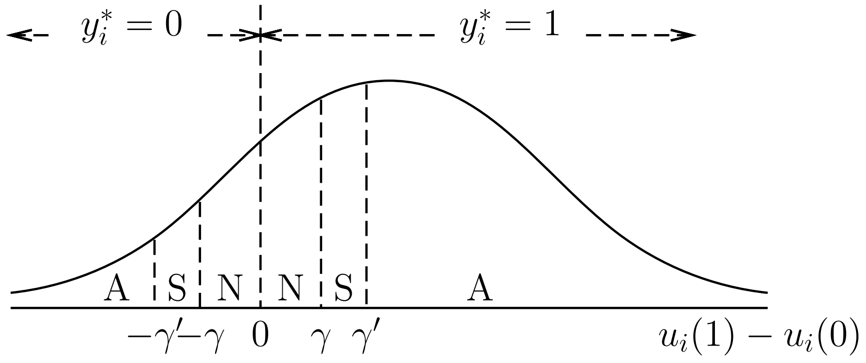
Nevertheless, additional structure can make the statistics on ordinal preferences studied in the body of the paper sufficient for optimal policy. Consider a utilitarian social planner choosing $d \in \{d_0, d_1\}$ to maximize

$$W(d) = \int_i u_i(y(d)) - \rho\gamma_i 1\{y_i(d) \neq d\} di.$$

where $\rho \in [0, 1]$ governs the normative relevance of "as-if" opt out costs. One can show that when 1) $\bar{u}_i$ and γ_i are independent and 2) the distribution of $\bar{u}_i$ is single-peaked and symmetric, that the optimal default for a utilitarian social planner is $d = 1$ if and only if

[11] $cov(y_i^*, c_i) = E[y_i^* c_{ii}] - E[y_i^*]E[c_{ii}] = p(y_i^* = c_{ii} = 1) - p(y_i^* = 1)\, p(c_{ii} = 1)$.

Figure 6: Sometimes Consistent Choosers in the Costly Opt-out Model



$E[y_i^*|c_i = 1] > \frac{1}{2}$.[12] Moreover, these sufficient conditions for $E[y_i^*|c_i = 1] > \frac{1}{2}$ to determine the optimal policy obtain regardless of whether γ is a real utility cost that enters the planner's objective or merely a normatively irrelevant "as-if" cost, i.e. it holds for any $\rho \in [0, 1]$.

Decision-quality instruments are immensely useful in this model, both for accounting for selection into the consistent subgroup and for identifying preference intensity. 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 transactions 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 and we will have:

$$(y_{i0}(\gamma), y_{i0}(\gamma'), y_{i1}(\gamma'), y_{i1}(\gamma)) = \begin{cases} (0,0,0,0) & \text{if } \bar{u}_i < -\gamma_i \\ (0,0,0,1) & \text{if } \bar{u}_i \in [-\gamma_i, -\gamma_i'] \\ (0,0,1,1) & \text{if } \bar{u}_i \in [-\gamma_i', \gamma_i'] \\ (0,1,1,1) & \text{if } \bar{u}_i \in [\gamma_i', \gamma_i] \\ (1,1,1,1) & \text{if } \bar{u}_i > \gamma_i \end{cases} \tag{27}$$

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 6 depicts the different cases in Equation (27), given two values of a decision-quality instrument.

Because of the two-sided, symmetric nature of selection into the consistent subgroup in this model, we can identify the cardinal utility parameters governing the distribution of $\bar{u}_i$ and γ_i. With sufficient (observable) variation in γ_i and/or functional form assumptions on

[12] If the conditional distributions of γ_i and u_i have these properties when conditioning on given observable characteristics, using our matching on observables approach to tailor defaults by individual characteristics, as discussed in Section 4, would also lead to optimal policies.

the joint distribution of γ_i and $\bar{u}_i$, one can back out the underlying structural parameters using maximum-likelihood estimation or semi-parametric techniques. The setup of these estimation strategies is similar to that of the one in Section 3.3.2, except that selection is two-sided here and one-sided in Section 3.3.2.

Here we will focus on the situation where γ is homogeneous and known.[13] Figure 7 illustrates the recovery of preference intensity given arbitrarily rich (exogenous) variation in γ. The top panel depicts Equation (26) for some γ. From this panel we can see that Equation (26) implies that given some γ, $E[y_{i0}(\gamma)]$ tells us about the fraction of consistent choosers with $\bar{u}_i > \gamma$, while $E[y_{i1}(\gamma)]$ tells us about the fraction of consistent choosers with $\bar{u}_i < -\gamma$. The bottom panel shows how given rich variation in γ, so that we know $E[y_{i0}(\gamma)]$ and $E[y_{i1}(\gamma)]$ as a function of $\gamma \in [0, \infty)$, we can recover the full cumulative distribution function of $\bar{u}_i$, where the units of $\bar{u}_i$ are measured in the same units as γ.

C.3 Variants of the Costly Opt-Out Model

This section considers models in which default sensitivity arises for reasons apart from transaction costs but that can be formally modeled along the same lines. An important motivation for these models is the fact that default sensitivity is observed among decision-makers even when the stakes are large and opt-out is inexpensive (Bernheim, Fradkin and Popov, 2015; Carroll et al., 2009; Chetty et al., 2014), which suggests that pure neoclassical transaction costs are insufficient to explain behavior. Note that although the positive behavioral models are similar, the distinction between misperceived and real transaction costs does matter for setting the optimal policy, as misperceived costs should not be incorporated into the social welfare function (i.e. they should be excluded from the γ parameter in Appendix B).

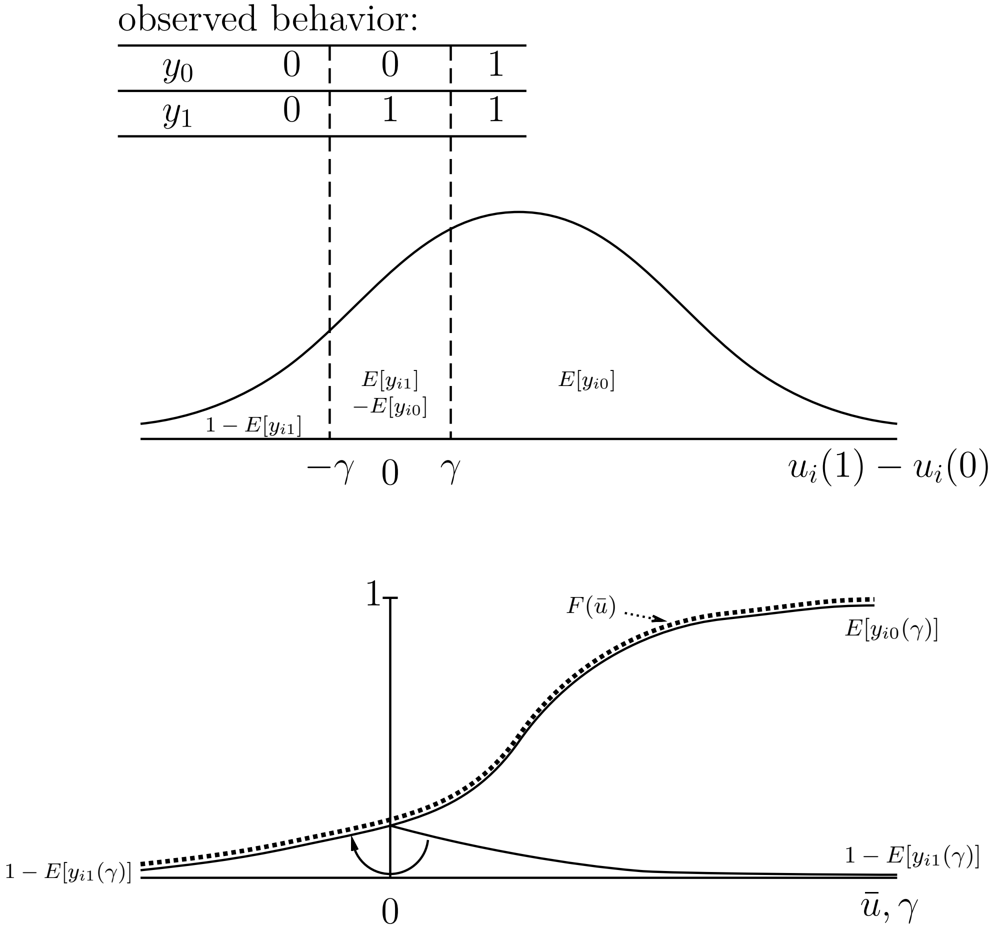
C.3.1 Procrastination

The model in this section shows how present bias can contribute to default sensitivity by inflating perceived opt-out costs, which must be presently-incurred, relative to a benefit that is realized in the future (Laibson, 1997). The model is a simplified version of the one presented in Carroll et al. (2009), which employs non-binary menus and a richer dynamic structure.

There are two time periods in the model, $t = 1, 2$. At $t = 1$, decision-makers choose between $y = 0$ and $y = 1$. As above, considering the non-default option requires incurring

[13] We defer to future work situations in which γ is heterogeneous and may have units that are difficult to quantify. Identifying a model with heterogeneous γ might utilize exogenous variation in $\bar{u}$. This variation could come from varying the relative price of $y = 0$ and $y = 1$, for example, but allowing for such variation leads to a considerably more sophisticated model than the binary choice model we employ here.

Figure 7: Identifying Preference Intensity with a Decision-Quality Instrument



some opt-out cost of c_i, for example, the disutility of expending mental energy for the decision-maker to figure out which option she prefers. At $t = 2$, the decision-maker receives the option she selected at $t = 1$ and realizes the associated utility. The DM's time preference is captured by $\delta_i \in (0, 1]$. Additionally, the DM may also be present-biased, in that she perceives the utility consequences of current period decisions to be greater than future ones (over and above her discount rate δ_i). We denote the DM's present bias by $\beta_i \in (0, 1]$. Thus, for a DM who prefers $y = 1$ facing a default of d_0 in $t = 1$, the (perceived) net utility effect of selecting the non-default option is given by

$$U_i(1, d_0) = \beta_i \delta_i \overline{u}_i - c_i$$

Let $\gamma_i \equiv \frac{c_i}{\delta_i}$ denote the current value of true utility costs from the perspective of the second period, which will be equivalent to opt-out costs in the previous section (since they are valued in the same period that utility over the menu options is realized). In this model, a decision-maker selects the non-default option if and only if the benefit to doing so exceeds the opt-out costs from the present-biased perspective:

$$y_{i0} = 1 \iff \overline{u}_i > \frac{\gamma_i}{\beta_i} \tag{28}$$

$$y_{i1} = 0 \iff \overline{u}_i < -\frac{\gamma_i}{\beta_i} \tag{29}$$

Comparing Equations (28) and (29) to Equation (26) reveals that the procrastination model is formally equivalent to the standard costly opt-out model if we re-define the "opt-out cost" to be the true current value of opt-out costs inflated by the degree of the DM's present-bias. Thus the key question for identifying the inconsistent choosers' preferences is understanding the empirical correlation between $\overline{u}_i$ and $\frac{\gamma_i}{\beta_i}$ in the population of decision-makers. From the perspective of social welfare, the only difference between this model and the previous one is that when we adopt the long-run view of welfare, only a fraction of the as-if opt-out costs are normatively relevant.

C.3.2 Status-Quo Bias

One candidate explanation for default effects is that decision-makers are more likely to choose whichever option represents a continuation of the status quo and perceive the default to be such a continuation. In this model, the frames denote which option represents a continuation of the status quo and which represents a change. A straightforward way to model status quo bias is to assume that individuals incur a utility cost when deviating from the status quo.[14] That is, if the status quo is for an individual to end up with $y = 0$, she will only select $y = 1$ when the utility gains to doing so exceed a positive threshold. The only difference between this model and the one in Appendix C.2 is that the utility loss associated with deviating from the status quo is non-neoclassical. However, such behavior can be modeled in exactly the same way as default sensitivity that arises in response to a neoclassical transaction cost. Good candidates for decision-quality instruments would be variations in the e*xtent* to which an option is framed as a continuation of the status quo, such as by emphasizing or downplaying that aspect of the option at the time of decision-making.

C.3.3 Anchoring

In this model designating some option as the default creates a psychological pull towards that option, due to an anchoring effect. Relative to utility over x and y, DM i chooses as if she values the psychological pull of the default option at δ_i. DM's decision utility of option $y \in X$ is given by

$$\hat{u}_i(y, d) \equiv u_i(y) + \delta_i 1\{y = d\}$$

where $1\{y = d\}$ indicates whether option y is the default. We assume that an individual maximizes her decision utility, i.e. that $y_i(d) = \arg\max\ \hat{u}(y, d)$. When $d = d_0$, the solution to this problem is given by $y_{i0} = 1 \iff \bar{u}_i > \delta_i$. When $d = d_1$, the solution is given by $y_{i1} 1 = y \iff \bar{u}_i > -\delta_i$. We can summarize the three distinct possibilities for the choices of individual i as follows:

$$(y_{i0}, y_{i1}) = \begin{cases} (0, 0) & \text{if } \bar{u}_i < -\delta_i \\ (0, 1) & \text{if } -\delta_i < \bar{u}_i < \delta_i \\ (1, 1) & \text{if } \bar{u}_i > \delta_i \end{cases} \tag{30}$$

Equation (30) reveals that although the psychological motivation is quite different, this model too is formally equivalent to the costly opt-out models described above. For the binary case considered here, the difference between the two models reduces to whether the effect of the frame on decision-making is modeled as a cost (as above) or as a benefit (as here).[15] Candidate decision-quality instruments might vary the psychological pull of the default option (δ), for example by varying its prominence relative to the rest of the menu.

[14] Researchers and psychologists disagree about whether a status quo bias is normatively relevant or not – that is, whether status quo effects arise from non-standard preferences as opposed to an optimization failure.

[15] Outside of the binary setting, Bernheim, Fradkin and Popov (2015) note that another way to distinguish between the models and the opt-out costs models is that anchoring should cause choices to cluster around the default, whereas opt-out costs should create a trough in the distribution of choices around the default.

C.3.4 Rational Attention

Here we consider the possibility that the opt-out costs are the psychological costs of paying attention to the non-default option. When the individual rationally incurs these costs, the model in Appendix C.1 becomes a model of rational inattention.[16] Formally, we can suppose that the individual faced with default d first chooses $\Gamma(d_0)$

$$d_0: \max_{\Gamma(d_0)\in\{\{0\},\{0,1\}\}} u_i(y(\Gamma(d_0))) - \gamma_i 1\{1 \in \Gamma(d_0)\} \text{ s.t. } y(\Gamma(d_0)) = \arg\max_{y\in\Gamma(d_0)} u_i(y) \tag{31}$$

$$d_1: \max_{\Gamma(d_1)\in\{\{1\},\{0,1\}\}} u_i(y(\Gamma(d_1))) - \gamma_i 1\{0 \in \Gamma(d_1)\} \text{ s.t. } y(\Gamma(d_1)) = \arg\max_{y\in\Gamma(d_1)} u_i(y) \tag{32}$$

The constraint in this optimal attention problem corresponds to the choice rule described by equation (24) in Appendix C.1, and the key assumption from that Section, that the individual always pays attention to the default option (Equation (25) before), is embedded in the possible choices of $\Gamma(.)$. When we endogenize attention in this way, the inattention model becomes exactly identical to the opt-out costs model, as characterized by Equation (26). This model is also a version of the planner-doer model of Fudenberg and Levine (2006), where the behavior of the "doer" is given by Equation (24) and the behavior of the "planner" – who regulates the behavior of the doer at some cost – by Equations (31) and (32).

C.4 Defaults as Advice

Here we present a simple model based on the idea that decision-makers who lack information about the utility of the available options may interpret the default as advice from the planner about which option is likely to be in their best interest. That is, decision-makers have some prior beliefs about the utility they would derive from either choice and they update that prior based on a signal in the form of the default. The model we develop here is a simplified version of that of Caplin and Martin (2012).

There are three potentially interesting types of preference information in this setting: (1) preferences baseed on individuals prior beliefs *before observing the default*, (2) preferences based on individuals beliefs after observing the default, and (3) ex post (full information) preferences. In this setting, our approach will identify information on the first of these types of preferences, based on individuals' priors. With more structure, one can also use the observed choice data to learn about the distribution of the second type of preferences (2),

[16] As with many models of rational inattention, this model is subject to an infinite regress critique (Conlisk, 1996). One might wonder, for instance, how the individual allocates attention to the problem of allocating attention among menu options. We assume the individual perfectly understands the costs and benefits of attention allocation, but acknowledge that this does not fully overcome the conceptual difficulty.

and subjective beliefs in the signal value of defaults – the main reason that (1) and (2) differ.

Formally, denote individual i's expected utility of $y = 1$ relative to $y = 0$ by $E[\bar{u}_i] = E[u_i(1) - u_i(0)]$. Upon observing the default, the individual updates this expected utility to $E[\bar{u}_i|d]$ and chooses $y = 1$ iff $E[\bar{u}_i|d] > 0$. For simplicity, we will assume the change in expected relative utility upon updating is symmetric across the two defaults and denoted by $v_i \geq 0$:[17]

$$E[\bar{u}_i|d_1] = E[\bar{u}_i] + v_i \tag{33}$$

$$E[\bar{u}_i|d_0] = E[\bar{u}_i] - v_i \tag{34}$$

Frame monotonicity will be satisfied in this model, since $E[\bar{u}_i|d_1] \geq E[\bar{u}_i|d_0]$.[18] In fact, this model is quite similar to the model of costly opt-out in Appendix C.2, with a different interpretation of the underlying parameters. Combining (33) and (34) yields

$$(y_{i0}, y_{i1}) = \begin{cases} (0,0) & \text{if } -E[\bar{u}_i] > v_i \\ (0,1) & \text{if -}E[\bar{u}_i] < v_i,\ E[\bar{u}_i] < v_i \\ (1,1) & \text{if } E[\bar{u}_i] > v_i \end{cases} \tag{35}$$

which closely resembles (26). The key difference between this model and the costly opt-out one is that instead of ex post preferences ($\bar{u}$) and cognitive costs (γ), our methods will identify information about prior beliefs ($E[\bar{u}_i]$) and the signal value of defaults (v_i). In particular, the preferences of consistent choosers will be be the fraction of consistent choosers for whom

[17] We can use a slightly different set-up to develop the same intuition and make the updating upon learning the default more explicit. Suppose each DM i has prior belief $P(\bar{u}_i > 0) = \alpha_i$, and believes the default is correct with probability $\beta_i > \frac{1}{2}$, regardless of whether the signal is d_0 or d_1 (which assumes some symmetry like in the other model considered here). The signal she receives is independent of her prior. An individual updates her subjective probability that y is preferred to x as follows:

$$P(\bar{u}_i > 0|d_1) = \frac{P(\bar{u}_i > 0, d_1)}{P(\bar{u}_i > 0, d_1) + P(\bar{u}_i < 0, d_1)} = \frac{\alpha_i\beta_i}{\alpha_i\beta_i + (1-\alpha_i)(1-\beta_i)}$$

$$P(\bar{u}_i > 0|d_0) = \frac{P(\bar{u}_i > 0, d_0)}{P(\bar{u}_i > 0, d_0) + P(\bar{u}_i < 0, d_0)} = \frac{\alpha_i(1-\beta_i)}{\alpha_i(1-\beta_i) + (1-\alpha_i)\beta_i}$$

Suppose that the individual chooses $y = 1$ iff $p(\bar{u}_i > 0|d) > \frac{1}{2}$, which would be true if the subjective distribution of $\bar{u}_i$ were symmetric for all i. Then the analogue to equation (35) is as follows:

$$(y_{i0}, y_{i1}) = \begin{cases} (0,0) & \text{if } \alpha < 1 - \beta_i \\ (0,1) & \text{if } 1 - \beta_i < \alpha < \beta_i \\ (1,1) & \text{if } \alpha > \beta_i \end{cases}$$

which conveys a similar intuition about using our approach to learn about prior beliefs (α_i in this model) and the signal value of defaults (β_i in this model).

[18] Note that this obtains even without the symmetry assumption employed above, so long as $E[\bar{u}]$ is shifted (weakly) upward by d_1 and downward by d_0. Frame monotonicity fails when some individual perceives the default-setter as providing *bad* advice, i.e. $v_i < 0$ for some i.

$u_i > 0$. Consistent choosers are those who choose the option favored by their prior beliefs regardless of the frame, which implies that their signal value was small relative to the strength of their prior. As in the costly opt-out model with arbitrary heterogeneity, decision-quality independence will not hold in general. One can imagine adding interventions analogous to decision-quality instruments here by attempting to manipulate the perceived signal value of the default option. For example, one could provide some decision-makers with a salient disclaimer that the default is *not* intended to convey advice.

Because preferences based on prior beliefs are not ex post preferences, the relationship between the parameters we represent and the optimal default is more nuanced than the one we explored in Section B. This should not be surprising: when the planner's choice of default affects beliefs about preferences because individuals assume that the planner has preference-relevant information they do not have, it should not be possible to infer their preferences directly from their choices under alternative default policies. Nevertheless, when individuals' priors reflect private information not available to the planner, a planner who initially sets a policy based on his own information[19] could potentially improve the optimal policy by also recovering information about individuals' priors using the techniques proposed in our paper.

D Generalizations

D.1 Non-Binary Frames, Varying Intensity

Consider choice situations in which an individual i chooses from a binary menu $X = \{0, 1\}$ under one of multiple frames that vary in their intensity, $d \in \{d_0, d_1, ..., d_J\}$. For example, if $y = 0$ has greater up-front costs than $y = 1$, the frames might describe the extent to which those costs are made salient to the decision-maker, with d_1 denoting the frame in which the costs are least salient and d_J denoting the frame in which they are most salient. 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 presented (if decision-makers are subject to an anchoring effect, the larger the reference point the more they might be willing to pay).

The key assumption that will let us apply the tools from the rest of the paper to this

[19] When the planner has no information about preferences that does *not* come from observing choices under different defaults, there is a striking recursion embedded in the optimal framing problem for this model: the planner's optimal choice of default must depend on inferring preference information from choices, but default effects only occur in the first place because individuals assume the planner already has some preference information.

setting is that the frames can be ordered according to their intensity:

$$y_i(d_j) \geq y_i(d_{j'}) \ \forall i\,,\, j > j' \tag{36}$$

where the ordering of frames is without loss of generality. In words, (36) requires that if a decision-maker chooses an option under one frame, he or she will also choose that option under any frame that pushes more intensely in that option's direction. With (36), frame monotonicity applies with respect to any two frames.

We also assume a global *consistency principle*:

$$y_i(d_j) = y_i(d_{j'}) \ \forall j, j' \implies y_i(d_j) = y_i^* \tag{37}$$

This is a natural extension of the consistency principle discussed earlier: if a decision-maker would select the same option in each frame, we assume that choice is her preferred one.

When (36) and (37) hold, the multi-frame setting can be reduced to the binary one studied in the rest of the paper. In particular, one can apply Proposition 1 after setting $d_0 = d_0$ and $d_1 = d_J$ to obtain information about the set of decision-makers consistent with respect to all of the observed frames. Additionally, even when (37) does not hold, one can apply the matching estimator (Corollary 4.1) to recover information about the characteristics of consistent and inconsistent decision-makers at any two observed frames so that the researcher can investigate heterogeneity in which decision-makers are susceptible to low-intensity framing effect as opposed to both high- and low-intensity ones.

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. Formally, let $d_1 = (z_l, d_0)$, $d_2 = (z_h, d_0)$, $d_3 = (z_h, d_1)$, and $d_4 = (z_l, d_1)$ be the four possible frames. The consistency principle from Section 3.3, at z_l, is equivalent to the global consistency principle (37). 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_l: $Y_C(z^l) = E[y_i^*|c_{il} = 1]$. The consistency principle at z^h in Section 3.3also implies a second condition for consistency across d_2 and d_3, which will imply that $Y_C(z^h) = E[y_i^*|c_{ih} = 1]$. Frame separability in this model implies decision quality exclusion, and frame monotonicity implies decision quality monotonicity, so that all changes in behavior between d_1 and d_2, and between d_3 and d_4 tell us about the preferences of individuals consistent across (d_1, d_4) but not across (d_2, d_3), which allows us to recover the preferences of decision-makers whose choices depend on the intensity of framing,

$Y_S = E[y_i^*|c_{ih} > c_{il}]$.

D.2 Multi-Dimensional Frames

This sub-section considers choice settings that differ along multiple dimensions so that frames cannot be ordered by intensity. Consider choice situations in which an individual i chooses from a binary menu $X = \{x, y\}$ under a *frame vector* $d = (d_1, ..., d_J)$, so that each component of d encodes some feature of the choice environment. We assume each frame component d_j of d is discrete with two possible realizations: $d_j \in \{d_{0j}, d_{1j}\}$.[20] 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:

$$\forall i\, , \forall j \;\; y_i(d_{1j}, d_{\neg j}) \geq y_i(d_{0j}, d_{\neg j}) \tag{38}$$

where $d_{\neg j}$ is the vector consisting of all frame components other than j. Equation (38) implies that frame monotonicity holds for each component of the frame vector when all 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 $y = 0$ the default induces more decision-makers to select $y = 0$ when 0 is listed first but that making 0 the default induces more decision-makers to select $y = 1$ when 0 is listed second.

As above, we assume a global *consistency principle*:

$$y_i(d) = y_i(d')\ \forall d, d' \implies y_i(d) = y_i^* \tag{39}$$

This principle means that whenever the individual would choose the same option in every frame, she prefers the option that she chooses.

When (38) and (39) hold, we can proceed similarly to the previous subsection using data on the two most extreme frames. To do this we can apply Proposition 1 with $d_0 = (d_{01}, d_{02}, ..., d_{0J})$ and $d_1 = (d_{11}, d_{12}, ..., d_{1J})$ to recover the preferences of globally consistent decision-makers, those whose choice is insensitive to all framing effects. We use this approach in the empirical illustration in Appendix D.2.1. Further possibilities are generated by considering what additional structure might allow us to learn about preferences from the

[20] It is straightforward to combine this approach with the one in the previous section, extending this setup to settings in which each frame component has multiple possible realizations that vary in their intensity.

behavior of individuals who are consistent with respect to all frame components but one, all frame components but two, and so on.

D.2.1 Application to Privacy Controls

In this section we re-examine data collected by Johnson, Bellman and Lohse (2002), whose work inspired our running example of online privacy controls. In this study, the authors initially asked participants to complete an online survey about their health. Participants were then asked whether they would like to receive additional surveys.[21] There were two sources of framing effects for these solicited preferences. First the question was either framed negatively ("Do NOT notify me about more health surveys") or positively ("Notify me about more health surveys."). Second, the answers to the question were pre-selected to be either yes or no, so that the consumer would have to actively change their answer to avoid the default. We therefore employ the extension of our approach to multi-dimensional framing effects considered in Appendix D.2. In particular, y indicates whether the individual gave permission to be contacted for additional surveys, d_0 is the situation in which the question was framed positively *and* the default was non-participation ("opt-in"), and d_1 is the situation in which the question was framed as a loss *and* the default was participation ("opt-out"). We do not use data from the other two frame combinations. This exclusion drops 139 of the sample of 277 individuals in the original study, leaving 138 observations. Although we do have some demographic variables, our sample size makes results from dividing the population into demographic groups imprecise.

Table 7 presents the results from this application. Several interesting patterns emerge. First, framing affects the participation decisions of almost half (48.2 percent) of the survey-takers; only 51.8 percent of the population are globally consistent, meaning their choice depends neither on the positive or negative wording of the question nor on the default, pre-selected answer. Second, of the globally consistent subgroup, over 90 percent prefer to participate. The size of this estimate is attributable to the fact that, under d_1, very few individuals opt out of participation, while under d_0, many more individuals opt in. Given that a very high fraction of globally consistent individuals participate in the surveys, we conclude that under decision-quality independence, an opt-out policy would maximize the fraction of participants ending up with their preferred option.

[21] We focus on the data reported in "Study 1" of Johnson, Bellman and Lohse (2002).

Table 7: Participation in Online Surveys

	Total
Participation rate under d_1	0.9563 (0.0251)
Participation rate under d_0	0.482 (0.067)
Fraction consistent	0.518 (0.070)
Fraction of consistent who prefer participation	0.929 (0.039)

Note: standard errors are estimated using the delta method (see Appendix E.2).

D.3 Non-Binary Menus with Ordered Options

This section develops an approach for preference recovery over non-binary menus. There are many interesting possibilities for generalizations, but we focus here on choice situations $g \in G$ consisting of a fixed, finite menu of K ordered options $X = \{1, 2, ..., K\}$ and one of two frames, $d \in \{d_h, d_l\}$. Intuitively, one can think of d_h and d_l as a "high" frame and a "low" frame. For example, we might suppose that an individual chooses from a menu of insurance plans, ordered from low-cost, low-benefit plans to high-cost, high-benefit plans, and the frame either emphasizes or de-emphasizes the individual's risk of developing a serious illness. We will assume that we observe each individual i in exactly one frame, as before. Recall that in the binary case, the consistency principle and frame monotonicity imply that individuals who choose the "low" option in the "high" frame prefer the low option. We will use this same intuition to develop an identification strategy for the non-binary setting.

The preferences of agent i are represented by choice function $y_i^* \in X$. We continue to assume frame separability (A1), so that y_i^* does not depend on d.

We strengthen the frame monotonicity assumption as follows:

D1 (Frame monotonicity for many options) For all individuals, $y_i(d_h) \geq y_i(d_l)$.

This frame monotonicity assumption imposes an implicit ordering on the menu and assumes that all individuals are pushed in the same direction by the frames. We also strengthen the consistency principle with the following assumption

D2 (Partition-Consistency principle) For all individuals i and options $k \in X$,

$$\begin{aligned} y_i(d_l) \geq k &\implies y_i^* \geq k \\ y_i(d_h) \leq k &\implies y_i^* \leq k. \end{aligned} \tag{40}$$

The name of this assumption comes from the following: suppose that we partition the menu into $X' = \{J, J+1, ...K\}$ and $X'' = X \setminus X'$, for some J and $K \geq J$. If the individual consistently chooses within X' across both frames, so $y_i(d_h) \in X'$, and $y_i(d_l) \in X'$, then assumption (40) implies that $y_i^* \in X'$. Note also that the partition consistency principle implies the consistency principle used in previous sections: if $y_i(d_h) = y_i(d_l)$, then assumption (40) implies that $y_i(d) = y_i^*$. Finally, note that the partition consistency principle and frame monotonicity together imply that $\forall i,\ y_i(d^h) \geq y_i^* \geq y_i(d^l)$.

For each $k = 1, ..., K$, we define *partition consistency at* k, c_i^k, as follows

$$c_i^k \equiv 1\{y_i(d^h) \leq k \text{ and } y_i(d^l) \leq k\} + 1\{y_i(d^h) > k \text{ and } y_i(d^l) > k\}$$

Intuitively, c_i^k captures whether an individual consistently chooses an option above or below k. Note also that frame monotonicity implies that one of the conditions inside each indicator function will be implied by the other condition.

Proposition A2 *Let $G_j(k) \equiv P(y_i(d_j) \leq k | d_i = d_j)$ for $k = 1, ..., N$, $j = h, l$ and let $G_j(0) \equiv 0$. Let $Y_k \equiv \frac{G_h(k)}{G_h(k)+1-G_l(k)}$ for $k = 0, ..., K$. Frame separability (A1), frame monotonicity (D1), partition consistency (D2), and unconfoundedness (A4) imply that for $k = 1, ..., K$,*

(A2.1) *The fraction of partition-consistent individuals at k with $y_i^* \leq k$ is given by $P(y_i^* \leq k | c_i^k = 1) = Y_k$.*

(A2.2) *The fraction of partition-consistent individuals at k is given by $E[c_i^k] = G_h(k) + 1 - G_l(k)$.*

(A2.3) *The fraction of the population who prefer option k is bounded as follows: $p(y_i^* = k) \in [G_l(k) - G_h(k-1),\, G_h(k) - G_l(k-1)]$.*

(A2.4) *If we additionally assume strong decision-quality independence,$\forall k,\ y_i^* \perp c_i^k$, then the fraction of the population who prefer option k is $p(y_i^* = k) = Y_k - Y_{k-1}$.*

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 (A2.1) and (A2.2): Fix some $k \in \{1, ..., K-1\}$. Let $X' = \{x_1, ...x_k\}$and $X'' = \{x_{k+1}, .., x_K\}$ Note that we can write the many-choices problem into a binary menu choice problem between X' and X''. Similarly, note that frame separability (A1), frame monotonicity (D1), partition consistency (D2), and partition unconfoundedness (A4) imply the binary analogues to these assumptions A1-A4. As such, (A2.1) and (A2.2) follows directly from the application of Proposition 1 to this problem.

Proof of (A2.3): First suppose that $k = 1$. Applying Proposition 2 to the binary menu choice problem with $X' = \{1\}$ and $X'' = \{2, ..., K\}$ implies that

$$E[\phi_1] \in [G_l(1), G_h(1)] \tag{41}$$

Note that this confirms the desired result for $k = 1$ since $G_h(0) = G_l(0) = 0$ 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 (41), this implies

$$\overline{\phi}_2 \in [G_l(2) - G_h(1),\, G_h(2) - G_l(1)] \tag{42}$$

Similarly with $k = 3$, we have that $\overline{\phi}_1 + \overline{\phi}_2 + \overline{\phi}_3 \in [G_l(3),\, G_h(3)]$, and applying (41) and (42) implies that $\overline{\phi}_3 \in [G_l(3) - G_h(2),\, G_h(3) - G_l(2)]$. Proceeding recursively, suppose that for some k, we know that for any $k' < k$,

$$\overline{\phi}_{k'} \in [G_l(k') - G_h(k' - 1),\, G_h(k') - G_l(k' - 1)] \tag{43}$$

Then application of Proposition 2 to the binary menu choice problem with $X' = \{x_1, ..., x_k\}$ yields $\overline{\phi}_1 + \overline{\phi}_2 + ... + \overline{\phi}_k \in [G_l(k),\, G_h(k)]$, so $\overline{\phi}_k \in [G_l(k) - (\overline{\phi}_1 + \overline{\phi}_2 + ... + \bar{\phi}_{k+1}),\, G_h(k) - (\overline{\phi}_1 + \overline{\phi}_2 + ... + \bar{\phi}_{k+1})]$. Applying the lower and upper bounds from (43) and simplifying yields the desired result.

Proof of (A2.4): Along with (A2.1), strong decision-quality independence implies that for any k,

$$P(y_i^* \leq k | c_i^k = 1) = P(y_i^* \leq k) = Y_k \tag{44}$$

Applying (44) at $k = 1$ yields

$$\bar{\phi}_1 = Y_1 \tag{45}$$

Applying (44) at $k = 2$ yields $\bar{\phi}_1 + \bar{\phi}_2 = Y_2$ and substituting equation (45) yields

$$\bar{\phi}_2 = Y_2 - Y_1$$

As in the proof of (A2.3), we proceed recursively to obtain the desired result. Given some k, suppose that for any $k' < k$ we have

$$\bar{\phi}_{k'} = Y_{k'} - Y_{k'-1} \tag{46}$$

Applying (44) at k yields $\bar{\phi}_1 + \bar{\phi}_2 + ..., + \bar{\phi}_k = Y_k$. Applying (46) for $\bar{\phi}_1, ..., \bar{\phi}_{k-1}$ and simplifying yields the desired result. ■

Discussion of Proposition A2 If we partition the menu of choices into options above and below some option k, then frame monotonicity and the partition-consistency principle transform the problem to a binary problem, allowing us to use earlier propositions to identify individuals whose preferred choice is above or below k. The first two results, (A2.1) and (A2.2), are therefore the analogue of Proposition 1 in this setting.

Return to the insurance example described above, where the frame either emphasizes or de-emphasizes the risk of serious illness. When some individuals choose a low-benefit, low-cost plan under the frame that emphasizes the risk of serious illness, our assumptions imply that they prefer an option with costs and benefits at least as low as the ones they choose. The first two results in Proposition A2 allow us to estimate the fraction of decision-makers who consistently choose an insurance plan that is above or below some specified cost-benefit level, and among those people, how many prefer the low-cost plan.

As in Proposition 2, we can also bound population preferences, reflected in (A2.3). In this case, the many-options problem has a new and interesting structure. 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.

Finally, with a stronger version of the decision-quality independence assumption, we can recover the distribution of preferences for the full population. Strong decision-quality independence guarantees that the tendency to be partition-consistent for any partition is unrelated to an individuals' preferences.[22] Under strong decision-quality independence, obtaining the preferences of partition-consistent individuals will yield the distribution of preferred choices in the population. The equivalence of this problem to the binary problem implies that we could generalize other identification strategies from the binary case. For example, we can identify the preferences of the population using observables via a *conditional strong decision-quality independence* assumption (the generalization of the matching approach), and in the absence of any decision-quality independence assumptions we can study variation induced by a decision-quality instrument.

E Estimating Asymptotic Variance in Finite Sample

The body of the paper ignores finite sample concerns, but any empirical application, including the ones we undertake in Section 4, should account for finite sample concerns and report standard errors for the estimators we propose. This section derives estimators of asymptotic

[22] This assumption implies our earlier definition of decision-quality independence, because individuals who are consistent across frames will be partition-consistent for all partitions. If we were to assume that individuals are partition-consistent only if they are fully consistent across frames, which is trivially true in the binary case, then the two assumptions about decision-quality independence would be equivalent.

variance, which may be used to construct standard errors of the estimators of the population parameters of interest. Because all variables we work with in this paper are discrete and the data are presumed to come from independent random sampling of the population, our derivations here rely only on the characteristics of the binomial and multinomial distributions, along with the delta method. The standard errors we derive are incorporated into Stata program files that are available upon request from the authors.

E.1 Standard Errors for Proposition 1

To facilitate compact exposition, let us introduce some statistical notation. Note that y_{i0} and y_{i1} are two-valued random variables. Let Y_0 and Y_1denote the population moments we wish to estimate from data as before. Let $\bar{y}_0$ and $\bar{y}_1$ denote the sample averages under d_0 and d_1, respectively. From the properties of the binomial distribution,

$$\sqrt{n}\begin{pmatrix} \bar{y}_0 - Y_0 \\ \bar{y}_1 - Y_1 \end{pmatrix} \overset{a}{\sim} N\left(0, \Sigma\right)$$

where

$$\Sigma = \begin{pmatrix} \frac{1}{\alpha_0}Y_0(1-Y_0) & 0 \\ 0 & \frac{1}{\alpha_1}Y_1(1-Y_1) \end{pmatrix}$$

and $\alpha_j = \frac{n_j}{n}$ for $j \in \{0,1\}$, where n_j denotes the number of individuals observed under frame d_j.

The first statistic from Proposition 1 is Y_c, which expressed as a function of the primitive parameters of the model is

$$Y_c = \frac{Y_0}{Y_0 + 1 - Y_1} \tag{47}$$

which we estimate consistently via

$$\hat{Y}_c = \frac{\bar{y}_0}{\bar{y}_0 + 1 - \bar{y}_1}$$

Using the delta method, we know that $V(\hat{y}_c) \simeq \frac{1}{n}\nabla Y_c' \, \Sigma \, \nabla Y_c$, where ∇Y_c is evaluated at $(Y_0,\, Y_1)$. Taking derivatives of (47) yields

$$\nabla Y_c = \begin{pmatrix} \frac{1-Y_1}{(Y_0+1-Y_1)^2} \\ \frac{Y_0}{(Y_0+1-Y_1)^2} \end{pmatrix}.$$

Simplifying the expression for $V(\hat{y}_c)$ yields

$$V(\hat{y}_c) \simeq \frac{(1-Y_1)^2 Y_0(1-Y_0)}{\alpha_0(Y_0+1-Y_1)^4 n} + \frac{Y_0^2 Y_1(1-Y_1)}{\alpha_1(Y_0+1-Y_1)^4 n}$$

which, letting $\bar{c} = Y_0 + 1 - Y_1$ and using that $V(\bar{y}_j) = \frac{Y_j(1-Y_j)}{n_j}$ for $j = 0, 1$, simplifies to

$$V(\hat{y}_c) = \left(\frac{1 - Y_c}{\bar{c}}\right)^2 V(\bar{y}_0) + \left(\frac{Y_c}{\bar{c}}\right)^2 V(\bar{y}_1) \tag{48}$$

We can consistently estimate the variance of the asymptotic distribution of our estimator $\hat{y}_c$ by replacing all the terms in (48) with the corresponding sample means, i.e. replacing Y_0 with $\bar{y}_0$ and Y_1 with $\bar{y}_1$.

Note also that we can estimate the variance of the asymptotic distribution of our estimator for $\bar{c} \equiv E[c_i]$. Write the estimator itself as:

$$\hat{\bar{c}} = \bar{y}_0 + 1 - \bar{y}_1$$

and the variance of this estimator is simply:

$$V(\hat{\bar{c}}) = V(\bar{y}_0) + V(\bar{y}_1) \tag{49}$$

E.2 Standard Errors for Matching Estimator

The standard errors for the matching-on-observables estimator are considerably 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 an estimation strategy imposing parametric restrictions for how observable characteristics are related to choices under either frame, and another adopting a fully non-parametric approach.

First, one can simply estimate our model via

$$E[y_i|d, w] = f(w_i, \theta) + 1\{d_i = d_1\}g(w_i, \theta')$$

where $f()$ and $g()$ are specified up to vectors of parameters θ and θ', which are estimated from data. For example, we could implement a linear model with uni-dimensional w:

$$E[y_i|d, w] = \alpha + \beta w_i + 1\{d_i = d_1\}(\gamma + \delta w_i)$$

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

$$\hat{Y}_c(w) = \frac{\hat{\alpha} + \hat{\beta} w_i}{1 - \hat{\gamma} - \hat{\delta} w_i}$$

$$\hat{E}[c_i|w] = 1 - \hat{\gamma} - \hat{\delta} w_i$$
$$\hat{E}[c_i] = \frac{1}{n}\sum_i (1 - \hat{\gamma} - \hat{\delta} w_i)$$

$$\hat{E}[y_i^*] = \frac{1}{n}\sum_i \hat{Y}_c(w)$$

$$\hat{E}[y_i^*|c_i = 0] = \hat{Y}_c(w) = \frac{1}{n}\sum_i \frac{1 - \hat{E}[c_i|w = w_i]}{1 - \hat{E}[c_i]}\hat{Y}_c(w_i)$$

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

When w_i is discrete, taking values $w_1, ..., w_J$ one may derive straightforward delta-method-based standard errors. We provide analytical 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, used for the illustration in the body of the paper, is also available upon request from the authors.

The primitive parameters of the discrete-characteristics model are, for each w,Y_{0w}, Y_{1w}, and $p_w = p(w_i = w)$. When the w'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. We denote the estimators of these quantities by $\bar{y}_{0w}$, $\bar{y}_{1w}$, and $\hat{p}_w$. Now we construct the variance covariance matrix of the vector primitive parameters. Letting $\theta = (Y_{0w_1}, Y_{0w_2}, ..., Y_{0w_J}, Y_{1w_1}, ..., Y_{1w_J}, p_{w_1}, ..., p_{w_J})'$, and $\hat{\theta} = (\bar{y}_{0w_1}, \bar{y}_{0w_2}, ..., \bar{y}_{0w_J}, \bar{y}_{0w_1}, ..., \bar{y}_{0w_J}, \hat{p}_{w_1}, ..., \hat{p}_{w_J})$, we know that

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

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

$$\Sigma(\theta) = \begin{pmatrix} M_0 & & \\ & M_1 & \\ & & P \end{pmatrix}$$

where M_0 is a dianogal matrix with entries of the form $Y_{0w_j}(1 - Y_{0w_j})/\alpha_{0w_j}$, M_1 is a diagonal matrix with entries of the form $Y_{1w_j}(1 - Y_{1w_j})/\alpha_{1w_j}$, and P is given by:

$$P = \begin{pmatrix} p_{w_1}(1 - p_{w_1}) & -p_{w_1}p_{w_2} & ... & -p_{w_1}p_{w_J} \\ -p_{w_1}p_{w_2} & p_{w_2}(1 - p_{w_2}) & ... & -p_{w_2}p_{w_J} \\ ... & ... & ... & ... \\ -p_{w_1}p_{w_J} & -p_{w_2}p_{w_J} & ... & p_{w_J}(1 - p_{w_J}) \end{pmatrix}.$$

All blank entries of the Σ matrix are zeroes.[23]

The matching approach employs many different combinations of these primitive parameters. We begin with the weights for the subset of inconsistent choosers,

$$s_w \equiv p(w_i = w|c_i = 0) = \frac{Y_{1w} - Y_{0w}}{\sum_{w=v}(Y_{1v} - Y_{ov})p_v} p_w$$

Carefully taking derivatives of this function and simplifying using the definition of s_w, we obtain the following for any w and w':

$$\frac{\partial s_w}{\partial Y_{0w'}} = -\frac{1\{w = w'\} - s_w}{\sum_{w=v}(Y_{1v} - Y_{ov})p_v} p_{w'}$$
$$\frac{\partial s_w}{\partial Y_{1w'}} = \frac{1\{w = w'\} - s_w}{\sum_{w=v}(Y_{1v} - Y_{ov})p_v} p_{w'}$$
$$\frac{\partial s_w}{\partial p_{w'}} = \frac{1\{w = w'\}s_w - S_{w'}s_w}{p_{w'}}$$

were $1\{\}$ 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_w.

Next we consider the weights for the subset of consistent choosers, $q_w \equiv p(w_i = w|c_i = 1) = \frac{p_{xw}+1-p_{yw}}{p_x+1-p_y} p_w$. Proceeding similarly to before, rewrite R_w as

$$q_w = \frac{Y_{0w} + 1 - Y_{1w}}{\Sigma_{w=v}(Y_{0v} + 1 - Y_{1v})p_v} p_w$$

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

$$\frac{\partial q_w}{\partial Y_{0w'}} = \frac{1\{w = w'\} - R_w}{\Sigma_{w=v}(Y_{0v} + 1 - Y_{1v})p_v} p_{w'}$$
$$\frac{\partial q_w}{\partial Y_{1w'}} = -\frac{1\{w = w'\} - R_w}{\Sigma_{w=v}(Y_{0v} + 1 - Y_{1v})p_v} p_{w'}$$
$$\frac{\partial q_w}{\partial p_{w'}} = \frac{1\{w = w'\}R_w - R_w R_{w'}}{p_{w'}}$$

Next we consider the estimators for the preferences of various subgroups. First, define the preferences of the inconsistent subgroup

$$Y_N = \Sigma_w s_w Y_{cw}$$

Taking derivatives of this – which may be done more easily using several expressions derived

[23] To be specific, we know that the off-diagonal elements in the first $2 * J$ rows and the first $2 * J$ columns, which govern the covariance of the various Y_{kj} estimates, are zero because the estimation sample for each one is distinct. We know that the entries of Σ governing the covariance of Y_{kj} and p_w, for some k and j, are zero because of the unconfoundedness assumption.

above – we obtain the following for any w:

$$\begin{aligned}\frac{\partial Y_N}{\partial Y_{0w}} &= s_w \frac{1 - Y_{cw}}{\bar{c}_w} + \frac{p_w}{1 - \bar{c}}(Y_N - Y_{cw})\\ \frac{\partial Y_N}{\partial Y_{1w'}} &= s_w \frac{Y_{cw}}{\bar{c}_w} - \frac{p_w}{1 - \bar{c}}(Y_N - Y_{cw})\\ \frac{\partial Y_N}{\partial p_w} &= \frac{s_w}{p_w(1 - \bar{c})}[Y_{cw} - (1 - \bar{c})Y_N]\end{aligned}$$

where $\bar{c}_w \equiv Y_{0w} + 1 - Y_{1w} = E[c_i|w]$ and $\bar{c} = \Sigma_{w=v}(Y_{0v} + 1 - Y_{1v})p_v$. From these three expressions we construct the gradient of Y_N.

Proceeding similarly for the full population, $Y_{FP} = \sum_w p_w Y_{cw}$, we obtain, for any w

$$\begin{aligned}\frac{\partial Y_{FP}}{\partial Y_{0w}} &= p_w \frac{1 - Y_{cw}}{\bar{c}_w}\\ \frac{\partial Y_{FP}}{\partial Y_{1w}} &= p_w \frac{Y_{cw}}{\bar{c}_w}\\ \frac{\partial Y_{FP}}{\partial p_w} &= Y_{cw}\end{aligned}$$

which allows us to construct the gradient of Y_{FP}.

We can also obtain the gradient of $Y_c = \sum_w q_w Y_{cw}$ in terms of the primitive parameters of this model.[24]Taking derivatives of the expression for Y_c yields

$$\begin{aligned}\frac{\partial Y_c}{\partial Y_{0w}} &= q_w \frac{1 - Y_{cw}}{\bar{c}_w} + \frac{p_w}{\bar{c}}(Y_{cw} - Y_c)\\ \frac{\partial Y_c}{\partial Y_{1w}} &= q_w \frac{Y_{cw}}{\bar{c}} - \frac{p_w}{\bar{c}}(Y_{cw} - Y_c)\\ \frac{\partial Y_c}{\partial p_w} &= \frac{q_w}{p_w \bar{c}}(Y_{cw} - \bar{c} Y_c)\end{aligned}$$

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 gradient matrix of the form:

$$G(\theta) = (\nabla s_{w_1}, ..., \nabla s_{w_J}, \nabla q_{w_1}, ..., \nabla q_{w_J}, \nabla Y_N, \nabla Y_{FP}, \nabla Y_c)$$

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.

[24] This part is not necessary to obtain a standard error on Y_c, because we know how to obtain a simpler formula for the asymptotic variance of our estimator of Y_c using the result in the previous section of this Appendix. Doing it the hard way here yields an 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 Y_c and Y_N, which is necessary for the statistical test of the null hypothesis of decision-quality independence against the alternative hypothesis of conditional decision-quality independence.

We can now estimate the full variance-covariance matrix of all the parameters given 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 correct 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 in this framework.

E.3 Standard Errors for Proposition 4

Using similar notation to before, let $Y_{jk} \equiv E[y_{ijk}]$ for $j = 0, 1$ and $k = h, l$, and denote the estimator for each population moment by $\bar{y}_{jk}$. Similarly to the previous section, we begin by noting that

$$\sqrt{n} \left(\begin{pmatrix} \bar{y}_{0h} \\ \bar{y}_{0l} \\ \bar{y}_{1h} \\ \bar{y}_{1l} \end{pmatrix} - \begin{pmatrix} Y_{0h} \\ Y_{0l} \\ Y_{1h} \\ Y_{1l} \end{pmatrix} \right) \overset{a}{\sim} N(\overrightarrow{0}, \Sigma)$$

where Σ is a diagonal matrix with entries of the form $\frac{1}{\alpha_{jk}} Y_{jk}(1 - Y_{jk})$.

The new statistic in Proposition 4 is

$$Y_s = \frac{Y_{0h} - Y_{0l}}{Y_{1l} - Y_{0l} - (Y_{1h} - Y_{0h})} \tag{50}$$

which we can estimate consistently with:

$$\hat{Y}_s = \frac{\bar{y}_{0h} - \bar{y}_{0l}}{\bar{y}_{1l} - \bar{y}_{0l} - (\bar{y}_{1h} - \bar{y}_{0h})}$$

Using the delta method, we obtain $V(\hat{Y}_s) \simeq \frac{1}{n} \nabla Y_s' \, \Sigma \, \nabla Y_s$, where ∇Y_S is evaluated at $(Y_{0h}, Y_{0l}, Y_{1h}, Y_{1l})$. Taking the gradient of (50) gives

$$\nabla Y_s = \left(\frac{Y_{1l} - Y_{1h}}{(\Delta\bar{c})^2}, \frac{Y_{1h} - Y_{1l}}{(\Delta\bar{c})^2}, \frac{Y_{0h} - Y_{0l}}{(\Delta\bar{c})^2}, \frac{Y_{0l} - Y_{0h}}{(\Delta\bar{c})^2} \right)'$$

where $\Delta\bar{c} = Y_{1l} - Y_{0l} - (Y_{1h} - Y_{0h})$. Plugging this into the formula for $V(\hat{Y}_S)$ and simplifying yields

$$V(\hat{Y}_s) = \left(\frac{1 - Y_s}{\Delta\bar{c}} \right)^2 [V(\bar{y}_{xh}) + V(\bar{y}_{xl})] + \left(\frac{Y_s}{\Delta\bar{c}} \right)^2 [V(\bar{y}_{yh}) + V(\bar{y}_{yl})]$$

where $V(\bar{y}_{jk}) = \frac{1}{n_{jk}} Y_{jk}(1 - Y_{jk})$. Replacing each Y_{jk} with the estimator $\bar{y}_{jk}$, we obtain a consistent estimate of the asymptotic variance of $\hat{Y}_S$. Note that when $\Delta\bar{c}$ is small, the variance of this estimator can be quite large, reflecting a familiar facet of instrumental variables estimation.

References

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

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.

Caplin, Andrew, and Daniel Martin. 2012. "Framing effects and optimization." Working Paper.

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

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics*, 1(1): 451–488.

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

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

Fudenberg, Drew, and David K Levine. 2006. "A Dual-Self Model of Impulse Control." *The American Economic Review*, 1449–1476.

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.

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

Manzini, Paola, and Marco Mariotti. 2014. "Stochastic choice and consideration sets." *Econometrica*, 82(3): 1153–1176.

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

Thaler, Richard H, and Cass R Sunstein. 2003. "Libertarian Paternalism." *American Economic Review*, 175–179.

Appendix to Chapter 3 (Taxes and Mistakes)

A Derivations and Proofs

Budget Adjustment Rules

The third budget adjustment rule in Section 3 corresponds to assuming that a consumer mentally reduces his wealth by the amount of excess tax paid, and then chooses an optimal bundle given this level of wealth and the perceived relative price. This is the same as stating that the consumer chooses the bundle on the true budget line where the marginal rate of substitution equals the perceived relative price. Behavior will be given by

$$u'(x) = (p + \hat{t})v'(y) \tag{24}$$

$$(p + t)x + y = Z \Leftrightarrow (p + \hat{t})x + y = Z - (t - \hat{t})x \tag{25}$$

We wish to understand this behavior as an attempt to choose based on perceived marginal prices and a budget adjustment rule. Before the consumer adjusts to meet the true budget constraint, we have equations (1) and (2) from Section 3:

$$u'(\hat{x}) = (p + \hat{t})v'(\hat{y}) \tag{26}$$

$$(p + \hat{t})\hat{x} + \hat{y} = Z \tag{27}$$

Recall that the budget adjustment rule ρ, defined as the share in the total reduction in expenditure falling on x as we move from $(\hat{x}, \hat{y})$ to (x, y), is given by equation (3):

$$(p + t)x = (p + t)\hat{x} - \rho(t - \hat{t})\hat{x} \tag{28}$$

Proposition 0. *The behavior described by equations* (24) *and* (25) *corresponds to a budget adjustment rule*

$$\rho_3 \simeq (p + t) \left.\frac{\partial x}{\partial Z}\right|_{t=\hat{t}} \frac{x}{\hat{x}} = \omega_x\big|_{t=t}\, \eta_{x,z}\big|_{t=\hat{t}}$$

where the approximation ignores third-order and higher derivatives of the utility function.

Proof. Taking a Taylor Series expansion of $u'(x)$ and $v'(y)$ we can write

$$u'(x) \simeq u'(\hat{x}) + u''(\hat{x})(x - \hat{x})$$

$$v'(y) \simeq v'(\hat{y}) + v''(\hat{y})(y - \hat{y})$$

Applying equation (26) and the above approximations, we have that equation (24) becomes approximately

$$u''(\hat{x})(x-\hat{x}) \simeq (p+\hat{t})v''(\hat{y})(y-\hat{y}) \tag{29}$$

Subtracting equation (27) from equation (25) and re-arranging yields

$$y-\hat{y} = (p+\hat{t})\hat{x} - (p+t)x = (\hat{t}-t)\hat{x} - (p+t)(x-\hat{x})$$

Plugging this into equation (29) and re-arranging gives

$$x-\hat{x} \simeq \frac{-(p+\hat{t})v''(\hat{y})}{u''(\hat{x})+(p+\hat{t})^2v''(\hat{y})}(t-\hat{t})x \tag{30}$$

Differentiating equations (26) and (27) with respect to Z and re-arranging gives

$$\left.\frac{\partial\hat{x}}{\partial Z}\right|_{t=\hat{t}} = \left.\frac{\partial x}{\partial Z}\right|_{t=\hat{t}} = \frac{(p+\hat{t})v''(\hat{y})}{u''(\hat{x})+(p+\hat{t})^2v''(\hat{y})} \tag{31}$$

The first equivalence comes from the fact that $\hat{x}=x$ when $\hat{t}=t$. Now we can write equation (30) as:

$$x-\hat{x} = -\frac{\partial x}{\partial Z}(t-\hat{t})x \tag{32}$$

From equation (28) we now know that $(p+t)(x-\hat{x}) = -\rho(t-\hat{t})\hat{x}$ so equation (32) becomes

$$-\rho(t-\hat{t})\hat{x} = -(p+t)\frac{\partial x}{\partial Z}(t-\hat{t})x \tag{33}$$

which simplifies to the desired result. Apply the definition of ω_x and $\eta_{x,Z}$, taking care to evaluate the former at $t=t$ and latter at $t=\hat{t}$, to obtain the second characterization of ρ_3. □

The Compensated Tax Effect

Lemma 1 (The Compensated Tax Effect with No Debiasing)**.** *When individuals behave in the manner described by equations* (1) *through* (4) *and there is no debiasing, then the compensated effect of a change in the tax rate starting from any point where $\hat{t}=t$ is given by*

$$\frac{\partial x^c}{\partial t_j} = \frac{\partial x^c}{\partial p} * \frac{\partial\hat{t}}{\partial t_j} + \left[x\frac{\partial x}{\partial Z} - \rho\frac{x}{p+t}\right]\left(1-\frac{\partial\hat{t}}{\partial t_j}\right) \tag{34}$$

where all expressions are evaluated at the current tax rate, $t=\hat{t}$.

Proof. Equation (34) is derived by taking the derivative with respect to t_j of equation (3), which we can re-write as

$$x = \hat{x}(1-\rho\frac{t-\hat{t}}{p+t}) \tag{35}$$

Taking the derivative with respect to t_j and re-arranging gives

$$\frac{\partial x}{\partial t_j} = \frac{\partial\hat{x}}{\partial\hat{t}}\frac{\partial\hat{t}}{\partial t_j}\left(1-\rho\frac{t-\hat{t}}{p+t}\right) - \rho\frac{\hat{x}}{p+t}\left(1-\frac{\partial\hat{t}}{\partial t_j}-\frac{t-\hat{t}}{p+t}\right) \tag{36}$$

When we evaluate this derivative at $\hat{t} = t$, $\frac{t-\hat{t}}{p+t} = 0$. Note that $\frac{\partial \hat{x}}{\partial \hat{t}} = \frac{\partial x}{\partial p}$ when $t = \hat{t}$. We then apply the definitions of $\frac{\partial x^c}{\partial t_j}$ and $\frac{\partial x^c}{\partial p}$ so that equation (36) becomes

$$\frac{\partial x^c}{\partial t_j} = \left(\frac{\partial x^c}{\partial p} - x\frac{\partial x}{\partial Z}\right)\frac{\partial \hat{t}}{\partial t_j} - \rho\frac{\hat{x}}{p+t}\left(1 - \frac{\partial \hat{t}}{\partial t_j}\right) + x\frac{\partial x}{\partial Z} \tag{37}$$

Re-arranging terms then yields the result. □

An application in Section 5.2 uses that the compensated tax effect is zero even in the presence of misperceived taxes—i.e. when $\hat{t} \neq t$—in the case where 1) $\partial \hat{t}/\partial t_j = 0$ 2) $\rho = \rho_3$, and 3) there are no income effects on demand for x under full optimization, so $\partial \hat{x}/\partial Z = 0$. This may be seen using equation (36) where $\partial \hat{t}/\partial t_j = 0$, plugging in for $\rho = 0$ and $\partial \hat{t}/\partial t_j = 0$, and then add $\hat{x}\partial \hat{x}/\partial Z = 0$ to obtain the compensated tax effect (which is the same as the uncompensated effect because there are no income effects).

Lemma 2 (The Compensated Tax Effect with Debiasing). *Suppose individuals behave in the manner described in Section 3 and the government has recourse to two taxes t_E and t_I. Suppose also that*

1. *Consumers only notice t_I, $\hat{t} = t_I$.*
2. *The budget adjustment rule is $\rho = \rho_3$.*
3. *There are no income effects on the demand for x under full optimization.*
4. *A fraction $F(G)$ of individuals debias*

Then the compensated effect of an increase in t_E of Δt_E for an initial tax rate t_{E0} is given by

$$\left.\frac{\partial X^c}{\partial t}\right|_{t_{E0}} = f(G)\frac{dG}{dt_E}(x^* - x) + F(G)\frac{\partial x^{*c}}{\partial p} \tag{38}$$

where X is aggregate demand for good x, $G = G(t_{E0})$.

Proof. We can write aggregate demand as

$$X(p,t,Z) = F(G(t-\hat{t}))x^* + [1 - F(G(t-\hat{t}))]x(p,t,Z)$$

where x^* is demand under full optimization and x is demand as described by proposition 1. Differentiating with respect to t we obtain

$$\frac{\partial x}{\partial t_E} = f(G)\frac{\partial G}{\partial t}(x^* - x) + F(G)\frac{\partial x^*}{\partial p} + [1 - F(G)]\frac{\partial x}{\partial t}$$

Because there are no income effects in the positive model with debiasing, we can write

$$\frac{\partial x^c}{\partial t_E} = f(G)\frac{\partial G}{\partial t}(x^* - x) + F(G)\frac{\partial x^{*c}}{\partial p} + [1 - F(G)]\frac{\partial x^c}{\partial t}$$

Equation (34) implies that when $\rho = \omega_x\eta = 0$ and $\partial \hat{t}/\partial t_E = 0$, we will have $\frac{\partial x^c}{\partial t_E} = 0$. This proves the result.

□

Excess Burden and Budget Adjustment Rules

Proposition 2 (How Does Excess Burden Depend on the Budget Adjustment Rule?)**.** *Assume that the government introduces a small misperceived tax t_j into an untaxed market. Behavior is characterized by the function $\hat{t}(t_j)$, ρ, and equations* (1) *through* (4)*. The excess burden of this tax is approximately*

$$EB(t_j) \simeq -\frac{1}{2}t_j^2\left(\frac{\varepsilon^c_{x,q|p}\frac{\partial\hat{t}}{\partial t_j} + [\omega_x\eta_{x,Z} - \rho]\left(1 - \frac{\partial\hat{t}}{\partial t_j}\right)}{\varepsilon^c_{x,q|p}}\right)^2 x\frac{\varepsilon^c_{x,q|p}}{p} \tag{39}$$

where $\xi = (\omega_x\eta_{x,Z} - \rho)\left(1 - \frac{\partial\hat{t}}{\partial t_j}\right)$ and all expressions are evaluated at $t = 0$.

Proof. The result follows directly from lemma 1 and proposition 1. Re-write the equation for excess burden from proposition 1 in the presence of no existing taxes, using the definition of θ^c, as

$$\begin{aligned} EB(t_j) &\simeq -\frac{1}{2}t_j^2\theta_j^c\frac{\partial x^c}{\partial t_j} && (40)\\ &= -\frac{1}{2}t_j^2(\theta_j^c)^2\frac{\partial x^c}{\partial p} && (41)\\ &= -\frac{1}{2}t_j^2\left(\frac{\frac{\partial x^c}{\partial t_j}}{\frac{\partial x^c}{\partial p}}\right)^2\frac{\partial x^c}{\partial p} && (42)\end{aligned}$$

Then insert the expression for $\frac{\partial x^c}{\partial t_j}$ from lemma 1 to obtain the result. □

Proposition 3 (How does the desirability of low-salience taxes depend on the budget adjustment rule?)**.** *Suppose that the government has recourse to two types of taxes, t_I and t_E, and that*

1. *Consumers only notice t_I: $\hat{t} = t_I$.*
2. *No individuals debias (see section 5.2).*

Then the introduction of a small tax individuals ignore, t_E, causes less excess burden than the introduction of a rate-equivalent tax that individuals perceive fully, t_I, if and only if $\rho < -\varepsilon_{x,q|p}$, where $\varepsilon_{x,q|p}$ is evaluated at $t_I = t_E = 0$.

Proof. The excess burden of a price inclusive tax will be

$$EB(t_I) \simeq -\frac{1}{2}t_I^2\frac{\partial x^c}{\partial p} \tag{43}$$

The excess burden of a price exclusive tax will be

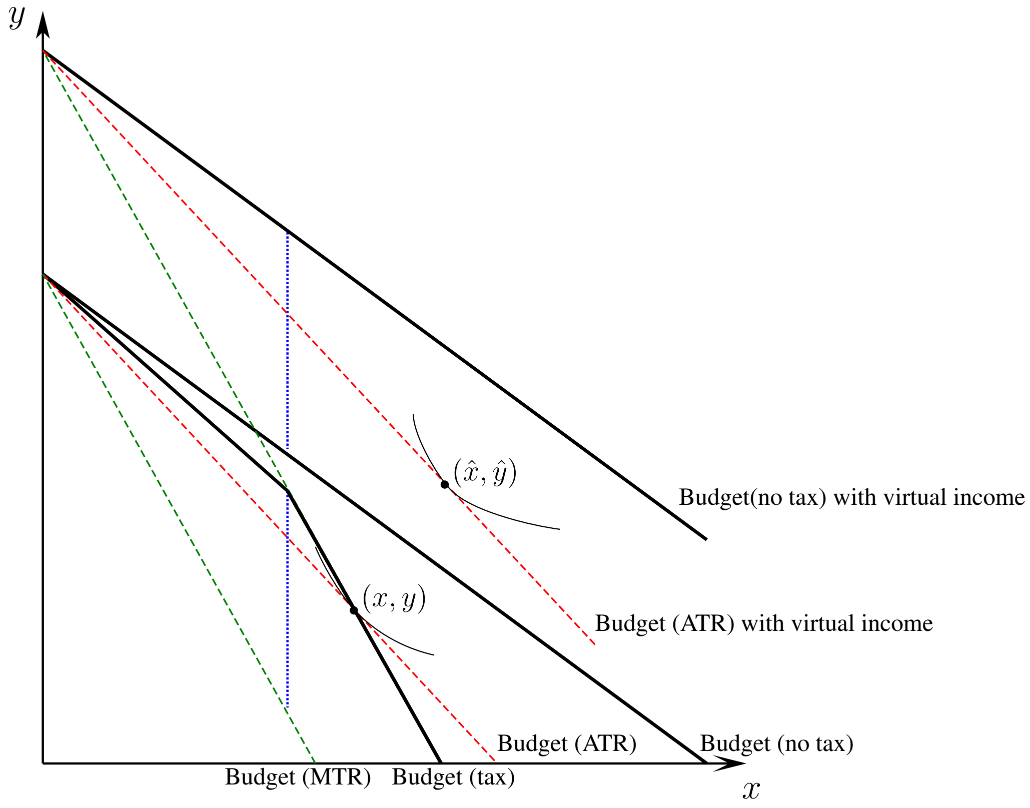
$$EB(t_E) \simeq -\frac{1}{2}t_E^2(\theta^c)^2\frac{\partial x^c}{\partial p} \tag{44}$$

When the taxes are rate-equivalent, we will have $EB(t_E) < EB(t_I) \iff (\theta^c)^2 < 1 \iff \theta^c < 1 \iff -\varepsilon^c_{x,q|t} < -\varepsilon^c_{x,q|p}$. (Note that we know $\theta \geq 0$). Using the elasticity formulation of proposition 1 and that $\frac{\partial\hat{t}}{\partial t_E} = 0$, we have that

$$EB(t_E) < EB(t_I) \iff -[\omega_x\eta_{x,Z} - \rho] < -\varepsilon^c_{x,q|p} \tag{45}$$

Re-arrange this equation and use the Slutsky equation $\varepsilon_{x,q|p} = \varepsilon^c_{x,q|p} - \omega_x\eta_{x,Z}$ to obtain the result. □

Figure 8: Budget Adjustment Under Ironing



Notes: The budget line with tax is drawn for a two-bracket linear tax on x with a higher marginal tax rate in the top bracket. We assume for comparison with figure 3 that $u'(x) > 0$, i.e. that x is a good and not income. An ironing individual assuming she faces a linear tax equal to the ATR she ends up paying will consume (x, y). Virtual income, the income not lost to taxation due to a lower marginal tax rate for some levels of x is equal to the length of the blue dotted line. This is calculated by comparing the budget line under the non-linear tax with the budget line for a linear tax with marginal tax rate equal to the top marginal tax rate (note that the price of y is normalized to 1). We add virtual income to the budget and assume the individual faces a linear tax equal to ATR to find the planned consumption bundle $(\hat{x}, \hat{y})$. Movement from $(\hat{x}, \hat{y})$ to (x, y) is determined by the income elasticity and x's share of the budget, just like in the third panel of figure 3.

Ironing

Figure 8 describes how the budget adjustment rule ρ_3 works in the Schmeduling model, where there are non-linear taxes. The figure illustrates why assuming that an individual assumes she faces a linear tax with rate equal to her average tax rate (and no exemption) is the same as assuming that she accounts for the impact of taxes on her virtual income but misperceives relative prices. The same techniques could be used to examine other budget adjustment rules.

Proposition 4. *Assume that*

1. *Individuals engage in ironing, so $\hat{\tau} = ATR \equiv \int_0^x \tau(s)ds/x$.*
2. *The budget adjustment rule is $\rho = 0$.*
3. *There are no income effects on demand for x.*

4. No individuals debias (see section 5.2).

Then the excess burden of a non-linear tax $\tau(x)$ will be given by

$$\begin{aligned} EB(\tau(x)) &\simeq -\frac{1}{2}\tau^2 \frac{ATR}{\tau}\frac{\partial x}{\partial \tau} && (46) \\ &= \frac{1}{2}\tau ATR x_0 \frac{\varepsilon_{x,1-\tau|\tau}}{1-ATR} && (47) \\ &= \frac{1}{2}ATR^2 x_0 \frac{\varepsilon_{x,1-\tau|ATR}}{1-ATR} && (48) \end{aligned}$$

where $\varepsilon_{x,1-\tau|\tau}$ is the elasticity of taxable income with respect to the net-of-tax rate, and $\varepsilon_{x,1-\tau|ATR}$ is the elasticity of taxable income under full optimization, i.e. the elasticity of demand with respect to the net of tax rate following the introduction of a linear tax equal to ATR (which also equals the elasticity of x with respect to the price). Derivatives and elasticities are evaluated at the no-tax equilibrium $\tau = ATR = 0$.

Proof. Applying proposition 2, adapted to the income tax case, normalizing to $p = 1$ so $t = \tau$, and assuming $\rho = \omega_x \eta = 0$ gives[42]

$$EB(\tau) = -\frac{1}{2}\tau^2 \left(\frac{\partial \hat{\tau}}{\partial \tau}\right)^2 x_0 \frac{\varepsilon^c_{x,q|p}}{1+\hat{\tau}} \qquad (49)$$

Note that $\tau^2 \left(\frac{\partial \hat{\tau}}{\partial \tau}\right)^2 \simeq \hat{\tau}^2$ (see footnote 25). Note also that $\varepsilon^c_{x,q|p} = \varepsilon^c_{x,q|\hat{\tau}} = -\varepsilon^c_{x,q|1-\hat{\tau}}$. Plugging in $\hat{\tau} = ATR$ yields equation (48).

Noting that $\varepsilon^c_{x,q|\hat{\tau}} = \varepsilon^c_{x,q|\tau}\frac{\partial \hat{\tau}}{\partial \tau}$ we can rewrite equation (49) as

$$EB(\tau) \simeq \frac{1}{2}\tau^2 \frac{\partial \hat{\tau}}{\partial \tau} x_0 \frac{\varepsilon^c_{x,q|\tau}}{1+\hat{\tau}} \qquad (50)$$

Note that when we introduce a nonlinear tax we will have $\theta = \frac{\partial \hat{\tau}}{\partial \tau} \simeq \frac{ATR}{\tau}$, so we will have

$$EB(\tau) \simeq \frac{1}{2}\tau ATR x_0 \frac{\varepsilon^c_{x,q|\tau}}{1+\hat{\tau}}$$

□

[42] One difference between this and the formulation in proposition 2 is that the denominator of the elasticity is evaluated at $\hat{\tau}$ instead of at $\tau_0 = 0$. We are free to use the latter, to be consistent with Liebman and Zeckhauser (2004), because to a second order approximation this will not matter. To see why note that

$$\frac{1}{1-\hat{\tau}} = 1 - \hat{\tau} + \hat{\tau}^2 - \hat{\tau}^3 + ...$$

so multiplying by τ^2 and ignoring third order terms we will have

$$\tau^2 \simeq \frac{\tau^2}{1-ATR}$$

Debiasing

Proposition 5 (Extension to Non-Linearities in Tax Rates)**.** *Assume there is only one type of misperceived tax, t_j. Under assumptions **A1** and **A2**, and the assumption that demand is locally linear in $(p+\hat{t})$ and Z but not in tax rates, the excess burden at a tax rate t_j is*

$$\begin{aligned}
EB(t_j) &= -\frac{1}{2}t_j^2\Theta^c\frac{1}{t-\hat{t}}\int_{\hat{t}}^{t}\left.\frac{\partial x^c}{\partial t_j}\right|_{t=t'}dt' && (51)\\
&= -\frac{1}{2}t_j^2(\Theta^c)^2\frac{\partial x}{\partial p} && (52)\\
&= -\frac{1}{2}t_j^2(\Theta^c)^2 x_0\frac{\varepsilon_{x,q|p}}{p}, && (53)
\end{aligned}$$

where $x_0 = x(p,\mathbf{0},Z)$ and $\frac{\partial x}{\partial p}$ is evaluated at $t = 0$. The excess burden of a small tax increase Δt_j starting from an initial tax rate t_0 is approximately:

$$\begin{aligned}
EB(\Delta t_j|t_0) &\simeq -\frac{1}{2}(\Delta t_j)^2\theta^c_{j,t_0}\left.\frac{\partial x^c}{\partial t_j}\right|_{t=t_0} - \Delta t_j\left.\frac{\partial x^c}{\partial t_j}\right|_{t=t_0}[\hat{t}_0+\Theta^c(t_0-\hat{t}_0)] && (54)\\
&= \frac{1}{2}(\Delta t_j)^2\theta^c_{j,t_0}x_0\frac{\varepsilon^c_{x,q|t_j}}{p+t_0}+\Delta t_j x_0\frac{\varepsilon^c_{x,q|t_j}}{p+t_0}[\hat{t}_0+\Theta^c(t_0-\hat{t}_0)] && (55)
\end{aligned}$$

where $\frac{\partial x^c}{\partial t_j}$ is evaluated at $t = t_0$, and $\frac{\partial x^c}{\partial p}$ at $t = \hat{t}_0$.

Proof. The proof is virtually identical to the proof in Chetty (2009), but instead of approximating the difference between fully optimizing demand and actual demand by $x^* - x \simeq (1-\theta_j)\frac{\partial x}{\partial p}(t-\hat{t})$, I approximate it, using that there is only one type of misperceived tax, as

$$\begin{aligned}
x^*(p,\mathbf{t},Z) - x(p,\mathbf{t},Z) &= (x^*(p,\mathbf{t},Z) - x(p+\hat{t},\mathbf{0},Z)) - (x(p,\mathbf{t},Z) - x(p+\hat{t},\mathbf{0},Z))\\
&\simeq \left(\frac{\partial x}{\partial p} - \frac{1}{t-\hat{t}}\int_{\hat{t}}^{t}\left.\frac{\partial x}{\partial t_j}\right|_{t=t'}dt'\right)(t-\hat{t})\\
&= (1-\Theta)\frac{\partial x}{\partial p}(t-\hat{t})
\end{aligned}$$

to allow for nonlinearity in tax rates. The rest of the proof follows the one in Chetty (2009), almost to the letter.

For the marginal excess burden approximation, note that by the fundamental theorem of calculus,

$$\theta_{j,t_0} = \left.\frac{\partial\Theta}{\partial t_j}\right|_{t_0}.$$

So for a small tax change Δt_j, we will have

$$(x^*(p,\mathbf{t_0}+\Delta t_j,Z) - x^*(p,\mathbf{t_0},Z)) - (x(p,\mathbf{t_0}+\Delta t_j,Z) - x(p,\mathbf{t_0},Z)) \simeq (1-\theta_{j,t_0})\frac{\partial x}{\partial p}\Delta t_j$$

Taking a Taylor series expansion of equation (51) about t_0, applying this approximation and ignoring third-order or higher terms yields the desired result. □

Proposition 6 (The Curse of Debiasing)**.** *Suppose that the government has recourse to two taxes, t_I and t_E, and that*

1. *Consumers respond perfectly to changes in t_I, so $\theta^c_{t_I} = 1$.*

2. *Consumers initially under-respond to changes in t_E, so $\theta^c_{t_E} < 1$ when $t_E = 0$.*
3. *Assumptions **A1**, **A2**, and **A3** are true.*
4. *Optimal demand x^* is strictly positive at all prices and tax rates.*

Then there exists a tax rate t'_E at which an increase in t_E causes higher *marginal excess burden than an increase in the noticed tax t_I.*

Proof. The first step of the proof shows that whenever the marginal degree of error for a change in t_E, $\theta_{t_{0E}} > 1$, an increase in t_E causes higher marginal excess burden than an identical increase in t_I. The second step proves the existence of a tax rate t_{0E} at which $\theta_{t_{0E}} > 1$ using the mean value theorem.

Step 1. Write the difference between the marginal excess burden of a small increase in t_E minus the marginal excess burden of a small increase in t_I, given initial tax rates t_{I0} and t_{E0} as follows, using proposition 4/ 5 and the fact that $\partial x/\partial t_I = \partial x/\partial p$[43]

$$D = EB(\Delta t_E | t_{E0}, t_{I0}) - EB(\Delta t_I | t_{E0}, t_{I0}) \tag{56}$$

$$= \begin{array}{l} -\frac{1}{2}(\Delta t_E)^2 \theta^c_{t0E} \frac{\partial x^c}{\partial t_E} - \Delta t_E \frac{\partial x^c}{\partial t_E}[t_{0I} + \Theta^c t_{0E}] \\ -\{-\frac{1}{2}(\Delta t_I)^2 \frac{\partial x^c}{\partial t_I} - \Delta t_I \frac{\partial x^c}{\partial t_I}[t_{0I} + \Theta^c t_{0E}]\} \end{array} \tag{57}$$

where derivatives are evaluated at (t_{0I}, t_{0E}). Now imposing that the increase in the per-unit tax rate be the same for both taxes, $\Delta t_E = \Delta t_I = \Delta t$, and recalling that $\theta^c_{t_j} = \frac{\partial x^c}{\partial t_j} \Big/ \frac{\partial x}{\partial p}$, we write equation (57) as:

$$-\frac{1}{2}\Delta t^2[(\theta^c_{t_{0E}})^2 - 1]\frac{\partial x^c}{\partial p} - \Delta t[t_{0I} + \Theta^c t_{0E}][\theta^c_{t_0 E} - 1]\frac{\partial x^c}{\partial p} \tag{58}$$

We therefore know that $D > 0$ iff $\theta^c_{t_{0E}} > 1$.

Before the second part of the proof can be completed, we need to derive a relationship between $\theta^c_{t_{0E}}$ and $\theta_{t_{0E}}$. Write

$$\theta_{t^c_{0E}} = \frac{\theta \frac{\partial x}{\partial p} + x\frac{\partial x}{\partial Z}}{\frac{\partial x}{\partial p} + x\frac{\partial x}{\partial Z}}$$

and use the negativity of the compensated price effect to derive that for non-Giffen goods (which are ruled out by additive separability), we will have

$$\theta^c_{t_{0E}} > 1 \iff \theta\frac{\partial x}{\partial p} + x\frac{\partial x}{\partial Z} < \frac{\partial x}{\partial p} + x\frac{\partial x}{\partial Z} \iff \theta\frac{\partial x}{\partial p} < \frac{\partial x}{\partial p} \iff \theta_{t_{0E}} > 1$$

Step 2 Write the difference between optimal demand and true demand as

$$x^*(p + t_{0I}, t_{0E}, Z) - x(p + t_{0I}, t_{0E}, Z) \simeq (1 - \Theta)\frac{\partial x}{\partial p}t_{0E} \tag{59}$$

When $t_{E0} = 0$, $x^* - x = 0$ by **A2**. By **A3**, there is a tax rate t' such that when $t_E = t'_E$, $x^* - x = 0$. Take a derivative

[43] Applying proposition 4 in this way requires that we let $\hat{t}$ be any components of the tax for which the individual optimizes fully at all tax rates. That is $x(p, \hat{t}, Z) = x^*(p + \hat{t}, Z)$.

of equation (59) to obtain[44]

$$\frac{\partial(x^* - x)}{\partial t_E} = (1 - \theta_{t_{0E}})\frac{\partial x}{\partial p} \tag{60}$$

By the second condition of the proposition, we can take a small tax rate t_E such that $\theta_{t_{0E}} < 1$. At this tax rate, $x^* - x < 0$. The mean value theorem therefore implies the existence of a tax rate t_E^* such that

$$\frac{\partial(x^* - x)}{\partial t_E} > 0 \Leftrightarrow \theta_{t_E^*} > 1 \Leftrightarrow \theta^c_{t_E^*} > 1$$

□

Finally, we need to derive θ and Θ for the positive model in section 5.2. We assume $\partial\hat{t}/\partial t = 0$, and $\rho = \omega_x\eta = 0$. Specifically, we will see that

$$\Theta^c_{t_{E0}} = F(G(t_E)),$$

and

$$\theta^c_{t_{E0}} \simeq F(G) + f(G)\frac{\partial G}{\partial t_E}t_{E0}.$$

First consider the marginal degree of error:

$$\theta^c_{t_{0E}} \equiv \frac{\left.\frac{\partial x^c}{\partial t_E}\right|_{t=t_{E0}}}{\left.\frac{\partial x^c}{\partial p}\right|_{t=\hat{t}_0}}.$$

Equation (38) implies that

$$\left.\frac{\partial X^c}{\partial t_E}\right|_{t_{E0}} = F(G)\frac{\partial x^{*c}}{\partial p} + f(G)\frac{dG}{dt_E}(x^* - x)$$

Applying our simplifying assumptions again, we can approximate $(x^* - x)$ by

$$x^* - x \simeq (\partial x/\partial p)t_{E0} = (\partial x^c/\partial p)t_{E0}$$

So we will have

$$\left.\frac{\partial X^c}{\partial t_E}\right|_{t_{0E}} = F(G)\frac{\partial x^{*c}}{\partial p} + f(G)\frac{dG}{dt_E}t_{E0}\frac{\partial x^c}{\partial p}$$

Dividing by $\partial x^c/\partial p$ gives the desired expression for $\theta^c_{t_0}$.

Second, consider the *average degree of compensated error* at some tax rate t_0:

$$\begin{aligned}
\Theta^c &\equiv \frac{\frac{1}{t_0 - \hat{t}_0}\int_{\hat{t}_0}^{t_0} \left.\frac{\partial x^c}{\partial t_E}\right|_{t'} dt'}{\left.\frac{\partial x^c}{\partial p}\right|_{\hat{t}_0}} \\
&= \frac{1}{t_0 - \hat{t}_0}\int_{\hat{t}_0}^{t} \theta^c_{t'} dt' \\
&= \frac{1}{t_{0E}}\int_0^{t_E} \theta^c_{E,t'} dt'
\end{aligned}$$

[44] This uses the fact that

$$\frac{\partial\Theta}{\partial t} = \frac{1}{t_{0E}}\left(\theta_{t_{0E}} - \Theta\right)$$

Using the expression we just derived for $\theta_{t_{E0}}$ we observe that the anti-derivative of the argument of the integral $\int_{\hat{t}_0}^{t} \theta_{t'}^{c} dt'$ will be $F(G)(t-\hat{t}) = F(G)t_{0E}$, and by evaluating this expression at the limits of integration, noting that $F(G(\hat{t}-\hat{t})) = F(G(0)) = 0$ and dividing by $t - \hat{t} = t_{0E}$ we obtain that $\Theta = F(G)$.

Note that Proposition 6 is proven in the text of the paper, prior to the statement of the proposition.