THE YALE LAW JOURNAL

JACOB GOLDIN

Which Way To Nudge? Uncovering Preferences in the Behavioral Age

ABSTRACT. Behavioral Law and Economics has created a dilemma for policymakers. On the one hand, research from the field suggests a wide range of unconventional policy instruments ("nudges") may be used to shape people's voluntary choices in order to lead them to the option they most prefer. On the other hand, the very nature of these new instruments precludes researchers from measuring people's preferences in the traditional way, i.e., by evaluating which option people choose from the set of available choices. As a result, policymakers often lack the information they need to design nudges that will make people better off. To tackle this dilemma, I propose a new framework that focuses on the distinction between those decision makers who respond to nudges and those who do not. The framework highlights that existing methods for designing nudges come up short—none accounts for what I argue is the crucial piece of information: the preferences of the nudge-sensitive decisionmakers. After exploring this dilemma, the Essay describes two new approaches for uncovering the preferences of this group and argues that they hold promise for informing the design of nudges in a wide range of policy settings.

AUTHOR. Assistant Professor of Law, Stanford Law School (beginning Fall 2016). Many of the ideas in this Essay were developed in joint work with Daniel Reck. For helpful discussions and suggestions, I am grateful to Ian Ayres, Daniel Deacon, Joshua Fischman, Barbara Fried, Jed Glickstein, Christine Jolls, Yair Listokin, Leah Litman, Jonathan Masur, Ariel Porat, Maria Ponomarenko, Jeffrey Rachlinski, Erin Scharff, and Cass Sunstein. Jack Boeglin provided excellent editorial suggestions.

ESSAY CONTENTS

INTRODUCTION

Governments face frequent and inescapable decisions about choice architecture—the features of legal and institutional design that affect people's voluntary choices.[1] For example:

- Should a bank be allowed to automatically enroll new account holders into "overdraft protection"[2] unless they opt out, or should customers have to opt in to receive this type of service?
- In what order should a government website list the prescription drug plans that are available for Medicare Part D beneficiaries to enroll in?[3]
- Should the packaging for a medical drug disclose the product's effectiveness in terms of its success rate or its failure rate?[4]
- Should the government compel companies like Facebook and Google to adopt default privacy settings that prohibit advertisers from using customers' personal data to predict which advertisements the customer will enjoy?[5]

Historically, law and economics scholars would answer each of these questions the same way: it doesn't matter. Legal rules shape behavior by

1. Cass R. Sunstein & Richard H. Thaler, *Libertarian Paternalism Is Not an Oxymoron*, 70 U. CHI. L. REV. 1159, 1164-65 (2003).

2. Overdraft protection involves a bank honoring a customer's attempt to transfer funds out of the account in excess of the account's balance—in exchange for a fee. *See* Lauren E. Willis, *When Nudges Fail: Slippery Defaults*, 80 U. CHI. L. REV. 1155, 1174-1200 (2013) (discussing how choice architecture affects whether bank account holders choose to enroll in overdraft protection).

3. *See* Jason Abaluck & Jonathan Gruber, *Choice Inconsistencies Among the Elderly: Evidence from Plan Choice in the Medicare Part D Program*, 101 AM. ECON. REV. 1180 (2011) (identifying systematic errors in Medicare Part D plan selection); Jonathan Levav et al., *Order in Product Customization Decisions: Evidence from Field Experiments*, 118 J. POL. ECON. 274 (2010) (documenting an interaction between default effects and the order in which options are presented).

4. *See* Donald A. Redelmeier et al., *Understanding Patients' Decisions: Cognitive and Emotional Perspectives*, 270 JAMA 72, 73 (1993) (reporting that more patients chose to select a treatment when presented with a 90% survival rate than a 10% death rate).

5. *See* Eric J. Johnson et al., *Defaults, Framing and Privacy: Why Opting In-Opting Out*, 13 MARKETING LETTERS 5 (2002) (reporting experimental findings that defaults shape decisions about online privacy). Of course, one option is for the government to leave the design of choice architecture to the company. In that case, the issues discussed here remain relevant as long as the company seeks to promote its customers' well-being when designing its product. *See* Daniel G. Goldstein et al., *Nudge Your Customers Toward Better Choices*, HARV. BUS. REV., Dec. 2008, at 99, 99-103 (discussing the design of choice architecture by private firms and the required balancing of "an array of interests, including customers' wishes and the company's desire to maximize profits and minimize risk").

changing people's incentives; policies that don't affect incentives don't affect behavior.

But the old view is strikingly out of date. Research conducted in behavioral law and economics over the past few decades has found that even features of a decision that don't affect the chooser's incentives – such as which option is the default or what order the choices are presented in – may nonetheless shape what people choose, and often in substantial ways.[6] A growing chorus of academics has called on governments to utilize these findings when designing choice architecture to "nudge" people's choices in directions that will improve the chooser's own welfare.[7] And policymakers have listened, placing advocates of this approach in powerful administrative positions and even creating entire "nudge squads" within government devoted to the idea.[8]

In this Essay, I argue that efforts to design choice architecture to improve people's welfare clash head-on against a tension rooted in the foundations of behavioral law and economics itself. The tension is this: using nudges to make people better off requires knowing people's preferences,[9] but the very fact that

6. *See, e.g.*, RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS 3 (2008) ("[S]mall and apparently insignificant details can have major impacts on people's behavior. A good rule of thumb is to assume that 'everything matters.'").

7. *See, e.g.*, Colin Camerer et al., *Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism,"* 151 U. PA. L. REV. 1211, 1212 (2003) (proposing an approach to policymaking in which a regulation is adopted if it "creates large benefits for those who make errors, while imposing little or no harm on those who are fully rational"); Sunstein & Thaler, *supra* note 1.

8. *See, e.g.*, CASS R. SUNSTEIN, SIMPLER: THE FUTURE OF GOVERNMENT (2013) (describing numerous examples of behaviorally informed regulation developed while the author served in the Obama Administration); Katrin Bennhold, *Britain's Ministry of Nudges*, N.Y. TIMES (Dec. 7, 2013), http://www.nytimes.com/2013/12/08/business/international/britains-ministry-of-nudges.html [http://perma.cc/F8ZZ-VHEV]; René van Bavel et al., *Applying Behavioural Sciences to EU Policy-Making*, EUR. COMMISSION (2013), http://ec.europa.eu/dgs/health_consumer/information_sources/docs/30092013_jrc_scientific_policy_report_en.pdf [http://perma.cc/9HQX-CH7J] (describing the advantages of policymakers with knowledge of how people behave); Mike Dorning, *Obama Adopts Behavioral Economics*, BLOOMBERG BUSINESSWEEK (June 24, 2010), http://www.businessweek.com/magazine/content/10_27/b4185019573214.htm [http://perma.cc/8HYM-KARF]; Maxim Lott, *Gov't Knows Best? White House Creates 'Nudge Squad' To Shape Behavior*, FOX NEWS (July 30, 2013), http://www.foxnews.com/politics/2013/07/30/govt-knows-best-white-house-creates-nudge-squad-to-shape-behavior [http://perma.cc/JAB9-34BV].

9. As I will use the term, one's *preferences* measure the degree to which choosing an option furthers the decision makers' objectives, whatever those may be. Welfare is defined in terms of preference satisfaction: a nudge increases welfare when it induces people to make choices that better reflect their preferences. It is important to emphasize that I am not defining preferences in terms of people's choices, as economists typically do, because that would (tautologically) rule out the possibility that people ever make mistakes, and thus the potential for nudges to affect welfare. For excellent discussions of this issue, see KAUSHIK

people respond to nudges means that one can't learn their preferences simply by observing their choices. Previous theories for designing nudges have overlooked the magnitude of this problem by focusing on the wrong piece of information—the majority preferences of all decision makers—when what really matters for designing nudges are the preferences of the *subgroup* of decision makers whose choices are affected by the nudge. And it is the decision makers in this subgroup for whom the preference identification problem is most extreme. One cannot learn which option a person prefers by looking to see which option she chooses when what she chooses varies according to preference-irrelevant factors—i.e., which way she is nudged. Because of this dilemma, the case for nudging often founders because it cannot answer a basic question: which option should people be nudged toward? The goal of this Essay is to clarify the nature of the challenge and to propose some concrete ways of overcoming it.

To illustrate a situation in which the problem of *behavioral preference identification* (as I label the tension described above) is likely to arise, consider the choice people face when deciding what type of car insurance to buy.[10] Suppose a state requires auto insurers to offer both limited- and full-liability insurance to customers, and the state must decide which type of insurance should be the default.[11] That is, the state could either require auto insurers to present limited insurance as the default and give consumers the option of upgrading to full insurance for an added fee, or instead present full insurance as the default and give consumers the option of downgrading to limited insurance in exchange for a price discount. Which insurance is the default is likely to affect which insurance some of the customers decide to purchase,[12] so the state's choice matters. Moreover, customers are likely to differ in which

BASU, PRELUDE TO POLITICAL ECONOMY: A STUDY OF THE SOCIAL AND POLITICAL FOUNDATIONS OF ECONOMICS (2003); and Amartya K. Sen, *Rational Fools: A Critique of the Behavioral Foundations of Economic Theory*, 6 PHIL. & PUB. AFF. 317, 322-24 (1977). For a defense of an alternative approach, which does define preferences in terms of voluntary choices in a behavioral context, see B. Douglas Bernheim, *Behavioral Welfare Economics* (Nat'l Bureau of Econ. Research, Working Paper No. 14622, 2008).

10. This example comes from Eric J. Johnson et al., *Framing, Probability Distortions, and Insurance Decisions*, 7 J. RISK & UNCERTAINTY 35, 48 (1993), and is discussed in Camerer et al., *supra* note 7, at 1226-27; and Sunstein & Thaler, *supra* note 1, at 1175-76.

11. During the 1990s, both New Jersey and Pennsylvania were faced with exactly this decision. New Jersey selected limited insurance as the default whereas Pennsylvania set the default to be full insurance. Johnson et al., *supra* note 10, at 48. With limited-liability insurance, the consumer accepts a restricted right to sue in exchange for a cheaper price. *Id.* Of course, a state could simply allow each company to select its own default, although doing so would be worse for consumers than mandating the default that maximizes consumer welfare (assuming that option can be identified).

12. *Id.*

type of insurance they prefer—some will value the affordability of limited insurance while others will care more about the reduced risk associated with the full insurance option. But because customers' insurance choices vary based on a preference-irrelevant feature of the decision (the default), policymakers cannot look to their choices to learn which option they actually prefer.[13] As a result, the state lacks the information it needs to determine whether a default of limited or full insurance will make people better off.

To be sure, there are two cases in which determining the best direction to nudge will not be so difficult. The first is when policymakers have strong (possibly paternalistic) views of their own about which of the available options would make decision makers better off. For instance, many experts believe that people should save more for retirement than they currently do.[14] If that view is correct, the best way to design the choice architecture that governs retirement savings decisions is to adopt whichever nudge moves saving rates closest to the level that the experts believe is optimal, such as by implementing automatic enrollment into savings plans with a high default contribution rate.[15] Similarly, if policymakers are confident that people would be better off eating healthier foods or smoking fewer cigarettes, they can design nudges to promote these objectives, such as by requiring calorie labeling at fast food restaurants[16] or eye-catching health warnings on cigarette packages.[17]

13. In their discussion of auto insurance defaults, Camerer and colleagues conclude, "This example reveals little about what the default ought to be but clearly illustrates the powerful effects defaults can have, suggesting the need to choose defaults carefully." Camerer et al., *supra* note 7, at 1227.

14. *See, e.g.*, Alicia H. Munnell et al., *The National Retirement Risk Index: An Update*, CTR. FOR RETIREMENT RES. BOS. C. (2012), http://crr.bc.edu/wp-content/uploads/2012/11/IB_12-20-508.pdf [http://perma.cc/5ZKV-UJH7]; Nari Rhee, *The Retirement Savings Crisis: Is It Worse Than We Think?*, NAT'L INST. ON RETIREMENT SECURITY (2013), http://www.nirsonline.org/storage/nirs/documents/Retirement%20Savings%20Crisis/retirementsavingscrisis_final.pdf [http://perma.cc/3BE3-A68B].

15. Automatic enrollment into savings plans with low default contribution rates can actually reduce total savings, if the increase in participation is more than offset by the reduction in the amount saved by those who adopt the default contribution rate. Ryan Bubb et al., *A Behavioral Contract Theory Perspective on Retirement Savings*, 47 CONN. L. REV. (forthcoming 2015) (manuscript at 36-37), http://papers.ssrn.com/abstract=2626069 [http://perma.cc/3JPN-EPR9].

16. *See* Patient Protection and Affordable Care Act, Pub. L. No. 111-148, § 4205, 124 Stat. 119, 573-74 (2010) (codified at 21 U.S.C. § 343(q)(5)(A) (2012)).

17. *See* R.J. Reynolds Tobacco Co. v. FDA, 696 F.3d 1205 (D.C. Cir. 2012); Required Warnings for Cigarette Packages and Advertisements, 21 C.F.R. § 1141 (2015). Similarly, in some settings policymakers will not know which option people prefer but will believe that the option people pick in one choice architecture design is more likely to reflect their preferences than the option they select in the other choice architecture design—e.g., when one decision is presented in a less confusing way than the other. *See* Christine Jolls & Cass R. Sunstein,

Second, the problem of behavioral preference identification is lessened when the decision under consideration involves substantial externalities to other members of society.[18] For example, nudges are frequently advocated for in the context of organ donation, where the strong positive externalities associated with there being additional organ donors provide grounds for implementing an opt-out organ donation default rule, quite apart from the preferences of the affected decision makers.[19] Similar arguments might inform the design of nudges to promote environmental goals[20] or the collection of tax revenue[21]—areas where one person's decisions have an effect on others in society.[22]

Although policymakers can circumvent the problem of behavioral preference identification when one of these conditions is met, settings like these represent the exception rather than the norm. Absent knowledge of people's choices, policymakers will typically lack reliable beliefs about which option will best promote people's welfare.[23] Consider the auto insurance example discussed above—is there any reason to predict that most drivers would be better off with full as opposed to limited insurance, or vice-versa?[24]

Debiasing Through Law, 35 J. LEGAL STUD. 199 (2006) (describing policy interventions along these lines).

18. *See* Russell Korobkin, *Libertarian Welfarism*, 97 CALIF. L. REV. 1651, 1671-83 (2009) (arguing that nudges should be used to shape decisions that produce externalities).

19. *See* Eric J. Johnson & Daniel Goldstein, *Do Defaults Save Lives?*, 302 SCI. 1338 (2003).

20. *See, e.g.*, Cass R. Sunstein & Lucia A. Reisch, *Green by Default*, 66 KYKLOS 398 (2013) (discussing the use of defaults to encourage pro-environmental behavior); Tatiana A. Homonoff, *Can Small Incentives Have Large Effects? The Impact of Taxes Versus Bonuses on Disposable Bag Use* (Princeton Univ. Indus. Relations Section, Working Paper No. 575, 2013) (finding that grocery shoppers are more likely to reduce disposable bag use in response to a small tax than to an equally sized bonus).

21. *See, e.g.*, Kathleen Delaney Thomas, *Presumptive Collection: A Prospect Theory Approach to Increasing Small Business Tax Compliance*, 67 TAX L. REV. 111, 117-18 (2013).

22. Note that even in such settings, assessing which nudge best promotes social welfare can require accounting for the preferences of the affected choosers, especially when those preferences conflict with the other policy goal being pursued.

23. In general, there is little reason to expect policymakers to be able to successfully divine people's preferences using information untethered to those people's voluntary choices. *See, e.g.*, JOHN STUART MILL, ON LIBERTY 80 (David Bromwich & George Kateb eds., Yale Univ. Press 2003) (1859) (expressing skepticism that governments will do a better job advancing people's welfare than what the people would do themselves). Russell Korobkin has put it well: "If I can't figure out whether I would be better off owning a car equipped with expensive airbags or a slightly more dangerous (and cheaper) automobile and the concomitant ability to purchase some other goods and services, why should I be confident that a state functionary can do any better?" Korobkin, *supra* note 18, at 1652.

24. Of course, one form of insurance might be associated with lower expected costs than the other, but which option a customer prefers depends on her degree of risk aversion—there is no one correct answer.

And in most cases, the externalities associated with a decision will be secondary in importance to the effect of the decision on the decision maker herself. Indeed, it seems likely that one of the reasons the retirement savings and organ donation examples have received so much attention in the literature on nudges is that they represent two of the rare major policy issues in which determining the right direction to nudge is relatively uncontroversial.[25]

The problem of behavioral preference identification thus creates a practical dilemma for nudging enthusiasts. Although behavioral law and economics has shown that important policy questions about choice architecture are ubiquitous, it has so far failed to provide any satisfactory tools for answering those questions when they arise. As a result, even policymakers who would like to implement welfare-enhancing nudges often lack the information needed to do so.[26] Without a principled method for choosing the best direction in which to nudge, policymakers are left to make decisions about choice architecture in an ad hoc manner or based on their own intuitions about what people prefer. Neither of those options is appealing; poorly designed choice architecture results in nudges that make people worse off.

25. No state has yet adopted opt-out organ donation, but the proposal that employees should be automatically enrolled into retirement savings plans has met with greater policy success. *See* Patient Protection and Affordable Care Act, Pub. L. No. 111-148, § 1511, 124 Stat. 191, 252 (2010) (codified at 29 U.S.C. § 218A (2012)) (requiring large firms to automatically enroll new employees into employer-provided health plans); 2014 Ill. Legis. Serv. 98-1150 (West) (to be codified at 30 ILL. COMP. STAT. 105/5.855) (Illinois law creating an automatic enrollment IRA funded by payroll deductions). More broadly, many of the recent policies that have been fashioned as nudges have as their goal furthering some pre-identified government goal—such as increasing retirement savings, promoting environmental conservation, encouraging healthy eating, or collecting tax revenue. Note that one cannot escape the problem of behavioral preference identification simply by confining nudges to settings like these where it is clear which of the available options will best promote a particular social goal. After all, even policies that aren't meant to be nudges still end up shaping choice architecture—and hence behavior. THALER & SUNSTEIN, *supra* note 6, at 10.

26. There are numerous areas of the law where nudging is inevitable but the best option to nudge toward is not obvious. To list just a few examples, consider: information disclosure (risk of accident from using a product, design of nutrition labels); consumer protection (default product features, digital privacy settings, likelihood of success from a medical drug or procedure); tort law (assignment of entitlements for property or liability rules); public benefits programs (option presentation on health insurance exchanges or Medicare Part D plans, timing of food stamps, welfare, or disability benefits disbursal); property law (intestacy rules); tax law (choice of entity defaults, assignment of tax dollars to presidential election campaign fund on Form 1040); federal student loan programs (default length of repayment plan; whether the IRS automatically shares borrower tax data with loan servicers for income-driven repayment plans); contract law (default rules for interpreting ambiguous provisions); and election law (order in which candidates are listed, alternative titles or descriptions for referenda).

Part I of the Essay sets out the problem of behavioral preference identification. I begin by identifying what information policymakers need to determine the optimal direction in which to nudge. To keep things simple I assume that people are choosing between two options and that the government's goal is to maximize the number of individuals who end up with the option they most prefer.[27] Those who have considered this question in prior work have concluded that the optimal nudge is majoritarian, in that it is best to nudge toward whichever option more people prefer.[28] In contrast, I argue that the optimal nudge is *subgroup majoritarian*: the best direction to nudge depends on the preferences of those individuals who are inconsistent, i.e., whose choices vary according to the direction of the nudge. Intuitively, it is the preferences of this subgroup—rather than the full population—that matter for determining the optimal nudge, precisely because only inconsistent people's choices are affected by the nudge's design. The consistent people don't enter into the equation because they end up choosing the same option regardless of the nudge's direction.[29]

After identifying what information matters for determining which way to nudge, the next question to ask is how that information can be obtained. And this is where the problem arises. The traditional method for identifying

27. My focus is on settings in which the government must design choice architecture in a uniform way for all decision makers, but the same issues arise when governments or private companies tailor the choice architecture based on the characteristics of the choosers. *See* Ariel Porat & Lior Jacob Strahilevitz, *Personalizing Default Rules and Disclosure with Big Data*, 112 MICH. L. REV. 1417 (2014) (arguing that default rules should be customized for different choosers); Cass R. Sunstein, *Deciding by Default*, 162 U. PA. L. REV. 1, 48 (2013) ("In principle, choice architects could design default rules for every one of us."). This possibility also arises in contract law. *See* Ian Ayres, *Preliminary Thoughts on Optimal Tailoring of Contractual Rules*, 3 S. CAL. INTERDISC. L.J. 1, 4 (1993) ("Untailored ('off-the-rack') defaults are rule-like because they are contingent on fewer variables, while tailored defaults are standard-like because they are contingent on more variables concerning the attributes or conduct of the particular contracting parties.").

28. *See* Camerer et al., *supra* note 7, at 1226 ("One consideration toward this end is determining the likely best option for most people—what is generally referred to as a majoritarian default."); Sunstein, *supra* note 27, at 31 ("The preferred approach is *to select the default rule that reflects what most people would choose if they were adequately informed*."); N. Craig Smith et al., *Smart Defaults: From Hidden Persuaders to Adaptive Helpers* 16 (INSEAD, Working Paper No. 2009/03/ISIC, 2009) ("[I]f the default is set to the preference most people would make when faced with making an active choice, the greatest number benefit."). As discussed *infra* Part II.D, the subgroup majoritarian rule will depart from the simple majoritarian rule when people's preferences over the available options are correlated with their susceptibility to the nudge.

29. As discussed below, in some cases the choice architecture imposes real transaction costs to selecting one of the available options and in such cases the consistent choosers' preferences are also relevant for designing the optimal policy. Even then, however, the preferences of the inconsistent choosers remain a separate input into the solution. *See infra* Part I.B.

people's preferences—the "revealed preferences" approach—imputes people's preferences directly from their voluntary choices. If I choose an apple when both apples and oranges are available, my choices reveal a preference for apples over oranges. The problem is that the traditional approach breaks down in settings where people respond to nudges; it would imply that the people who choose differently based on the direction of the nudge have preferences that are contingent on arbitrary features of the decision. In many settings, this conclusion is simply not realistic. Do we really think many people's goals and objectives about how much to save for retirement depend on whether their employers happen to design the company savings plan with opt-in versus opt-out enrollment?[30] If not, we need some alternative method for identifying preferences that can inform the design of choice architecture.

To be sure, many others have recognized the difficulty of recovering information about people's preferences in a behavioral context.[31] However, this Essay is the first to demonstrate the starkness of the tension, i.e., that the optimal policy depends on precisely the piece of information—the preferences of the inconsistent choosers—that the traditional approach for uncovering preferences is unable to provide.

After Part I sets out the problem of behavioral preference identification, Part II critically assesses the existing methods for overcoming it. Despite the importance of the problem—it comes up in nearly every setting where nudges affect behavior—the issue has received surprisingly little attention in the behavioral law and economics literature. In fact, despite the dozens of articles that have been written about the nudging approach in recent years, Cass Sunstein's and Richard Thaler's original article remains the most systematic effort to address how policymakers should go about choosing which direction to nudge.[32] In that article, Sunstein and Thaler suggest two "rules of thumb"

30. Granted, for some decisions an individual's preferences over the available options could actually depend on how the decision is framed, such as when the satisfaction one experiences from ending up with an option depends on whether a similar (but inferior) alternative to that option was also available. *See* BARRY SCHWARTZ, THE PARADOX OF CHOICE: WHY MORE IS LESS 117-35 (2004). I discuss such issues *infra* Part I.A when setting out the definition of a frame.

31. For a classic discussion of the issue, see Amartya K. Sen, *supra* note 9. For updated discussions that consider the findings of behavioral economics, see Russell B. Korobkin & Thomas S. Ulen, *Law and Behavioral Science: Removing the Rationality Assumption from Law and Economics*, 88 CALIF. L. REV. 1051 (2000); George Loewenstein & Peter A. Ubel, *Hedonic Adaptation and the Role of Decision and Experience Utility in Public Policy*, 92 J. PUB. ECON. 1795 (2008); and Mario J. Rizzo & Douglas Glen Whitman, *The Knowledge Problem of New Paternalism*, 2009 BYU L. REV. 905.

32. Sunstein & Thaler, *supra* note 1, at 1190-95. The other seminal paper on nudge theory acknowledges the preference identification issue in specific examples but concludes that the problem is similar to other "boundary-drawing" problems in policymaking. Camerer et al.,

for determining the direction in which to nudge: minimizing opt-outs and following majority choices.[33] Although both of these approaches seem sensible at first glance, I argue that neither actually incorporates the preferences of the inconsistent choosers—the key piece of information needed for nudging in the right direction. Instead, I show that both approaches turn out to select the nudge according to the preferences of the *consistent* choosers, the very group whose choices are unaffected by the nudge's design. Because there is generally no reason to think that consistent and inconsistent choosers share the same preferences (and for many decisions there are good reasons to suspect the preferences of the two groups will differ[34]), following Sunstein and Thaler's approaches for designing nudges can result in significant welfare costs.

Outside of the legal scholarship, economists have most commonly attempted to overcome the behavioral preference identification problem by obtaining preference information from settings where observers can be confident that decision makers are choosing according to their true preferences. I will refer to such a setting as a *revelatory frame*.[35] For example, forcing people to make active choices without any default to fall back upon may yield accurate information about which option they really prefer. However, I argue that, in most cases, these revelatory frame approaches also fail because they provide

supra note 7, at 1251. Here, I argue that the problem is substantially more serious than Camerer and colleagues acknowledge. Outside of the legal scholarship, John Beshears et al., *How Are Preferences Revealed?*, 92 J. PUB. ECON. 1787 (2008), addresses related issues, albeit with a different focus—their discussion is not specific to the optimal choice of frame question, and hence they are not concerned with identifying the preferences of the inconsistent decision makers. A related literature proposes methods for recovering the preferences of contracting parties for the implementation of optimal contract defaults. *See, e.g.*, Yair Listokin, *The Meaning of Contractual Silence: A Field Experiment*, 2 J. LEGAL ANALYSIS 397 (2010).

33. Sunstein & Thaler, *supra* note 1, at 1190-95. Sunstein and Thaler's preferred approach for designing nudges is to conduct a cost-benefit analysis, but as they acknowledge, this does not answer the question of how policymakers can learn decision makers' preferences—the key information needed for measuring the associated costs and benefits. *Id.* at 1190. Sunstein and Thaler also suggest a third rule of thumb—requiring people to make active choices—but this is more of an alternative policy that governments might impose than a method for uncovering people's preferences. I return to this issue *infra* Part II.C.

34. For example, those with more experience making the decision at hand may be less susceptible to the frame and also have different preferences than those making the decision for the first time. *See infra* Part II.D.

35. *See, e.g.*, Beshears et al., *supra* note 32, at 1791-92 (discussing approaches that include observing active choices, informed choices, and self-reported preferences). Legal scholars have utilized variations of this approach as well. *See, e.g.*, Sunstein & Thaler, *supra* note 1, at 1190-91 (arguing that many employees would be better off enrolled in retirement savings plans on the grounds that many of those who are not automatically enrolled nevertheless join such plans eventually).

preference information aggregated over the population as a whole rather than for the inconsistent choosers alone.

The final part of the Essay introduces two new methods for overcoming the problem of behavioral preference identification.[36] The first method builds on the revelatory frame approach described above. I show how combining information about people's choices from a revelatory frame with information about the consistent choosers can yield the preferences of the inconsistent choosers—the key unknown for choosing which way to nudge. The second method works even when revelatory frames are unavailable. It identifies the preferences of the consistent choosers and then extrapolates those preferences to the inconsistent choosers after adjusting for observable differences between the two groups. After explaining the idea behind the two methods, I illustrate how they would be applied and explore the range of conditions in which they work.

I. THE PROBLEM OF BEHAVIORAL PREFERENCE IDENTIFICATION

This Part describes the challenge that the rest of the Essay attempts to resolve. Part I.A sets out the basic terminology and assumptions. Part I.B argues that, from a behavioral law and economics perspective, the key information on which the optimal nudge policy depends is the preferences of the people whose choices are susceptible to nudges. Part I.C argues that this essential piece of information cannot be obtained using the traditional method for identifying people's preferences.

A. Setup: Nudges and Frames

Consider a group of people choosing between two available options, x and y. Each person has preferences over x and y, based on the extent to which each option furthers the person's goals and objectives, whatever those may be.[37]

The traditional starting point for economic analysis is that people make decisions according to the incentives and constraints they face. However, a key insight from behavioral economics is that other, seemingly arbitrary features of the decision may also affect which option people choose. In recent years,

36. The two methods are based on joint work with Daniel Reck. For a technical exposition, refer to Jacob Goldin & Daniel Reck, Preference Identification Under Inconsistent Choice (Mar. 25, 2015) (unpublished manuscript), http://scholar.princeton.edu/sites/default/files/jgoldin/files/goldin_and_reck_3_25_15.pdf [http://perma.cc/56DJ-FGB6].

37. As discussed *supra* note 9 and accompanying text, choice and preference are analytically distinct concepts. Someone may choose x but prefer y.

economic theorists have incorporated such insights into decision-making models by utilizing the concept of a "frame."[38]

For present purposes, I define a frame to be any feature of the decision that satisfies the following two properties. First, the frame must affect which option at least some people choose. That is, a feature of the decision is not a frame if it does not affect anyone's behavior. Second, to be a frame, a feature of the decision must be arbitrary from the point of view of people's preferences; that is, it must be irrelevant to whether the chooser prefers *x* or *y*. This means that the way a frame affects behavior cannot simply be by changing which option people prefer. To illustrate, suppose that Bob must choose whether to eat ice cream or hot chocolate. Bob enjoys ice cream more than hot chocolate in the summer but enjoys hot chocolate more than ice cream in the winter. If Bob's choices reflect this preference he will choose hot chocolate in the winter and ice cream in the summer. Although the season affects which option Bob selects, it is not a frame because it also affects Bob's true preference. In contrast, if Bob actually preferred hot chocolate throughout the year, but for some reason only selected hot chocolate during the winter, the season would then qualify as a frame.[39]

As illustrated by this example, outside observers cannot characterize features of the decision as frames in a purely objective manner. That is, if Ashley observes that Bob's choice between ice cream and hot chocolate varies by season, she can conclude that the season is a frame only by assuming that Bob's preferences do not themselves depend on the season. Similarly, if a decision maker is observed to choose *x* when *x* is the default and *y* when *y* is the default, the observer may only conclude that the default is a frame under the assumption that which option is the default does not affect which option the decision maker prefers. The general point is that because each observed choice is unique in some respect (e.g., the point in time at which it is made), outside observers seeking to identify frames must necessarily determine whether two observed choices are different in any preference-relevant respects. And making

38. For a formal exposition of this approach, see Yuval Salant & Ariel Rubinstein, *(A, f): Choices with Frames*, 75 REV. ECON. STUD. 1287 (2008). Careful readers will note my definition of a frame is slightly broader than theirs; they require the frame to be "irrelevant in the rational assessment of the alternatives," which rules out the presence of any transaction costs. *Id.* at 1287. Hence their definition of a frame coincides with what I will refer to later as a "pure nudge."

39. Similarly, a policy that changes the *content* of one's preferences along the lines described in Oren Bar-Gill & Chaim Fershtman, *Law and Preferences*, 20 J.L. ECON. & ORG. 331 (2004); and Maggie Wittlin, Note, *Buckling Under Pressure: An Empirical Test of the Expressive Effects of Law*, 28 YALE J. ON REG. 419 (2011), would, by definition, be preference-relevant, and therefore would not constitute a frame.

that determination requires taking a stance on which features of the decision are relevant to the affected decision makers.[40]

The concept of a frame described here is sufficiently general to accommodate a large number of findings that have been documented in the behavioral economics literature. Some well-known examples include: (1) the "stickiness" of default options (people's tendency to follow the default),[41] (2) the order in which choices are presented,[42] (3) whether an option's consequences are framed as a loss or a gain,[43] (4) the presence or absence of irrelevant alternatives,[44] (5) the salience or prominence of various features of the decision,[45] and (6) the point in time at which a choice is made.[46] As noted above, whether any of these constitutes a frame in a particular setting will depend on whether they are relevant to the preferences of the affected decision makers.

In considering policy decisions about which frame to implement, I will generally restrict my focus to *unidirectional frames*, which I define to be frames that tend to push people's choices in a uniform direction. This definition does not require that a frame affect the behavior of every decision maker; rather, it requires that for all decision makers who are affected, each must be pushed in the same direction. For example, suppose that decision makers are choosing whether to enroll in their employer's health insurance plans and that more

40. *See* AMARTYA SEN, RATIONALITY AND FREEDOM 123-32 (2004). Compare this to the conclusion drawn by B. Douglas Bernheim & Antonio Rangel, *Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics*, 124 Q.J. ECON. 51, 53 (2009), which claims, "We . . . develop a generalized welfare criterion that respects choice directly, without reference to the decision maker's underlying objectives." For a discussion of closely related issues, see Barbara H. Fried, *But Seriously, Folks, What Do People Want?*, 65 STAN. L. REV. 1249 (2013).

41. *See, e.g.*, Brigitte C. Madrian & Dennis F. Shea, *The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior*, 116 Q.J. ECON. 1149 (2001) (documenting that employers' retirement savings plan defaults shape employees' enrollment decisions).

42. *See, e.g.*, Levav et al., *supra* note 3, at 274 (finding that the order in which options are presented can affect what decision makers choose).

43. *See, e.g.*, Daniel Kahneman et al., *Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias*, 5 J. ECON. PERSP. 193, 200-02 (1991) (reporting evidence of loss aversion in a series of classic lab experiments). For a recent policy example, see Homonoff, *supra* note 20, which found that grocery shoppers are more likely to reduce disposable bag use in response to a small tax than to an equally sized bonus.

44. *See, e.g.*, Ian Ayres, *Menus Matter*, 73 U. CHI. L. REV. 3 (2006); Mark Kelman et al., *Context-Dependence in Legal Decision Making*, 25 J. LEGAL STUD. 287 (1996).

45. *See, e.g.*, Raj Chetty et al., *Salience and Taxation: Theory and Evidence*, 99 AM. ECON. REV. 1145 (2009) (finding that consumers are less responsive to taxes that are designed so that the taxed good's after-tax price is not prominent).

46. *See, e.g.*, David Laibson, *Golden Eggs and Hyperbolic Discounting*, 112 Q.J. ECON. 443 (1997) (setting out a theoretical model to explain time-inconsistent decision making).

people choose to enroll when participation is opt-out rather than opt-in. For the enrollment default to constitute a unidirectional frame, there must not be any employees who would choose to enroll if the plan were opt-in but would not choose to enroll if the plan were opt-out.[47]

To keep things simple, I will mostly focus on settings in which there are two possible frames that the government can implement with respect to people's choices between *x* and *y*. It will be convenient to refer to the frames as frame-*x* and frame-*y*, where the former is the frame that induces more people to select *x* and the latter the frame that induces more people to select *y*.

Broadly speaking, the government's choice of frame can affect people's choices through two types of channels: neoclassical or behavioral. Neoclassical channels affect behavior by imposing transaction costs on decision makers who select the option other than the one favored by the frame. These costs may be pecuniary—such as an administrative fee—or psychic—such as having to go through the hassle of filling out a form. The defining feature of a neoclassical channel is that choosing against the frame reduces the decision maker's welfare: someone who selects *x* under frame-*x* is better off than someone who selects *x* under frame-*y*.[48]

Frames may also affect people's choices through behavioral channels. Unlike neoclassical channels, behavioral channels don't change how the available options affect the decision maker's welfare. Instead, they induce decision makers to select a particular option through mechanisms that are unrelated to the decision maker's preferences. Prime examples of behavioral channels are psychological heuristics or biases, such as overvaluing losses

47. This restriction is primarily for expositional convenience. In most cases, the results generalize to settings in which the frame is not unidirectional (although such generalizations may require additional information that is not always available, as I discuss below).

Apart from unidirectional frames, the main type of frame considered in the behavioral law and economics literature are features of choice architecture that affect the quality of decision makers' choices, but do not systematically shift those choices in one direction or another. For example, simplifying a decision might affect what people choose, but it is likely to affect different people's choices in different directions. Absent other considerations such as externalities or the cost of implementation, the optimal design of such decision-quality frames is to implement the frame that induces individuals to make the higher-quality decisions. For a discussion of related issues, see Camerer et al., *supra* note 7, at 1219; and Jolls & Sunstein, *supra* note 17.

48. Note that a frame that affects behavior through neoclassical channels does not affect which option the decision maker would prefer to choose if all transaction costs were eliminated—if it did, it would not be a frame.

relative to gains[49] or favoring purchases when part of the price is not prominent at the time of sale.[50]

In many settings, the frame may influence decision makers through both behavioral and neoclassical channels. For example, setting an option to be the default may result in decision makers sticking with that option partly due to a psychological bias such as the status quo effect[51] and partly to avoid paying a fee that would be required to opt out. Similarly, placing an option earlier on a list of choices may result in more people selecting that option partly because they are biased to think more highly of whichever option they consider first and partly to avoid spending time reading the options further down the list.

In this Essay I will define a *nudge* as the government's decision to implement either frame-*x* or frame-*y*, and where the frame operates primarily through behavioral channels.[52] That is, a nudge affects behavior without changing the extent to which selecting *x* or *y* is consistent with the decision maker's preferences. Thus, a policy that influenced behavior by imposing a large tax on *x* or *y* would not be a nudge because its effect on people's choices would result from the imposition of a neoclassical cost. In contrast, altering which option was presented in a more positive light would constitute a nudge, as would labeling one of the two options as the default—so long as the welfare cost of opting out was small relative to the amount of welfare at stake in the decision for the affected choosers. Under this definition, a government *nudges toward x* when it implements frame-*x* and *nudges toward y* when it implements frame-*y*.

Because my focus in this Essay is on policies that affect choices through behavioral channels, for the most part I restrict the analysis to *pure nudges*—those for which there are no transaction costs associated with selecting the option other than the one being nudged toward. This assumption implies that a nudge affects people's well-being only to the extent it shapes which option

49. *See, e.g.*, Kahneman et al., *supra* note 43, at 201-02 (reporting evidence of loss aversion in a series of classic lab experiments).

50. *See, e.g.*, Chetty et al., *supra* note 45 (finding that consumers are less responsive to taxes that are designed so that the taxed good's after-tax price is not prominent).

51. The "status quo bias" refers to a propensity for the decision maker to select whichever option she perceives to be the continuation of the status quo. *See* Kahneman et al., *supra* note 43, at 194.

52. This definition of a nudge is similar to the one in Sunstein, *supra* note 27, at 5 ("[Nudges are] interventions that maintain freedom of choice, that do not impose mandates or bans, but that nonetheless incline people's choices in a particular direction."); and *id*. at n.15 ("The assumption here is that opting in and opting out are both easy and essentially costless."). *See also* Cass R. Sunstein, *The Storrs Lectures: Behavioral Economics and Paternalism*, 122 YALE L.J. 1826, 1859-60 (2013) (distinguishing "hard paternalism" that attempts to improve people's welfare by imposing material costs on them and "soft paternalism," which attempts the same goal but without imposing material costs).

they end up choosing. That is, someone who chooses x under frame-x is just as well off as someone who chooses x under frame-y. I briefly discuss the implications of relaxing this assumption at the end of the next Part.

Because governments must ultimately implement one frame or the other, it stands to reason that policymakers should compare the welfare effects of the available frames when deciding which direction to nudge. The next Part clarifies what information is required for making this determination.

B. The Optimal Choice of Frame

This Part considers what information is necessary for determining the optimal direction in which to nudge assuming that the government's objective is to maximize the number of people who end up selecting the option they prefer.[53] The question is: given that the nudge will affect the option that (at least some) people choose, should the government nudge toward x or toward y?

To answer this question, it is helpful to divide the decision makers into two groups: (1) those whose choices are shaped by the frame (the *inconsistent choosers*), and (2) those whose choices are not (the *consistent choosers*). Dividing the population this way greatly simplifies the optimal frame question. To see why, note that the well-being of the consistent choosers is unaffected by the government's choice of frame. By definition, individuals in this group end up

53. For a formal derivation of the results in this Part, refer to Goldin & Reck, *supra* note 36, at 50–53.

When the decision under consideration involves positive or negative externalities, determining the optimal nudge requires accounting for that externality in addition to the preferences of the decision makers whose choices the nudge affects (my focus here). In an ideal world, the government's objective function would also account for variation in the intensity of decision makers' preferences. I abstract from this consideration for a practical reason: even in non-behavioral settings, recovering information on preference intensity from choice data requires much stronger assumptions than what I impose here. For example, observing someone choose x over y suggests that the chooser prefers x to y but conveys no information regarding the extent to which she does so. One approach for recovering information on preference intensity in settings of inconsistent choice is described by Goldin & Reck. *Id.* Another option is to turn to non-choice data, as in Daphna Lewinson-Zamir, *Identifying Intense Preferences*, 94 CORNELL L. REV. 1391 (2009). Note that extending the analysis to account for preference intensity would not affect the basic result derived in this Part that it is the preferences of the inconsistent decision makers (rather than the full population) that are relevant for setting the optimal nudge. All that would change is that the inconsistent choosers with intense preferences would be counted more than the inconsistent choosers with weak preferences.

with the same option regardless of the frame,[54] and under the assumption that the nudge involves no transaction costs, the option that a decision maker ends up with is all that matters for determining her welfare.[55] Consequently, the government can ignore the preferences of the consistent choosers when deciding which direction to nudge.

Turning to the inconsistent choosers, the preferences of this group enter into the government's optimal frame problem in a straightforward way. By virtue of how this group is defined, we know that each inconsistent person's choice is affected by the frame: he or she will choose one option under frame-*x* and a different option under frame-*y*. Additionally, because the frame is unidirectional, we know that each inconsistent person's choice will be affected by the frame in the same direction: each will choose *x* under frame-*x* and *y* under frame-*y*. As a result, figuring out the optimal policy boils down to answering the following question: is it better to assign every inconsistent chooser with *x* or with *y*?

Table 1.
CHOICES BY FRAME AND CONSISTENCY

	Consistent Choosers		*Inconsistent Choosers*	
	Prefer *x*	Prefer *y*	Prefer *x*	Prefer *y*
Frame-*x*	*x*	*y*	*x*	*x*
Frame-*y*	*x*	*y*	*y*	*y*

By posing the question this way, it is clear that the optimal policy depends on the preferences of the inconsistent choosers, and in particular, on whether the majority of them prefer *x* or *y*. This follows from the fact that the government's goal is to maximize the fraction of decision makers who end up with their preferred outcome. We can therefore conclude that the government should nudge toward *x* when a majority of the inconsistent choosers prefer *x* to *y*, and toward *y* when a majority of the inconsistent choosers prefer *y* to *x*. Put

54. For expositional purposes, I'll assume that the option (consistently) selected by a consistent chooser is in fact the option she prefers. This assumption isn't required to reach the results of this Part but will be important later. *See infra* Part III.

55. In contrast, when a frame affects behavior through both neoclassical and behavioral channels, the government's choice of frame will also affect the well-being of the consistent decision makers. As discussed below, in such settings the optimal policy accounts for the preferences of both the consistent and inconsistent decision makers, where the weight on each group depends on the size of the transaction cost relative to the utility at stake in the underlying decision. Even then, determining the optimal policy still requires isolating the preferences of the inconsistent decision makers.

differently, the optimal policy is *subgroup majoritarian*: it depends on the majority preferences of a subgroup of decision makers—the ones who are inconsistent.[56]

Before concluding this Part, let us revisit for a moment the assumption that there are no transaction costs associated with resisting the nudge, i.e., with choosing the option other than the one favored by the frame. Thus far, I have focused on pure nudges—those for which the frame affects people's choices exclusively through behavioral channels. In practice, however, many policies that work primarily through behavioral channels are still associated with some amount of transaction costs. For example, it might entail some hassle for an employee to switch her retirement plan to the non-default option, even though the hassle is likely to be quite small relative to the ultimate difference in welfare between the two plans. In this setting, the rational choice for most decision makers will still be to incur the transaction costs because the welfare gains from switching insurance plans exceed the costs. Even so, there is no question that the costs associated with opting out of the default are normatively relevant—they affect the decision maker's welfare—and so should enter into the optimal policy analysis.

How does the presence of transaction costs affect the determination of the optimal frame?[57] It turns out that modifying the previous results to account for transaction costs is (conceptually) straightforward. On the one hand, the transaction costs don't affect the inconsistent choosers. Because individuals in that group always choose in the direction of the nudge, they never incur transaction costs. On the other hand, the presence of transaction costs implies that the preferences of the consistent choosers enter the equation as well. Because consistent choosers who prefer x will select x regardless of the default, they will be better off when the default is x—which allows them to avoid the transaction costs—than when the default is y—which results in them paying the transaction cost. The larger the transaction costs are, the worse it will be for the consistent choosers who resist the nudge. As a result, the presence of

56. This result, that the optimal nudge may diverge from the majority preferences of the population of decision makers, is related to an established literature on "minoritarian" default options in contract law. That literature provides an efficiency rationale for such defaults when the contracting parties have asymmetric information, Ian Ayres & Robert Gertner, *Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules*, 99 YALE L.J. 87 (1989), or in the presence of heterogeneous contracting costs, Ian Ayres & Robert Gertner, *Majoritarian vs. Minoritarian Defaults*, 51 STAN. L. REV. 1591 (1999). The present result highlights that differences in the preferences of consistent and inconsistent choosers can create an additional circumstance in which the optimal default is non-majoritarian, even absent incomplete information, or variation in contracting or other transaction costs.

57. For simplicity, I assume that the transaction costs are the same for all decision makers being considered. Decision makers may still vary in other characteristics that affect their psychological susceptibility to the nudge.

transaction costs counsels for taking the consistent choosers' preferences into account when setting the optimal nudge.

Given that the preferences of both the consistent and the inconsistent choosers can be relevant to determining the optimal nudge in the presence of transaction costs, what should policymakers do when these two groups prefer different options? Importantly, choosing based on the majority preferences for the population is still incorrect. Rather, one can show that the optimal direction for the nudge depends on the preferences of both the consistent and inconsistent decision makers, with the weight accorded to each group depending on the size of the transaction costs, the intensity of the inconsistent choosers' preferences, and the fraction of the population that is consistent.[58] As a result, even when transaction costs are present, the optimal policy will still depend primarily on the inconsistent choosers' preferences, so long as the costs are small relative to the welfare stakes for the inconsistent choosers. This will often be the case since by definition nudges affect decision making through primarily behavioral rather than neoclassical channels.[59] And even when the transaction costs are not small, the preferences of the inconsistent choosers remain a necessary input into the optimal policy determination; hence, even choosing between frames that impose significant transaction costs requires tools for preference identification along the lines considered here.[60]

This Part has demonstrated that identifying the optimal nudge requires knowing the preferences of the inconsistent choosers.[61] Unfortunately, as the

58. Note that the intensity of the consistent choosers' preferences does not enter into the equation. Because decision makers in this group end up selecting their preferred option regardless of the frame, all that the frame affects is whether they must incur the transaction cost. Hence, what matters for shaping the optimal nudge is not the relative intensity of the two groups' preferences, but rather the intensity of the inconsistent choosers' preferences relative to the transaction costs. For a formal derivation, refer to Goldin & Reck, *supra* note 36, at 50–53.

59. For example, nudges that involve manipulating the presentation of the available options would typically involve only trivial transaction costs.

60. Note that in purely neoclassical settings, choosing the optimal frame based only on the consistent choosers' preferences is likely to yield a reasonable approximation of the optimal policy when the magnitude of transaction costs dwarfs the amount of welfare at stake for the inconsistent choosers. The optimal nudge therefore remains subgroup-majoritarian in such settings, but with the relevant subgroup composed of the consistent rather than the inconsistent choosers.

61. Although the preferences of the inconsistent choosers are most important for choosing which way to nudge, it is worth noting that in other contexts it may be the preferences of the consistent choosers that matter most. For instance, if some voters in a shareholder election are swayed by arbitrary factors such as the order in which candidates are listed or the wording of a proposal, the corporation may wish only to count the votes of those shareholders whose votes do not turn on such factors. The framework developed below can be readily adapted for identifying the preferences of the consistent choosers. *See infra* Part

next Part shows, uncovering the preferences of this group requires overcoming an important hurdle.

C. *Limitations of the Traditional Approach*

The standard approach that economists take for identifying preferences is to look at people's voluntary choices. For this to work, one must assume that decision makers are *instrumentally rational*, in that the option they select is the one they most prefer. When people's preferences are reflected in their observable choices, backing out preferences from behavior is (at least conceptually) a straightforward task. If an instrumentally rational decision maker chooses x when x and y are both available (and there are no transaction costs to choosing y), it follows that she prefers x to y; otherwise she would have chosen y.

Having described the standard approach for identifying people's preferences, it should be apparent why it cannot be applied in the context of choices with frames. Recall that the defining characteristic of a frame is that it affects which option a decision maker chooses without being relevant to which option the decision maker truly prefers. But instrumental rationality—the key assumption behind the standard approach—requires people to always select whichever option is most consistent with their preferences. And for a preference-irrelevant feature of the decision (i.e., a nudge) to affect what people choose, it must be the case that in at least one of the frames the decision maker is not selecting her most-preferred option. Intuitively, if your choices are sensitive to the frame but your preferences are not, your choice under one of the two frames must be a mistake.[62]

The inability of the revealed preference approach to identify the preferences of the inconsistent choosers gives rise to what I call the *problem of behavioral preference identification*: the difficulty in uncovering people's preferences when

III. Whether it would be legal for a corporation to adopt procedures so that its elections turned only on consistent voters is an entirely different question.

62. To make the point concrete, return to the model described above (in which the decision maker faces a choice between x and y) and suppose that the decisions of the inconsistent choosers were instrumentally rational. As noted above, we know that an inconsistent chooser chooses x in frame-x. By instrumental rationality, we can conclude that this chooser prefers x to y, at least in the circumstances under which the choice was made (i.e., under frame-x). And because the frame is (by assumption) unrelated to the chooser's preferences, we know that the chooser would prefer x to y under frame-y as well, which, along with instrumental rationality, implies the chooser would select x under frame-y. But this establishes a contradiction, for we know that inconsistent choosers select y under frame-y. Consequently, we may conclude that the inconsistent choosers are not instrumentally rational and therefore that the standard approach cannot be used to identify their preferences.

what they choose varies according to arbitrary features of the decision. The problem of behavioral preference identification creates a dilemma for policymakers seeking to use nudges to make people better off. On the one hand, behavioral economics offers a range of new tools for influencing behavior, namely tools that affect people's choices without affecting their incentives or constraining what they may choose. On the other hand, the very fact that such tools can influence behavior calls into doubt the traditional methods for determining which available policy will best promote people's well-being. Choosing the best direction in which to nudge thus requires new techniques for identifying people's preferences, and it is to this subject that Part II turns.

II. EXISTING METHODS FOR CHOOSING WHICH DIRECTION TO NUDGE

Despite how frequently the issue arises, few solutions to the problem of behavioral preference identification have been proposed in the literature. In this Part, I review the more prominent ones that have. Of the three methods I consider here, two (minimizing the number of opt-outs and following the majority's preferences) were proposed by Sunstein and Thaler in their initial article setting out the case for nudging, and the third (utilizing revelatory frames) has been frequently applied by economists when discussing related issues.[63] Although all three of these approaches provide an answer to the question of which direction to nudge, I show that each does so in a way that is importantly flawed; the first two turn exclusively on the preferences of the consistent choosers and the third turns on the preferences for the full population, even though it is the preferences of the *inconsistent* choosers upon which the optimal nudge depends. As a result, each of the approaches described in this Part implicitly relies on the assumption that the average preferences of the consistent and inconsistent choosers are equal to one another for the decision being considered. But because individuals' preferences are

63. Although beyond the scope of this Essay, others have proposed more radical methods for policy analysis that measure people's well-being in terms of objective indicators such as happiness or subjective well-being. *See, e.g.*, John Bronsteen et al., *Well-Being Analysis vs. Cost-Benefit Analysis*, 62 DUKE L.J. 1603 (2013) (proposing the use of "well-being analysis" based on self-reported survey data for policy analysis). Such approaches avoid the problem of behavioral preference identification altogether by decoupling people's welfare from their observed choices. In contrast, capability-type approaches to policy analysis, *see, e.g.*, Martha Nussbaum, *Capabilities and Human Rights*, 66 FORDHAM L. REV. 273 (1997); Amartya Sen, *Development as Capability Expansion*, 19 J. DEV. PLAN. 41 (1989), do not provide much guidance for designing choice architecture because nudges influence behavior without changing the set of options available to decision makers.

often correlated with whether they are consistent, these approaches are unreliable guides for setting the optimal nudge.

A. Minimizing Opt-Outs

One intuitively appealing approach for selecting the direction in which to nudge is to minimize the number of people who opt out of the nudge—that is, to minimize the number of people who choose the option other than the one they are being nudged toward.[64] For example, if policymakers nudge people toward saving more for retirement by requiring employers to offer savings plans with automatic enrollment, an employee opts out of that nudge by choosing not to enroll. Although the "opting-out" language brings default rules to mind, the idea behind the approach can be extended to other types of nudges as well.

To illustrate how the minimizing opt-outs approach would be applied, consider the following example.

A sandwich shop offers customers a free piece of fruit—either an apple or an orange—after completing their purchase. Customers get to choose which type of fruit they want to receive. Both types of fruit are kept in baskets near the register, but because space at the shop is tight, one basket must be kept closer to the register and appears more prominently to customers as they are checking out.

Suppose that after several years of giving out fruit, the kindly shop owner notices that customers tend to select more of whichever fruit happens to be more prominently displayed that day. After analyzing his records, the owner finds that 80% of customers choose apples when apples are more prominent but that only 40% choose apples when oranges are more prominent.

Assuming that both types of fruit cost the store the same amount to provide and that the store's only goal in providing the fruit is to make its customers as well off as possible, which fruit should the sandwich shop place in the more prominent location? That is, should the shop nudge toward oranges or apples?

Answering this question by minimizing opt-outs is easy: one simply computes the fraction of people choosing against the nudge under both possible frames and selects the one that results in the lower fraction. Using the numbers from this example, 20% of customers opt out when the store nudges toward apples—i.e., 20% choose oranges when apples are prominent. In contrast, 40% of customers opt out when the store nudges toward oranges.

64. Sunstein & Thaler, *supra* note 1, at 1195.

Thus, the minimizing opt-outs approach suggests the store should nudge toward apples.

Table 2.
CHOOSING NUDGES BY MINIMIZING OPT-OUTS

	Choose Apples	*Choose Oranges*
Apples Prominent	80%	**20%**
Oranges Prominent	**40%**	60%

Note: Bolded figures denote the fraction of customers opting out.

Admittedly, designing a nudge by minimizing opt-outs seems like a reasonable approach. As Sunstein and Thaler put it, when fewer people opt out of a nudge, it suggests that "more people are sufficiently satisfied to leave it in place."[65] However, analyzing this approach using the framework developed in Part I reveals an important flaw.

The problem with designing nudges by minimizing opt-outs is that although opt-outs do provide information about whether decision makers are satisfied with a particular option, the information is for the wrong group of decision makers! Recall that the key information for choosing which way to nudge is the preferences of the inconsistent choosers—the people whose choices will actually be affected by the nudge's direction. Minimizing opt-outs tells us nothing about the preferences of this key group because the very definition of an inconsistent chooser is someone who does not opt out; an inconsistent chooser selects oranges when oranges are prominent and apples when apples are prominent.

To illustrate the extent of the problem, return to the grocery store example and consider the behavior of the consistent and inconsistent choosers under the two possible nudges. When apples are prominent, who selects apples? Two groups do so: (1) the inconsistent choosers (because they select whichever fruit is prominent), and (2) the subset of consistent choosers who prefer apples to oranges. The only people who select oranges are those not in either of these categories—those among the consistent choosers who prefer oranges to apples.

Next, consider who opts out when the store nudges toward oranges. Again, we know that two groups will select oranges when oranges are prominent: (1) the inconsistent choosers, and (2) the subset of consistent choosers who prefer oranges to apples. Consequently, the only people who opt out under this

65. *Id.*

nudge (i.e., who select apples) are those among the consistent choosers who prefer apples to oranges.

Table 3.
WHO OPTS OUT?

	Consistent Choosers		*Inconsistent Choosers*	
	Prefer Apples	*Prefer Oranges*	*Prefer Apples*	*Prefer Oranges*
Apples Prominent	No	Yes	No	No
Oranges Prominent	Yes	No	No	No

Note: Table cells denote whether decision makers opt out of the nudge based on their preferences and whether they are consistent.

By isolating the groups of decision makers who opt out under each nudge, we can see that minimizing opt-outs is really comparing (1) the number of consistent choosers who prefer oranges with (2) the number of consistent choosers who prefer apples. Consequently, deciding which way to nudge by minimizing opt-outs boils down to following the majority preferences of the consistent choosers. The preferences of the inconsistent choosers—the group that should be central for setting policy—do not even enter into this comparison. This point bears emphasis: minimizing opt-outs implies that the sandwich shop should nudge toward apples whenever 51% of the consistent choosers prefer apples, even if every single person who will actually be affected by the nudge would be better off with oranges!

My discussion in this Part has leaned heavily on the assumption that the nudge under consideration is unidirectional, e.g., that making apples more prominent doesn't induce anyone to buy oranges if they would have otherwise bought apples. However, the main result of this Part—that minimizing opt-outs is equivalent to following the majority of the consistent choosers—does not rely on this assumption; it holds even when frames are not unidirectional.

To see why, suppose now that there are two types of inconsistent choosers: those who follow the direction of the nudge and those who always resist the nudge. Borrowing some language from a different field of economics,[66] we can refer to the first group as the *compliers* and the second group as the *defiers*. For example, a defier would be someone who always chooses whichever option is not the default, regardless of which option that is.

66. Joshua D. Angrist et al., *Identification of Causal Effects Using Instrumental Variables*, 91 J. AM. STAT. ASS'N 444, 448 (1996).

Extending the analysis in this Part to account for defiers is not too difficult. Previously, when we assumed that nudges were unidirectional, it implied that the only people opting out were the consistent choosers. Here, that's no longer true because defiers opt out as well. But because defiers (by definition) opt out regardless of the nudge's direction, they don't affect which nudge direction minimizes opt-outs. As shown in Table 4, the number of people who opt out when apples are prominent is equal to the number of defiers plus the number of consistent choosers who prefer oranges. Similarly, the number of people who opt out when oranges are prominent is equal to the number of defiers plus the number of consistent choosers who prefer apples. Comparing these quantities results in the defiers dropping out of the equation, and as before, we obtain the result that the direction that minimizes opt-outs is the one that accords with the consistent choosers' preferences.

Table 4.
WHO OPTS OUT WHEN FRAMES ARE NOT UNIDIRECTIONAL?

	Consistent Choosers		*Compliers*		*Defiers*	
	Prefer Apples	*Prefer Oranges*	*Prefer Apples*	*Prefer Oranges*	*Prefer Apples*	*Prefer Oranges*
Apples Prominent	No	Yes	No	No	Yes	Yes
Oranges Prominent	Yes	No	No	No	Yes	Yes

Note: Table cells denote whether decision makers opt out of the nudge based on their preferences and whether they are consistent.

B. Following Majority Choices

The second approach that Sunstein and Thaler propose for choosing the direction in which to nudge is nudging toward the option that the majority prefers, when that information is known. Sunstein and Thaler acknowledge that in many cases the preferences of the majority will not be known and that in those cases this approach won't work. For example, in the sandwich shop hypothetical summarized in Table 2, the majority of customers choose apples under one frame and oranges under the other, and so there is no majority choice to follow.

But when the option receiving a majority of choices is the same under both possible frames, Sunstein and Thaler propose that this option is a sensible one

for policymakers to nudge toward.[67] For example, suppose we modified the numbers in the sandwich shop example so that 80% of customers choose apples when apples are prominent and 60% do so when oranges are prominent, as in Table 5. With this data, we know that a majority of customers choose apples under either frame. The majority-choice approach would then imply that the store should place the apples in the more prominent location.

Table 5.
FOLLOWING MAJORITY CHOICES

	Fraction Choosing Apples	*Fraction Choosing Oranges*
Apples Prominent	80%	20%
Oranges Prominent	60%	40%

Note: A majority of customers choose apples under both frames.

Sunstein and Thaler acknowledge that one shortcoming of the majority-choice approach is that it does not provide guidance in settings where there is no consistent majority. However, the deeper problem with nudging toward the majority's choices is that, like the minimizing opt-outs rule, the majority's preferences turn out to depend exclusively on the preferences of the consistent choosers. This is true despite the fact that the rule is expressed in terms of the majority of the entire population. In particular, it turns out that the approach will never suggest nudging toward a particular option unless a majority of the consistent choosers prefer that option. Moreover, the preferences of the inconsistent choosers—the key group for setting policy—do not enter into the analysis at all.

To see why, note that for a majority of customers to choose apples under both frames, it must be the case that a majority selects apples even when oranges are more prominent. But recall that under the pro-orange frame, the only people who select apples are the subset of consistent choosers who prefer apples. And for the people in this group to constitute more than 50% of *all* customers, it must be the case that they also constitute more than 50% of the *subset* of customers who are consistent.[68] Thus, the fact that a majority of

67. Sunstein & Thaler, *supra* note 1, at 1194.

68. That is, if the majority of a group is made up of tall men, it must also be the case that a majority of the men in the group are tall.

people choose one option under both frames necessarily implies that a majority of the consistent choosers prefer that option as well.[69]

Not only does the majority-choice rule incorporate the wrong information (the preferences of the consistent choosers), it entirely ignores the information that matters most (the preferences of the inconsistent choosers). Now, it is true that the raw number of inconsistent choosers does play a role in the analysis—the more inconsistent people there are, the less likely it will be that either option will receive a consistent majority. For example, if a majority of the decision makers are inconsistent, then neither option can ever receive a consistent majority. But even though the rule incorporates information about the quantity of inconsistent choosers, it still fails to take into account the key information for setting policy, which is not the *number* but the *preferences* of inconsistent choosers. For example, if 20% of the population is inconsistent, it could be that everyone in that group prefers apples (in which case apples should be prominent) or that all of them prefer oranges (in which case oranges should be prominent). Because the majority-choice approach doesn't take into account the inconsistent choosers' preferences, it will be an unreliable guide for choosing the best direction in which to nudge.

C. Revelatory Frames

The third approach I'll consider for choosing the direction in which to nudge is to obtain preference information gleaned from the choices people make in settings that an observer believes will be unbiased—settings I'll refer to as *revelatory frames*. A frame is revelatory if the choices people make in it reflect their true preferences.[70] For example, suppose people make some decision differently depending on which option is the default. Confronting those people with the same decision but without any default might constitute a revelatory frame. That is, forcing people to make an "active choice"—without any default

69. This result implies that the majority-choice approach never disagrees with the minimizing opt-outs approach: either the former gives no answer, or it gives the same answer as the latter. Put differently, minimizing opt-outs is simply a more general method for following majority choices as opposed to a distinct approach.

70. Another way to put it: a frame is revelatory if the choices it induces people to make are instrumentally rational. *See supra* Part I.C. What I am calling a revelatory frame is closely related to what B. Douglas Bernheim and Antonio Rangel, *supra* note 40, at 82-91, refer to as a "refinement." When people make conflicting choices between two frames, Bernheim and Rangel propose that observers might make judgments about which one more reliably reveals people's true preferences and discard the choice information obtained from the other. The revelatory frame is what is left after the researcher has completed the refinement process.

to fall back upon—would be revelatory if the choices made in this setting reveal the decision makers' true preferences.[71]

Before proceeding, it is important to distinguish between two ways policymakers might utilize revelatory frames: (1) they might observe the choices that people make in a revelatory frame and use that information as a guide to choosing the direction for a nudge, or (2) they might design choice architecture in a way that implements the revelatory frame itself.[72] Consider the second of these possibilities, in which the government implements the revelatory frame itself. Under the revelatory frame, all of the inconsistent people choose their preferred outcome and the consistent people make the same choices that they do in every other frame; thus, compared to any other frame, no one is worse off and some people are better off. All else equal, when a revelatory frame is available for policymakers to implement, doing so is typically the optimal policy.[73]

But frequently, all else is not equal. Even when a revelatory frame is available for policymakers to implement, doing so may be associated with substantial costs to either the government or to decision makers themselves.[74] For example, decision makers may wish to avoid having to devote the time and mental effort to making an active choice and instead may wish to rely on a default option. Many types of computer software have hundreds of minor details that can be customized, but it's hard to imagine that most users would prefer having to make active decisions about each of these settings themselves. And in many contexts, no revelatory frame will even be available for the policymaker to implement, such as when an incentive policy may be framed as either a loss or a gain.

71. *See* Gabriel D. Carroll et al., *Optimal Defaults and Active Decisions*, 124 Q.J. ECON. 1639 (2009) (formally deriving conditions in which active choice frames enhance social welfare); Sunstein, *supra* note 27, at 38-46 (discussing the costs and benefits of requiring active choice).

72. Sunstein and Thaler consider the second of these as one of their proposed "rules of thumb." *See* Sunstein & Thaler, *supra* note 1, at 1194-95.

73. *But see* Jacob Goldin, Note, *Sales Tax Not Included: Designing Commodity Taxes for Inattentive Consumers*, 122 YALE L.J. 258 (2012) (noting that maximizing social welfare may require designing choice architecture in ways that induce decision makers to err when making decisions that produce externalities).

74. For a good discussion of some of these costs, see Cass R. Sunstein, *Choosing Not To Choose*, 64 DUKE L.J. 1 (2014), which argues that forcing choice is a form of paternalism; and Sunstein, *supra* note 27, at 46-47, which identifies the drawbacks of active choosing. The question of how strong or weak of a default (or other nudge) to impose parallels the issues relating to the design of altering rules in contract law. *See* Ian Ayres, *Regulating Opt-Out: An Economic Theory of Altering Rules*, 121 YALE L.J. 2032 (2012) (setting out considerations that might shape the design of contract altering rules, i.e., the rules that govern whether a contracting party has successfully opted out of the default option).

Even when policymakers don't wish to implement the revelatory frame itself, such frames may still be useful for shedding light on people's preferences. For example, although requiring everyone to make active choices might be prohibitively costly, one could randomly assign a fraction of the population to a revelatory frame, observe the choices made by those people, and use the resulting preference information to inform the design of choice architecture for everyone else.[75] In the sandwich shop example described above, it might be costly or infeasible to permanently store both apples and oranges in an equally non-prominent location, but doing so for one week and observing the resulting choices might provide valuable information.

In considering what can be learned about people's preferences from the revelatory frame approach, it is important to keep in mind that the approach only works if the people choosing under the frame actually select the option that is most consistent with their preferences—that is, if the frame is actually revelatory. Otherwise the observed choices don't reveal the preferences of the choosers. The likelihood this condition will hold varies greatly from context to context.

Although there is no general formula for finding a revelatory frame, the following approaches can be promising in a number of settings. First, one can attempt to eliminate the source of the bias, e.g., by forcing active choice or by making the costs and benefits of each option equally salient. For example, the sandwich shop could remove both baskets of fruit and simply ask customers which one they would prefer.[76] Second, an outside observer might intervene to

75. *See* Porat & Strahilevitz, *supra* note 27 (proposing the use of human "guinea pigs" in which the decisions of a small random sample of the population are observed under favorable conditions to shed light on the optimal default); Sunstein, *supra* note 27, at 33 ("[S]election of a default rule might well be preceded by a period of active choosing as a way of assembling that information.").

76. One danger with this approach may be that although the bias is weakened, it is obscured rather than eliminated. For example, eliminating any default option may not eliminate other status quo biases—decision makers might still be drawn toward whichever option strikes them as a continuation of the status quo. The danger is that this new form of status quo bias can be harder for an outsider to detect. For example, a new employee may interpret the option of creating a new 401(k) plan to constitute a departure from the status quo (in which she had no 401(k) plan). Alternatively, a different new employee who had a 401(k) at her old job might perceive enrolling in the retirement plan to be a continuation of the status quo. The general point is that in practice it may be quite difficult to entirely eliminate a status quo bias because some option is always going to be more associated with the status quo than others. And in such cases, defaults may serve a useful informational role in that they help observers know which way the status quo bias runs.

A related point is that it may in practice actually be quite difficult to remove a default option. For example, in most cases a firm would not be willing to fire an individual for refusing to make a choice about which plan to enroll in, and even if it did, that would imply that the default is actually no plan, in the sense that that is the option associated with failing

improve the quality of the choice being made, such as by providing decision makers with additional time, information, decision-making aids, or by warning them of potential biases.[77] For example, a large firm seeking to design its default savings plan might provide a random sample of its employees with a financial counselor and observe the resulting enrollment decisions. A third possibility is to focus on decision makers' long-term behavior, based on the theory that people's choices tend to improve over time.[78] This seems most likely to work when decision makers are provided with regular feedback about the quality of their decisions.

Even when a frame is revelatory, we must still ask whether it provides the right preference information for choosing the direction in which to nudge. Recall that the key information concerns the preferences of the inconsistent choosers. Because everyone chooses correctly in the revelatory frame, it would seem that observing the choices made in such frames would reveal everyone's preferences—including those belonging to the inconsistent choosers. Unfortunately, the story is rarely this simple in practice.

The problem with using revelatory frames to identify the preferences of the inconsistent choosers is that doing so requires knowing which choosers are inconsistent. But all that the revelatory frame provides is what each person's true preference is; it does not provide a way of separating the consistent from the inconsistent choosers. If the researcher were able to observe each person's choice under all three frames—the two nudges under consideration plus the revelatory frame—then our task would indeed be done. That is, suppose Abby is a customer in the sandwich shop who chooses apples when apples are prominent, oranges when oranges are prominent, and oranges when neither apples nor oranges are prominent (a frame we'll assume is revelatory). From this data, we would be able to conclude that Abby is inconsistent and that she prefers oranges. Similarly, if we could observe Boris make choices under all three frames, and saw that he chose apples under each, we could conclude that Boris is consistent and that he prefers apples. In settings where we can observe

to choose. For example, Carroll and colleagues, *supra* note 71, discuss a retirement savings plan in which the employer required employees to make an active choice about whether to enroll. But even there, a default option still existed, in the sense that employees who failed to make a choice were not enrolled in a plan. *Id.* at 1641.

77. *See, e.g.*, James N. Druckman, *Using Credible Advice To Overcome Framing Effects*, 17 J.L. ECON. & ORG. 62, 73 (2001) (finding that framing effects can be reduced by providing advice to decision makers); Hunt Allcott & Dmitry Taubinsky, *The Lightbulb Paradox: Evidence from Two Randomized Experiments* (Nat'l Bureau of Econ. Research, Working Paper No. 19713, 2013) (employing information interventions to assess the welfare effects of subsidies and mandates regarding the use of energy-efficient lightbulbs).

78. *See, e.g.*, Beshears et al., *supra* note 32.

choices for each person under all three frames, it is straightforward to identify which people are inconsistent and what those people's preferences are.

In practice, however, the researcher usually will not be able to observe how any one person chooses in multiple frames. Instead, one is likely to observe something like Abby choosing when oranges are prominent, Boris choosing when apples are prominent, and Calvin choosing under the revelatory frame. Of course, the researcher could ask people to repeat their choices over multiple frames, but doing so is not always feasible and even when it is, asking people to choose multiple times may introduce a new set of problems. In particular, people may choose differently in one frame depending on whether they've previously been exposed to another frame.[79] For example, a person who initially makes an active choice between two options may be less sensitive to defaults when choosing between the same two options in the future because she will have already incurred the time and mental effort needed to decide on the choice at hand. When researchers are only able to observe a single choice for each person, a revelatory frame will reveal each person's preference but won't reveal which people are consistent and which are not.

Another way to put the problem is that if (as others have assumed) the optimal frame is simply majoritarian, revelatory frames would provide all the information that is required (the average preferences for the population as a whole). But because, as I argued in Part I, the optimal frame depends on the average preferences of the inconsistent choosers, determining which way to nudge requires information not only on what decision makers prefer but also on whether or not they are consistent. For example, suppose two-thirds of sandwich shop customers are consistent, 70% of the consistent ones prefer apples, and 40% of the inconsistent ones do so as well. In this case, observing choices in a revelatory frame would imply that 60% of all customers prefer apples;[80] what it would not reveal is that the majority (60%) of the inconsistent choosers prefer oranges. As such, revelatory frames (on their own) do not provide the necessary information for determining the best direction to nudge.[81]

D. Preferences and Consistency

All three of the approaches considered in this Part suffer from a similar limitation. Each provides guidance for designing nudges, but each fails to

79. Robyn A. LeBoeuf & Eldar Shafir, *Deep Thoughts and Shallow Frames: On the Susceptibility to Framing Effects*, 16 J. BEHAV. DECISION MAKING 77, 85-86 (2003).

80. $(2/3) * 0.70 + (1/3) * (0.40) = 0.60$.

81. In Part III.A, I provide a method for supplementing the information from revelatory frames to accomplish this task.

account for what should be the key piece of information—the preferences of the inconsistent choosers. The methods of minimizing opt-outs and following majority choices mistakenly focus on the preferences of consistent choosers, while revelatory frames reveal the preferences of the population as a whole. Consequently, for any of these approaches to provide the correct answer it must be the case that the average preferences of the consistent choosers are equal to the average preferences of the inconsistent choosers.[82] In other words, each of the three approaches described in this Part implicitly assumes that knowing whether or not someone chooses consistently would tell an observer nothing about the person's likely preferences over the options being considered. Unfortunately, the assumption that people's preferences are not systematically related to whether they are consistent is a strong one, and one that is likely to fail in many of the settings in which nudges are being considered.

To illustrate why, consider the factors that make it easier or harder for decision makers to resist the pull of a nudge. The answer to this question likely varies widely by situation and the empirical research in the area is still nascent. Still, several patterns seem likely. First, experienced decision makers may be more likely to choose consistently. Those who have made a decision in the past may have already spent time thinking about the consistency of the available options with their preferences and hence may face lower attention costs when faced with the same decision a second time around.[83] Second, more highly educated people tend to be more consistent.[84] Especially for more complicated decisions, or decisions requiring a higher level of mathematical sophistication, one might imagine that it is easier for better-educated decision makers to pay attention to both options. Finally, for many decisions, susceptibility to a nudge is likely to vary by income. As Sendhil Mullainathan and Eldar Shafir have argued, attention is a scarce resource for everyone, but it is especially scarce for

82. Since the full population consists of consistent and inconsistent choosers, the average preferences of the full population will only be equal to the average preferences of the inconsistent choosers when the consistent and inconsistent choosers' average preferences are equal.

83. *See, e.g.*, John A. List, *Does Market Experience Eliminate Market Anomalies?*, 118 Q.J. ECON. 41 (2003) (finding that experienced traders are less susceptible to the endowment effect); Sumit Agarwal et al., Learning in the Credit Card Market (Apr. 14, 2013) (unpublished manuscript), http://scholar.harvard.edu/files/laibson/files/learning_credit_042413.pdf [http://perma.cc/T348-J5X7] (finding that experienced credit card users tend to accumulate fewer fees). *See generally* Jeffrey Rachlinski, *Cognitive Errors, Individual Differences, and Paternalism*, 73 U. CHI. L. REV. 207, 219-21 (2006) (discussing how training and experience affect decision-making quality).

84. *See* Syngjoo Choi et al., *Who Is (More) Rational?*, 104 AM. ECON. REV. 1518 (2014) (reporting that better-educated participants in a decision-making experiment were less likely to make internally inconsistent choices).

those under greater financial pressure;[85] such individuals may find it especially costly to expend their mental energy resisting the pull of a nudge.[86]

The reason that consistency and preferences are often related is that the same individual characteristics likely to be associated with consistency—e.g., experience, education, and income—tend also to be associated with people's preferences. Consider automatic enrollment into employer-sponsored retirement savings plans. One could imagine that higher-income employees would be more likely to choose consistently—perhaps they have better financial education or are more likely to consult a financial planner when making the decision. Moreover, it's also quite likely that higher-income employees have different retirement savings goals than their lower-income colleagues—e.g., because of differing expectations about their lifetime income. Or, to take another example, someone with substantial experience purchasing auto insurance may be less susceptible to whether the default option is full- or limited-liability. But at the same time, drivers with lots of experience purchasing auto insurance probably tend to have different insurance coverage preferences than drivers without such experience. Experienced purchasers may tend to be older and hence might have different risk preferences or beliefs about the likelihood of getting into a collision.

Granted, it's not impossible to imagine settings in which people's preferences are not systematically related to whether they choose consistently. Consider the sandwich shop hypothetical, for example. There, one might imagine that the sensitivity of customers to the prominence of the fruit is unrelated to their preferences between apples and oranges. The notion is that one's preferences between apples and oranges is something so arbitrary that it could plausibly be unrelated to the characteristics about a person that shape whether or not they are consistent. On the other hand, even preferences over fruit may be correlated with demographic characteristics like income and education—factors that could in turn be correlated with susceptibility to various framing effects.[87] Hence, although preferences and consistency might

85. *See* SENDHIL MULLAINATHAN & ELDAR SHAFIR, SCARCITY: THE NEW SCIENCE OF HAVING LESS AND HOW IT DEFINES OUR LIVES (2014).

86. Note, though, that this reasoning suggests that low-income decision makers are *more* likely to choose consistently when making decisions with financial stakes that would be unimportant to higher-income decision makers. *See* Jacob Goldin & Tatiana Homonoff, *Smoke Gets in Your Eyes: Cigarette Tax Salience and Regressivity*, 5 AM. ECON. J. 302 (2013).

87. *See, e.g.*, Elling Bere et al., *Why Do Parents' Education Level and Income Affect the Amount of Fruits and Vegetables Adolescents Eat?*, 18 EUR. J. PUB. HEALTH 611 (2008) (documenting the correlation between adolescents' fruit and vegetable choices and parental income and education).

turn out to be unrelated in certain applications, this assumption should not be accepted uncritically.[88]

III. OVERCOMING THE PROBLEM OF BEHAVIORAL PREFERENCE IDENTIFICATION

Having highlighted some problems with existing methods for uncovering inconsistent choosers' preferences, this Part introduces two novel techniques to overcome these concerns.[89] The first (the Augmented Revelatory Frame approach) builds on the revelatory frame approach described in Part II.C. It uses a revelatory frame to obtain the average preferences of the full population and from this information discerns the preferences of the inconsistent choosers by isolating and then removing the average preferences of the consistent choosers. The second (the Demographic Extrapolation approach) identifies differences in the characteristics associated with consistent and inconsistent choosers and then extrapolates the consistent group's preferences to the inconsistent group after adjusting for those differences in characteristics.

A. The Augmented Revelatory Frame Approach

The Augmented Revelatory Frame approach (ARF) takes information obtained from a revelatory frame and augments it with information about consistent choosers' preferences in order to isolate the preferences of the inconsistent choosers. As discussed in Part II.C, a revelatory frame identifies the average preferences for the population as a whole but cannot isolate the average preferences for the inconsistent choosers unless each person is observed in multiple frames. The main improvement of the ARF approach over the revelatory frame approach is that it provides a way to obtain the average preferences of the inconsistent choosers without having to observe any one person making choices in multiple frames.

The ARF approach accomplishes this by exploiting a fundamental tool of probability theory known as the Law of Iterated Expectations. The theorem relates the average value of a variable in a population to the average value of the variable among distinct subgroups in that population. Here, the Law of Iterated Expectations states that the average preferences of the full population

88. A related point is that even when this assumption does turn out to be reasonable in a particular application, the assumption should at least be made explicit in the analysis so that readers can evaluate its plausibility for themselves.

89. These methods are based on joint work with Daniel Reck. For a technical exposition, see Goldin & Reck, *supra* note 36, at 16-17, which derives the Augmented Revelatory Frame approach (ARF); and *id.* at 18-22, which derives the Demographic Extrapolation approach.

are equal to the weighted average of the preferences of the consistent and inconsistent choosers, where the weights are the fraction of the population in each group. Thus, we can combine the average preferences of the full population with our knowledge of the average preferences for the consistent choosers as well as information regarding what fraction of the population is consistent; together, this lets us back out the average preferences for the inconsistent choosers.

The application of the Law of Iterated Expectations to this context is summed up in the following formula:

$$\text{Avg. Prefs of Population} = (\text{Avg. Prefs of Consistent}) * (\text{Fraction Consistent}) + (\text{Avg. Prefs of Inconsistent}) * (\text{Fraction Inconsistent})$$

Implementing the ARF approach therefore requires knowing three quantities: the average preferences for the full population, the average preferences for the consistent choosers, and the fraction of the population that is consistent. The revelatory frame provides the first of these. The trick is uncovering the other two.

To see how information about the consistent choosers can be obtained, return to the sandwich shop example described in Part II and suppose that customers there are randomly assigned to one of three frames: one in which apples are prominent, one in which oranges are prominent, and one in which the two fruits are equally prominent (e.g., both are placed far from the cash register). Next, suppose that 1,000 customers are assigned to each frame. Assume that 525 choose oranges when apples are prominent, 825 choose oranges when oranges are prominent, and that 600 choose oranges when the two fruits are equally prominent. This data is summarized in Table 6.

Table 6.
LEARNING ABOUT THE CONSISTENT CHOOSERS

	Choose Oranges	*Choose Apples*
Oranges Prominent	825	175
Apples Prominent	525	475
Apples and Oranges Equally Prominent	600	400

First, consider how to uncover the fraction of choosers that are consistent. It turns out that the quantity is simply given by the sum of the fractions of people opting out under the two (nonrevelatory) frames. To see why, recall from Part II.A that the fraction of people opting out of a nudge is equal to the

fraction of choosers who are consistent and who prefer the option other than the one they are being nudged toward. Since there are only two possibilities for which option people prefer, adding these fractions together yields the total fraction of consistent choosers in the population.[90]

We can summarize this result in a general formula. Let $\beta(x)$ denote the fraction of individuals who choose against the nudge when nudged toward x. Similarly, let $\beta(y)$ denote the fraction who choose against the nudge when nudged toward y. We may then conclude:

$$\text{Fraction of Consistent Choosers} = \beta(x) + \beta(y)$$

Here, 525 people opt out when the store nudges toward apples, or 52.5%. Similarly, 175 customers (17.5%) opt out when the store nudges toward oranges. Adding up these fractions implies that 70% (52.5% + 17.5%) of the total customers are consistent. Note also that this implies that the remaining 30% of customers (i.e., everybody else) is inconsistent.

Identifying the fraction of consistent choosers provides only half of the information we need for implementing the ARF approach; we also need the average preferences for those people who are consistent—i.e., the fraction of the consistent choosers who prefer apples to oranges. Luckily, we can obtain that information using a similar technique. Recall that the 525 people who opt out when the store nudges toward apples represent the subset of consistent choosers who prefer oranges. Similarly, the 175 customers who opt out when the store nudges toward oranges represent the subset of consistent choosers who prefer apples. Thus, for a randomly selected sample of 1,000 customers, we can predict that of the 700 of them likely to be consistent, 525 will prefer oranges to apples. As such, we can conclude that 525/700, or 75%, of the consistent choosers prefer oranges, and that 175/700, or 25%, prefer apples.

Again, this approach can be stated in general terms. Using the above notation, the preferences of the consistent choosers can be found with the following formula:

$$\text{Fraction of Consistent Choosers Who Prefer } x = \beta(y) \,/\, (\beta(x) + \beta(y))$$

The final piece of information for implementing the ARF approach is the average preferences of the population as a whole. But this is where the revelatory frame comes in. Because 600 out of 1,000 customers select oranges

90. Formally, the result follows because P(prefers x & consistent) + P(prefers y & consistent) = P(consistent) since each person prefers either x or y. For a similar analysis, see Rachlinski, *supra* note 83, at 211-13, which interprets the data from a classic study of framing effects.

when the two fruits are equally prominent, we can conclude that 60% of the full population prefers oranges.

Now that we have all the required ingredients, we can combine the information together to recover the preferences of the inconsistent choosers. Substituting the results from the last few paragraphs into the Law of Iterated Expectations equation, we can solve for the average preferences of the inconsistent choosers:

$$\begin{aligned}&\text{Avg. Prefs of Inconsistent}\\ &= \frac{\text{Avg. Prefs of Population} - (\text{Fraction Consistent} * \text{Avg. Prefs of Consistent})}{\text{Fraction Inconsistent}}\\ &= \frac{0.60 - (0.70 * 0.75)}{0.30}\\ &= 0.25\end{aligned}$$

Thus, the ARF approach implies that only 25% of the inconsistent choosers prefer oranges to apples, suggesting that the sandwich shop should nudge its customers toward apples. Note that this is the opposite conclusion that would have been arrived at by the minimizing opt-outs, majority preferences, or revelatory frame approach.

The preceding analysis is summed up in Table 7.

Table 7.
THE AUGMENTED REVELATORY FRAME APPROACH

	Fraction Choosing Oranges	*Fraction Choosing Apples*
Oranges Prominent	0.825	0.175
Apples Prominent	0.525	0.475
Equally Prominent	0.60	0.40
Fraction Consistent	0.525 + 0.175 = 0.70	
Fraction of Consistent that Prefer Oranges	0.525 / 0.70 = 0.75	
Fraction of Population that Prefer Oranges	0.60	
Fraction of Inconsistent that Prefer Oranges	[0.6 − (0.7 * 0.75)] / 0.3 = 0.25	

Combining all of the results in this Part yields the following general formula:

$$\text{Fraction of Inconsistent Choosers that Prefer } y = \frac{\text{Fraction Choosing } y \text{ if Revelatory Frame} - \text{Fraction Choosing } y \text{ if Nudged Toward } x}{\text{Fraction Choosing } y \text{ if Nudged Toward } y - \text{Fraction Choosing } y \text{ if Nudged Toward } x}$$

Before proceeding, it is worth highlighting several assumptions that were implicit in the foregoing analysis and upon which the ARF approach relies. Most importantly, one must assume that the revelatory frame was actually revelatory, in that the choices observed in that frame actually reveal the choosers' true preferences. For example, if it turned out that 70% of the full population preferred oranges to apples, rather than the 60% implied by the revelatory frame, this would imply that 58% of inconsistent choosers prefer oranges to apples (rather than the 25% we had previously determined).

A second assumption relied on by the ARF analysis is that the nudge in question is unidirectional (as described in Part I.A). For example, in the sandwich shop context, this assumption rules out the possibility that some customers always choose whichever fruit is least prominent. The assumption is important because whether a nudge is unidirectional affects what fraction of decision makers we estimate to be consistent as well as the average preferences of that group. In particular, falsely assuming that a nudge is unidirectional would cause us to bias the estimated preferences of the consistent choosers, which would then result in an incorrect estimate for the inconsistent choosers' preferences.[91]

The ARF analysis's third assumption is that the consistent people actually prefer the option that they (consistently) select. That is, it must be that those customers who always choose apples (even when oranges are prominent) actually prefer apples to oranges. This assumption would be violated in contexts where people are subject to multiple biases, some of which are unrelated to the nudge that policymakers are designing. For example, if employees at a company are present-biased, some of those who opt out might actually have been better off sticking with the default option. Although there are undoubtedly many settings in which consistent choosers make (consistent)

91. When the unidirectional nudge assumption fails, researchers will still be able to identify a range of possible values for the preferences of the consistent choosers, and can apply the ARF approach to each possibility in that range to obtain a range of possible values for the inconsistent choosers' preferences as well. *See* Goldin & Reck, *supra* note 36, at 30. Additionally, when researchers have information about the degree to which the assumption is violated (e.g., that 10% of the inconsistent choosers select whichever option is not the default), the approach described here can be corrected in a straightforward way. *See id.* at 31.

mistakes, there is little reason to infer that this will generally be the case in settings where people respond to nudges. For example, in the auto insurance example described in the introduction, there is no obvious reason to assume that those customers who consistently select full insurance would actually be better off with the limited insurance option.[92]

Finally, a key assumption of the ARF approach is that the assignment of decision makers to frames is as good as random. If the grocery store assigned only weekend shoppers to the pro-orange frame and only weekday shoppers to the pro-apple frame, observed differences in choices between the two frames could reflect differences in average preferences between weekend and weekday shoppers rather than any effect of the frame. Similarly, when assignment to the revelatory frame is not random, there is no guarantee that the choices observed in that frame will actually reflect those of the population.[93]

When these four assumptions are satisfied, the ARF approach provides a useful guide for choosing the best direction in which to nudge.[94] Unfortunately, revelatory frames are unavailable in many settings and policymakers must design choice architecture in such settings as well. The next approach I consider substitutes a different type of information for the revelatory frame.

92. Moreover, absent evidence to the contrary, maintaining a presumption that consistent choices reveal preferences is normatively appealing on grounds of nonpaternalism. When choices are inconsistent, policymakers lack coherent preference information upon which to base policy. In contrast, consistent choices—although potentially at odds with preferences—may at least form the basis for policy determinations. As such, there is a sense in which inferring preferences from consistent behavior avoids "overriding" choice in the formation of policy. *See* Bernheim & Rangel, *supra* note 40, at 53 (defending a related assumption in the context of their approach).

93. For instance, if only male choosers were observed in the revelatory frame, the choices from that frame would perfectly reveal the choices of men rather than the choices of the full population. When random assignment to frames is not possible, a second-best solution is for the researcher to control for differences between the groups assigned to each frame. *See infra* Part III.B.

94. Another convenient property of the ARF approach is that it can be readily extended to settings with more than two frames. In particular, with N frames, it remains the case that only one revelatory frame must be observed to identify the preferences of those who choose inconsistently. To see why, note that for any pair of frames, one can identify the consistent choosers' preferences as above, where consistency is defined with respect to those two frames only, and subsequently apply the ARF approach to recover the preferences of the choosers who are inconsistent with respect to those two frames. One can then repeat this process for each distinct pair of frames.

B. The Demographic Extrapolation Approach

The Demographic Extrapolation approach provides a method for overcoming the problem of behavioral preference identification in settings where revelatory frames are unavailable. Like the ARF approach, Demographic Extrapolation begins by recovering the preferences of the consistent choosers. Whereas ARF then combines that information with preference information for the full population, without a revelatory frame the researcher won't have a way of obtaining the preferences of the full population. And, as discussed in Part II.D, the preferences of the consistent choosers on their own is typically not enough for choosing which way to nudge since the choosers in that group may have systematically different preferences from the choosers who are inconsistent.

The Demographic Extrapolation approach attempts to overcome this problem by extrapolating the preferences of the consistent choosers to the inconsistent choosers only after controlling for the differences in observable characteristics between the two groups. For example, in the retirement savings context, Part II.D discussed the possibility that employees' preferences about savings as well as their propensity to make consistent choices could both be correlated with whether they have graduated college. However, it could be that employees' preferences and their consistency are uncorrelated once one controls for educational attainment. That is, if one were to divide up employees based on their education—e.g., college graduates, high school graduates, no high school, and so on—it could be that within those groups, people's preferences over retirement savings are not systematically related to their susceptibility to defaults. In this case, we can learn about the inconsistent choosers' preferences by extrapolating the preferences of the consistent choosers in the corresponding educational group.[95]

The general idea here is the same one that underlies multivariate regressions or "matching on observables" techniques commonly employed in other areas of empirical research: as long as we can observe the factors that induce a correlation between preferences and consistency, we can account for those factors to recover the preferences of the inconsistent choosers.

Implementing the Demographic Extrapolation approach consists of three steps. First, one divides the population into groups based on the available

95. A different way to think of the Demographic Extrapolation approach is as an extension of the minimizing opt-outs approach discussed *supra* Part II.A, in which the opt-outs are weighted according to the demographic similarity between the person opting out and the inconsistent choosers. That is, if the inconsistent choosers tend to be younger and disproportionately female, an opt-out by a twenty-four-year-old woman will be weighted more heavily in selecting which way to nudge than would an opt-out by a seventy-five-year-old man.

demographic information. The number of groups could range anywhere between two to ten to hundreds, depending on how much information is available and the type of choice under consideration.

The second step is to identify the preferences of the consistent choosers within each group. To do so, one can employ the same formula that we described in the ARF section (restricted to data on the choices made by the particular demographic group in question). As in the ARF approach, the main appeal of this method for identifying the preferences of the consistent choosers is that it does not require observing individual decision makers in multiple frames.

Finally, one must combine the preferences of the inconsistent choosers within each demographic group into a measure for the average preferences of the inconsistent choosers as a whole—the quantity that the optimal nudge depends on. This task is made more difficult by the fact that because one cannot typically observe individual decision makers in multiple frames, the researcher is not able to learn who is consistent and who is not. Nonetheless, doing so is possible by comparing the distribution of the demographic characteristics among the consistent choosers and the full population.[96] By combining this mess of information together, the researcher can recover the average preferences of the inconsistent choosers and hence the best direction in which to nudge.

To illustrate the Demographic Extrapolation approach in practice, consider the following hypothetical numbers for the retirement savings plan example described above. Suppose that 70% of employees choose to participate with opt-out enrollment and that 40% do so when enrollment is opt-in. Because we suspect that both preferences and consistency may be related to employees' education, we will apply the Demographic Extrapolation approach based on this variable. To keep things simple, we will divide employees into only two groups: those who have graduated college (60%) and those who have not (40%). Assume that the employees' choices are described by the following table.

96. The basic idea is to compute the fraction of each demographic group that is consistent as well as the prevalence of that demographic group in the full population. Using the Law of Iterated Expectations, one can then back out the fraction of the prevalence of that demographic group among the inconsistent choosers. *See* Goldin & Reck, *supra* note 36, at 18-19 (deriving a formula for identifying the prevalence of demographic groups among the populations of consistent and inconsistent choosers).

Table 8.
THE DEMOGRAPHIC EXTRAPOLATION APPROACH

	College Graduates	*Nongraduates*	*Total*
Participation with Opt-In Enrollment	0.56	0.16	0.40
Participation with Opt-Out Enrollment	0.66	0.76	0.70
Fraction of Population	0.60	0.40	1.00

After dividing the population into the demographic groups (college graduates versus nongraduates), the next step is to identify the average preferences of the consistent choosers in each group. Applying the formula for the consistent choosers' preferences derived in the previous Part tells us that 62% of the consistent graduates prefer enrollment,[97] but that only 40% of the consistent non-graduates do.[98] Assuming that consistent and inconsistent employees with the same college graduation status also have the same preferences (on average), we can conclude that the respective fractions of inconsistent graduates and non-graduates that prefer enrollment are also given by 62% and 40%.

Finally, we need to combine these numbers to obtain the average preferences of the inconsistent employees. Whereas college graduates make up 60% of all employees, these figures imply that graduates constitute only 20% of the inconsistent employees; 80% of the inconsistent employees did not graduate from college.[99] As a result, we can conclude that 44.4% of inconsistent

97. In particular, 90% of the college graduates are consistent [0.56 + (1 − 0.66) = 0.90], and of those who are consistent, 62% prefer enrollment [0.56 / (0.56 + 0.34) ≈ 0.62].

98. The reasoning is the same as in the preceding footnote: 40% of nongraduates are consistent [0.16 + (1 − 0.76) = 0.40] and of that group 40% prefer enrollment [0.16 / (0.16 + 0.24) = 0.40].

99. This result follows from a formula derived in Goldin & Reck, *supra* note 36, at 18. In particular, the fraction of inconsistent choosers in demographic group g is given by

$$P(g|inconsistent) = \frac{1 - \beta_g(y) - \beta_g(x)}{1 - \beta(y) - \beta(x)} P(g)$$

where $P(g)$ denotes the fraction of group g in the population, $\beta(x)$ and $\beta(y)$ denote the fraction of the population choosing against the frame under frame-x and frame-y

employees prefer enrollment.[100] Consequently, the Demographic Extrapolation approach as applied to this data implies that the employees will be better off (on average) without automatic enrollment. Note that this is the opposite conclusion of what one would draw by following either of the approaches suggested by Sunstein and Thaler.[101]

As with other empirical techniques, the credibility of the Demographic Extrapolation approach varies depending on how much demographic information about the decision makers is incorporated into the analysis. The more detailed the demographic information that the researcher can observe and account for, the more likely it is that the extrapolation will be valid. For example, extrapolating based on gender alone will be less reliable than extrapolating based on gender, age, income, education, and state of residence. When the choosers' gender is all that is available, employing the Demographic Extrapolation approach requires assuming that among men, those who are consistent have on average the same preferences as those who are inconsistent, and that the same is true among women. In contrast, when one includes information based on gender, age, income, education, and state of residence, the comparison groups are much narrower; a group might consist of twenty-seven-year-old high-school-educated women earning $60,000 a year. This approach will be valid when the comparison groups are defined in a sufficiently detailed way so that remaining differences in whether people choose consistently are not systematically related to their preferences.[102]

(respectively), and $\beta_g(x)$ and $\beta_g(y)$ denote the corresponding quantities restricted to choosers in group g. To apply this formula to this example, let y denote participation in the savings plan, so that frame-x corresponds to opt-in enrollment and frame-y corresponds to opt-out enrollment. For college graduates, we have $P(g) = 0.60$, $\beta_g(x) = 0.56$, and $\beta_g(y) = 0.34$, and for the full population, we have $\beta(x) = 0.40$ and $\beta(y) = 0.30$. Applying the formula yields that 20% of inconsistent choosers are college graduates.

100. This result comes from taking the weighted average of the preferences of the inconsistent choosers with and without college degrees, where the weights reflect the relative prevalence of each group: $(0.2) * (0.62) + (0.8) * (0.4) = 0.444$.

101. *See supra* Part II.A-B.

102. In some cases, this assumption may constitute an important difficulty for the approach, especially when both preferences and sensitivity to the nudges are related to hard-to-measure psychological factors. In particular, people who have identical characteristics (with respect to the information the researcher can observe) may nonetheless differ in a systematic way in both their consistency and their preferences. Note that although I have focused on demographics, any information about individuals can theoretically be incorporated into the analysis, such as the results of psychological testing, detailed genomic information, brain imaging scans, or information about past choices.

CONCLUSION

By highlighting how choice architecture shapes behavior, recent scholarship has drawn attention to a wide range of new policy instruments. At the heart of this new approach lies a dilemma: governments attempting to use nudges to promote welfare must have some way of identifying which nudge will make decision makers better off. But by the very nature of a nudge, one cannot answer this question in the traditional way, i.e., by looking to see which option decision makers freely choose.

Although daunting, the problem of behavioral preference identification does not require giving up the goal of designing choice architecture to help people achieve their objectives. Perhaps the most important lesson from the nudge literature is that a government cannot help but nudge: policymakers end up shaping choice architecture whether they intend to or not. Choosing not to nudge is simply not an option. Because nudging is inevitable, a crucial goal for research in this area is to develop tools to ensure that the nudges that are employed are ones that help rather than hurt people's welfare.

This Essay constitutes a starting point in this research agenda. It shows that overcoming the problem of behavioral preference identification is possible, at least under certain conditions. Which of the proposed approaches is best to apply will depend on the setting under consideration. When one can observe choice data from a revelatory frame, the ARF approach will typically be more reliable because it doesn't require one to make assumptions about which characteristics shape people's preferences and their susceptibility to nudges. On the other hand, in settings where the revelatory nature of a frame is suspect but individual characteristics can be observed, applying the Demographic Extrapolation approach will often make more sense. Although neither approach is a silver bullet, together they provide a straightforward set of tools for designing effective nudges under a range of circumstances.