NBER WORKING PAPER SERIES

BEHAVIORAL PUBLIC ECONOMICS

B. Douglas Bernheim
Dmitry Taubinsky

Working Paper 24828
http://www.nber.org/papers/w24828

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2018

This chapter is forthcoming in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson (eds.), Handbook of Behavioral Economics, Volume 1, New York: Elsevier. In preparing this chapter, we have benefitted from the thoughtful and insightful comments we received from a number of individuals, including Hunt Allcott, Sandro Ambuehl, Stefano DellaVigna, David Laibson, Benjamin B. Lockwood, and Juan Rios Rivera as well as seminar participants at the Stanford Institute for Theoretical Economics and UC Berkeley. William Morrison and Josh Kim have provided excellent research assistance. Taubinsky thanks the Sloan Foundation for support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Behavioral Public Economics
B. Douglas Bernheim and Dmitry Taubinsky
NBER Working Paper No. 24828
July 2018
JEL No. H0

ABSTRACT

This chapter surveys work in behavioral public economics, emphasizing the normative implications of non-standard decision making for the design of welfare-improving and/or optimal policies. We highlight combinations of theoretical and empirical approaches that together can produce robust qualitative and quantitative prescriptions for optimal policy under a range of assumptions concerning consumer behavior. The chapter proceeds in four parts. First, we discuss the foundations and methods of behavioral welfare economics, focusing on choice-oriented approaches and the measurement of self-reported well-being. Second, we examine commodity taxes and related policies: we summarize research on optimal corrective taxes, the efficiency costs of sales taxes that are not fully salient, the distributional effects of sin taxes, the use of non-price policies such as nudges, the tax treatment of giving, and luxury taxes. Third, we examine policies affecting saving, including capital income taxation, commitment opportunities, default contribution provisions for pension plans, financial education, and mandatory saving programs. Fourth, we detail the manner in which under-provision of labor supply and misunderstandings of policy instruments impact optimal labor income taxation and social insurance. We close with some recommendations for future work in behavioral public economics.

B. Douglas Bernheim
Department of Economics
Stanford University
Stanford, CA 94305-6072
and NBER
bernheim@stanford.edu

Dmitry Taubinsky
University of California, Berkeley
Department of Economics
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
dmitry.taubinsky@berkeley.edu

1 Introduction

The standard economic approach to policy evaluation relies on the assumption of rational "revealed preferences," which holds that people always choose what is best for them, that their choices do not depend on seemingly inconsequential "frames," and that the preferences revealed by their choices are transitive and complete. This assumption may seem stringent from a psychological perspective. Nevertheless, it is at the heart of modern Public Economics because it directly connects theory and data. Within the rational choice paradigm, economists can often quantify the welfare effects of policies involving commodity taxes, income taxes, unemployment insurance benefits, and savings incentives using only a few measurable, high-level statistics, such as the elasticity of consumption or labor with respect to the tax rate. In a field often concerned with quantitative evaluation of real-world policies, revealed preferences is a powerful and seemingly crucial identifying assumption.

Even so, the ostensible purpose of many important public policies is to address the concern that people do not always choose what is best for them, and that the determinants of consumer behavior extend beyond narrow self-interested optimization. For example, many countries have established government bureaus that offer "consumer protection" to guard against the possibility that firms may attempt to exploit unsophisticated buyers.[1] A number of countries have also created "behavioral insights" teams, the role of which is to leverage findings from psychology and Behavioral Economics to formulate more effective government policies.[2] Policy makers often justify otherwise standard policies such as "sin taxes" on cigarettes, alcohol, sugary drinks, and similar goods on the grounds that they discourage harmful behaviors. Motivations for consumer-facing energy policy include the possibility that people may undervalue energy-efficient goods and overvalue energy-inefficient ones due to a "defective telescopic faculty" (Hausman, 1979). Arguments for mandatory retirement savings programs often reference consumer myopia (Feldstein and Liebman, 2002).

The existence of such policies, combined with a large and growing body of empirical work in Behavioral Economics, suggests that the standard approach in Public Economics to policy evaluation may yield misleading conclusions about the welfare effects of some policies, and are simply inapplicable to other policies that influence behavior through framing effects, such as those that determine salience. The rapidly expanding literature in "Behavioral Public Economics" (henceforth BPE) combines the methods and insights from Behavioral Economics and Public Economics to extend the public-economics toolbox, thereby allowing for more robust evaluations of real-world policies, to develop innovative policy tools, and to explain why consumers' responses to policy incentives are sometimes anomalous (Chetty, 2015).

This handbook chapter summarizes the emerging field of BPE. A comparison with Bernheim and Rangel (2007), which assessed the state of the nascent field roughly a dozen years ago, reveals that progress has been dramatic. Our focus is on the *normative* questions that have historically

[1] Examples include the Consumer Financial Protection Bureau in the U.S., the Federal Ministry of Justice and Consumer Protection in Germany, the Competition and Consumer Commission in Australia, and the Financial Conduct Authority in the U.K. See the OECD 2017 report for a summary of over 100 applications of "behavioral insights" by government bureaus and sectors across the world.

[2] As of the writing of this chapter, such countries include the U.K., U.S., Australia, and Singapore.

played central roles in the field of Public Economics; we are not primarily concerned with research that only aims to describe the *positive* effects of government policies. There are at least two ways to organize such a chapter: we could focus on substantive policies, considering relevant behavioral phenomena in each instance, or on behavioral phenomena, describing the various policy implications in each instance. Consistent with our substantive focus, we adopt the first of these approaches. The challenge for BPE that we seek to highlight throughout this chapter is the need to maintain the tight link between empirically measurable statistics and welfare estimates, while moving beyond the revealed preferences assumption.

The merger of Behavioral Economics and Public Economics has required the formulation and refinement of new paradigms for evaluating economic welfare. Accordingly, the chapter begins in Section 2 with a review of recent developments involving the foundations of Behavioral Welfare Economics. We distinguish between two main schools of thought, one that employs choice-oriented methods, another that relies on measures of self-reported well-being. We articulate the foundations for each approach, explain strategies for implementation, and discuss limitations. Additional topics include paternalism and alternatives to welfarism.

While Section 2 focuses on foundational conceptual issues such as the nature of economic welfare and the definition of a "mistake," as well as on classes of empirical strategies for quantifying mistakes, the next three sections examine concrete aspects of policy design and evaluation, as well as empirical implementation. The general conceptual framework usefully disciplines the applications, sometimes in subtle and surprising ways, and it clarifies their interpretation. However, readers whose interests lie in concrete policy analysis will find that it is possible to read Sections 3 through 5 without first absorbing all of Section 2.

In Section 3 we summarize research on policies targeting commodities. We use a simple model to illustrate how changes in the commodity tax affect social welfare when a bias arises either from consumption "internalities" (i.e., people over- or under-consume a particular good) or from a lack of tax salience. We also explain how to incorporate redistributive concerns, as many sin taxes are regressive. We summarize existing empirical estimates and empirical approaches that facilitate robust implementation of the commodity tax formulas. We end by mentioning some implications of social preferences for commodity taxation, and by reviewing the potential roles of non-tax policy instruments, such as information provision and graphic warning labels.

Section 4 reviews research concerning policies that target personal saving. We begin by summarizing two behavioral themes that have played important roles in this literature: imperfect self control, and limited financial sophistication. From there we turn to capital income taxation, which we explore as an application of the principles developed in Section 3. Other policy instruments include features of special savings accounts, such as opportunities for commitment and default options. We use simple models to explore the use of each instrument, and discuss strategies for deploying them in combination. We close the section with discussions of other related policies, such as financial education, choice simplification, and mandatory saving.

In Section 5 we turn to policies targeting earnings. Analogously to Section 3, we provide a simple formula for optimal income taxation in the presence of biases that lead to either under- or over-provision of labor, or that foster inattention to, or misperception of, the tax. We use the formula to guide a discussion of theoretical work involving more complex models, as well as related research on social insurance programs designed to address medical needs, unemployment, and other adverse developments. We summarize empirical studies that yield estimates of the key parameters appearing in the formula, and point out that many of the rationality failures documented in the literature can be *good* for social welfare. We also discuss the feasibility of using the mechanism design approach to optimal income taxation when consumers are behavioral, as well as the possibility of motivating labor supply through non-tax instruments.

Section 6 concludes with a discussion of challenges for future work.

Despite the length of this chapter, we have not attempted to canvas the field comprehensively. Rather, our object has been to provide a somewhat unified perspective on a reasonably large collection of themes that we regard as important. We could make a strong case in favor of covering many other papers and topics. To the authors of those papers, we offer our apologies.

2 Behavioral Welfare Economics

Normative questions are central to the field of public economics. For well over half a century, the dominant approach to those questions was rooted in the paradigm of revealed preference, which instructs us to infer objectives and welfare from choices. But behavioral economics teaches us that choices are not always consistent. While we have achieved some insight into the sources of that inconsistency, many puzzles and controversies remain. How can we make coherent statements about welfare when the choices to which we look for guidance are inconsistent for reasons we do not fully understand? In this section, we briefly review the leading approaches to welfare analysis in settings with behavioral agents. For more complete discussions of these issues, see Bernheim (2016; 2018).

2.1 What is welfare?

Meaningful measurement requires a clear conceptual understanding of what one is trying to measure. Accordingly, we begin with a foundational question: what is economic welfare? To be clear, our focus here is on the definition of individual well-being. We address the important issue of social aggregation below in Section 2.2.7.

Accounts of well-being. Philosophers often divide accounts of well-being into three broad classes. The following labels and one-line summaries are from Kagan (1998); see also Parfit (1984) and Griffin (1986).

1. *Welfare hedonism*: "Well-being consists solely in the presence of pleasure and the absence of pain." Classical economists such as John Stuart Mill and Jeremy Benthan advocated forms of welfare hedonism. To the extent modern economists sympathize with this view, they are

usually drawn to a variant called *mental statism*, which holds that well-being is exclusively a reflection of mental states.

2. *Preference theory*: "Well-being consists in having one's preference satisfied." To be clear, the question here is whether preferences are satisfied in reality – in other words, whether the world is as the individual would like it to be, rather than whether she believes this to be the case. However, generalized versions of preference theory allow for the possibility that the individual's preferences encompass their own mental states, which may depend on their understanding of outcomes. Modern economics firmly embraces preference theory.

3. *Objective theories*: "Well-being is a matter of having certain goods in one's life, goods that are simply worth having, objectively speaking," irrespective of whether one prefers them or not. The classic statement of this perspective is due to Aristotle (2012, translation). For a more contemporary expression, see Sen (1985), who defines welfare in terms of basic "functionings," such as nourishment.[3] Objective theories have received considerably less attention in behavioral public economics than the alternatives.

The following example illustrates why it is important to think through foundational issues concerning the definition of welfare when practicing behavioral welfare economics.

The parable of the oblivious altruist. A small town in Arkansas experiences massive flooding, leaving many families homeless. To provide financial assistance for the impacted families, the government raises taxes, including a $100 levy on Norman. As a general matter, Norman thinks government spending is wasteful, but he is also an altruist, and would gladly contribute $100 to the fund if he knew about it. However, he never learns about the flood or the relief effort. Does the government's policy make him better off or worse off?

According to welfare hedonism, "external" states such as the true status of impacted families in Arkansas matter to Norman only insofar as they affect his "internal" states. Because he assumes his incremental taxes fund low-value government projects, the relief effort degrades the quality of his internal states. Welfare hedonists must therefore conclude that the aid initiative reduces his well-being.

According to preference theory, the true state of the world determines Norman's well-being. Because the government actually uses the incremental taxes to assist impacted families, and because Norman would approve of this expense if he understood it, those adhering to simple versions of preference theory must conclude that the initiative makes him better off.

The example is instructive because neither conclusion is entirely satisfactory. On the one hand, welfare hedonism elevates perceptions over truth and applauds happy delusions. On the other hand, simple preference theory fails to account for the genuine psychic costs that may result from Norman's misconceptions.

[3] In some respects, Sen's discussion of functionings is preference-theoretic, inasmuch as he argues that people likely have similar preferences over functionings.

A more satisfactory account of well-being follows from a generalized version of preference theory in which desires encompass both external and internal states. Imagine, for example, that Norman knows legitimate needs arise from time to time, such as those of the Arkansas flood victims. According to this theory, if he would prefer to live in a world where the government addresses those needs when they come up as a matter of policy regardless of his awareness, then the hypothesized initiative enhances his well-being. However, if he would prefer to live in a world where the government addresses those needs only when he is aware of them, then the same initiative reduces his well-being. Under this theory, Norman's own preferences determine the relative weights attached to his internal mental states versus external reality.[4]

The relevance of this example to behavioral economics should be clear. We are frequently concerned with settings in which people may misunderstand the consequences of their choices. In those cases, does well-being depend on the imagined state of affairs, the real state of affairs, or both? The answer to this question fundamentally shapes the conclusions that follow from normative economic analyses.

2.2 Choice-oriented methods

Implementation of preference theory requires us to identify empirical expressions of consumers' desires. In classical welfare analysis, choices serve this role.[5] Naturally, there are other potential windows into preferences, and we address them in the course of the discussion below.

2.2.1 The behavioral critique of standard welfare economics

Bernheim (2016; 2018) articulates the preference-theoretic premises for standard welfare economics as follows (see also Hausman, 2012):

- *Premise 1*: Each individual is the best judge of their own well-being.
- *Premise 2*: A single coherent, stable preferences governs each individual's judgments.
- *Premise 3*: Each individual's preferences determine their choices: when they choose, they seek the greatest benefit according to their own judgment.

Significantly, these premises do not require one to take a rigid philosophical stand on the precise nature of well-being. Instead, one can leave such matters to the individual. For the parable of

[4] While the generalized version of preference theory offers more satisfactory normative prescriptions than the simple version, implementation is especially challenging. For example, it is difficult to see how one elicits preferences over deluded states of mind without identifying and hence removing the delusions. As a result, simple preference theory often provides the implicit philosophical foundations for practical exercises in behavioral welfare economics.

[5] Confusion can arise, however, because philosophers and economists sometimes use the word "preference" differently. To illustrate, imagine Norman chooses a sour apple over a pear, believing incorrectly that the apple is sweet. Some philosophers would say that, by virtue of his choice, Norman demonstrates a preference for the sour apple over the pear. This perspective leads to certain criticisms of preference theory (see, e.g., Hausman, 2012). An economist would distinguish between preferences and beliefs: Norman prefers a sweet apple to a pear, and falsely believes the sour apple to be sweet. According to that perspective, the problem lies in Norman's beliefs, not in his preferences.

the oblivious altruist, we can be philosophically agnostic as to whether the true and/or imagined state of affairs contributes to welfare, and defer to each individual's own judgment, as reflected in appropriate choices.[6] Some see this agnosticism as an advantage of the preference theory approach.

Behavioral economics arguably calls for a new welfare paradigm because it challenges the validity of these premises. *Fallibility critiques* call Premise 1 into question on the grounds that people do not or cannot reliably exercise good judgment. *Consistency critiques* highlight the sensitivity of our choices to apparently irrelevant contextual features of decision problems, a phenomenon that implies either a lack of coherent and stable objectives (contrary to Premise 2), or a loose connection between preferences and choices (contrary to Premise 3). Aggressive versions of consistency critiques raise the possibility that the concepts of "true preferences" and aggregate "experienced utility" are fictions – that we do not aggregate the many diverse aspects of our experience until we are called upon to do so for a given purpose, such as making a choice or answering a question about our well-being, at which point, instead of accessing and applying pre-existing preferences, we "construct" (or "assemble") our judgments (see, e.g., Lichtenstein and Slovic, 2006).[7] This perspective attributes context-dependent choice to the vagaries of aggregation: different circumstances may render different aspects of experience more or less salient, and thus change the weights attached to them during the process of preference construction.

2.2.2 Behavioral Revealed Preference

Many economists are reluctant to relinquish the core assumption that people have coherent, stable preferences, or the normative dictum that those preferences ought to govern welfare analyses. Accordingly, they attribute the phenomena animating the fallibility and consistency critiques to features of decision processes that ostensibly distort true preferences. To construct formal theories of decision making, they supplement standard models with additional elements representing the "cognitive biases" that arguably give rise to those distortions.

Unfortunately, choice data can shed only so much light on the parameters of such models. Accordingly, if one hopes to recover preferences, one must adopt a reasonably parsimonious representation of the pertinent biases. Bernheim (2016; 2018) summarizes the core principle underlying this approach, known as *behavioral revealed preference* (or sometimes *model-based behavioral welfare economics*) as follows:

- *The Principle of Behavioral Revealed Preference (BRP):* If enough is known about the process mapping preferences to choices, then one can invert it conditional on its unknown parameters, and recover both those parameters and preferences from choice data.

[6] These choices may be unconventional and difficult to implement, but one can visualize them in principle. For Norman, we might seek to elicit the compensating variation for learning the true disposition of the incremental taxes, stipulating that the memory of the decision would be erased upon making the choice.

[7] The hypothesis that people construct their judgments contextually may help to explain why "stated preferences" differ systematically from actual choices; see, for example, Harrison and Rutstrom (2008). If it were possible simply to access preexisting preferences, consumers would presumably be able to access and state those preferences accurately, even in the absence of choice. Instead, it appears that people do not actually know what they will choose until they choose it.

As an example, analyses positing *biased beliefs* fall within this paradigm. The typical study of this type supplements the standard von Neumann-Morgenstern model of decision making under uncertainty with an account of systematic divergences between beliefs and objective probabilities. Under appropriate assumptions, and with sufficient data, one can both measure the bias and estimate the other parameters of the utility function. Substituting the objective probabilities for the distorted beliefs, one then obtains "true preferences." Koszegi and Rabin (2008b) illustrate this approach by modeling a particular bias (the gambler's fallacy) in a setting where a decision maker bets on repeated flips of an objectively fair coin, and showing that one can in principle recover both beliefs and risk preferences from choices. See Sections 2.2.3 and 3.2.3 for further discussions of biased beliefs. We discuss many other examples in subsequent sections of this chapter.

An attraction of this approach is that it accommodates behavioral economics by departing only modestly from the underlying perspectives of standard welfare analysis. However, the apparent simplicity of the approach can be deceptive. We turn our attention next to the main complications and challenges encountered when applying the behavioral revealed preference paradigm.

The nature of consumers' limited concerns. All choice-oriented welfare methods require the practitioner to take a stand on the aspects of experience that contribute to well-being. The very concept of a delimited consumption bundle implicitly distinguishes between experiences that intrinsically matter to the individual and those that do not. The dimensions of that bundle provide the analyst's answer to the question, what do people care about?

The BRP paradigm allows for the possibility that decisions depend on conditions that have no direct bearing on well-being, but that instead impact biases. Once the analyst takes a stand on the aspects of experience that contribute to well-being, the identity of these conditions, known as *decision frames*, follows as an implication.

As an illustration, suppose we ask Norman to order his lunch for a scheduled meeting one week in advance. Whether he selects a sandwich or a salad may depend on whether we require him to decide at 1pm after he has just eaten, or at 4pm when he's hungry (Read and van Leuwen, 1998). Here, the natural assumption is that Norman's concerns, and hence his consumption bundle, only encompass food items, in which case the decision frame consists of the time at which he makes his choice. A BRP model might account for the framing effect by positing that hunger (or alternatively the absence thereof) induces a cognitive bias.

Bernheim (2016; 2018) points out that the BRP paradigm inextricably links the notion of a framing effect to the concept of a bias. If the choices of a consumer with a coherent and stable preference relation vary across decision frames (as in our motivating example), then bias must of necessity infect some of those choices. Conversely, whenever a choice suffers from a hypothesized bias, one can imagine a reframed version that removes the cause.[8] In some applications, the reframed

[8] As a purely logical matter, one can of course imagine environments in which cognitive processes *always* distort choices. However, if there is no context within which an individual expresses a judgment consistent with the "optimal" choice according to a BRP model, then there is no empirical foundation for claiming that the model correctly captures his true preferences. As an example, imagine that there are no conceivable circumstances under which Norman would

choice problem has an obvious empirical counterpart. If Norman's hunger is the source of bias when choosing at 4pm, then moving his decision time to 1pm facilitates an unbiased choice. In other applications, the reframed choice problem is merely a potentiality. For example, if a consumer holds biased beliefs concerning events that occur with known probabilities, one could imagine replacing the naturally occurring (and potentially confusing) information structure with a transparent alternative, such as drawing balls from an urn. One can interpret the welfare-optimal alternative according to a BRP model of biased beliefs as the choice the consumer would make in this reframed setting (assuming it successfully removes the cause of the bias).

Different assumptions about the scope of consumers' concerns lead to different implications about the nature of framing effects and biases. To appreciate this point, notice that our motivating example admits a second interpretation: Norman's well-being depends not only on the food he eats, but also on what he orders and when he orders it. In that case there are no decision frames, and arguably no biases: Norman acts on his true preferences at all points in time, despite making time-dependent selections.

This alternative interpretation of Norman's behavior suggests a variant of the BRP approach, wherein the analyst expands the assumed boundaries of the consumer's concerns until all inconsistencies disappear, and then proceeds as if there are no biases. Gul and Pesendorfer's (2001) analysis of temptation preferences fall within this category. They account for various patterns commonly associated with time inconsistency by assuming that consumption bundles consist not only of the items consumed, but also of the menus from which consumers select them. Both of these applications place aspects of the decision problem, rather than merely the selected item, within the scope of consumers' concerns, thereby raising a complication that we discuss momentarily (the Non-comparability Problem).

In practical applications, finding objective criteria for drawing lines between decision frames and elements of the consumption bundle can prove challenging. Because one can in principle rationalize virtually any behavior as a reflection of either framing effects or exotic preferences, valid justifications for drawing the lines one way rather than another inherently hinge on non-choice evidence. We mention some possible empirical approaches in Section 2.2.4.

Unfortunately, as we discuss next, intuition concerning consumers' concerns can sometimes steer the analysis into conceptually treacherous waters.

The Non-comparability Problem. In some applications, it may seem natural to assume that the experience of *choosing* falls within the scope of the consumer's concerns. For example, if Norman chooses a sandwich when salad is available, he may feel guilty, and if he chooses salad when a sandwich is available, he may enjoy greater self-respect. Unfortunately, these possibilities raise conceptual challenges for choice-based welfare analysis.

The following example illustrates how seemingly sensible assumptions about consumers' concerns

order salad for lunch. An economist theorizes that Norman actually prefers salad, but suffers from a pervasive cognitive bias. While this theory is logically consistent, it is also untethered from the facts.

can lead to difficulties.[9] Suppose we task Norma with dividing a sum of money between herself and a friend. Norma is averse to bearing the responsibility for leaving her friend with nothing when other options are available. Consequently, no matter how the task is framed, she divides the money equally. However, she is inherently selfish and fervently wishes someone would take the decision out of her hands and give her the entire prize. Plainly, none of Norma's choices can reveal this preference. In particular, if we ask her to choose between the original choice problem and a setting in which a third party decides to give her everything, she will still feel responsible for the outcome, and consequently choose to divide the money herself, splitting it equally.

Bernheim (2016, 2018) conceptualizes the general problem as follows.[10] When a planner faces a decision involving various potential courses of action, choice-based welfare analysis makes a prescription by asking what the affected consumer would choose if offered the *same alternatives.* But in situations where consumers' concerns encompass the experience of choosing, the planner's task and the consumer's task are inherently *non-comparable.* In particular, presenting the planner and the consumer with (ostensibly) the same menu does not mean that the alternatives (correctly defined) are actually the same. For instance, if Norma's well-being depends not only on what she orders but also on what she personally chooses to forego, her choices cannot shed light on the best course of action for a planner who makes the decision for her, because she personally chooses to forego nothing when the planner makes the selection.

We can avoid the non-comparability problem completely if we are willing to assume that consumers' concerns do not encompass conditions pertaining specifically to the experience of choosing (*conditions of choice*, as opposed to *conditions of consumption*). Another possibility is to assume that consumers only care about conditions of choice under well-defined circumstances. For example, choice-based welfare analysis becomes possible in Gul and Pesendorfer's (2001) theory of temptation, which is otherwise susceptible to the non-comparability critique, if we assume that people care about the conditions of choice (e.g., experience temptation) only when decision tasks have immediate material consequences.[11] Objectively justifying such assumptions can prove challenging, however, because justifications must hinge on non-choice evidence rather than on choice patterns.

[9] We have adapted this example from Koszegi and Rabin (2008a).

[10] The following is a more formal statement of the non-comparabilty problem. Let (X, f) denote the decision task consisting of the opportunity set X presented with framing f. To allow for the possibility that the consumer's concerns may encompass the experience of choosing, we assume preferences are defined over objects of the form (x, X, f). If the consumer chooses $x^*(X, f)$ when presented with the problem (X, f), we can conclude only that $(x^*(X, f), X, f) \succ (x, X, f)$ for all $x \in X$. For two distinct decision problems, (X, f) and (X', f'), the consumer's choices provide us with no basis for determining whether she is better off with $(x^*(X, f), X, f)$ or $(x^*(X', f'), X', f')$. Consequently, we can never say whether a policy that changes the decision problem facing a consumer helps or hurts her. Presenting her with a choice between two decision problems does not by itself resolve the issue, since the metachoice simply creates a new choice problem of the form $(X \cup X', f'')$ (where the new frame, f'', captures the fact that the decision is now structured as a choice between frames). Without additional assumptions, there is no reason to think that the choices in this new setting reveal the consumer's preferences between an unchosen assignment to one decision problem or the other.

[11] Implicitly, Krusell et al. (2010) make this assumption when evaluating welfare using Gul and Pesendorfer's (2001) model of temptation preferences, which otherwise implicates the non-comparability problem.

The identification of biases. In any given application, once we settle issues pertaining to the boundaries of consumers' concerns, we confront another equally vexing question: when choices in two frames conflict, how can we tell which (if either) accurately reflects preferences, and which is biased? In Norman's case, hunger might cloud his judgment or focus his attention. How do we tell the difference? As with any economic question, researchers should resolve these issues based on objective, generally applicable criteria informed by pertinent evidence. It (almost) goes without saying that "I know it when I see it" is not a sound methodological principle. We discuss empirical strategies for making these judgments in Section 2.2.4. Here and in Section 2.2.3 we examine the conceptual foundations for those strategies.

A common practice among practitioners of the BRP paradigm is to posit the existence of a utility function, $U(x, f)$ (where x is the chosen item and f is the decision frame) that rationalizes decisions. This function summarizes all positive knowledge about choice. For obvious reasons, many behavioral economists call it *decision utility* (or sometimes *ex ante* utility). Another common practice is to posit the existence of a welfare function, $V(x)$. In this framework, welfare depends only on the chosen item because, by definition, the frame lies outside the scope of the consumer's concerns. In any frame, f, bias then consists of the (ordinal) discrepancies between $U(\cdot, f)$ and $V(\cdot)$. The literature offers three alternative interpretations of V: first, that it captures *true preferences*, second that it reflects *experienced utility* (also known as *ex post* utility), and third, that it is simply a function that rationalizes choices within a special subset of decision frames (and hence is also a form of decision utility). Here we focus on the first two interpretations, noting some conceptual difficulties. The third interpretation, which emerges from the Bernheim-Rangel framework (discussed in Section 2.2.3), provides an attractive alternative for those who find the following issues problematic.

The circularity trap. A common but problematic idea is to define a biased choice as one that is contrary to true preferences. Unfortunately, that approach can lead to circularity: we identify bias by looking for choices that conflict with true preferences, while inferring true preferences from unbiased choices. A key challenge in behavioral welfare economics is to find a conceptually sound escape route from this circularity trap. In Section 2.2.3, we describe an approach that involves focusing on whether particular decisions reflect correct perceptions of available actions and the outcomes they yield (conditional on the available information), rather than on whether particular objectives are "true," and we detail strategies for empirical implementation in Sections 2.2.4 and 3.2.3.

Sometimes economists attempt to recover true preferences by estimating structural models of choice. While this approach can prove invaluable, it cannot provide the needed escape route. Such models always have multiple normative interpretations; see, for example, the discussion of quasi-hyperbolic discounting in Section 2.2.5. Using them for welfare analysis therefore requires an assumption concerning the component of the model that represents true preferences. In the absence of some other objective foundation for inferring bias, labeling a model one way rather than another amounts to resolving normative issues by assumption. It is simply too much to hope that choices themselves can reveal which choices are unbiased.[12] Consequently, the identification of bias generally requires consideration of non-choice evidence. That said, in some contexts, evidence favoring minimalistic structural assumptions will suffice; see Goldin and Reck (2015) and Benkert and Netzer (forthcoming) for theoretical treatments.

The trouble with experienced utility. The interpretation of $U(x, f)$ and $V(x)$ as, respectively, decision utility and experienced utility has gained traction among some economists; see, for example, Chetty (2015). Even setting aside important questions regarding empirical implementation, this interpretation raises some conceptual concerns.

First, the assumption that people derive welfare only from experience is limiting because it excludes legitimate non-experiential objectives, and consequently leads to some problematic conclusions. Recall the case of the oblivious altruist: a policy of routinely assisting flood victims reduces Norman's experienced utility because he is never aware of the flooding and is always upset about the associated taxes. Suppose Norman's preferences favor the policy even in light of these consequences. Are we nevertheless prepared to say that the policy makes him worse off?[13]

The following example illustrates how such considerations can drive a wedge between decision utility and experienced utility even in the absence of a bias. Every day, Norma eats vegetables sautéed in olive oil. She actually thinks vegetables taste better when sautéed in butter, but she is vegan and believes it is immoral to consume animal products. She is also forgetful: if she deviated from her routine and used butter, she would not remember, and would attribute the better taste to the freshness of the vegetables.[14] She is fully aware of her forgetfulness, but still chooses olive oil over butter. In this example, experienced utility ranks the options differently than decision utility (butter over olive oil rather than olive oil over butter). Yet Norma's decisions are clearly consistent with her preferences.

Second, even if people only care about hedonic experience, there are natural and important settings in which that experience cannot logically include the welfare evaluation V. To illustrate, suppose a consumer's decisions in period 1 determine her consumption in periods $t = 1, ...T$. We will assume that welfare V depends on a collection of hedonic sensations $(h_1, ..., h_T)$ that span all periods – in other words, that every period's experience matters to some degree. We allow for the possibility that h_t may be a vector of sensations, but it does not have to include all sensations experienced in period t. To apprehend $V(h_1, ..., h_T)$ as a coherent hedonic sensation, the consumer would have to experience it in at least one period, t, either as an element of h_t, or as an additional sensation.[15] Letting $\tilde{h}_s^t$ denote the perception of period-s sensations as of period t (either a memory for $s < t$ or an anticipation for $s > t$), the consumer can in principle experience $V(\tilde{h}_1^t, ..., \tilde{h}_{t-1}^t, h_t, \tilde{h}_{t+1}^t, ..., \tilde{h}_T^t)$

[12] Sen (1993)makes a version of this point: "there is no way of determining whether a choice function is consistent or not without referring to something external to choice behavior (such as objectives, values, or norms)."

[13] To be clear, welfare hedonism embraces this implication, but we suspect most readers will reject it.

[14] The purpose of assuming she is forgetful is to eliminate the possibility that knowledge of her unethical behavior might degrade her ex post experience.

[15] Indeed, if V represents an ex post evaluation, she would have to experience it in period T, which is potentially problematic in an infinite-horizon setting, but we will not impose that restriction.

as a hedonic sensation in period t, but she cannot experience the true value of aggregate welfare, $V(h_1, ..., h_T)$, unless all period-t recollections and expectations are accurate. Thus, when we assume the consumer hedonically experiences aggregate welfare, V, we exclude a broad swath of behavioral economics.

Analogous issues arise in the context of settings with uncertainty. To illustrate, suppose the welfare function is $V(x_1, .., x_S) = p_1 v(x_1) + ... + p_S v(x_S)$, where x_s is the payment received in state s and p_s is the associated (objective) probability. To experience V ex post (that is, after the realization of s), the consumer's sensations would have to include regret and/or relief associated with unrealized outcomes, based on a correct understanding not only of the alternative outcomes and the sensations they would have induced, but also of the probabilities. When a consumer chooses an action and then experiences an outcome, she does not actually experience any of the other outcomes, nor does she experience the associated probabilities (inasmuch as only one outcome materializes). While her experience may correct ex ante misconceptions concerning x_s or $v(x_s)$ for the realized state s, it does not inherently correct misconceptions about x_r of $v(x_r)$ for any unrealized state r, nor about the probabilities (biased beliefs). Accordingly, the consumer cannot plausibly apprehend actual aggregate welfare, V, as an ex post hedonic sensation in most behavioral settings with uncertainty,

A satisfactory interpretation of the welfare function V therefore requires a clearer conceptual foundation for the concept of bias. We discuss foundations in Section 2.2.3, and address empirical implementation in Section 2.2.4.

A rigid consistency requirement. Because the BRP paradigm adheres rigidly to the core assumption that people have coherent, stable preferences, it *requires* one to define the scope of consumers' concerns, and then to identify decision frames that induce "bias," in a manner that yields an internally consistent set of "unbiased" choices.[16] This inflexible consistency requirement can compel one to make assumptions about consumers concerns, and about bias, that lack objective supporting evidence and go beyond our actual understanding of choice processes.[17] As an example, several studies have found that decisions with no immediate consequences are sensitive to the weather at the moment of choice (Busse et al., 2015; Meier et al., 2016). Yet as far as we know, there is no objective foundation for declaring that rain induces a bias while sunshine does not, or vice versa. Even more fundamentally, the requirement is sensible only if people make decisions by attempting to access pre-existing, coherent preferences. If instead they construct preferences contextually (as strong versions of the consistency critique maintain), one cannot claim that "bias" is the only possible source of inconsistency.[18] In that case, the BRP paradigm can require the adoption of models that are too simplistic given the underlying decision processes.

[16] Formally, the set of unbiased choices must satisfy the Weak Axiom of Revealed Preference (WARP) to ensure the existence of a "true preference" representation.

[17] Goldin and Reck (2015) show that it is sometimes possible to recover the consumer's preferences without such assumptions, but the applicability of their methods is limited.

[18] Notably, attempts to "clean" choice data through the application of objective criteria do not generally remove all significant inconsistencies (Benjamin et al., 2016).

One potential solution is to introduce the possibility that each consumer acts upon multiple "true" preference relations, which they harmonize inefficiently (for example, by expressing different preferences in different frames). To make welfare statements, one must aggregate over the preference relations. As an example, Laibson et al. (1998) interpret the standard model of quasi-hyperbolic discounting as implying that the consumer has a distinct "true" preference relation at each moment in time; their welfare analysis employs the Pareto criterion. Despite some initial interest, this approach is not currently in widespread use. For further discussion, see the Appendix.

Model uncertainty. The BRP approach is also demanding on analysts because it presupposes that they can successfully identify correct behavioral models. Because behavioral economists operate within a domain that offers abundant degrees of freedom, many distinct models of choice processes can potentially account for the same or similar choice mappings. Experience teaches us that building a professional consensus for the "right" model can be extremely difficult, even when the choice mapping is known.

2.2.3 The Bernheim-Rangel framework

The absence of a conceptual framework for identifying biases based on objective evidence represents a serious gap in the BRP paradigm. Unfortunately, attempts to fill that gap collide head-on with the paradigm's rigid consistency requirement. There is simply no guarantee that general principles for diagnosing biased choices will, in any given application, reduce the choice domain to an internally consistent subset, and indeed no hope of success if people construct their preferences contextually. It is possible, for example, that upon applying a set of sound principles, one would conclude that Norman's choices at 1pm and 4pm both reflect legitimate perspectives, or that choices made on rainy and sunny days are equally valid, even though they differ. (See also the discussion of time inconsistency and welfare in Section 2.2.5). What then?

The framework for behavioral welfare economics proposed by Bernheim and Rangel (2009), and refined by Bernheim (2016; 2018), eliminates this tension. As a result, it opens the door to principled evidence-based methods for identifying the scope of consumers' concerns and diagnosing decision-making errors. It dispenses with the need to make strong assumptions concerning the nature of preferences and decision mechanisms simply to satisfy the rigid consistency requirement, while at the same time permitting such stands where there is adequate foundation. Accordingly, as we explain in greater detail below, it nests BRP, as well as other approaches (see Sections 2.3 and 2.4).[19]

The overall structure As emphasized in Section 2.2.2, all choice-oriented welfare methods require the practitioner to take a stand on the scope of consumers' concerns. Conditional on that stand, the Bernheim-Rangel approach involves two steps.

[19] In this respect, we disagree with the characterization of these methods in Chetty (2015), who sees them as competing rather than nested.

- *Step 1*: Identify all decisions that merit deference (the welfare-relevant domain)
- *Step 2*: Construct a welfare criterion based (at least in part) on the properties of choice within that domain.

These steps are implicit in the BRP approach. As explained in Section 2.2.2, in settling on a positive model and adopting a particular normative interpretation, we effectively identify collections of decision problems for which choices ostensibly express undistorted "true" preferences. BRP amounts to conducting standard revealed preference analysis on those restricted domains.

The BRP approach entails serious challenges because it places demanding restrictions on the inputs for the second step: we cannot "recover preferences" unless welfare-relevant choices are mutually consistent. In contrast, a key feature of the Bernheim-Rangel framework is that the second step employs a criterion that flexibly accommodates inconsistencies among the choices that merit deference. That feature fundamentally alters the nature of the first step. We can in principle identify welfare-relevant choices by entertaining the same evidence, arguments, and modeling strategies as in the BRP framework. However, unlike BRP, the Bernheim-Rangel framework does not compel the analyst to settle on welfare-relevant domains within which all choices are internally consistent. This difference is particularly important in contexts where there is skepticism about the evidence used to identify biases. If the application of objective and appropriate criteria for evaluating whether any given choice merits deference fails to yield a set of internally consistent choices, the analyst does not need to "try harder." The Bernheim-Rangel framework also allows one to perform welfare analysis provisionally under different views of which choices do and do not merit deference, and thereby provide a more thorough understanding of the assumptions upon which particular normative conclusions depend.

Revised premises for choice-oriented welfare analysis To derive defensible general principles for diagnosing decision-making errors and constructing welfare criteria, one needs to build on sound conceptual foundations. As discussed in Section 2.2.1, behavioral economics offers various critiques that call the foundations of standard welfare economics into question. Bernheim (2016; 2018) argues that certain essential features of the main premises nevertheless survive. He distinguishes between *direct judgments*, which are opinions that pertain to outcomes we care about for their own sake, and *indirect judgments*, which involve alternatives that lead to those outcomes. He then reasons that while behavioral economics and psychology provide a foundation for questioning certain types of indirect judgments, they do not impugn direct judgments. With respect to the latter, standard arguments for deference to individual judgment continue to apply. One such argument invokes justification for self-determination in the tradition of classical liberalism: my views about my life are paramount because it is, after all, *my* life. A second entails the Cartesian principle that subjective experience is inherently private and not directly observable, which renders each of us uniquely qualified to assess our own well-being.[20] Neither of those arguments presupposes the independent

[20] Modern libertarian philosophers such as Nozick (1974) describe self-determination as a fundamental right rather than a means to an end, and construe that principle as constraining the legitimate scope of government.

existence of "true preferences" or of aggregate "experienced utility." Nor do they assume that an individual always reaches exactly the same judgment. Objections to direct judgments entail nothing more than a difference of opinion between the analyst and the consumer as to what constitutes a good or fulfilling life. Thus there is no objective foundation for overturning the presumption in favor of a direct judgment and declaring the analyst's perspective superior. The same argument applies to indirect judgments for which the consumer properly understands the connection between actions and consequences.

The question remains, why draw the line at choices? Why not accord equal status to other types of judgments, such as evaluations of happiness and life satisfaction? Obviously one cannot assert the primacy of choice based on a presumed connection with "true preference" if the latter does not actually exist. If choice is simply a constructed judgment, then one could argue that other types of constructed judgments, such as self-reported well-being, should be equally admissible for the purpose of evaluating welfare. The answer given in Bernheim (2016; 2018) is that deference to a constructed judgment in the course of analysis is warranted only if the purposes of the analysis and the judgment are conformable. He argues that economists usually see normative analysis as a tool for guiding policy makers when they select among alternatives, under the assumption that the objective is to promote the well-being of those affected by the selection. When people make choices for themselves, they aggregate over the many dimensions of their experience for precisely the same reason. Accordingly, when advising policy makers on the selection of an alternative that affects a particular consumer, we may justifiably defer to that consumer's choices because they reveal the alternatives that, in her judgment, would provide her with the greatest overall benefit if selected. In contrast, other types of constructed judgments aggregate experience for different purposes; see in particular the discussion of self-reported well-being in Section 2.3, below.

These considerations lead to the following revised premises:

- *Premise A*: With respect to matters involving either direct judgment or correctly informed indirect judgment, each of us is the best arbiter of our own well-being.
- *Premise B*: When we choose, we seek to benefit ourselves by selecting the alternative that, in our judgment, is most conducive to our well-being.

To formulate a welfare framework based on these revised premises, one has to grapple with two main issues. First, how does one distinguish between choices that reflect correctly and incorrectly informed judgments? Second, how does one accommodate inconsistencies among the judgments that merit deference? The next two sections describe the answers provided in Bernheim and Rangel (2009) and Bernheim (2016; 2018).

Welfare-relevant choices In principle, the two-step structure allows analysts to define welfare-relevant domains however they wish, but forces them to make these restrictions explicit so others can evaluate them. Despite this flexibility, only certain types of restrictions on the welfare-relevant domain are consistent with the underlying philosophical foundations set forth above. Those foundations justify the exclusion of a choice that expresses an incorrectly informed indirect judgment, but not one that is correctly informed. Indeed, Bernheim (2016, 2018) takes the position that, absent clear evidence that a judgment is incorrectly informed, or that choices and judgments diverge, the presumption in favor of deference to individual choice should stand. Under this view, proper exclusions from the welfare-relevant domain should be limited to identifiable mistakes. Others may take a broader view. For example, some argue against deference to sadistic or immoral choices (Harsanyi, 1978; Sen, 1980-1981).

The challenge, of course, is to identify mistakes without presupposing a knowledge of preferences, and thereby encountering the circularity described in Section 2.2.2. Bernheim (2009, 2016, 2018) classifies a decision as a mistake if it has two distinctive features. First, there must be some unchosen option in the opportunity set that the decision maker would select over the chosen one in some other decision problem, where either the menu or the framing differs (i.e., a *choice reversal*). If, on the contrary, the decision maker robustly stands by her choice irrespective of menus or framing, then we have no empirical basis for claiming that another option in the opportunity set is superior according to her judgment. A choice reversal is, however, neither helpful in identifying which choice is mistaken, nor even sufficient for establishing the existence of mistake, inasmuch as it could reflect contextually constructed judgments. Thus we look for a second feature: a mistaken choice is predicated on a characterization of the available options and the outcomes they imply that is inconsistent with the information available to the decision maker (*characterization failure*). In other words, it reflects an incorrectly informed indirect judgment. By itself, characterization failure raises the possibility that a mistake may have occurred, but does not guarantee that outcome, because one can make the right decision for the wrong reason. However, as long as characterization failure infects only one of two decision problems associated with a choice reversal, we can declare the infected choice a mistake. Because this definition avoids any reference to divergences between choices and preferences, it avoids circularity.

To identify a mistake under this definition, one requires both rich choice data and information concerning the decision maker's understanding of the available options and the outcomes they imply. We discuss possible empirical strategies in Section 2.2.4.

To illustrate the principles discussed above, suppose we are concerned that a consumer makes mistakes due to biased beliefs. As we explained in Section 2.2.2, one can interpret the welfare-optimal alternative according to a structural model of biased beliefs as the choice the consumer would make in a reframed, transparent setting. Thus, choice reversals are implicit, and one could verify their existence by implementing the corresponding decision problems. In settings with objective probabilities, one can demonstrate characterization failure in a given frame by showing that people misunderstand the mathematical rules governing the derivation of pertinent probabilities from the available information (e.g., conditioning), or that they do not notice, retain, or properly understand pertinent facts governing the probabilities. (See Spinnewijn, 2015, for an empirical example involving excessive optimism concerning reemployment prospects.) Significantly, some of the strategies for demonstrating characterization failure discussed in subsequent sections are equally applicable in settings with subjective probabilities, where objective odds are either unknown or unmeasurable absent debatable assumptions about the underlying data-generating processes. Accordingly, in contrast to the BRP method of Koszegi and Rabin (2008b), this approach allows one to analyze the impact of biased beliefs even in settings where one cannot rule out any particular belief as objectively irrational.

The aforementioned notion of a mistake has parallels in the literature on the philosophical foundations of paternalism. New (1999) separates failures of reasoning into two general categories: "those pertaining to judgments about the appropriate course of action and those related to the actual choices made to achieve a given object" (see also Dworkin, 1971). The first category includes "technical inability," which prevents the individual from properly using the available information to understand the nature of the available options – in short, characterization failure. The second category includes phenomena such as "weakness of will," which ostensibly cause choices and judgments to diverge. A question arises as to whether this second category defines an additional class of mistakes involving *optimization failure*.

To make this discussion more precise, imagine the consumer responds to a particular decision task by attempting to solve the following problem:[21]

$$\max_{x\in X} u(g(x)), \tag{1}$$

where x is an action, g maps actions to outcomes that matter to her intrinsically (e.g., mental states), and u captures her judgments. Objections to the consumer's choice must fall into one of the following four categories: (i) she misunderstands the set of available actions, X, (ii) she misunderstands the mapping from actions to outcomes, g, (iii) she fails to perform the "max" operator correctly, or (iv) she uses an inappropriate objective function, u.[22] Choice-oriented welfare analysis rejects (iv) as a source of mistakes. Characterization failure subsumes (i) and (ii), while optimization failure subsumes (iii).

With respect to optimization failure, the critical question is whether one can detect it using systematic evidence-based criteria without knowing the objective function. Discussions of optimization failure generally assume not only that the individual has a single coherent objective function (which is debatable), but also that it is known. "Weakness of will" is a good example. If we are amenable to assuming that the consumer has a well-defined unitary objective that reliably guides her choices only when all consequences are delayed, then it is sensible to say that optimization failure can occur when some consequences are immediate. Yet it is also possible that the consumer embraces an objective that guides the choices she makes when actions have immediate consequences. Moreover, this may be her "true" objective, or she may embrace different objectives in different contexts (see the discussion in Section 2.2.5). In that case, references to "weakness of will" reflect disagreements about proper objectives rather than problems with optimization; in other words, the objection to

[21] To be clear, in adopting this formulation, we do not intend to imply that the consumer employs the same objective function u for different decision problems.

[22] We thank Sandro Ambuehl for suggesting this categorization.

the consumer's choices actually falls within category (iv), which choice-oriented methods disavow, rather than category (iii).

In principle, evidence on decision processes could establish that consumers choose their actions by applying algorithms that cannot logically maximize any objective function. Such evidence would obviously establish the existence of optimization failure. Whether this strategy proves useful in practical applications remains to be seen.

The welfare criterion Upon completing Step 1, the analyst may find that the welfare-relevant domain is "too large" in the sense that inconsistencies among choices remain, "too small" in the sense that certain choice-based comparisons are impossible, or "just right" in the sense that choices are comprehensive and consistent. Here we focus on the case where the domain is "too large" (which is especially pertinent if there is skepticism about evidence of bias), and then comment on the case in which it is "just right." We take up the possibility that the domain is "too small" in Section 2.4.

In the Bernheim-Rangel framework, a normative criterion is a binary relation. If W is a generic normative relation, and if x and y are outcomes, then "xWy" means that outcome x is better than outcome y. Bernheim (2016, 2018) argues that any sensible criterion should satisfy the following three properties:[23]

- *Property #1 (coherence)*: W is acyclic.[24]
- *Property #2 (respect for unambiguous choice)*: If, within the welfare-relevant domain, y is never chosen when x is available, then xWy.
- *Property #3 (consistency with the welfare-relevant domain)*: If x is chosen in some decision problem with opportunity set X within the welfare-relevant domain, then x is not welfare-improvable within X according to W.

The justification for the first two requirements is transparent, but the third may require some explanation. To declare x welfare-improvable within X would mean that choosing x in the specified problem is a mistake. But a central purpose of Step 1 is to weed out all identifiable mistakes, and no data or inferential methods in Step 2 are excluded from Step 1. Therefore, if one can legitimately classify the selection of x as a mistake in Step 2, one should already have deleted it from the welfare-relevant domain in Step 1.

Bernheim and Rangel (2009) demonstrate that there exists one and only one candidate for the welfare relation W satisfying these three properties: the unambiguous choice relation, P^*.[25] This result makes our lives fairly simple: if one endorses the three requirements, then P^* is the only game

[23] The second and third properties reference choices made within the welfare-relevant domain. To be clear, nothing in the framework requires direct observation of those choices. The analyst can use the usual methods to fill in missing data, including interpolation, extrapolation, and structural modeling.

[24] Acyclicity is generally regarded as the weakest possible coherence requirement, in the sense that it is necessary and sufficient for the existence of maximal elements.

[25] Formally, xP^*y if and only if the welfare-relevant domain contains no decision problem in which x is chosen but y is available.

in town. When there are inconsistencies within the welfare-relevant domain, P^* is an incomplete relation. Intuitively, it instructs us to respect choice whenever choice provides clear normative guidance, and to live with whatever ambiguity remains. Thus, it allows analysts to exploit the coherent aspects of behavior, which feature prominently in virtually all behavioral theories, while embracing the normative ambiguity implied by any lack of coherence.

In settings where choice inconsistencies within the welfare-relevant domain are pervasive, P^* may not be very discerning. Whether the resulting ambiguity undermines our ability to draw useful welfare conclusions depends on the context; for an example, see the discussion of Bernheim et al. (2015a) in Section 4.5.2. When a lack of discernment proves problematic, one can attempt to sharpen one's conclusions by returning to Step 1 and focusing on the theoretical and empirical issues governing the definition of the welfare-relevant domain.

What if it turns out that step 1 yields a welfare-relevant domain that is "just right," in the sense that it is both comprehensive and internally consistent, rather than "too large"? In that case, P^* coincides with the normative criterion obtained from the familiar principles of revealed preferences. We therefore arrive at the third interpretation of the welfare function, V, introduced in Section 2.2.2: it is simply a function that rationalizes choices within a special subset of decision frames (the welfare-relevant domain), and hence is actually a form of decision utility. Thus the framework provides a true generalization of both standard welfare economics and BRP.[26] Seeing BRP exercises through the lens of the Bernheim-Rangel framework is useful because it highlights the fact that welfare analysis hinges on the properties of the naturally occurring and welfare-relevant choice mappings, rather than on the cognitive models one invokes to rationalize those mappings. The importance of deriving welfare implications from choice mappings implied by models of cognition, rather than from the models themselves, is a theme of Section 3.

Applying the criterion The analytic implementation of the aforementioned welfare criterion is reasonably straightforward. The framework yields intuitive counterparts for the standard tools of welfare analysis, including equivalent variation, compensating variation, and consumer surplus. To take a simple case, suppose Norman has two tickets to a college football game, and is wondering whether he should use them or sell them. His willingness-to-accept differs across decision frames, but is never less than \$50 and never more than \$60. In that case, we can say that having and using the tickets improves his welfare by \$50 to \$60. That range reflects the ambiguity implied by his choices. In many instances, applications of the framework simply involve evaluating a policy from the perspective of the most favorable and least favorable frames using otherwise conventional methods; see Section 4.5.2 for an example.

[26] Bernheim (2016; 2018) argues that apparent counterexamples reflect a failure to envision the entire choice domain. As an example, consider the following BRP model of a masochist: instead of maximizing utility, $u(x)$, the consumer minimizes it. The Bernheim-Rangel interpretation of this model is that, if u truly represents the consumer's objectives, then it must be possible to envision an alternative decision frame in which the consumer acts on those objectives; absent *any* setting that is free from an alleged distortion, we ought to question whether the associated conception of preference lacks a foundation and is merely a contrivance.

Discussion The Bernheim-Rangel framework draws a stark distinction between choices that fall within and outside the welfare-relevant domain. In principle, one could imagine an alternative approach that admits uncertainty concerning the correct classification while simultaneously reviving the core BRP assumption that people have coherent, stable preferences. Under this view, inconsistencies remain after eliminating all "obvious" mistakes in step 1 simply because some errors are difficult to detect. Accordingly, one might hope to compute expected welfare effects based on posterior probabilities concerning the likelihood of error for each decision problem. In Norman's case, the expected improvement in his welfare from having and using the football tickets would be a single value between \$50 and \$60.

Any approach that assumes the existence of "true preferences" is obviously vulnerable to the criticism that our judgments may be contextually constructed. Even setting that objection aside, implementation of the alternative approach described in the preceding paragraph is challenging due to the difficulty of devising an objective method for recovering the posterior probabilities of error. No such methods currently exist and, unfortunately, it is hard to imagine an implementation that avoids arbitrary and problematic assumptions. For example, one could build and implement a structural model based on the assumption that choices tend to cluster around preferred options, in which case outliers are likely mistakes. However, if the frames that induce error arise far more frequently than those that do not, the outliers may be the best guides to welfare.

2.2.4 Empirical Implementation of Choice-Oriented Methods

In this section, we discuss general empirical strategies for conducting behavioral welfare analyses using choice-oriented methods. Because we view the Bernheim-Rangel framework as a generalization of the Behavioral Revealed Preference paradigm, our discussion will employ the vocabulary of the former.

Core methods. As we have emphasized, all applications of choice-oriented behavioral welfare economics implicitly or explicitly specify the scope of consumers' concerns and define a welfare-relevant domain. The ideal application also performs the following three tasks, in each case by marshaling appropriate evidence:

- *Task 1*: Estimate choice mappings within the naturally occurring domain, and within the welfare-relevant domain.
- *Task 2*: Justify assumptions concerning the boundaries of the welfare-relevant domain by providing evidence that inconsistencies between choices in naturally occurring and welfare-relevant frames are attributable to characterization failures in the latter and not the former.
- *Task 3*: Justify assumptions about the scope of consumers' concerns.[27]

[27] To be clear, it is impossible to perform the first two tasks without assumptions about the scope of consumers' concerns. We do not mean to suggest otherwise by listing the task of justifying these assumptions third.

As discussed in subsequent sections, most applications are more attentive to task 1 than to tasks 2 and 3. We recommend addressing each task with equal seriousness where there is legitimate scope for controversy. Here we elaborate on general approaches to each task.

Task 1: Estimating choice mappings Essential inputs for choice-oriented welfare methods include rich descriptions of behavior within both naturally occurring and welfare-relevant decision frames (choice mappings). The task of estimating a choice mapping over a naturally occurring domain is entirely standard. In cases where welfare-relevant choices are also observed, the same methods apply. Here we are concerned with the frequently encountered problem of estimating the choice mapping for the welfare-relevant domain when data on welfare-relevant choices are either sparse or nonexistent. Applicable methods fall into the following four categories.

The first method is to create the data by presenting people with appropriately reframed decision problems. Examples include Allcott and Taubinsky (2015) on the demand for lightbulbs, discussed in Section 3.2.3, and Ambuehl et al. (2017) on the quality of financial decision making, discussed in Section 4.6.2. An important advantage of this approach is that one can deduce welfare implications directly from the discrepancies between the original and reframed choices without the need for restrictive assumptions about behavioral and cognitive processes.

When psychologists use this first method, they call it *debiasing*; for a recent survey, see Soll et al. (forthcoming). In effect, the objective of debiasing is to reframe the decision so that it lies within the welfare-relevant domain. That said, the normative superiority of the supposedly debiased choices is not always justified as carefully as it should be. As an example, one procedure is to point out inconsistencies across choices, and then ask subjects to rethink their decisions. The intent is to improve decision making, but the effect could be to prompt spurious resolutions of legitimate normative ambiguity by inducing an experimenter demand effect.[28]

The second method is to extrapolate the missing welfare-relevant choices from other types of decisions using structural models. As an example, in applications involving the "$\beta\delta$" model (quasi-hyperbolic discounting), many analysts have assumed, in effect, that the welfare-relevant domain consists of choices with no immediate consequences (see Section 2.2.5). Even if no such choices are observed for the application of interest, one can in principle recover the model's parameters either from the naturally occurring choices, or from time-preference experiments. Setting $\beta = 1$, one can then use the model to infer choices within the welfare-relevant domain. The approach to biased beliefs described at the outset of Section 2.2.2 has a similar structure, and falls within the same category. We provide many examples of this method in subsequent sections.

Unlike the first method, structural modeling requires one to make restrictive assumptions about behavior and decision processes. However, minimalistic structural assumptions suffice for some applications. Suppose, for instance, that consumer demand for a product depends on two types of fees, one transparent and the other shrouded. If we assume only that the response to the shrouded fee ought to be the same as the response to the transparent fee, we can reconstruct unbiased demand.

[28] One can design alternative protocols that minimize those demand effects, but then significant normative ambiguities may remain. See, for example, Benjamin et al. (2016).

For applications of this idea, see the discussions of Allcott and Wozny (2014) and Busse et al. (2013) in Section 3.2.3.

The third method is to extrapolate the missing welfare-relevant choices from the decisions of similar individuals who ostensibly avoid characterization failure (the "rational consumer benchmark"). For instance, one could attempt to deduce sensible portfolio decisions from the choices of financial professionals, or sensible medical decisions from the choices of doctors; see the discussion of Bronnenberg et al. (2015) and Allcott et al. (2018b) in Section 3.2.3. Studies employing this approach must address the possibility that the tastes of the "experts" differ systematically from those of the target population, or that the experts suffer from other sources of characterization failure (e.g., medical training may tend to induce hypochondria).

The fourth method is to extrapolate the missing welfare-relevant choices from non-choice data. One variant of this approach assumes that a properly informed consumer would choose the option that leads to the greatest happiness or life satisfaction; see Section 2.3.1. One could also attempt to draw such inferences from data on hypothetical choices and "stated" preferences (Shogren, 2005; Carson and Hanemann, 2005; Carson, 2012). A related strategy is to rely on statistical models that treat these types of subjective responses as predictors, instead of taking them at face value as predictions. With that approach, there is no need to resolve which of two or more SRWB measures is "correct" because one can use them as co-predictors of choice, potentially along with other subjective reactions and even biometric measurements.[29] Experimental evidence indicates that this strategy can dramatically reduce both mean-squared error and bias when predicting choice out of sample; see Bernheim et al. (2015b).

We return to these methods in Section 3.2.3, where we discuss empirical applications.

Task 2: Justifying the welfare-relevant domain While the conceptual principles governing the identity of the welfare-relevant domain are reasonably straightforward (see Section 2.2.2), justifying particular assumptions within the context of an application can be challenging, and indeed this task often receives short shrift. Potential strategies include the following. First, one can evaluate whether people properly understand concepts central to the proper characterization of certain choice problems. See the discussion of Ambuehl et al. (2017) in Section 4.6.2 for an illustration. Second, one can examine evidence concerning the processes of observation, attention, memory, forecasting, and/or learning, with the object of determining the contexts in which certain types of facts are systematically ignored or processed incorrectly. See the discussion of Bernheim and Rangel (2004) in Section 2.2.5 for an illustration. Finally, one can evaluate whether people understand particular decision problems by posing factual questions with verifiable answers. Potential evidence includes ex post acknowledgements by decision makers that they ignored or misunderstood pertinent facts. See Benjamin et al. (2016) for an implementation.

[29] See Smith et al. (2014) for an application involving biometric reactions.

Task 3: Justifying assumptions about the scope of consumer's concerns It is worth reiterating that all choice-oriented methods require one to take a stand on the aspects of experience that contribute to well-being – in other words, to specify the component dimensions of consumption bundles. How does one determine what people care about, and thereby draw a line between consumption bundles and frames? The most common approach is to assume, as in conventional analyses, that consumers care only about "standard" consumption items such as traded goods, and to blame framing effects for most patterns that appear anomalous under that assumption. Yet there are also applications in which consumer's ostensible concerns extend to non-standard considerations such as internal emotional states.

Justifications for assumptions about the scope of consumers' concerns necessarily invoke non-choice evidence, inasmuch as one cannot describe choice data prior to settling on the definition of the consumption bundle.[30] Formal methods for executing this task remain underdeveloped. Sometimes one can rely on information about the mechanism through which a given condition affects choice – for example, whether it demonstrably leads to confusion. Another strategy is to ask people what they care about, or to introspect.

Next we turn to some alternative approaches that do not fit neatly within these core methods (in the sense that they do not systematically address each of the three tasks), but that can nevertheless inform choice-based welfare evaluations.

Dominated choices. An alternative empirical approach is to evaluate changes in the quality of decision making by monitoring the frequency of dominated choices; for applications, see the discussion of financial decision making in Section 4.6.2. One variant of this approach focuses on decision tasks with dominant options; see, for example, Bhargava et al. (2017), who find that the majority of employees in a large U.S. firm choose dominated health care plans.[31] A second variant examines decisions with non-degenerate efficient frontiers.

Dominance methods allow one to proceed with minimalistic assumptions, but they are not assumption-free. To justify these methods within the Bernheim-Rangel framework, one must assume that direct judgments respect monotonicity. Upon observing a dominated choice, one can then infer the existence of frames in which the consumer notices the dominance relation and makes a different selection. Thus, choice reversals are implicit, and one could verify their existence by implementing the corresponding decision problems. The same assumption also removes all potential explanations for the dominated choice other than characterization failure. According to this reasoning, this approach offers an important potential advantage: one does not need to identify the nature of characterization failure or provide direct evidence of its existence in order to classify a dominated choice as a mistake.

[30] This observation poses a logical difficulty for those who argue against the use of non-choice data in economics, such as Gul and Pesendorfer (2008).

[31] The firm in question offered a large menu of options that differed only with respect to financial cost-sharing and premium. High-deductible plans often dominated low-deductible plans because the premium differentials exceeded the deductible differentials.

Unfortunately, dominance methods also have their limitations. The first variant of the approach – studying decision tasks with dominant options – removes personal preferences from the mix. Each decision effectively boils down to solving a math problem that has one and only one correct answer. In contrast, the vast majority of real-world decisions are not simply math problems: the "right" choice almost always depends on preferences. This difference is important because consumers may be more susceptible to characterization failure when preferences come into play. Posing a problem that has no objectively correct answer may reduce the resemblance to textbook examples, making the applicable objective principles harder to recognize.[32] People may be less likely to deploy mathematical tools when mathematics potentially govern only one amongst several aspects of evaluation. Preferences may also activate specialized heuristics or psychological mechanisms, such as motivated reasoning (Kunda, 1990), that sweep relevant principles into the background, even if they are invoked.

The second variant of this method – studying decision tasks with non-degenerate efficient frontiers – avoids this criticism by keeping preferences in the mix. However, in that case, a reduction in the frequency of dominated choices does not necessarily go hand-in-hand with a welfare improvement. One can make unambiguous welfare statements only in special cases where there happens to be a dominance relation between the actions a given consumer takes with and without the intervention of interest.

Consistency with revealed preference axioms. Another alternative approach is to evaluate changes in the quality of decision making by monitoring the frequency of WARP or GARP violations. For an example involving financial decision making, see Choi, Kariv, Mueller and Silverman (2014). Unlike the dominance method, this strategy presupposes the existence of coherent, stable preferences, and cannot accommodate the possibility that inconsistencies may reflect the vagaries of preference construction. However, conditional on that assumption, WARP/GARP violations plainly imply that some choices are mistaken. As with the dominance method, the analyst avoids the need to identify the source of the characterization failure or provide direct evidence of its existence. However, this method offers no basis for determining which choices are mistaken. Instead, it provides a way to quantify the overall prevalence and severity of decision errors.[33]

A conceptual problem with this method is that decision-making errors do not necessarily give rise to WARP/GARP violations. A consumer who exhibits a consistent misunderstanding of a principle governing the relation between choices and outcomes will nevertheless respect such axioms. For example, suppose Norman prefers oranges to bananas and bananas to apples, but consistently mistakes apples for oranges in naturally occurring decision frames. In that case, his choices will

[32] Along these lines, Enke and Zimmermann (2015) show that many people tend to neglect correlations even in simple settings, despite knowing how to account for them. Likewise, Taubinsky and Rees-Jones (forthcoming) find that many consumers underreact to sales taxes, even though they can properly compute tax-inclusive prices.

[33] To be clear, some measures of non-conformance with GARP, such as the Afriat (1972) critical cost efficiency index, do have efficiency interpretations; see, e.g., Choi et al. (2014) for a related application. Moreover, Echenique et al. (2011) provide a measure of non-conformance that is interpretable as the maximal amount of money one can extract from a decision maker with specific violations of GARP.

satisfy WARP: he will consistently choose apples over bananas, and will never choose bananas when apples are available. It follows that a reduction in WARP/GARP violations is neither necessary nor sufficient for an improvement in the quality of decision making.

Metachoices. A final choice-oriented empirical strategy for evaluating welfare involves the use of metachoices – that is, choices among decision problems. In principle, we could attempt to discover which of two decision frames leads to a better outcome when paired with the same menu by asking the consumer to choose between them. Likewise, we could try to determine the value of a choice situation by assessing the consumer's willingness-to-pay either to obtain it or to avoid it.

When evaluating such methods, the reader should bear in mind that a metachoice between two decision problems constitutes a third decision problem, in which the menu of options is the union of the menus for the two component problems, and the framing subsumes the sequential nature of the choice. Therefore, choices and metachoices are not different types of objects. While there are circumstances in which choices that happen to be framed as metachoices are informative, they do not automatically resolve welfare questions, for at least two reasons.

First, bias that infects either or both of the component problems may also infect the metachoice. To illustrate, suppose the presence of objects in shiny wrappers causes Norman to ignore all other options. In that case, decision problems that present a subset of the options in shiny wrappers (type S problems) generally leave him worse off than ones that present all options in dull wrappers (type D problems). If Norman's bias manifests itself only when he actually sees objects in shiny wrappers, a metachoice will correctly reveal the superiority of the type D problems. However, if merely thinking about objects in shiny wrappers triggers his bias, he may express a preference for type S problems. The metachoice is then misleading because the bias infects it. For a more consequential illustration of these issues, see the discussion of metachoices in Section 2.2.5, which concerns time inconsistency.

Second, metachoices can also introduce new biases. To illustrate, suppose we offer Norma a metachoice between two 100-question multiple choice tests, one on history, the other on biology. Either way, she will receive \$1 for each correct answer. Objectively, the probability that Norma answers the typical question correctly is 80% for history and 85% for biology. Abstracting from risk aversion, she should therefore be willing to pay \$5 to switch from the history test to the biology test. Yet if she incorrectly believes she answers 70% of history questions and 95% of biology questions correctly, she will overpay by \$20 to switch tests. Here, the new bias infecting the metachoice is poor *metacompetence* (i.e., an inaccurate assessment of competence).

Some economists have attempted to use metachoices to assess the welfare effects of changes in the conditions of choice. For example, DellaVigna et al. (2012) assess the willingness to pay to avoid face-to-face charitable solicitations. Because these solicitations do not create new opportunities to give, their only effect is to change a condition of choice. As noted in Section 2.2.2, assuming that conditions of choice fall within the scope of consumers' concerns potentially introduces the Non-comparability Problem. The welfare analysis in DellaVigna et al. (2012) is nevertheless valid under the additional assumption that the effects of social pressure do not influence the metachoice.

However, that assumption may be mistaken. For example, the existence of a solicitor may create social pressure to allow the solicitation (i.e., a perceived obligation). In that case, the welfare effects of solicitation do not necessarily coincide with the measured willingness to pay, and indeed may not be identified. See also the discussion of Allcott and Kessler (forthcoming) in Section 3.5.

2.2.5 An application to time-inconsistency

Even after settling on the conceptual foundations for choice-based welfare analysis, normative judgments can remain challenging and controversial. To illustrate some of the difficulties that can arise, we will examine the problem of evaluating welfare for a time-inconsistent consumer.

For concreteness, suppose Norma must choose between eating pizza or salad for lunch. She enjoys pizza more than salad but recognizes that salad is healthier. Prior to lunchtime, she prefers salad overall because she prioritizes health. However, when lunchtime arrives, she prefers pizza because she prioritizes immediate gratification. Assuming Norma cares only about the identity of the selected lunch item, this pattern constitutes time inconsistency. One could also say that she manifests imperfect self-control in the sense that she hopes and intends to order salad for lunch, but ends up with pizza.

Here we are concerned with welfare: Is Norma better off with salad or pizza? According to one prominent school of thought, the present-focused tendencies that emerge in each moment reflect a cognitive bias (see, for example, O'Donoghue and Rabin, 1999). In Norma's case, this perspective favors salad.

To apply the Behavioral Revealed Preference paradigm, we require a model of Norma's behavior. The most widely used framework for modeling time inconsistency posits quasi-hyperbolic discounting (QHD preferences, or, more colloquially, the $\beta\delta$ model).[34] The period-t objective function for a QHD consumer is $u_t + \beta \sum_{s=t+1}^{T} \delta^{s-t} u_s$, where $(u_t, ..., u_T)$ represents flow utility, and β is assumed to lie between 0 and 1. The judgment articulated in the previous paragraph associates "true preferences" with δ discounting (the *long-run criterion*), and construes $1 - \beta$ as parameterizing the magnitude of "present bias."

A difficulty with this approach is that the QHD model admits a large number of disparate normative interpretations (Bernheim, 2009). For example, one could take the position that people achieve true happiness by living in the moment, but that they suffer from a tendency to over-intellectualize when making decisions about the future. Relabeling the model, one arrives at a different account of true preference and cognitive bias consistent with this alternative perspective. The model itself provides no guidance as to which account is right and which is wrong.

Ideally, economists should rely on objective evidence-based criteria to justify using value-laden labels such as "bias" and "true preference" for elements of the model. The Bernheim-Rangel framework provides structure for such inquiries. In Norma's case, we identify two decision frames, differentiated according to whether she chooses in the moment (contemporaneous framing) or in ad-

[34] This formulation was popularized by David Laibson (1997; 1998), who borrowed it from a related experimental literature in psychology (Chung and Herrnstein, 1961).

vance (forward-looking framing). If we exclude contemporaneous choices from the welfare-relevant domain, then salad is optimal, but if instead we exclude forward-looking choices, pizza is optimal. If we define the welfare-relevant domain to include all choices, Norma's best choice between pizza and salad is ambiguous.[35] In the latter case, we may still be able to make statements such as: Norma is definitely better off with salad plus $0.50 than with pizza, and definitely worse off with salad than with pizza plus $0.75.[36]

As noted in Section 2.2.3, Bernheim (2009, 2016) argues that an evidence-based inquiry into welfare-relevance should focus on characterization failure. As an example, Bernheim and Rangel (2004) marshal evidence on the neurobiology of addiction to support their contention that the welfare-relevant domain should not include contemporaneously framed choices made in the presence of substance-related environmental cues. In brief, neurobiological research shows that a specific mechanism (the mesolimbic dopamine system, or MDS) measures correlations between environmental cues and subsequent rewards. The use of addictive substances causes the MDS to malfunction in a way that exaggerates those correlations in the presence of use-related environmental cues. As a result, the system effectively supplies the addict's brain with inflated forecasts of available rewards. The pertinent mechanism is, however, specific to addiction, and does not justify the general practice of classifying present focus as a cognitive bias.

Despite widespread use of the phrase "present bias" rather than the more neutral and descriptively accurate "present focus," the literature offers little in the way of general evidence (not pertaining specifically to addiction) of characterization failure in contemporaneously framed decisions. Bernheim (2016) offers several cautionary observations, including the fact that many cultures emphasize the importance of living in the moment, as well as the popular adage that deathbed regrets rarely include having spent too few hours at the office. These observations raise the possibility that support for the long-run criterion among some economists is a consequence of "type-A paternalism" – that is, successful workaholics imposing their own personal values on others.

The case of time inconsistency underscores the limitations of the metachoice method, discussed in Section 2.2.4, Because a metachoice must temporally precede the component choices, any attempt to officiate between the contemporaneous and forward-looking perspectives based on this method would involve a metachoice made in the forward-looking frame. It follows that the normative validity of the metachoice hinges on the assumption that the forward looking frame is free from bias. Using it therefore assumes the conclusion.

[35] In this example, one reaches the same conclusion by treating Norma as two separate individuals ("selves") and applying the Pareto criterion. For a more elaborate application of the multi-self Pareto criterion involving life-cycle planning, see Laibson et al. (1998). However, in the context of life-cycle planning problems, the multi-self Pareto criterion does not generally coincide with the Bernheim-Rangel unambiguous choice criterion under an unrestricted welfare-relevant domain. Indeed, Bernheim and Rangel (2009) argue that the multi-self Pareto criterion lacks a conceptually sound foundation. Perhaps most problematically, it assumes that each self is indifferent with respect to all past experience. That assumption is empirically vacuous, inasmuch as choices cannot shed light on backward-looking preferences.

[36] For example, this conclusion would follow if, prior to lunchtime, Norma is indifferent between salad and pizza plus $0.74, but at lunchtime is indifferent between salad plus $0.49 and pizza.

2.2.6 The problem of the second best

The fields of psychology and behavioral economics have identified a wide assortment of broadly applicable framing phenomena which analysts generally examine one at a time in narrowly delimited contexts. Unfortunately, welfare analyses that abstract from the pervasiveness and multiplicity of framing effects and biases arguably overlook critical second-best considerations (in the sense of Lipsey and Lancaster, 1956-57) that could overturn their implications.

To illustrate this concern, suppose consumers initially overestimate the benefits of compound interest. Imagine in addition that the government could eliminate this bias by adopting a financial education program, T. Ignoring the possibility that consumers suffer from other biases, the program is plainly beneficial. But what if consumers also suffer from severe present bias,[37] so that, on balance, they initially save too little? Considering all sources of inefficiency, the policy is likely harmful. Indeed, formal welfare analysis might favor an alternative "educational" intervention, D, that misleads consumers into exaggerating the benefits of compound interest even further.

Matters are even worse if one acknowledges the possible existence of unknown biases outside the immediate scope of analysis. If behavioral welfare economics defies compartmentalization (i.e., considering biases one, or a few, at a time) and instead requires a comprehensive account of human decision making, the prospects for useful progress are remote.

Fortunately, there is a coherent case to be made for compartmentalization. Returning to our example, the indictment of policy T and justification for policy D arguably follow from a conceptual error: the analysis attempts to treat sources of inefficiency comprehensively, but does not treat policy options comprehensively.[38] Distorting policies that target consumers' understanding of compound interest in order to address concerns arising from present bias makes little sense if other policy tools are better suited for the latter purpose. Suppose the optimal comprehensive policy consists of T combined with measures that create appropriate commitment opportunities. Then one can arrive at the optimum by compartmentalizing policies and the concerns that motivate them in parallel. A compartmentalized evaluation of financial education would focus on welfare effects involving comprehension, and would treat concerns about present bias as if they will be (but are not yet) fully resolved through appropriate commitments. Likewise, a compartmentalized evaluation of commitment opportunities would focus on welfare effects involving present bias, and would treat concerns about comprehension as if they will be (but are not yet) fully resolved through appropriate education.

Ambuehl et al. (2017) refer to this approach as *idealized welfare analysis*, to indicate that it treats sources of inefficiency outside the scope of the analysis as if other policies will provide ideal resolutions. The main advantage of the approach is that it provides a coherent justification for compartmentalization, at least in cases where there are good solutions for each compartmentalized problem: the planner can focus on one problem at a time, and still achieve the overall optimum.

[37] For the purpose of this example, we assume that present focus constitutes a mistake, as is often assumed.

[38] One could also object to policy D based on concerns about the ethics of spreading misinformation, or about the government's long-term credibility. Those considerations are orthogonal to our current focus.

That said, compartmentalization obviously involves compromises. If there is no good way to address a source of inefficiency outside the scope of analysis, the approach will overlook potentially important second-best considerations.

At first, it might appear that idealized welfare analysis requires a deep understanding of all decision-making flaws and their solutions, because it references judgments made in an idealized setting, rather than actual decisions. However, Ambuehl et al. (2017) show that it is sometimes possible to approximate idealized welfare effects using actual choice data, even if one has no information concerning the existence or nature of other biases that may affect those choices.

A simple example helps to illustrate the preceding point. Suppose a financial instrument z yields a future payoff $f(z)$, which the consumer mistakenly perceives as $g(z, \theta)$, where θ is an educational policy. For simplicity, the consumer expects to spend income when it is received, and evaluates outcomes according to the utility function $c_1 + \gamma u(c_2)$, where c_1 is current consumption, c_2 is future consumption, and γ is a discount factor. In that case, the consumer is willing to pay $\gamma u(g(z, \theta))$ for the instrument, but should be willing to pay $\gamma u(f(z))$. Thus, the measured valuation error is $\gamma\left(u(g(z, \theta)) - u(f(z))\right)$. Now suppose that, unbeknownst to the analyst, the consumer discounts the future excessively due to "present bias," and that true time preferences are governed by a discount factor $\delta > \gamma$. To conduct idealized welfare analysis of the educational policy, we would construct the valuation error that would result from the discrepancy between $f(z)$ and $g(z, \theta)$, assuming a full resolution of present bias through some other policy (e.g., one involving commitments). Under our assumptions, the idealized valuation error is $\delta\left(u(g(z, \theta)) - u(f(z))\right)$. Notice that the measured valuation error equals the idealized valuation error up to a factor of proportionality (here, γ/δ). Accordingly, the measured valuation error has the right sign, ranks policies (θ) in the correct order, and provides a valid gauge of their proportional costs and benefits. Notice also that the factor of proportionality does not depend on the instrument under consideration, z. Ambuehl et al. (2017) prove under much more general conditions that these properties hold to a first-order approximation for small instruments (e.g., even when true preferences involve an arbitrary function v that differs from u).

2.2.7 Social aggregation

The thorny problem of social aggregation has fascinated and perplexed economists for decades. The same challenges are present in behavioral economics, and similar solutions are available. For instance, the Bernheim-Rangel framework lends itself to generalizations of aggregate consumer surplus, the Pareto criterion, and various methods of making interpersonal comparisons. A complete discussion of these issues would consume many pages; we refer the reader to Bernheim and Rangel (2009), Bernheim et al. (2015a) (also discussed in Section 4.5.2), and Fluerbaey and Schokkaert (2013).

2.3 Self-reported well-being

The past two decades have witnessed an explosion of interest in various measures of self-reported well-being (SRWB).[39] Perhaps the most visible application in economics has involved the construction and refinement of "national happiness accounts" (see, for example, Helliwell et al., 2014; Kahneman et al., 2004). The literature is far too vast to survey here; see Helliwell and Barrington-Leigh (2010); Fujiwara and Dolan (2016); Graham (2016).

As noted in Section 2.1, one can potentially provide conceptual foundations for SRWB through either mental statism or preference theory. Unfortunately, the intended foundations for particular applications are sometimes unclear. We will begin the preference-theoretic perspectives because, in our view, it provides the strongest foundation for SRWB analysis.

2.3.1 SRWB as an implementation of preference theory

There are two possible routes to justifying SRWB as an implementation of preference theory. The first, which we already discussed in Section 2.2.4, construes SRWB as an adjunct to choice-oriented methods. Instead of taking SRWB at face value as a generally reliable measure of overall well-being, we interpret it instead as an indicator of what people would likely choose. This distinction has important practical implications because it recasts the object of the exercise as accurate prediction (of choice) rather than accurate measurement (of well-being). Such indicators may be particularly useful when pertinent choice data are unavailable, or when we have reason to believe the associated choices reflect misconceptions. As an example, Frey et al. (2010) use SRWB data to infer the willingness to pay for environmental goods. Likewise, Stutzer and Frey (2008) hypothesize that people make faulty decisions about where to live because they systematically misunderstand how they will feel about lengthy commutes; the study uses SRWB data to fill the resulting evidentiary gap concerning preferences. See also Benjamin et al. (2012, 2014).

The second preference-theoretic route to justifying SRWB assumes that the domain of preferences is limited to the decision maker's mental states. Answers to questions about overall well-being arguably express preferences over those states, as do choices. Of course, respecting the decision maker's preferences over mental states also qualifies as a form of mental statism, and consequently some of the challenges facing mental statist interpretations of SRWB, discussed in the next subsection, apply.

2.3.2 SRWB as an implementation of mental statism

In many ways, mental statist interpretations of SRWB methods seem more natural than preference-theoretic interpretations. The objective of these methods is to elicit the mental states a person actually experiences as the result of pursuing a particular course of action. As the parable of the oblivious altruist illustrates (Section 2.1), a decision to adopt mental statism is highly consequential

[39] The phrase "subjective well-being" (abbreviated SWB) is more commonly used in the literature. We prefer the phrase "self-reported well-being" (SRWB) because it avoids the incorrect implication that subjective experience is directly observable.

for the many settings in behavioral economics wherein people are assumed to hold incorrect beliefs. In effect, one must embrace the adage that "what you don't know can't hurt you."

Justifying SRWB as an implementation of mental statism is, however, more challenging than one might think. The following is a brief summary of the conceptual issues discussed in Bernheim (2016, 2018).

There are two distinct schools of thought about the nature of "aggregate utility" (AU). The first holds that we go through life experiencing disaggregated hedonic sensations, and aggregate only when we are called upon to express judgments or make choices.[40] According to this view, AU does not exist until we have reason to construct it. The second holds that AU exists as a continuous hedonic sensation that we can access and report when asked about our well-being. Obviously, the second perspective is more favorable to the use of AU as a welfare measure.

Case 1: Aggregate utility as a constructed judgment. If AU is merely a constructed judgement, then efforts to formulate a sound conceptual foundation for using SRWB to measure welfare within the mental statist paradigm encounter significant obstacles. Suppose we can describe hedonic experience as a vector $h = (h_1, .., h_N) \in H$. We can think of a judgment as a binary relation $\succcurlyeq$ that orders potential experiences (elements of H) either partially or fully. If people care only about their own mental states, then choice reflects one such judgment, call it $\succcurlyeq_C$. SRWB embodies another judgment, $\succcurlyeq_S$. If aggregate hedonic experience implies a true preference relation, $\succcurlyeq_E$, then one is free to argue that $\succcurlyeq_S$ serves as a better proxy than $\succcurlyeq_C$. However, if the consumer does not experience aggregate utility hedonically, the justification for respecting $\succcurlyeq_C$ cannot reference a relation such as $\succcurlyeq_E$; the criterion must then stand on its own, as must $\succcurlyeq_C$.

As noted in Section 2.2.1, the argument for $\succcurlyeq_C$ is that the purpose of choice is conformable with the purpose of normative economic analysis: in each case, the objective of the judgment is to promote the well-being of those affected by the selection and implementation of an alternative. In contrast, the purpose of any judgment underlying SRWB is simply to answer a question. Granted, arriving at an answer is itself a choice, but it is a choice of words rather than of the particular alternative or outcome the words describe. An honest respondent aggregates over the dimensions of h according to her understanding of the words and phrases that comprise the SRWB elicitation question. In the best possible scenario, those words have a precise meaning – for example, they may instruct the subject to score experiences according to a particular function, $f(h)$ – in which case the analyst's choice of wording, rather than the subject's judgment, dictates the principles of aggregation. Using vague words and phrases such as "happiness" and "satisfaction" that do not precisely specify the function f only magnifies these concerns. If, in response to her own idiosyncratic experiential associations, consumer i has learned to equate the word "happiness" with the value $f_i^H(h)$ for some aggregator f_i^H, and the word "satisfaction" with the value $f_i^S(h)$ for some aggregator f_i^S, the analyst's choice of wording will continue to dictate the principles of aggregation,

[40] The notion that life consists of highly disaggregated subjective experiences has a long philosophical tradition: see, for example, Aristotle (2012, translation), Mill (2012, reprinted), and more recently Sen (1980-1981), who advocates a vector view of utility.

but in a more haphazard way.

Case 2: Aggregate experienced utility as a continuous hedonic sensation. If instead we assume that aggregate experienced utility exists as a continuous hedonic sensation, then the object of an SRWB question is to elicit it. Here we also encounter several conceptual challenges.

As discussed in Section 2.2.2, there are natural and important settings in which hedonic experience cannot logically include the aggregate welfare evaluation, V, for example because experience is distributed across time or states of nature. If the consumer does not hedonically experience V, it must reflect a judgment concerning experience. But then we are effectively back in Case 1. In principle, one can speak of eliciting the *true* momentary AU at each point in time and under each state of nature because, by definition, the individual can only have one aggregate hedonic experience at any given moment. Yet there is no single "true" version of overall welfare, V, to elicit: because different judgments can have different purposes (e.g., evaluating satisfaction versus evaluating happiness), the consumer can simultaneously subscribe to multiple judgments about the same state-and-time-contingent profiles of hedonic experiences. Thus, in attempting to justify a particular version of V, one cannot reference the experiential "truth;" rather, the criterion must stand on its own, as in Case 1.

There are two possible paths forward. One is to take the view that our objective is to elicit some particular V (apparently other than the objective function that rationalizes choice), the justification for which remains unclear. The other is to focus on measuring the stream of momentary hedonic sensations, h_t, as in the Experience Sampling Methods of Kahneman et al. (2004). A limitation of this second approach is that one must supplement it with some other criterion for aggregating sensations across time and states of nature; otherwise, one has no basis for comparing two momentary AU trajectories, $(h_0, h_1, ...)$ and $(\hat{h}_0, \hat{h}_1 ...)$, except in rare cases of dominance.

Regardless of which analytic path one chooses, elicitation raises a separate set of conceptual challenges. To measure AU, we have to ask a question about it. But the phrases that economists, psychologists, and philosophers use to describe normative ideals, such as "experienced utility," are terms of art. People construe natural language according to their own experiential associations, rather than the rigorous principles the analyst intends. As an illustration, consider the problem of eliciting momentary AU at time t rather than V, or vice versa, in a setting where people may have memory utility and anticipatory emotions. What phrasing would allow respondents to understand that we want them to account for certain types of feelings about the past and future, but not others?

Another elicitation issue concerns motivations. People may not feel obliged to answer questions about well-being truthfully, or based on careful introspection. Answers may have incidental consequences that provide respondents with incentives to misreport their true feelings. For example, some responses may speak well of the subject's character, others poorly.[41] Also, because deliberation is costly, people may give SRWB questions only cursory consideration, particularly if they are averse to contemplating negative sensations. Even a preference for honesty cannot resolve these issues if

[41] See, for example, List et al. (2004). Another possibility is that I may have an incentive to exaggerate my preferences if I think the resulting SWRB analysis will be politically impactful; see Frey and Stutzer (2007).

respondents talk themselves into believing answers that sustain self-serving personal narratives, or if they truthfully report superficial judgments.

Further challenges arise from the fact that we always measure SRWB on a unitless scale. As a result, respondents have to decide what the numbers mean, and their interpretations may vary with context. For example, the respondent might treat 4 as "typical" because it is in the middle of the 1-to-7 range, and then renormalize the scale subsequent to an event that changes what is typical. Celebrated results in the literature concerning hedonic adaptation, such as the Easterlin paradox, may be attributable to confounding changes in normalizations.[42] Bernheim (2009) argues that there is no objective way to distinguish between changes in underlying well-being and changes in the way people interpret the scale – in other words, that these two effects are not independently recoverable, in the sense that we cannot identify their separate effects even with ideal data.[43]

In defense of SRWB as an implementation of mental statism, one could argue that the appropriate standard for evaluating a welfare measure is not whether it is perfect, but rather whether it reasonably approximates a consumer's well-being. Some economists find this defense convincing because they believe that, as a practical matter, answers to questions about states of mind such as happiness and satisfaction must correlate with any reasonable notion of true welfare. Others find this defense problematic for at least two reasons.[44] First, even if our objective is approximation, we are still obliged to identify the ideal we seek to approximate, and to explain why it provides an appropriate standard. Thus the preceding discussion continues to apply. Second, since "true" (as opposed to reported) AU is unobservable, there is no way to validate the elicitation process and gauge the accuracy of the approximation.[45] Without the possibility of validation, debates about normative methodology inevitably devolve into unprovable assertions. From the perspective of a skeptic, a justification for a welfare measure that relies on its relationship to some unknowable "underlying truth" is no justification at all; if the pertinent truth is not knowable, the measure must stand on its own, exactly as in Case 1.

A point that potentially favors SRWB over choice-oriented methods is that, in some contexts, it may more easily accommodate the possibility that consumers' concerns include conditions of

[42] See Easterlin (1974), or Stevenson and Wolfers (2008) for some contrary evidence.

[43] While the SRWB literature acknowledges the possibility that changes in the interpretation of the well-being scale may confound comparisons, those commentaries usually do not address the question of recoverability; see, for example, the discussion of scaling in Dolan et al. (2011). There are exceptions such as Lacy et al. (2008), who claim to measure rescaling separately from effects on happiness. However, that study relies on supposedly intuitive assertions rather than rigorous accounts of identification, and close examination reveals that its conclusions hinge on unstated and potentially unprovable assumptions (in particular, that people use the same scale when rating their own experiences and others' hypothetical experiences).

[44] The same argument is more persuasive in the context of discussions of national accounts, where alternatives such as GDP are also intended as rough proxies, rather than as rigorous welfare measures. Here we are concerned instead with the conceptual foundations of microeconomic welfare analysis. The question is whether it is possible to provide rigorous foundations for a mental statist interpretation of SRWB.

[45] Some argue that correlations between self-reported well-being, biometric variables, and neural measurements corroborate the use of such objects as indicia of well-being (see, e.g., Larsen and Fredrickson, 1999). But that argument is circular: it demonstrates only that the variables in question have something in common, not that they individually or collectively embody true well-being. Nor does it tell us much about the accuracy of the purported approximation.

choice.[46] To illustrate, suppose Norma cares both about the item she chooses, x, and the set from which she chooses it, X. Without further restrictions, all we can infer from her choice is that she prefers (x, X) to (x', X) for all x' in X. This type of information does not allow us to determine whether she is better off with a policy that mandates x (thereby giving her $(x, \{x\})$), or one that mandates y (thereby giving her $(y, \{y\})$). The SRWB method potentially avoids this difficulty because – setting aside other concerns – it ostensibly allows us to gauge well-being under each type of mandate.

As with choice-oriented methods, social aggregation poses important challenges. The issues are largely similar. The common practice of reporting simple summary statistics such as average SRWB responses resolves these issues somewhat arbitrarily, and makes the implied welfare weights dependent on how different consumers happen to use the scale. To illustrate, suppose a consumer who initially rates her happiness as $r(h)$ when experiencing sensations h switches to reporting $\widetilde{r}(h) = 4 + \alpha(r(h) - 4)$, with $\alpha > 1$ (where 4 is the midpoint of the scale). By virtue of reinterpreting the unitless scale in this way, the consumer would effectively increase the weight she receives in social welfare analyses. For discussions of other aggregation issues, see Ng (1997); Nordhaus (2009); Frey and Stutzer (2007).

2.4 Flavors of paternalism and justifications for government intervention

Few people would argue that deference to consumers' judgments should be absolute. Obvious exceptions include the treatment of children and the cognitively impaired, who arguably lack the capacity required to understand the consequences of their actions. However, the scope of paternalistic policymaking is far broader in practice. Dworkin (1971) lists a wide range of examples, such as laws that require motorcyclists to wear safety helmets, forbid swimming at public beaches when no lifeguards are on duty, criminalize suicide, and preclude contracts for perpetual involuntary servitude. In each case, the primary rationale for these policies is arguably to protect the decision maker, rather than to limit harm to others.

In Section 2.2.3, we identified four classes of reasons for objecting to a consumer's choices, and thus potentially for intervening in their decisions (see equation (1)): (i) the consumer misunderstands the set of available action, (ii) she misunderstands the mapping from actions to outcomes, (iii) she fails to perform the "max" operator correctly, or (iv) she uses an inappropriate objective function.. Those who subscribe to welfare hedonism or to objective accounts of welfare can potentially justify paternalistic interventions based on (iv). However, preference theory limits us to (i), (ii), and (iii) (Dworkin 1971; New 1999); it is consonant with a weak form of paternalism that defers to the consumers' underlying objectives, but finds fault with their execution. Within a choice-oriented framework, a paternalistic planner can improve welfare by proscribing or compelling particular

[46] This advantage is not always present. Recall, for example, our discussion of DellaVigna et al. (2012), which examined the willingness to pay to avoid charitable solicitation. We observed in Section 2.2.4 that, if people feel socially obligated to hear out charitable fundraisers, social pressure may infect the metachoices that define the willingness to pay. A similar phenomenon could arise in the context of SRWB: people may feel a social obligation to report high well-being despite charitable solicitation.

actions whenever naturally occurring decision frames lie outside the welfare-relevant domain, the object being to achieve outcomes the consumers would choose in welfare-relevant frames. This consideration motivates the various corrective policies we consider in the subsequent three sections of the chapter.

Preference-theoretic approaches encounter conceptual difficulties in cases where the individual in question suffers from endemic characterization failure (as with children and the cognitively impaired), so that the welfare-relevant domain is either empty or too sparse to permit useful comparisons. While this problem may initially strike the reader as severe, it is important to remember that economists rarely observe rich choice data for any particular individual, and that we routinely impute vast portions of the choice correspondence from the behavior of people we deem similar according to statistical models. The current problem is no different. Thus the choice-oriented framework yields a disciplined recipe for implementing paternalism in cases with endemic characterization failure: fill out the sparsely populated welfare-relevant domain based on the choices of consumers who avoid characterization failure but resemble the individual of interest in all other respects. This approach to paternalism represents an application of the "rational consumer benchmark" method discussed in Section 2.2.4; see also Section 3.2.3.

A more recent strand of literature explores the notion of *libertarian paternalism* (Thaler and Sunstein, 2003). It focuses on a class of policies known as *nudges*, defined as non-coercive changes in "choice architectures" that minimally impact opportunities, but nevertheless incline people toward "good" decisions. Such policies are arguably libertarian in the sense that choice is left to the individual, but they are paternalistic in the sense that the government intervenes with the objective of improving outcomes, on the grounds that people have cognitive limitations and suffer from biases. Just as with paternalism, each account of welfare offers a different route to rationalizations of this perspective. For example, within choice-oriented (preference-theoretic) framework, the planner can improve welfare by modifying the framing of a decision problem so that it falls within, rather than outside, the welfare-relevant domain. Because changes in framing do not alter opportunities, they are interpretable as nudges. We return to the topic of nudges in Section 3.5.

2.5 Other perspectives on social objectives

It is important to acknowledge that normative analysis is not limited to welfarist perspectives. Alternatives arise for both practical and conceptual reasons.

As a practical matter, economists are rarely given carte blanche to design policies from the ground up with the objective of promoting consumers' interests. More commonly, we respond to specific directives from policy makers. For example, a government official or agency may adopt the normative view that more saving is better, and ask economists to devise strategies for increasing rates of saving at the smallest possible cost to the government. Directives can reflect carefully thought-out welfarist objectives, or they may be simple-minded proxies. Alternatively, they may reflect the personal objectives of the pertinent officials, such as maximizing the odds of reelection. Usually, one can reformulate such directives as formal problems that are amenable to economic analysis.

A variation of this theme is present in the optimal tax literature. Consider the problem of setting income tax rates to optimally balance redistribution from rich to poor against the costs of discouraging labor supply. What weight should an economist attach to the redistributive motive? A common approach is to formulate the problem in terms of parameters measuring the marginal social benefits of increasing each individual's income (Saez and Stantcheva, 2016). Implicitly, these welfare weights reflect the preferences of the policymaker.

There are also conceptual alternatives to welfarism. For instance, some have argued that policy evaluation should focus on opportunities rather than outcomes (Sen, 1992; Arrow, 1995; Roemer, 1998; Sugden, 2004), while others emphasize the importance of process (Frey et al., 2004). To date, these perspectives have gained relatively little traction within behavioral public economics.

3 Policies targeting commodities

Our focus now shifts from the general principles of welfare analysis to specific classes of applications. The role of taxes or subsidies as a means of correcting consumer mistakes is one of the first questions explored in BPE.

The first wave of the literature has focused on particular biases in specific markets. The motivation was simple: conventional interpretations of standard behavioral models implied that consumers would not spend their money optimally due to decision making failures such as imperfect self-control. The research agenda was then to formulate a model with these features and examine its implications for, e.g., taxes on cigarettes or potato chips. This work highlighted a variation of the Pigouvian principle for externalities: the optimal tax should offset the average degree of over- or under-estimation of the marginal utility from the good in question.

Because the Pigouvian principle is not limited to any one particular bias, the next wave of papers provided richer analyses by deriving more general optimal tax formulas that envision a variety of biases. These papers fleshed out the modified Pigouvian principle in greater generality. We derive this principle in Section 3.2 for the simple framework introduced in Section 3.1. In addition to illuminating the forces behind the optimal commodity tax formulas, in Section 3.2 we also survey the empirical strategies that economists use to implement the formulas.

In practice, policymakers often worry that taxes on sin goods such as cigarettes or sugary drinks are regressive. The literature has therefore advanced beyond the simple Pigouvian principle by incorporating motives for redistribution. We discuss this work and derive some basic lessons in Section 3.3.

Alongside the literature on corrective commodity taxes, recent work has begun to explore the welfare implications of inattention to taxes that are not fully salient or misunderstood because, for example, they are not included in the posted prices of products. We survey the core theoretical principles, as well as the strategies for empirical implementation in Section 3.4.

Finally, we briefly discuss non-financial policy instruments, such as "nudges," in Section 3.5, and the implications of consumers' social preferences for commodity taxes in Section 3.6.

3.1 A stylized model of consumer choice

To organize our discussion, we study a simple model based on the general framework of Farhi and Gabaix (2015). We consider an economy in which consumers choose to divide their wealth between two goods, x and y. Firms produce x at a constant marginal cost c and sell it in a competitive market at a price p (which equals c in equilibrium), where it is also subject to a sales tax t. The second good, y, is the numeraire. We use $x_\theta(p,t)$ to denote a type-θ consumer's demand for x at a price p and tax t, and $D(p,t)$ to denote the total demand. The measure over types is $\mu(\theta)$.

We let V_θ denote the objective (or welfare) function that a type θ consumer "should" maximize. See Sections 2.2.2 and 2.2.3 for commentary on interpretations of this function. By positing the existence of a well-defined welfare function V_θ, we focus on settings in which the analyst entertains no normative ambiguity.

For simplicity, we assume there are two types of consumers, $\theta \in \{s, b\}$. Type s (for "standard") consumers always maximize V_θ. Type b (for "behavioral") consumers may follow a different behavioral rule owing to some cognitive bias. The following three biases have attracted particular attention within the literature on corrective commodity taxation:

1. *Limited attention* or *salience bias.* Consumers may be inattentive to features of decision problems that are insufficiently salient. In certain contexts, sales taxes and energy costs may fall into this category. Consumers may also ignore health costs that accrue slowly over the course of time. See, for example, Gabaix and Laibson (2006); DellaVigna (2009); Gabaix (2014); Bordalo et al. (2013); Koszegi and Szeidl (2013); Koszegi and Szeidl (2013).

2. *Incorrect beliefs.* Consumers may have incorrect beliefs about certain attributes of a good, such as its calorie content, its future health implications, or its energy efficiency. See, for example, Allcott (2013); Attari et al. (2010); Bollinger et al. (2011); Bordalo et al. (2013).

3. *Imperfect self-control.* Consumers who place excessive weight on immediate gratification will tend to overconsume goods with immediate benefits and delayed costs, and underconsume goods with immediate costs and delayed benefits. Delayed consequences can be particularly important for activities with implications for health; see, for example, Gruber and Kőszegi 2001, 2004; Bernheim and Rangel 2004. We discuss theory, evidence, and normative issues pertaining to self-control in Section 6.

For simplicity, we assume utility is quasilinear: $V_\theta(x_\theta(p,t), y) = y + v_\theta(x_\theta(p,t))$. The budget constraint requires $y = z_\theta - (p+t)x_\theta(p,t)$, where z_θ is the initial endowment of type θ, which we assume is large enough such that $x_\theta(p,t) < z_\theta$ in the range of taxes we consider. Accordingly, we can write utility as $\widetilde{V}_\theta(x_\theta(p,t)) = z_\theta - (p+t)x_\theta(p,t) + v_\theta(x_\theta(p,t))$. This formulation allows for heterogeneity with respect to behavior (x_θ), normative objectives (V_θ), and income (z_θ), but imposes no restrictions on the manner in which these characteristics are related.

3.2 Corrective taxation for behavioral consumers: Basic implications

3.2.1 Basic theory

The policymaker sets t to maximize aggregate welfare $W(t) = \sum_\theta \mu(\theta)\tilde{V}_\theta(x(p,t))$, accounting for the fact that all revenues are returned to consumers through lump-sum distributions. (Thus, $z_\theta = \bar{z}_\theta + T$, where $\bar{z}_\theta$ is the exogenous endowment and T is the lump-sum distribution.) Although here we have in mind commodity taxes addressing "internalities," rather than inattention to or misperceptions of the tax t itself as in Section 3.4, our derivations do not require x to depend only on the tax-inclusive price $p+t$.

A small increase in the tax, dt, has three effects:

1. It lowers consumers' utility by $D(p,t)dt$ through a direct wealth effect, but mechanically increases revenue, and hence lump-sum distributions, by Ddt. With quasilinear utility and no differences in the marginal social value of a dollar across potential recipients, these two changes cancel out. However, once we relax quasilinearity, differences in the distributions of revenue collections and lump-sum transfers will affect aggregate welfare.

2. Consumers substitute away from good x, causing tax revenue to fall by $tD_t(p,t)dt$.

3. Behavioral consumers alter their purchases, causing their utility to change by $\mu(b)\left(v_b'(x_b) - p - t\right)\frac{d}{dt}x_b(p,t)dt$.

The third effect is not present for standard consumers. This conclusion follows from the Envelope Theorem: because x_s maximizes V_s, it satisfies the first-order condition $v_s'(x_s) = p+t$, which means $\mu(s)\left(v_s'(x_s) - p - t\right)\frac{d}{dt}x_s(p,t) = 0$. The presence of behavioral consumers introduces the term $\gamma_b(p,t) := p + t - v_b'(x_b(p,t))$, henceforth called the *price-metric measure of bias*, into optimal tax formulas.

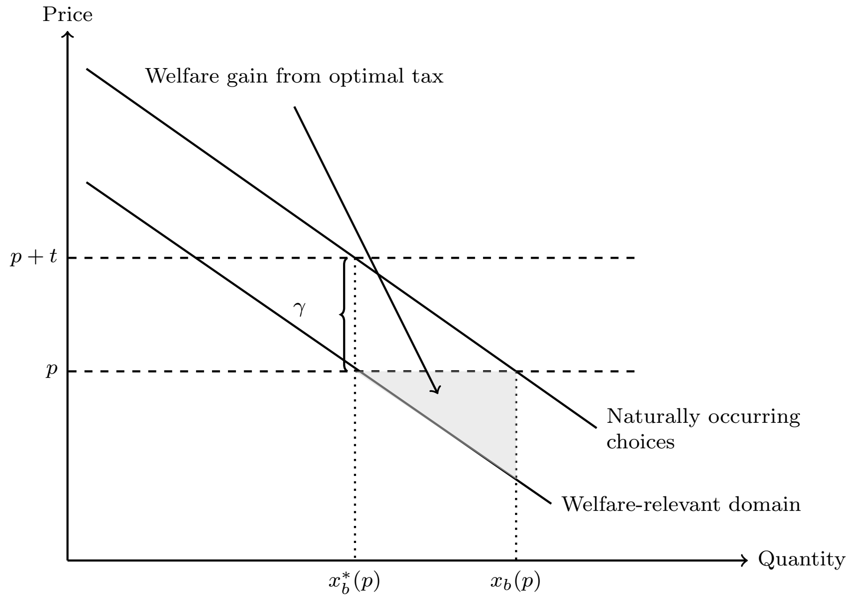
The term $\gamma_b(p,t)$ has a concrete empirical interpretation. Let $x_b^*(p,t)$ be the demand relation implied by maximization of V_b. Using the associated first-order condition, it is straightforward to verify that $x_b^*(p - \gamma_b(p,t), t) = x_b(p,t)$. Intuitively, $\gamma_b(p,t)$ "prices out the bias": it is the amount by which the price of x would have to fall to bring optimal purchases in line with actual purchases at consumer price of $p+t$. As discussed later, a number of studies including Chetty et al. (2009), Allcott and Taubinsky (2015), Taubinsky and Rees-Jones (forthcoming), and Allcott et al. (2018b) have used this price-metric approach.

The welfare effects of a change in the tax rate depend critically on the price-metric measure of bias. Putting the three effects together, we find:

$$\begin{aligned} W'(t) &= tD_t(p,t) - \gamma_b(p,t)\mu(b)\frac{d}{dt}x_b(p,t) \\ &= (t - \bar{\gamma}(p,t))D_t(p,t) \end{aligned} \quad (2)$$

where $\bar{\gamma}(p,t) = \frac{\gamma_b(p,t)\mu(b)\frac{d}{dt}x_b(p,t)}{D_t(p,t)}$ is the *average marginal bias;* i.e., it is the average degree to which consumers over- or under-estimate the net benefits of the marginal purchases stimulated by a change in the tax rate t, weighted by their demand responses. This statistic is the critical determinant of

Figure 1: Graphical illustration of the optimal tax rule



the optimal tax rate because biases only matter insofar as people with those biases adjust their consumption in response to variations in the tax-inclusive price.

Because $W'(t) = 0$ at the optimum, equation (2) immediately yields a simple formula for the optimal commodity tax:

$$t^* = \bar{\gamma}(p,t) \tag{3}$$

The parallel to Pigouvian taxation is clear: the planner sets the commodity tax to offset the "marginal internality" (i.e., the average wedge induced by consumers' cognitive biases), instead of the marginal externality.

Figure 1 illustrates the optimal tax rule for the case of homogeneous consumers under the additional simplifying restriction that we can write $x_b(p,t)$ and $x_b^*(p,t)$ as $x_b(p+t)$ and $x_b^*(p+t)$, respectively. The figure plots the naturally-occurring demand curve and the welfare-relevant demand curve. At market prices, individuals overconsume the good by $x_b(p) - x_b^*(p)$, because they perceive its marginal utility to be γ higher than it actually is (for simplicity, γ does not vary with p in the figure). A tax equal to γ decreases the quantity consumed from $x_b(p)$ to the optimal $x_b^*(p)$, because $x_b(p+\gamma) = x_b^*(p)$. The welfare-gain from the optimal tax is given by the shaded triangle below the market price p and above the demand curve $x_b^*(p)$.

3.2.2 Applications

The literature contains variations of formula (3) that are specific to particular combinations of products and cognitive biases. Some examples follow.

Unhealthy foods: O'Donoghue and Rabin (2006) study the taxation of unhealthy foods such as potato chips under the assumption that present-focused tendencies render consumers time-inconsistent. They adopt the normative perspective that present focus constitutes a bias. It follows that people overconsume unhealthy foods because they place too little weight on future health costs. Based on a variant of (3), they conclude that, as long as the operative cognitive biases lead to overconsumption and behavioral consumers are at least somewhat price-sensitive, the optimal tax is positive.[47]

Smoking: Gruber and Kőszegi (2001) study the taxation of cigarettes using the same model of time inconsistency as O'Donoghue and Rabin (2006), and adopt the same normative perspective. Their model is at once simpler because they consider a homogeneous population with no standard consumers, and more complicated because they account for the types of intertemporal complementarities in preferences commonly associated with addictive tendencies. Their analysis demonstrates that the main insights from (3) carry over to settings with this type of dynamic preference structure, provided one adjusts the definition of the marginal bias γ to account for the effect of current consumption on future overconsumption.

Energy-using durables: Allcott and Taubinsky (2015) and Allcott et al. (2014) study the subsidization of energy-using durable goods under the assumption that consumers underweight future energy costs, and derive variants of (2) and (3). Allcott et al. (2014) also provide extensions to settings with multiple policy instruments (both taxes on energy consumption and subsidies for energy-efficient products), as well as externalities. A key result in Allcott et al. (2014) is that, in the presence of heterogeneous behavioral biases and externalities, the optimal policy mix involves a subsidy for the energy efficient durable good and a tax on energy that is less than the marginal externality of energy use. The intuition builds on the observations that standard consumers will over-purchase the subsidized energy-efficient durable good, and will have larger reactions to the energy tax (which behavioral consumers discount because it is in the future). Making the tax lower than the marginal externality for energy use is welfare-enhancing because it offsets the subsidy's distortionary effect on standard consumers while affecting behavioral consumers to a lesser degree. See Farhi and Gabaix (2015) for a related and more general analysis of the violation of the "principle of targeting." See also Heutel (2015) and Tsvetanov and Segerson (2013) for other applications to energy-using durables.

General frameworks. Mullainathan et al. (2012) provide a general treatment of commodity taxation similar to the one presented here, with the exception that consumers have unit demand for the good in question. Farhi and Gabaix (2015) examine a more general framework that encompasses continuous demand for multiple products with arbitrary patterns of complementarities, substitutabilities, and biases.

[47] See the Appendix for a discussion of the pertinent behavioral theory, and Section 2.2.5 for a critical discussion of the normative standard.

3.2.3 Empirical measurement and implementation

The naturally occurring and ideal demand functions for behavioral consumers, $x_b(p,t)$ and $x_b^*(p,t)$, are the key inputs for the optimal commodity tax formula.[48] Applications require empirical estimates of these functions. The studies discussed in this section undertake empirical applications using the approaches discussed in Section 2.2.4, and we reference the three core tasks discussed therein throughout.

Calibrating or estimating models with "bias" parameters. Gruber and Kőszegi (2001) and O'Donoghue and Rabin (2006) derive optimal tax formulas for the case of present-biased consumers, and use evidence from behavioral economics and public health to calibrate the parameters of structural models of choice. If h represents the future health costs of smoking or eating unhealthy food (calibrated from public health studies), and if consumers improperly discount those costs by the factor β (calibrated from estimation in other domains), then the magnitude of the bias is $\gamma = (1-\beta)h$.[49]

How do these studies address the three tasks set forth in Section 2.2.4? Both assume implicitly that the welfare-relevant domain consists of decisions with no immediate consequences, so that present focus (β) exerts no influence on ideal behavior. Choices in those frames are not actually observed. Consequently, task 1 is accomplished by inferring x_b^* from x_b based on a structural model of preferences, with the key parameters, β and h, identified from decisions in other domains. Tasks 2 and 3 receive less attention and are implicitly addressed through assumptions. These assumptions raise both conceptual and empirical issues; see Section 2.2.5 for a discussion of time inconsistency and the welfare-relevant domain (task 2), and the Appendix for comments on competing models of self-control, which make different assumptions about consumers' concerns (task 3).

A limitation of the aforementioned studies is that, by focusing on a particular model of bias, they assume away other plausible biases. For example, in addition to being present-focused, consumers may also hold incorrect beliefs about the health consequences of unhealthy foods or addictive substances. Or consumers may over-indulge in addictive substances because they underestimate how addictive those substances actually are. As noted in 2.2.6, focusing on biases one at a time ignores second-best issues arising from the potential existence of multiple biases.

As these applications illustrate, an advantage of using a parametrized structural model of behavior and bias is that the problem of recovering the choice mapping becomes tractable even when the analyst cannot directly observe choices in the welfare-relevant domain.

Belief elicitation. Allcott (2011a; 2013) and Rees-Jones and Taubinsky (2018a) study the welfare effects of biased beliefs concerning energy costs and income taxes, respectively. As we explained in Section 2.2.3, the welfare-relevant domain for any setting with purportedly biased beliefs (implicitly) consists of choice problems in which simple and transparent framing of pertinent information ensures proper comprehension of the consequences following from each potential action. In an ideal implementation, the analyst would perform task 1 by observing the naturally occurring and

[48] Within the Bernheim-Rangel framework, one views both of these objects as manifestations of a more general demand function, $\widehat{x}_b(p,t,f)$, where f is the decision frame.

[49] This formulation of bias also plays a crucial role in the types of contracts offered by profit-maximizing firms to present-biased consumers (DellaVigna and Malmendier, 2004).

welfare-relevant choices directly. Task 2 requires corroborating evidence that consumers misunderstand consequences (e.g., that they ignore, misinterpret, or misuse information pertaining to likelihoods) in the naturally occurring problems, but not within the welfare-relevant domain. Task 3 raises standard issues, but the scope of consumers' concerns is usually not controversial in these applications.

However, the ideal implementation is rarely feasible, because the hypothesized welfare-relevant choices are typically artificial and generally difficult to implement. One must therefore find another strategy for recovering the full choice mapping (task 1). The practical alternative used in the cited studies is to formulate a structural model relating choices to beliefs, attempt to measure those beliefs as directly as possible, and then substitute objective probability for subjective beliefs to extrapolate choices within the welfare-relevant domain.[50]

The most direct route to justifying the belief-elicitation approach, and the one most applied economists likely have in mind (at least implicitly), is to treat the expected utility model (or some variant thereof) as a *literal* depiction of cognitive processes. In other words, upon writing the consumer's objective function as $\sum_{i=1}^{n} \pi_i u(x_i)$, we assume that $u(x_i)$ and π_i actually exist within the decision maker's mind, the former representing her actual hedonic evaluation of the outcome x_i, the latter representing an actual subjective belief that takes the form of a mathematical probability, and that – at least to an approximation – the cognitive process maximizes the expectation of the hedonic reward. We are then free to ask decision makers about the values $(\pi_1, ..., \pi_n)$, or to elicit these parameters in some other manner, and to infer unbiased choices by substituting objective probabilities for the subjective values.

Many economists prefer an interpretation of the theory under which the model of decision making is an "as-if" representation, rather than a literal depiction of cognition. This alternative view has many attractions, including its ability to accommodate the realistic possibility that people actually act on qualitative assessments of likelihoods rather than quantitative notions of subjective probabilities. However, once one adopts this perspective, a conceptual gap potentially opens up between the elements of the theory and their ostensible empirical counterparts. Moreover, the foundation for inferring "unbiased" choices by replacing the as-if "subjective probability" parameters with objective probabilities becomes murky. The belief-elicitation approach may or may not be valid under an as-if interpretation of the theory; in any given case, the question is amendable to empirical investigation, and merits closer attention. See Bernheim (2018) for further discussion.

One important limitation of the belief-elicitation strategy is that it cannot accommodate settings in which objective probabilities are either unknown or controversial. Some have argued that realistic economic settings rarely admit objective probability assessments; see, for example, Kurz (1994) on the diversity of rational beliefs.

Another important limitation of this strategy is that it assumes one particular bias—here, incorrect beliefs—while ignoring all others. For example, if we derive $x_b^*(p,t)$ by adjusting $x_b(p,t)$

[50] For the purpose of this discussion, one can think of a belief as a probabilistic assessment. Elicitation methods that induce people to state beliefs as point estimates rather than probability distributions are problematic for these purposes, because they abstract from subjective uncertainty, which may well affect behavior.

to account for false beliefs in a setting where x_b also reflects present focus and inattention, the normative standard will likely be misleading. See again the general discussion of second-best issues in Section 2.2.6.

An additional challenge encountered when taking this approach relates to equation (2): ultimately, what matters are the mistaken beliefs of individuals who respond on the margin to the policy in question, and not those of the overall population. Unfortunately, surveys that elicit beliefs generally aim to do so for the latter and not the former. In their study of misperceptions concerning income taxation, Rees-Jones and Taubinsky (2018a) address this issue by performing robustness checks using the elicited beliefs of population subgroups that are more responsive to changes in tax rates, such as those in the labor force and the self-employed.

Comparing analogous demand responses. Allcott and Wozny (2014) and Busse et al. (2013) reason that consumer demand for vehicles should be equally responsive to the present value of gasoline costs and up-front prices. Upon finding that the sensitivity to gasoline costs is in fact much lower, they attribute the differential to biases affecting the evaluation of future costs, such as inattention.

In such settings, the definition of the welfare-relevant domain depends on the nature of the assumed bias. For the case of vehicle demand, it presumably consists of settings in which conditions putatively favor attentiveness to all components of cost. These studies accomplish task 1 by inferring $x_b^*(p,t)$ from $x_b(p,t)$ based on the observed responses to changes in vehicle prices and the net present value of gasoline costs. In effect, they fill out the choice mapping by imposing a weak structural assumption. This extrapolation hinges on a crucial statistic that Allcott and Wozny (2014) do not measure directly: the appropriate discount rate.[51] Accordingly, they present estimates for a range of discount rates between 0% and 15%.

To map the demand response estimates to the price-metric measure of bias, γ, using the Allcott and Wozny (2014) procedure, it is also necessary to assume that elasticities to salient costs are uncorrelated with the magnitude of the bias, and that the costs do not influence the bias. For example, if consumers are more attentive to gasoline costs when gasoline is more expensive, the ratio of the two demand responses would generate only a lower bound on the bias (Allcott et al., 2014). Knowing only the slopes of the demand curves $x_b(p,t)$ and $x_b^*(p,t)$ at the market price does not permit imputation ofγ_b; generally, one must measure these demand curves more comprehensively, much as in the next two strategies described below.

An evidence-based approach to task 2 does not appear in the aforementioned papers, but would require a demonstration that the difference in demand sensitivities is in fact traceable to selective attention or biased beliefs, rather than to some other contextual reaction such as exaggerated "sticker shock." The latter hypothesis could have diametrically opposed implications for the welfare-relevant domain.

Task 3 raises standard issues, but the scope of consumers' concerns is usually not controversial

[51] And the procedure assumes that this discount rate is homogeneous across consumers. But to the extent that uncertainty and liquidity constraints vary, the discount rate would as well.

in these applications.

Rational consumer benchmarks. Bronnenberg et al. (2015) show that doctors and pharmacists are less likely to choose branded drugs over generic alternatives that are cheaper and chemically equivalent. This finding suggests that imperfect information distorts the purchases of other consumers toward branded drugs. Here, the welfare-relevant domain presumably consists of settings in which the typical consumer putatively receives and correctly processes the same information as doctors and pharmacists.

For task 1, one extrapolates demand from the observed choices of the "unbiased" consumers. In practice, this extrapolation does not involve a simple comparison between the expert and non-expert consumers, as they may differ with respect to demographic characteristics that are correlated with tastes, or they may shop at different stores and thus see different presentations of the items. Bronnenberg et al. (2015) adjust for differences in the observable characteristics of consumers and the stores they frequent. Of course, analysts cannot control for *unobservable* taste differences between professions. An advantage of strategies that reframe decisions, discussed below, is that they avoid this potential confound by, in effect, experimentally *inducing* expertise.

For task 2, Bronnenberg et al. (2015) support their assumption concerning the welfare-relevant domain by showing that the doctors and pharmacists are indeed much more knowledgeable about their purchases than others. Their strategy is to survey a subset of consumers in their retail dataset, asking them to name the active ingredient in various national-brand headache remedies. They find that pharmacists, physicians, and surgeons answer 90 percent of these questions correctly, compared with only 59 percent for the general population. In principle, expertise might go hand in hand with other biases; for example, medical students are known to suffer from excessive anxiety concerning the conditions they study. However, Bronnenberg et al. (2015) also demonstrate that the knowledge gap accounts for most of the differences in the purchasing behavior of experts and non-experts.

Task 3 raises standard issues, but the scope of consumers' concerns is usually not controversial in these applications.

In another application, Allcott et al. (2018b) compute the optimal tax on sugar-sweetened beverages allowing for the possibility that consumers may suffer from both misinformation and imperfect self-control. They measure misinformation using the General Nutrition Knowledge questionnaire, and they measure domain-specific self-control using a combination of assessments (by respondents and their spouses) of the extent to which respondents consume sugary drinks more than they should. In effect, consumers who display high nutritional knowledge and claim (with their spouse's agreement) that they do not overconsume sugary drinks provide the rational consumer benchmark for this study.[52] The empirical methods used in this study involve several other notable features. First, the study addresses potential confounds associated with unobservable taste differences by exploiting survey questions that directly elicit the degree to which respondents like va-

[52] Implicitly, this benchmark assumes that, if a consumer who struggles with self-control were able to commit to decisions in advance, he would make the same choices as a consumer who does not struggle with self-control. It also assumes that the welfare-relevant domain consists of these advance commitments – in other words, it adopts the long-run criterion.

rious sugary drinks and the importance they attach to health. Second, it explicitly accounts for the possibility that the rational consumer benchmark yields noisy proxies for the decisions consumers would make within the welfare-relevant domain. Third, it directly quantifies the money-metric bias γ for each consumer by combining an estimate of overconsumption with an estimate of the price-elasticity of demand for sugar-sweetened beverages. To arrive to the money-metric measure, it utilizes a log-linearization of the demand function: $\ln x \approx \ln x^* + \zeta^c \gamma_b / p$, where ζ^c is the compensated elasticity and p is the market price. As an example of this approach, imagine that bias increases quantity demanded by 30%, and that the compensated demand elasticity is 1.5. Then the impact of bias is the same as a 20% price reduction: $\gamma_b = p \cdot 30\%/1.5 = 0.2p$.

Reframed decisions. Allcott and Taubinsky (2015) examine purchases of more vs. less energy-efficient lightbulbs by consumers who are potentially inattentive to, or misinformed about, the (relative) energy costs of the lightbulbs. They conduct a within-subject experiment that consists of three steps. First, they elicit consumers' initial willingness to pay for the lightbulbs. Second, they treat a subset of consumers with an intervention that "teaches" consumers about the total costs of the lightbulbs and helps them learn this information through a series of quiz questions. The control group receives statistical information that does not shed light on the relative value of the different lightbulbs. Third, they elicit willingness to pay for the lightbulbs a second time.

The foundational assumption behind this strategy is that the welfare-relevant domain consists of choices made after the informational treatment. Task 1 follows from a simple difference-in-difference comparison of the pre- versus post-willingness to pay between the treated and untreated consumers. The main challenge here lies in task 2: how does one demonstrate that that inconsistencies between the original and reframed choices are attributable to characterization failures in the former, and not in the latter? One potential confound for the reframing strategy is that the effects could be at least partially attributable to browbeating, social pressure, and/or the induction of guilt. Allcott and Taubinsky (2015) address this issue in three ways. First, they show that a measure of susceptibility to social pressure is not correlated with the treatment effect. Second, they obtain similar results based on cross-subject comparisons when the initial valuation round is eliminated. This result addresses the hypothesis that subjects might feel pressure to *change* their decisions. Third, they demonstrate that their results continue to hold when they add information to the main treatment that arguably obscures the experimenter's intent by highlighting negative aspects of energy-efficient bulbs (specifically, the fact that they take longer to warm up and contain mercury). This third strategy assumes that these negative features are important to consumers (otherwise their inferences about the experimenter's objectives would be unaffected).[53]

Another potential concern regarding task 2 is that some consumers may ignore or discount the informational treatment, in which case characterization failure will continue to infect some portion of the putative welfare-relevant domain. Requiring consumers to correctly complete a quiz guards

[53] Note that this treatment variation could also depress choice if these features were not already known. The fact that this does not occur thus additionally implies that consumers are familiar with features such as warm-up time. This conclusion is consistent with a theory of learning in which warm-up times are easily observable and memorable experiences, whereas the impact of various appliances on the total energy bill are difficult to infer and recall.

against this possibility to some degree, but does not ensure that subjects believe what they learn. Indeed, Allcott and Taubinsky (2015) find that some treated consumers do not have correct beliefs about the energy cost savings of efficient lightbulbs after the completion of the experiment. However, focusing more narrowly on consumers who do express correct beliefs (in effect, a refinement of the welfare-relevant domain), they find that the impact of the treatment on the willingness to pay is 30% larger. This finding provides the basis for alternative welfare estimates.

Task 3 raises standard issues, but the scope of consumers' concerns is usually not controversial in these applications.

Advantages and disadvantages of the approaches: An advantage of the last three empirical approaches we have discussed is that they do not require one to take a stand on a precise model of cognition. For example, when studying analogous demand responses, one does not need to know whether the differences between the responses of the two groups are attributable to inattention, incorrect beliefs, or present focus, provided one can justify the assumption that the responses ought to be identical. Similar comments apply to strategies involving rational consumer benchmarks and reframed choices.

The aforementioned approaches are, however, neither assumption-free nor psychology-free, as our critiques of particular applications highlight. Allcott and Taubinsky (2015), for example, take the stand that the relevant psychological mechanism involves inattention or incorrect beliefs. They note that their reframing intervention would not necessarily eliminate biases that might arise in a Koszegi and Szeidl (2013) model of focusing.

An additional advantage of the last two empirical approaches discussed above is that they are more direct. For example, even if the researcher has a very specific model and normative criterion in mind, such as quasi-hyperbolic discounting coupled with the long-run criterion, they permit *direct* recovery of the key empirical objects, $x_b(p,t)$ and $x_b^*(p,t)$; there is no need to infer those objects from a structural model based on estimated parameters. Either by analyzing consumers who demonstrably do not suffer from self-control problems (the fourth approach), or by asking consumers to make decisions with no immediate consequences (the fifth approach), the analyst can elicit the welfare-relevant demand curve directly. A comparison between the welfare-relevant and naturally occurring demand curves reveals the policy-relevant statistic γ_b, without the separate need to measure the present-focus parameter for the relevant consumption dimension, such as the the marginal health costs. See, e.g., the Allcott et al. (2018b) application of the rational consumer benchmark method to the case of over-consumption of sugar-sweetened beverages, discussed above.

Of course, direct measurement of the welfare-relevant demand function is not always possible, in which case stronger structural assumptions are needed to identify $x_b^*(p,t)$ from naturally occurring choices. When it is clear that a direct approach is infeasible, structural methods can be fruitful, provided the analyst clearly spells out and justifies the necessary assumptions. However, one can needlessly sacrifice robustness and generality by jumping *directly* to tightly parametrized psychological models, rather than focusing on recovering the key empirical objects of interest, $x_b(p,t)$ and $x_b^*(p,t)$, through the method that requires the least restrictive assumptions.

3.3 Distributional concerns

Section 3.2 focused exclusively on a behavioral "Pigouvian" principle, which holds that the object of taxes and subsidies is to correct "internalities," and thereby bring actual demand in line with "optimal" demand. In practice, taxes and subsidies also redistribute resources.

Concerns about redistribution include the common complaint that sin taxes are regressive. The poor consume disproportionate quantities of cigarettes and sugary drinks (see Gruber and Kőszegi (2004), Goldin and Homonoff (2013), and Lockwood and Taubinsky, 2017; Allcott et al., 2018b), while the rich benefit disproportionately from subsidies for energy efficiency (see Allcott et al., 2015, Davis and Borenstein, 2016, Davis and Knittel, 2016). These regressive patterns have fostered forceful opposition to "sin taxes" and "virtue subsidies" on the grounds of equity and fairness.

In settings with uncertainty, redistribution can either occur ex ante across individuals, or ex post across realizations for the same individual. The mathematics of these two settings are essentially identical, except that in the second case the "social welfare function" corresponds to the individual's ex ante preferences over outcomes in the various states of nature.

We now generalize the basic ideas of Section 3.2 to incorporate concerns about redistribution. Our setting is a stylized version of the Diamond (1975) generalization of the Ramsey model, which allows for heterogeneous consumers varying in their marginal utility of wealth.

3.3.1 Basic theory

Here we consider the same model as in Section 3.2, except we assume that $V_\theta = G(y + v_\theta(x_\theta(p,t)))$, where G is a concave and differentiable function. Notice that the introduction of G does not change the first-order condition that characterizes the demand function, x_θ. Let $g_\theta(t) = G'\left(z_\theta - (p+t)x_\theta(p,t) + v_\theta(x_\theta(p,t)\right.$ This term denotes the marginal utility of wealth for a type θ consumer, normalized by the value of public funds, $\lambda := \frac{dW}{dT} = \sum \mu(\theta) G'\left(\bar{z}_\theta + T - (p+t)x_\theta(p,t) + v_\theta(x_\theta(p,t))\right)$. By construction, $E[g_\theta(t)] = 1$.

A small increase in the commodity tax rate, dt, has the following four effects:

1. A direct effect on consumer welfare, $-\mu(s)x_s(p,t)g_s(t)dt - \mu(b)x_b(p,t)g_b(t)dt$
2. A direct effect on public funds, $D(p,t)dt$
3. An indirect effect on public funds, $tD_t(p,t)dt$
4. An indirect effect on consumer welfare, $-\mu(b)g_b\gamma_b\frac{d}{dt}x_b(p,t)dt$, where $\gamma_b = p + t - v_b'(x_b(p,t))$, as before.

Putting these effects together, we find that

$$\begin{aligned} W'(t)dt/\lambda &= -E[x_\theta^*(p,t)g_\theta(t)dt] + D(p,t)dt + tD_t(p,t)dt - \mu(b)g_b(t)\gamma_b\frac{d}{dt}x_b^*(p,t)dt \\ &= \underbrace{(t - \bar{\gamma}g_b(t))D_t(p,t)}_{\text{corrective benefits}}dt - \underbrace{Cov[x_\theta(p,t), g_\theta(t)]}_{\text{regressivity costs}}dt \end{aligned} \qquad (4)$$

Because $W'(t) = 0$ at the optimum, equation (4) immediately yields a simple formula for the optimal commodity tax:

$$t^* = \bar{\gamma} g_b(t) + \frac{Cov[x_\theta(p,t), g_\theta(t)]}{D_t(p,t)} \tag{5}$$

Formulas (4) and (5) lead to a few insights. First, it is crucial to account for the manner in which the propensity to consume x covaries with marginal utility from income. When low-income consumers are more likely to purchase the taxed good, the tax is regressive, and hence the optimal rate is lower. Conversely, when high-income consumers are more likely to purchase the taxed good, the tax is progressive, and hence the optimal rate is higher.

The second and more subtle insight is that the corrective benefits of the commodity tax no longer simply equal $\bar{\gamma}$, the money metric measure of the average bias of marginal consumers. To illustrate, suppose everyone purchases the same amount of the good x, but the behavioral consumers have lower income, so $g_b(t) > g_s(t)$. Then the optimal tax is higher than the pure Pigouvian benchmark, $\bar{\gamma}$. The intuition is as follows (see Lockwood and Taubinsky, 2017): When consumption is the same for both types, the direct effects on consumer welfare and public funds cancel out. The two remaining effects are the same as in the model with no distributional concerns, except that the indirect effect on consumer welfare is multiplied by the term g_b. This change reflects the fact that a planner with redistributive motives is willing to pay more, for example, to eliminate a \$1 mistake made by the poor than by the rich.

More broadly, if we view the marginal welfare weights $g_\theta(t)$ as reflecting the policymaker's redistributive preferences, the formulas show that, as a general matter, one cannot translate empirical measurements of bias into optimal policy prescriptions without taking those preferences into account. The only exception arises in the case where $g_\theta(t) \equiv 1$, which is sensible only if we assume quasilinear utility.

Third, the relative importance of corrective versus redistributive motives in shaping the optimal commodity tax depends on how price-responsive consumers are. When they are not very price-responsive ($|D_t(p,t)|$ is small), redistributive motives dominate corrective motives. When consumers are very price responsive ($|D_t(p,t)|$ is large), corrective motives dominate redistributive motives. To obtain intuition for why consumers' response to the tax is crucial, imagine the extreme case in which consumers are completely inelastic. In this case, the regressive tax simply shifts funds from low-income consumers to high-income consumers, without correcting their behavior.

3.3.2 Applications and related literature

Bernheim and Rangel (2004) consider a dynamic model of addiction in which consumers randomly encounter environmental cues that trigger compulsive tendencies to consume the addictive good. They assume that consumption in the triggered state is completely inelastic to the tax. Although the good is enjoyable, sustained consumption impairs health, thereby reducing both earnings and baseline well-being.

The authors restrict the welfare-relevant domain to the state-contingent choices consumers would make in advance, prior to being cued. They argue that this restriction is justified because characterization failure infects choices made in the presence of substance-related environmental cues, a proposition that finds support in the literature on the neurobiology of addiction (see Section 2.2.5).

A central conclusion of the Bernheim and Rangel (2004) analysis is that the optimal tax on addictive goods is negative; in other words, they should be subsidized. Our simple optimal tax formula, equation 5, anticipates this result. The inelastic response of behavioral consumers implies $\bar{\gamma} = 0$, which means the tax offers no corrective benefits. As a result, the covariance between the marginal utility of income and the consumption of x, $Cov[x^*_\theta(p,t), g_\theta(t)]$, determines the sign of the tax. If consumption reduces income, the covariance is positive. Because D_t is negative,the optimal tax is negative. Although Bernheim and Rangel's dynamic model is more complicated, the simple two-good model captures the essential economic forces.

While Bernheim and Rangel (2004) focus on a case where distributional concerns generate a "sin subsidy," in other cases the optimal tax can still be large and positive even when it appears to be regressive. This result will obtain when the term $\bar{\gamma} g_b(t)$ is sufficiently large; that is, under the assumption that behavioral consumers have lower incomes, and (contrary to the Bernheim-Rangel premise) that they respond elastically to the tax even when expressing their behavioral biases. These conditions may hold for at least some sin goods. Gruber and Kőszegi (2004) use the Consumer Expenditure Survey (CEX) to show that the aggregate demand for cigarettes among low-income consumers responds elastically to cigarette taxes. By assuming away the possibility, featured in Bernheim and Rangel's analysis, that present-focus is a cue-triggered state, and that its activation also suppresses demand elasticities, they show through numerical simulations that cigarette taxes can make low-income consumers better off even without accounting for the benefits of the additional revenue, provided present bias is sufficiently severe. The intuition is most easily understood for the case in which the magnitude of the price elasticity is greater than one. In this case, a 1% increase in price decreases demand by more than 1%, thus consumers' total expenditures on the sin good fall, and so they spend more money on the other goods. At the same time, if consumers are sufficiently biased toward over-consuming the sin good, then exchanging some of the sin good for even a little bit of another good makes them better off. Gruber and Kőszegi (2004) thus argue that cigarette taxes may not be regressive according to a comprehensive welfare metric. We emphasize, however, that their argument rests on the assumption that present focus is always active, and consequently that the high demand elasticity they measure applies to biased decisions. While Gruber and Kőszegi (2004) do not consider optimal tax implications, Farhi and Gabaix (2015) apply their framework to a two-type Ramsey model that generalizes the insight about the importance of the demand elasticity of low income consumers.

Bernheim and Rangel (2004), Gruber and Kőszegi (2004), and Farhi and Gabaix (2015) all study environments in which commodity taxes are the only means for redistribution. It is arguably inappropriate, however, to set the tax rate for any given commodity based on distributive implications without considering the full range of redistributive instruments at the government's disposal.

Far from being an abstract or technical consideration, this issue surfaces in practical discussions of "sin taxes" under the guise of "revenue recycling" – the idea that the government can use sin tax revenues to fund progressive initiatives that benefit low-income consumers. For example, some cities in the U.S. earmarked the revenue from taxes on sugar-sweetened beverages for progressive policy initiatives such as universal pre-K education.

Allcott et al. (2018b) address these considerations by studying the simultaneous optimization of commodity taxes and nonlinear income taxes. Their analysis builds on Saez's (2002) extension of Atkinson and Stiglitz (1976), in that they model an economy consisting of behavioral consumers with heterogenous earning abilities and tastes who choose labor supply and a consumption bundle that exhausts their after-tax income. The optimal policy depends on the relative importance of income and preference heterogeneity in driving the consumption of sin goods. When all differences in sin good consumption stem from income effects, the planner addresses distributional considerations entirely through the income tax, and commodity taxes depend only on their corrective benefits. When elasticities and biases are non-decreasing with income, the optimal tax is unambiguously higher than the Pigouvian benchmark. However, when preference heterogeneity plays a larger role, progressive income taxation offsets the distributional effects of commodity taxation imperfectly, creating labor supply distortions that outweigh the redistributive benefits. In that case, the optimal commodity tax rates depend on distributional effects.

3.4 Efficiency costs of misperceived commodity taxes

3.4.1 Basic theory

We now turn our attention to settings in which consumers misperceive taxes. Unless consumers also suffer from some other bias, they correctly understand the prices they pay when taxes are absent. Consequently, there is no *corrective* role for commodity taxation. Here, our focus is on measuring the efficiency costs of commodity taxes in settings where the government raises revenue for other purposes, and does not necessarily optimize the use of tax instruments. With quasilinear utility, the efficiency cost of a tax is identical to its impact on the consumer welfare function we defined in Section 3.2.

We focus here on the implications of imperfectly salient commodity taxation: consumers react to the tax t as if it is σt, where σ is a decision weight that could potentially depend on the tax but varies smoothly with it. This modeling strategy encompasses a number of related psychological biases such as exogenous inattention to the tax, so that consumers always react to the tax as if it is a constant fraction σ of its size (DellaVigna, 2009; Gabaix and Laibson, 2006); endogenous inattention to the tax, or boundedly rational processing more broadly (Chetty et al., 2007; Gabaix, 2014); certain types of rounding heuristics; or simply forgetting (in which case $\sigma = 0$).

For simplicity, assume throughout this discussion that $V_b = V_s$; that is, the welfare function is the same for behavioral and rational consumers. We continue to assume quasilinearity.

The behavioral consumer's first-order condition is $v_b'(x_b(p,t)) = p + \sigma_b t$. Defining the bias term γ_b as before, we have $\gamma_b(p,t) := p + t - v_b'(x_b(p,t)) = (1 - \sigma_b)t$. Equation (2) continues to apply.

However, in this special case, x_b^* depends only on the perceived tax-inclusive price, so we have $x_b(p,t) = x_b^*(p+\sigma t)$.

In settings where all consumers are behavioral, formula (2) implies:[54]

$$\begin{aligned} W'(t) &= (t-\gamma_b)D_t(p,t) \\ &= (t-(1-\sigma_b)t)D_t(p,t) \\ &= \sigma_b t D_t(p,t) \end{aligned} \tag{6}$$

Formula (6) appears in Chetty et al. (2009). Its key implication is that underreaction reduces efficiency costs through two separate channels: first, it reduces D_t, the sensitivity of demand to the tax rate; second, it reduces the efficiency costs for any fixed value of $D_t(p,t)$ (through the multiplicative term σ_b). An economist who overlooks the consumer's misperception, but who nevertheless correctly measures the sensitivity of demand to taxes, will capture the first effect but not the second, and as a result will overstate the welfare costs of the tax. The reason is that the consumer's marginal utility of consumption is only $v'(x) = p + \sigma t$ rather than $v'(x) = p + t$. Consequently, when the tax induces the consumer to purchase D_t fewer units, utility declines by $(p+\sigma t)D_t(p,t)$, and net social surplus falls by $[(p+\sigma t) - p]\, D_t(p,t)$, where the term $-pD_t(p,t)$ corresponds to the decrease in production costs that results when D_t fewer units are purchased.

Suppose next that the economy also includes some rational consumers, with $\sigma_s = 1$. The efficiency cost formula becomes:

$$\begin{aligned} W'(t) &= tD_t(p,t) - \mu(b)(1-\sigma_b)t\frac{d}{dt}x_b(p,t) \\ &= \mu(s)t\frac{d}{dt}x_s(p,t) + \mu(b)t\frac{d}{dt}x_b(p,t) - \mu(b)(1-\sigma_b)t\frac{d}{dt}x_b(p,t) \\ &= \mu(s)\sigma_s t\frac{d}{dt}x_s(p,t) + \mu(b)\sigma_b t\frac{d}{dt}x_b(p,t) \\ &= tE[\sigma_\theta]tD(p,t) + tCov\left[\sigma_\theta, \frac{d}{dt}x_\theta(p,t)\right] \end{aligned} \tag{7}$$

Equation (7) shows that the marginal efficiency costs depend not only on the average σ, but also on how σ covaries with the demand elasticity. Models of tax salience build in a negative covariance between bias and elasticities: a higher value of σ_θ (less bias) implies a larger demand response, $\frac{d}{dt}x_\theta(p,t)$. Suppose in particular that $\frac{d}{dp}x_s(p,t) \approx \frac{d}{dp}x_b(p,t)$ at the price-tax pair (p,t), and that σ_b does not depend on t (assumptions that are likely valid for low tax rates). Then $\frac{d}{dt}x_\theta(p,t) = \sigma_\theta\frac{d}{dp}x_\theta(p,t) \approx \sigma_\theta D_p(p,t)$, in which case equation (7) becomes

$$W'(t) \approx tE[\sigma_\theta]tD(p,t) + tVar[\sigma_\theta]D_p(p,t). \tag{8}$$

[54] Notice that the first-order condition for the optimal tax rate, $W'(t) = 0$, is satisfied for $t = 0$. This property reflects the fact that our model includes a lump-sum tax. Exclusive reliance on the lump-sum tax achieves the first-best because then the consumer perceives all prices correctly.

Equation (8) is a special case of the formulas derived in Taubinsky and Rees-Jones (forthcoming). It shows that the marginal efficiency cost of taxation depends not only on the average value of σ, but also on the variance: the higher the variance, the higher the efficiency costs. The broad principle driving this result is that an increase in the tax has a higher impact on welfare when the consumers who are most elastic to the tax are the most biased ones. See, e.g., equation (2) and our discussion of the "average marginal bias" below it. This principle is true for any kind of bias, and since bias here is given by $\gamma_b(p,t) = (1-\sigma_b)t$, a positive covariance between σ_b and the elasticity implies a negative relationship between the size of the bias and the elasticity.[55]

Figure 2 provides a graphical illustration of efficiency costs when consumers underestimate taxes to the same degree (the homogeneous case), and separately when consumers underestimate taxes to differing degrees (the heterogeneous case). Beginning with the homogeneous case, the demand curve $D(p_0, t)$ corresponds to how observed demand varies with the not-fully-salient tax. The demand curve $D(p_0 + t, 0)$ corresponds to how demand would vary with a fully salient tax (for example, one that is included in posted prices). The equilibrium quantity sold in the market is such that the marginal utility from the product is $p_0 + \sigma t$. Thus, the deadweight loss from taxation corresponds to the smaller triangle with height σt, rather than to the larger triangle under the demand curve $D(p_0, t)$ with height t. Turning to the heterogeneous case, we can reinterpret $D(p_0, t)$ as capturing the demand of the consumer with the mean salience parameter, $E(\sigma)$. As shown in the figure, there are additional efficiency costs beyond those the average consumer incurs. Again, this result follows because consumers with the highest values of σ have the most elastic responses to the tax, but also attach the greatest value to the good on the margin.

In light of the preceding analysis, both the mean and variance of misperceptions should affect the magnitude of optimal commodity taxes within a Ramsey framework. Farhi and Gabaix (2015) provide general optimal tax formulas showing that optimal taxes are indeed decreasing in $E[\sigma_\theta]$ and increasing in $Var[\sigma_\theta]$.

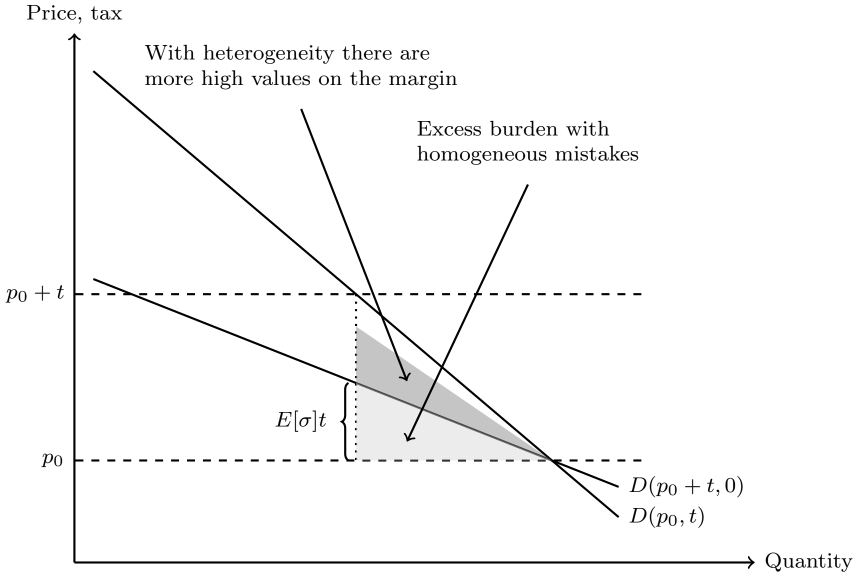
3.4.2 Empirical measurement and implementation[56]

Chetty et al. (2009) provide the first empirical estimates of underreaction to sales taxes, using two empirical strategies. The first involves a field experiment at a grocery store. The main finding is that posting new tags that highlight the tax and display tax-inclusive prices reduces demand, and that the magnitude of the effect is the same as that of a price increase equal to 65% of the tax. The authors infer that the average value of σ is 0.35. This experiment is perhaps the first example of the empirical strategy that we previously labeled "reframing decisions."

[55] To build intuition for this principle, recall that the marginal efficiency cost of taxation for a single consumer with misperception parameter σ is $\sigma t D_t$. For simplicity, assume for the moment that $D(p,t) = a - b(p + \sigma t)$. Then $\sigma t D_t = -b\sigma^2 t$. Notice in particular that this expression is negative and *concave* in σ. As a result, an increase in the variance of σ necessarily increases the population average of the marginal efficiency cost of taxation (as a consequence of Jensen's inequality).

[56] See also Gabaix (forthcoming) in this handbook for a discussion of measuring inattention in a variety of domains including sales taxes.

Figure 2: Efficiency costs and tax salience



Price, tax
With heterogeneity there are more high values on the margin
Excess burden with homogeneous mistakes
$p_0 + t$
$E[\sigma]t$
p_0
$D(p_0 + t, 0)$
$D(p_0, t)$
Quantity

The second empirical strategy employs naturally occurring data to measure demand responses to changes in excise taxes and sales taxes on alcohol using the method of differences-in-differences. Excise taxes are included in posted prices, while sales taxes are not. Based on the small observed responses to changes in sales taxes but large responses to changes in excise taxes, the authors infer that the average value of σ is 0.06. This empirical strategy is an example of the empirical strategy we labeled "analogous demand responses."

The tendency for people to underreact in response to taxes that are not included in posted prices has been replicated in laboratory experiments by Feldman and Ruffle (2015) and Feldman et al. (2015). Although these experiments were not designed to permit estimation of σ, they nevertheless corroborate the spirit of the Chetty et al. (2009) results in settings with cleaner identification of the behavioral effects. Finkelstein (2009) also provides related evidence that paying a toll electronically is less salient than paying it personally, which leads to an increase in tolls once electronic tolling is operationalized.

The Chetty et al. (2009) approach to welfare analysis is an application of the Bernheim-Rangel framework. Changing the presentation of information concerning taxes does not alter opportunities; hence it is an aspect of framing. A discrepancy between the quantities purchased in the two frames raises the possibility that consumers err when making decisions in either or both of them. Arguably, posting tax-inclusive prices makes the opportunities transparent, while computing them at the register does not. Consequently, characterization failure is most likely when posted prices are not tax-inclusive. The authors conduct welfare analysis based on that premise.

However, there are plausible reasons for thinking this restriction of the welfare-relevant domain may not be the right one. The first empirical strategy in Chetty et al (2009) may lead consumers to become especially "tax averse," for example because the new tags cause them to focus on their resentment of taxes. Alternatively, the tags may simply confuse consumers, who might interpret the after-tax prices as before-tax prices, and thus erroneously think the products are more expensive than they actually are.

One way to justify the paper's implicit restriction on the welfare-relevant domain would be to show that people are not aware of unposted taxes through surveys. But in fact, the authors demonstrate precisely the opposite using a survey administered to shoppers exiting the store.

The second empirical strategy in Chetty et al. (2009) addresses some of the confounds that could follow from the use of unusual tags in their experiment. Because naturally occurring posted prices include excise taxes, there is no problem with conspicuous highlighting. Consequently, this second strategy avoids potential experimental demand effects, as well as the consumer confusion that could arise in the experiment.

Taubinsky and Rees-Jones (forthcoming) conduct an experiment that directly varies both prices and taxes. Because their experimental design does not rely on tags that draw attention to the tax-inclusive vs. the tax-exclusive price, their estimates are not subject to framing effects that could have generated confounds in the Chetty et al. (2009) experiment.

A more important limitation of the Chetty et al. (2009) approach is that it does not shed light on individual differences in σ. Nor is it suitable for measuring how σ changes with the size of the tax. Taubinsky and Rees-Jones (forthcoming) estimate a lower bound for the variance of σ using a within-subject experimental design. They replicate the qualitative findings of Chetty et al. (2009) concerning underreaction to taxes. Their estimates place the average value of σ at roughly 0.25 with a tight confidence interval. At the same time, they estimate a large lower bound for the variance of σ. Using a generalization of formula (8) along with the estimated mean and variance of σ, they find that the representative-agent formula used in Chetty et al. (2009) underestimates the deadweight loss of taxation by a factor of three or more.

Taubinsky and Rees-Jones (forthcoming) also find that people underreact less when tax rates are higher. This finding is important because the distortionary effects of tax increases can be substantially greater if high tax rates stimulate attention, than if attention is exogenous. The intuition is straightforward: behavioral responses tend to be larger with endogenous attention because $\sigma(t') > \sigma(t)$ for $t' > t$ implies

$$\sigma(t')t' - \sigma(t)t = \underbrace{\sigma(t)\Delta t}_{\text{Effect given constant } \sigma} + \underbrace{(\sigma(t') - \sigma(t))t'}_{\text{Effect on } \sigma} > \sigma(t)\triangle t$$

. In words, when attention is endogenous, a higher tax increases *perceived* (after-tax) prices not only by mechanically making actual prices higher, but also by increasing attention to the tax.

3.5 Non-financial policy instruments

While we have focused primarily on corrective tax policy, academics and policy makers have also proposed using other non-standard policy instruments to achieve changes in behavior. These instruments include interventions that make information salient, such as visibly posting caloric content for foods (e.g., Bollinger et al., 2011) or requiring graphic cigarette warning labels (e.g., Chaloupka et al., 2014); disseminating information on social norms (e.g,. Allcott, 2011b; Allcott and Rogers, 2014; Ayres et al., 2013; Costa and Kahn, 2013); increasing the social visibility of consumers' behavior (e.g., Butera et al., 2018); offering commitment opportunities (e.g., Beshears et al., 2005); encouraging people to form concrete actions plans (i.e., "implementation intentions"; see, e.g., Milkman et al., 2011 or Carrera et al., 2018); and simply providing reminders (e.g., Karlan et al., 2016a). We call these policy instruments "non-standard" to distinguish them from more standard non-price instruments such as quantity regulation (e.g., Weitzman, 1974) and mandatory information disclosure (Grossman and Hart, 1980; Grossman, 1981; Milgrom, 1981).[57] For an extensive catalog of such policies, see OECD (2017).

3.5.1 What is a "nudge"?

Summarizing the perspectives articulated in Thaler and Sunstein (2003); Sunstein and Thaler (2003) and Thaler and Sunstein (2008), Sunstein (2014) refers to all such strategies as "nudges," which he defines as "liberty-preserving approaches that steer people in particular directions, but that also allow them to go their own way." In our view, it is inappropriate to group all these policies together under the "nudge" rubric. Implicit in the rationale for "libertarian paternalism" is the notion that nudges do not change opportunity sets. Yet most of the examples of non-price interventions cited above do change opportunities in meaningful ways. For example, providing people with information about social norms, or revealing their behavior to others, fundamentally changes the social and emotional costs and benefits of taking various actions; thus, it changes the nature of available consumption bundles. Similar remarks apply to interventions that manipulate the salience of certain types of information, such as graphic imagery on cigarette packs. One should not call an intervention a "nudge," which falsely suggests a minimal level of pressure, simply because the consequences are non-financial. On the contrary, social and/or emotional manipulation can be highly coercive. While it is worth knowing that certain types of non-price interventions can achieve desired changes in behavior at lower financial costs than traditional policies (e.g., Benartzi et al., 2017), one should not leap to the conclusion that these interventions are welfare improving without explicitly factoring in non-financial effects on well-being.

For the remainder of this section, we define a nudge more precisely as a non-price intervention that achieves a change in behavior by modifying the decision problem in a way that would not alter a consumer's perception of the opportunity set absent some error in reasoning. In other words, nudges exploit framing effects, defined as in Section 2.2.2: they may leave the consumer's perception of the opportunity set intact but change a contextually constructed judgment, or they may change the perceived opportunity set due to a cognitive error without actually altering the consumer's objective information. Whether we classify any given intervention as a nudge therefore depends on our assumptions about the scope of consumers' concerns, which may be controversial. As an example, if consumers' concerns are limited to conventional goods and services, then posting

[57] We contrast mandatory disclosure of otherwise non-available information with information saliency interventions that make otherwise available information more salient or easier to process.

tax-inclusive prices (in a setting where information on tax rates is generally available) is a nudge. However, it is not a nudge if calculating tax-inclusive prices entails non-trivial cognitive costs, or if salient reminders of tax rates cause aversive emotional reactions.

3.5.2 Justifications for non-price interventions

The most compelling case for non-price interventions involves a perfect nudge that "debiases" behavioral consumers by modifying the prevailing decision frame so that the task lies within the welfare-relevant domain rather than outside it. Standard consumers are unresponsive to the decision frame, and are therefore unaffected. In contrast, taxation is a blunt instrument that generally changes the actions of all consumers, benefitting some while hurting others. Thus, in some settings with heterogeneous agents, nudges can be more efficient than taxes because their effects are more appropriately targeted.[58]

Despite this potential advantage, the case for non-price interventions is more nuanced than it might at first seem. First, as we have already noted, the purported "cheapness" of these non-price interventions (i.e., the contention that they involve relatively low costs to governments and consumers), can be a misconception if they do change opportunity sets. While a commodity tax raises revenue, a non-price intervention does not. As an example, graphic images on cigarettes packs generate negative emotions that resemble a tax from the consumer's perspective, but they raise no revenue (Loewenstein and O'Donoghue, 2006; Glaeser, 2006). Of course, these considerations can also favor non-price interventions: a promotional campaign that makes consumers feel good about buying "green" products (rather than guilty about buying energy-inefficient ones) can potentially replicate the utility boost obtained through a subsidy without depleting public funds.

A second problem with non-price interventions, including nudges (or near-nudges), is that their impact on behavior may be either limited or temporary.[59] Even the typical information saliency intervention, which arguably entails more than a mere nudge, has only modest effects on behavior. Some degree of reliance on conventional policy instruments such as taxes may therefore be unavoidable.

A third issue arises in settings where suboptimal choices are actually desirable because they offset other distortions. Imagine, for example, that the government must raise revenue through a distortionary tax. Welfare will be higher if consumers ignore the tax. Nudging them to make better decisions by posting tax-inclusive prices increases distortions and reduces efficiency.

The preceding discussion suggests that the framework of libertarian paternalism is not particularly useful for rigorously evaluating the costs and benefits of non-price interventions, especially

[58] This advantage may not be present, however, if a nudge affects different subsets of biased consumers differently. For example, suppose all consumers display the same quantitative bias, γ, with respect to the consumption of cigarettes in the naturally occurring frame. Then a tax $t = \gamma$ achieves an efficient allocation, as shown in Section 3.2. In contrast, despite the homogeneity of γ, smokers' responses to nudges (such as warning labels) may be heterogeneous. Nudging some consumers to respond efficiently may cause others to underreact, and still others to overreact. In that case, nudges are less efficient than the optimal tax.

[59] See, e.g., Long et al. (2015) for a review of calorie labeling interventions, or Conn et al. (2016) for medication adherence interventions.

when it is applied to policies that are not true nudges (as defined above). Instead, the task of policy evaluation calls for economic analyses that embrace a defensible welfare criterion and use it to evaluate costs and benefits accounting for behavioral responses, direct effects on utility, and interactions with other policy instruments such as taxes (e.g., through fiscal externalities).

In this spirit, Farhi and Gabaix (2015) provide a theoretical analysis of optimal nudges that addresses a key question: whether it is better to nudge or to tax. They show that in a setting with redistributive motives, as in Section 3.3, a nudge tends to be more (resp. less) efficient than a tax if consumption patterns render the latter regressive (resp. progressive).

3.5.3 Empirical measurement of welfare effects for non-financial interventions

On the empirical side, Allcott and Kessler (forthcoming) attempt to analyze the welfare effect of a particular type of non-price intervention on consumers' utility. The intervention provides consumers with social comparisons about how well they conserve energy relative to their peers. They estimate welfare effects by eliciting each consumers' willingness to pay (WTP) for receiving information about social comparisons in the future. They find significant heterogeneity in elicited WTP, with the range encompassing both positive and negative values. The average is moderately positive.

A crucial assumption of the Allcott-Kessler analysis is rational expectations: for WTP to be a valid money-metric measure of welfare, consumers have to rationally anticipate how much energy they will use in the future. If consumers are overconfident about their ability to conserve in the future, or simply underestimate the energy consumption of their appliances, then the Allcott-Kessler method would yield upwardly-biased estimates of welfare, assuming consumers prefer to receive reports that show them doing well rather than poorly.

The conceptual validity of the Allcott-Kessler welfare analysis depends on the underlying behavioral mechanism. A negative willingness to pay (WTP) for information is inconsistent with standard theories of decision making under uncertainty. The prevalence of negative WTPs therefore calls for a behavioral theory that can explain why the possession of information concerning social comparisons is sometimes unpleasant. An obvious possibility is that the consumer enjoys hearing that she uses less energy than others and dislikes hearing that she uses more than others. If these feelings are independent of the circumstances that determined her consumption, then the Allcott-Kessler method is valid. However, others plausible possibilities merit consideration. For example, a consumer who discovers that she uses more (less) energy than others may be more likely to suffer guilt (feel virtuous) if she knows she had the option to conserve (squander). For similar reasons, a consumer who does not receive the usage comparison report may be more likely to feel guilty if she declined the opportunity than if the report were never available. Problems arise in such cases because the consumer's concerns encompass conditions of choice, which potentially implicates the Non-comparability Problem (see Section 2.2.2). For example, a consumer with a positive WTP for the report may feel a strong social obligation to seek it out and act on it when the opportunity arises, but may nevertheless abhor having that opportunity and feeling that obligation, and may fervently wish for Congress to enact legislation banning its dissemination (provided she avoids responsibility by playing no role in policymaking). As long as the consumer cares only about the conditions of energy choice and not the conditions of the metachoice, it is possible to rationalize a negative WTP without rendering welfare unrecoverable, but this assumption is debatable and unproven.

In a similar vein to Allcott and Kessler (forthcoming), Butera et al. (2018) develop a method for evaluating the welfare effects of a social recognition intervention, but one that avoids relying on the rational expectations assumption and arguably allows for a more robust welfare interpretation of WTP. They conduct a field experiment with the YMCA in which consenting individuals are enrolled in a "Grow and Thrive Program." During the program, a donor contributes $2 to the participants' local YMCA every time they attend it over a month-long period. Additionally, some individuals are assigned to a "social recognition" group in which YMCA attendance is revealed to all other group members at the end of the "Grow and Thrive" month. The assignment to the social recognition group is exogenous with 90% probability. With 10% probability the assignment is based on the individuals' choices. In particular, Butera et al. (2018) elicit from each individual the WTP to be in the social recognition group for each possible realization of his or her monthly attendance. This WTP elicitation is incentive compatible because with 10% chance, a Becker-DeGroot-Marschak (BDM) mechanism determines whether or not the individual's behavior is made public at the end of the "Grow and Thrive" month.

Because Butera et al. (2018) elicit the WTP to be in the social recognition group for every possible attendance pattern, they are able to measure welfare effects using only ex-post choices, rather than ex-ante expectations of behavior. This approach avoids the need for a rational expectations assumption. Moreover, by adopting a social signaling interpretation, which assumes that consumers' concerns extend to social image rather than to conditions of choice, they avoid the conceptual problems that potentially arise in Allcott and Kessler. In principle, an observer could draw a negative inference about individuals who are not part of the social recognition group because some of them are assigned based on realized WTPs, which are positively related to YMCA attendance. This consideration could generate a signaling incentive to express a higher WTP for joining the social recognition group. However, the experiment minimizes this effect by assigning groups based on WTPs with only 10% probability.

Consistent with previous work, Butera et al. (2018) find that social recognition is a significant motivator of behavior. And consistent with standard social signaling models, they find that low-attendance individuals are worse off in the social recognition treatment, while high-attendance individuals benefit significantly from it. They then study welfare in the aggregate, and show that because the social recognition utility function is modestly convex, social signaling is a modestly positive-sum game, and thus is more welfare-enhancing than financial incentives that achieve the same (distribution of) changes in behavior.[60]

Other empirical welfare analyses of non-financial interventions examine the effects of changes in

[60] This holds under the assumption of quasilinear utility, which they propose as a reasonable approximation for small to modest financial incentives.

default options, e.g., Bernheim et al. (2015a). We mention that work in Section 4.5.

As the literature progresses, careful empirical studies of non-price interventions that are grounded in basic economic principles will be crucial for assessing their role as potentially useful tools in the optimal policy mix. Close attention to the conditions needed to draw empirical conclusions about the welfare effects of these policies is essential (see, e.g., Benkert and Netzer, forthcoming).

3.6 Commodity taxation with social preferences

An important branch of Behavioral Economics concerns the existence and effects of social preferences. Here we briefly mention some implications for tax policy.

3.6.1 The taxation of giving

While other-regarding concerns do not generally entail failures of rationality, they can give rise to externalities. Consequently, they can also justify the use of corrective taxation. Perhaps the most obvious applications involve the tax treatment of giving, either to charities or to family members. The US tax system currently subsidizes charitable contributions because they are, to a degree, deductible for the purpose of calculating income taxes. In contrast, interpersonal transfers in the form of gifts and estates are subject to taxation.

An important property of giving is that it creates an externality that benefits the recipient, as well as those who care about the recipient. Because the giver does not account for these benefits, giving tends to be suboptimal. Kaplow (1995) cites this mechanism as providing the foundations for a general argument in favor of subsidizing charitable contributions and other giving. Others examine the form of the optimal subsidy. For example, Hochman and Rodgers (1977) argue that tax credits for charitable contributions are more efficient than charitable subsidies. More recent work explores the optimal treatment of contributions to privately provided public goods in the context of income taxation (Saez, 2004; Diamond, 2006). There is also a parallel literature on the optimal tax treatment of gifts and bequests, which inherently implicates concerns about distribution (e.g., Piket and Saez, 2013).

A notable theme emerging from this literature is that optimal policy depends on the particular motives that account for giving (Diamond (2006)). Leading alternatives include pure altruism, “warm glow” giving, and signaling. As an extreme illustration, Bernheim (1986) and Bernheim and Bagwell (1988) show that, if all giving is purely altruistic and everyone is connected either directly or indirectly through voluntary transfers, then ostensibly distortionary taxes have no effects on resource allocation. However, the authors intend that observation as a critique of models positing pure altruism, rather than as descriptive of actual tax policy. Signaling motives for giving introduce rather different types of externalities. Because signals are often socially excessive, taxing them can be efficient.

3.6.2 Luxury taxes

Social motives also play a significant role in the analysis of commodity taxes on luxury goods, the purchase of which may involve status-seeking. Ireland (1994) formulates a signaling theory of conspicuous consumption in which people overconsume certain goods to signal their wealth, and demonstrates that a tax on those goods can be welfare-improving. This result follows from the general property of signaling models noted at the end of the previous section. For a related analysis, see Corneo and Jeanne (1997).

Bagwell and Bernheim (1996) examine the effects of luxury taxes in a setting that generates Veblen effects, which are said to exist when consumers prefer to pay a higher price for the same conspicuous good in order to render it more "exclusive." They identify conditions under which people prefer to signal their wealth by paying too much rather than by consuming too much. In those settings, luxury brands earn strictly positive profits under conditions that would, with standard formulations of preferences, yield marginal-cost pricing. As a result, commodity taxes on luxury goods are equivalent to non-distortionary taxes on pure profits.

The following simple model illustrates the logic of the Bagwell-Bernheim conclusions concerning luxury taxation. Suppose each consumer chooses either one unit of a luxury good or none. Firms produce the good at cost c per unit. All versions are functionally identical, but they are nevertheless distinguishable (i.e., they are conspicuously branded). The prices of all brands are publicly observable and sellers cannot grant secret price concessions to individual customers. Preferences are given by $u(x) + v(R - px) + \hat{R}$, where $x \in \{0, 1\}$ denotes consumption of the luxury good, R is wealth (which takes on one of two values, R_L and R_H), $R - px$ is consumption of the non-conspicuous numeraire good, and $\hat{R}$ is perceived wealth. We assume that the function v is increasing and strictly concave. According to this formulation, greater perceived wealth entails greater status, which the consumer values.

For the moment, imagine that versions (brands) of the good are available at every price p (weakly) exceeding c. Low-wealth consumers would then buy the cheapest version ($p = c$). High-wealth consumers would choose to buy the good at a price $p_H > c$, chosen to satisfy the non-imitation constraint: $v(R_L - p_H) + R_H = v(R_L - c) + R_L$.[61] Thus, Veblen effects emerge: wealthy consumers prefer to pay a price above costs for the conspicuous good.[62] In a competitive market with free entry and Bertrand pricing (where consumers resolve indifference in favor of incumbent firms), entrants will provide "budget brands" at $p = c$, while incumbent firms will provide "elite" branded products at $p = p_H$.

Now suppose the government imposes a luxury tax – in other words, an excise tax t on the amount paid for the good above some threshold, where the threshold exceeds cost. High-wealth

[61] Consistent with the application of various standard equilibrium refinements, this condition characterizes the most efficient signaling equilibrium.

[62] In this simple setting, consumers do not have the option to signal with quantity rather than with price. Bagwell and Bernheim (1996) provide conditions under which their results generalize to settings in which consumers can choose any quantity $x > 0$. They show that Veblen effects emerge when indifference curves exhibit double crossing rather than single crossing, and they argue that double crossing arises naturally in settings with liquidity constraints.

consumers continue to prefer an all-in price of p_H; it is of no consequence to them whether they pay the markup to a firm or to the government. Therefore, competition among branded firms will drive the before-tax price of elite brands down by precisely t. The tax is therefore a non-distortionary levy on pure profits. Bagwell and Bernheim (1996) observe that, consistent with this implication, subsequent to the imposition of a substantial federal luxury tax on various conspicuous goods such as high-end automobiles and yachts in 1990, several automakers including Rolls Royce, BMW, and Jaguar advertised that they would reimburse customers for the full tax payment.

4 Policies Targeting Saving

4.1 Behavioral themes pertaining to saving

The literature on behavioral approaches to understanding household saving grew from concerns about the empirical validity of the classical Life Cycle Hypothesis (LCH) due to Ando and Modigliani (1963). During the 1980s and 1990s, questions arose as to whether the LCH could adequately account for basic facts about saving among U.S. households. Empirical investigations revealed that most households accumulate relatively little financial wealth (Diamond, 1977; Diamond and Hausman, 1984), a finding that proved difficult to reconcile with the ostensible life-cycle objective of sustaining pre-retirement living standards after retirement (Bernheim, 1993),[63] as well as with consumers' stated objectives and intentions (Bernheim, 1995; Laibson, 1998). Far from contriving smooth consumption profiles, households that accumulate little wealth often experience sharp declines in consumption at retirement, particularly in cases where Social Security and employer-based defined-benefit pension plans provide low income replacement (Bernheim et al., 2001b).[64]

Other work in this area called specific LCH assumptions into question. One important line of criticism emphasized imperfections in self-control. Two approaches to modeling self-control emerged, one emphasizing time inconsistency (Strotz, 1955-1956; Schelling, 1984; Laibson, 1997), which we have already touched upon in Section 2.2.5, the other positing the existence of internal goods (Thaler and Shefrin, 1981; Shefrin and Thaler, 1988; Gul and Pesendorfer, 2001; Fudenberg and Levine, 2006). A second important line of criticism explored the limits of consumer sophistication, documenting (i) deficiencies in the knowledge and skills necessary for sound financial planning (Bernheim, 1988, 1995, 1998; Gustman and Steinmeier, 2004, 2005; Lusardi and Mitchell, 2007; Lusardi, 2009; Lusardi and Mitchell, 2011, 2014b), (ii) the pervasive failure to consult financial experts or use planning tools (Bernheim, 1998; Lusardi, 2009; Lusardi and Mitchell, 2011), (iii) the superficiality of decision processes (Bernheim, 1994; Lusardi, 1999; Lusardi and Mitchell, 2007), and

[63] Subsequently, Scholz et al. (2006)argued that it is nevertheless possible to rationalize patterns of wealth accumulation using life-cycle models.

[64] See also Hamermesh (1984), Mariger (1987), and Banks et al. (1998). Based on a disaggregated analysis of changes in expenditures, Aguiar and Hurst (2013) argue that the reductions in consumption at retirement are consistent with declines in work-related expenses and increases in household production, but they fail to address the observed relationship between the decline in consumption and income replacement rates. Indeed, Olafsson and Pagel (2018) show that the patterns of personal financial choices around retirement are inconsistent with Aguiar and Hurst's explanation.

(iv) the prevalence of ostensibly problematic choice patterns.[65]

We briefly review the literatures on self-control and limited financial sophistication in the Appendix to this chapter. For related discussions, see the chapters on "Intertemporal Choice" (Laibson and Marzilli-Ericson, forthcoming) and "Household Finance" (Beshears et al., forthcoming) in this Handbook.

The case for collective action. An important question is whether self-control problems justify collective action. Profit-seeking companies have incentives to design financial contracts and informational products that appeal to consumers seeking better tools for exercising self-restraint. This principle presumably applies to employers as well, who are motivated to configure their pension plans so as to maximize the value of benefits to its employees. Where is the market failure?

Some justifications for government intervention hinge on consumers' lack of sophistication. Markets do not necessarily fix problems arising from misinformed decision making. On the contrary, instead of providing needed information and education, competitive firms may exploit consumers' limited comprehension of opportunities (Gabaix and Laibson, 2006).

Even if all consumers are sophisticated, government intervention may be warranted. As we note below in Section 4.3.3, efforts to design menus of options that optimally accommodate population heterogeneity with respect to behavioral biases potentially encounter constraints arising from asymmetric information. The asymmetries can give rise to adverse selection, a well-known source of market failure. See Section 4.4 for further discussion.

4.2 The tax treatment of capital income

A strictly positive (negative) capital income tax in period t implies that period-$(t+1)$ consumption is taxed at a higher (lower) effective rate than period-t consumption. Accordingly, zero capital income taxation is equivalent to a uniform system of commodity taxes applied to the elements of the time-dated consumption bundle. According to a classical result due to Judd (1985) and Chamley (1986), the optimal capital income tax rate is zero in the long run for economies with infinite-lived consumers.[66] One can reinterpret this statement as implying that the optimal solution to the equivalent commodity tax problem involves rates that converge to a constant for large t. Recently, Straub and Werning (2015) have argued that the proofs of the Chamley-Judd results are incorrect, and that in fact the optimal capital income tax rate is positive in the long run when the intertemporal elasticity of substitution is less than or equal to unity.

Any factor that distorts the allocation of consumption between consecutive periods can alter the character of the policy prescription. The optimal policy creates an offsetting wedge that reduces or

[65] One line of work identifies choice patterns that experts deem inadvisable, such as low rates of saving (Bernheim (1993)), low enrollment in pension plans that offer generous matches, naive diversification strategies, and the tendency for employees to invest in their employers' stock (Benartzi and Thaler (1999, 2001, 2007)). Another focuses on evidence of excessive inertia, suggestability, and intention (e.g., Madrian and Shea (2001), Bernheim et al. (2015a), and Karlan et al. (2016b)).

[66] In contrast, taxation and subsidization of capital income are both potentially optimal in the long run for economies with overlapping generations of consumers (Atkinson and Sandmo, 1980; Erosa and Gervais, 2002).

removes the distortion in every period, and hence is present even in the long run. In this section, we discuss the nature of appropriate wedges for settings involving imperfect self-control and limited financial competence.

4.2.1 Imperfect self-control and the case for capital income taxation

One school of thought holds that many people fail to save as much as they should because they lack sufficient self-control. Under this view, public policy can compensate to some degree for poor decision making by providing incentives to save more. To explore the validity of this intuitive prescription, one must first adopt a particular theory of self-control. In this section, we explore the implications of three theories, two of which support the intuition, and one of which does not. The contrast between these approaches underscores the importance of exploring nuances concerning the nature of the choice mapping.

Correcting internalities arising from time inconsistency. We begin with theories that attribute poor self-control to time inconsistency. Following Laibson (1996), we adopt the perspective that choices provide valid normative guidance only when their consequences are correctly anticipated and limited to future periods. As with the analysis of commodity taxation, one can think of the decisions made in normatively suspect frames as involving "internalities," in the sense that the consumer does not fully or properly internalize all the costs and/or benefits she imposes on herself. As we have noted, these internalities can interact in interesting ways with concerns about revenues and distribution. For the time being, we will defer all discussion of distribution, and focus on policies impacting a representative individual.

Relabeling the commodity tax model of Section 3.2, we can think of y and x as current and future consumption, respectively. Assuming the consumer undervalues the future, the optimal policy will involve subsidization of future consumption, which is achievable through capital income subsidies. Additional complications arise in settings with more than two periods. Altering the tax rate on capital income at time t changes the implied commodity tax rate on consumption in all future periods. It also has complex budgetary implications because it potentially alters the entire trajectory of wealth and hence impacts tax collections in all periods. Even so, the logic of the simple model continues to apply.

We illustrate this point through a simple model, versions of which appear throughout this section. Suppose the consumer lives for four periods, $t = 0, 1, 2, 3$. No consumption takes place in period 0, though for some purposes we will assume the consumer makes a decision affecting later opportunities. In each subsequent period ($t = 1, 2, 3$), she consumes c_t. Consumption yields flow utility $u(c_t)$, which she aggregates according to quasi-hyperbolic discounting, with $\delta = 1$ and $\beta \in (0, 1)$. For simplicity, we assume Cobb-Douglas flow utility $u(c_t) = \ln(c_t)$. The consumer also receives income Y in period 1 and, for the time being, nothing in later periods. She has access to a savings account that pays a gross return of 0, but the government subsidizes period t saving (for $t = 1, 2$) at the rate σ_t. She also pays a lump-sum tax, T, in period 1, which balances the government's budget. Under these assumptions, we can write her intertemporal budget constraint as $c_1 + \frac{c_2}{1+\sigma_1} + \frac{c_3}{(1+\sigma_1)(1+\sigma_2)} \leq Y - T$.

The normative standard mentioned above effectively equates welfare with the consumer's objective function as of period 0. According to this standard, the first-best allocation solves

$$\max \beta \left[\ln(c_1) + \ln(c_2) + \ln(c_3)\right]$$

$$s.t.\ c_1 + c_2 + c_3 = Y$$

The solution is plainly to consume $c^* = \frac{Y}{3}$ each period.

How will a time-inconsistent consumer behave subject to arbitrary policy parameters T and $\sigma = (\sigma_1, \sigma_2)$? Her decision in period 1 will depend on how she expects to deploy her remaining resources at the start of period 2. A naif will expect to maximize $\ln(c_2) + \ln(c_3)$, while a sophisticate will expect to maximize $\ln(c_2) + \beta \ln(c_3)$. Using the fixed expenditure property of the Cobb-Douglas function, we see that the naif expects to spend half of its income in period 2 and half in period 3, while the sophisticate expects to spend the fraction $\frac{1}{\beta+1}$ in period 2 and the balance in period 3. In either case, it is straightforward to check that the solution to maximizing the first-period objective function, $\ln(c_1) + \beta[\ln(c_2) + \ln(c_3)]$, subject to the second-period continuation rule, is:

$$c_1^U = \frac{1}{2\beta + 1}(Y - T).$$

Thus, first-period consumption is the same irrespective of whether the consumer is naive or sophisticated. (This property is a special feature of logarithmic utility, and is not generally true.) Note in addition that $c_1^U > c^*$ for the case of $T = 0$. In other words, first-period consumption is excessive regardless of whether the consumer is naive or sophisticated.

In the second period, the consumers allocates her remaining resources in the manner anticipated by a sophisticate (even if she is a naif). Accordingly, she consumes the following in periods 2 and 3:

$$c_2^U = (1+\sigma_1)\left(\frac{1}{1+\beta}\right)\left(\frac{2\beta}{1+2\beta}\right)(Y - T)$$

$$c_3^U = (1+\sigma_1)(1+\sigma_2)\left(\frac{\beta}{1+\beta}\right)\left(\frac{2\beta}{1+2\beta}\right)(Y - T)$$

For the case of $T, \sigma = 0$, it is straightforward to check that $c_2^U \gtreqq c^* \gtreqq c_3^U$ if and only if $\beta \in [\frac{1}{2}, 1)$. Accordingly, for empirically plausible parameters, the model predicts overconsumption in periods 1 and 2, and underconsumption in period 3.

We now claim that subsidies of $\sigma_1 = \frac{1-\beta}{2\beta} > 0$ and $\sigma_2 = \frac{1-\beta}{\beta} > 0$ achieve the first-best allocation for both naifs and sophisticates. This conclusion is easily verified by substituting these values into the formulas for consumption, yielding $c_1^U = c_2^U = c_3^U = \frac{1}{1+2\beta}(Y - T)$. Because there are no leakages of resources from the system, there is no need to solve for T explicitly; the only possible solution is $c_1^U = c_2^U = c_3^U = \frac{1}{3}Y = c^*$.

Intuitively, the positive subsidies for saving correct the "internality" arising from placing "too much" weight on the present by increasing future rewards commensurately. The first-period subsidy is lower than the second-period subsidy because the former generates a larger fiscal externality than the latter: more first-period saving leads to higher subsidy payouts in both periods, while more second-period saving only leads to higher subsidy payouts in the second period. If age-dependent subsidies are politically infeasible, the optimal (constant) rate will reflect a balance between the first- and second-period objectives.

Moderating the disutility from temptation. Similar conclusions also follow for some theories that associate imperfect self-control with internal goods rather than time inconsistency. Krusell et al. (2010) make this point in the context of temptation preferences (Gul and Pesendorfer, 2001). To illustrate, assume the consumer's preferences are defined over the consumption bundle $c = (c_1, c_2)$, as well as the menu X, and correspond to the utility function

$$U(c, X) = u(c) - \alpha \left(\max_{\hat{c} \in X} v(\hat{c}_1) - v(c_1) \right),$$

where $\max_{\hat{c} \in X} v(\hat{c}_1) - v(c_1)$ represents a "temptation penalty" incurred when choosing anything other than the most tempting alternative from the menu X. We assume v is strictly increasing, so that higher consumption is more tempting. If we interpret $\alpha \max_{c \in X} v(c_1)$ as representing an internal "bad" the decision maker experiences when selecting an option from the menu X, then it is natural to treat $U(c, X)$ as a measure of welfare; see the Appendix.

Now assume, as in the previous model, that the consumer receives net-of-tax income $Y - T$ in period 1, and has access to a savings account paying a rate of return of $1 + \sigma$, where σ is a subsidy. Her opportunity set is then given by $X = \left\{ c \mid c_1 + \frac{c_2}{1+\sigma} \leq Y - T \right\}$. In this setting, the optimal subsidy is strictly positive. To understand intuitively why this result holds, note first that there is no tension between the objectives of the consumer and planner with respect to the choice of consumption from any menu, which depends only on $u(c) + \alpha v(c_1)$, or with respect to the choice of the menu from a set of feasible menus, which also depends on $\max_{\hat{c} \in X} v(\hat{c}_1)$. Because the consumer optimizes over c, it follows from standard optimal tax principles that $\sigma = 0$ achieves the highest possible value of $u(c) + \alpha v(c_1)$, or equivalently that $\frac{d[u(c)+\alpha v(c_1)]}{d\sigma} \mid_{\sigma=0} = 0$. The sign of $\frac{dU(c,X)}{d\sigma} \mid_{\sigma=0}$ therefore depends entirely on whether $\max_{\hat{c} \in X} v(\hat{c}_1)$ is locally increasing or decreasing in σ – in other words, it depends on the manner in which a subsidy affects the most tempting option in the choice set. Because the consumer does not optimize over X, there is no reason to think we have $\frac{d[\max_{\hat{c} \in X} v(\hat{c}_1)]}{d\sigma} \mid_{\sigma=0} = 0$ as well. Indeed, using the fact that $\max_{\hat{c} \in X} v(\hat{c}_1) = v(Y - T)$ (the most tempting alternative is to consume everything in period 1), we see that small subsidies are better than small taxes: subsidies necessitate positive lump-sum taxes ($T > 0$), which reduce first-period disposable income and hence the level of temptation experienced at the consumer's optimal choice, while taxes have the opposite effect because they necessitate positive lump-sum subsidies ($T < 0$).

Insuring risks arising from state inconsistency. As the preceding discussion suggests, the classes of theories emphasized in the literature provide formal rationales for the intuitive proposition that the government should subsidize capital income in settings where consumers save too little as a result of challenges associated with exercising self-control. Yet that conclusion may be less robust than such analyses suggest.

Consider the following alternative model of self-control and capital accumulation, inspired by the Bernheim and Rangel (2004) theory of addictive behavior. Suppose the consumer lives for two periods, $t = 1, 2$, consuming c_1 and c_2. With probability $1-\pi$, she chooses first-period consumption to maximize the intertemporal utility function $u(c_1) + u(c_2)$. Following Loewenstein (1996), we will call this the "cold" decision state. With the complementary probability π, she enters a "hot" decision state and binges, consuming $\bar{c}$ in the first period, which is significantly more than she would choose in the cold state. These hot states are triggered by environmental cues outside the consumer's control. She also receives income Y in period 1 and nothing in period 2. As before, government policy consists of a saving subsidy, σ, and a first-period lump-sum tax, T, yielding the intertemporal budget constraint $c_1 + \frac{c_2}{1+\sigma} \leq Y - T$. The government budget constraint requires the budget to balance for a large population of ex ante identical consumers.

Let c_t^θ for $\theta \in \{C, H\}$ denote consumption in period t in either the cold or hot state, respectively. We will adopt a normative standard that equates welfare with the cold-state objectives, and that places no weight on any objectives that might rationalize the hot-state behavior.[67]To solve for the optimal subsidy, we employ a simple perturbation argument as in Section 3.3. For analytic convenience, we define $s = \frac{\sigma}{1+\sigma}$. Decreasing s by some small amount, ds, yields the following consequences:

1. Revenue rises by $(c_2^C + c_2^H)ds$.

2. Since the price of c_2 rises by ds, the utility of cold-state consumers falls by $u'(c_2^C)c_2^C ds$ (according to the envelope theorem), and the utility of hot-state consumers falls by $u'(c_2^H)c_2^H ds$ (because their consumption patterns are fixed).

3. Due to substitution effects among cold-state consumers, revenue increases by $\frac{dc_2^C}{ds} s ds$.

The marginal value of government revenue is $\lambda = (1-\pi)u'(c_2^C) + \pi u'(c_2^H)$. Accordingly, the net effect of this policy perturbation is

$$dW = -Cov[(u'(c_2^\theta)), c_2^\theta]ds + \lambda \frac{dc_2^C}{ds} s ds \tag{9}$$

To find the optimal subsidy, we use the first-order condition, $dW = 0$. Because $c_2^C > c_2^H$ and $u'(c_2^C) < u'(c_2^H)$ for $s \geq 0$, the covariance term is positive when $s \geq 0$. Thus the first-order condition implies $s < 0$.

[67] Bernheim and Rangel (2004) provide a neurobiological justification for a parallel assumption in the context of addiction. Whether there is sufficient evidence to support cold-state welfare analytics in the current context is an open question.

The preceding reasoning is a variation of the analysis in Section 3.3. Under the assumptions in our example, we have $\bar{\gamma} = 0$. It follows from equation 4 that the marginal welfare effect of taxing future consumption is governed by the correlation between future consumption and the marginal utility of income. Because binges reduce the former and increase the latter, the correlation is negative. Accordingly, small taxes on future consumption, and hence on capital income, are welfare-improving.

According to this theory, the benefit of a capital income tax is that it provides implicit insurance against the otherwise uninsurable risk of encountering environmental cues that trigger a spending binge. Our specific conclusions plainly depend on the assumption that first-period spending in the hot state is unresponsive to taxes and subsidies, and one can in principle overturn the main result by building in a sufficiently elastic response. Even so, the example provides reason to question the widespread presumption that capital income subsidization is desirable when low saving results from imperfect self-control.

Implications of population heterogeneity. We can alternatively interpret the preceding example as one in which there are two types of consumers: optimizers and undersavers who are inelastic to taxes and subsidies. Our result is then that the optimal policy does not actually induce the undersavers to behave in a socially optimal way. Generally speaking, when preferences vary across the population, it becomes impossible to optimize the policy for all consumers simultaneously. In light of this observation, it is important to ask whether conventional tax and subsidy instruments are too blunt for this policy application.

In Section 4.3, we consider an alternative policy approach involving the creation of opportunities for consumers to undertake commitments. As we explain, that approach accommodates population heterogeneity more effectively. However, it too has potentially important limitations. Most obviously, by reducing the consumer's flexibility, it magnifies the consequences of unanticipated expenses and income fluctuations. In contrast, taxes and subsidies preserve the consumer's flexibility to make adjustments as events unfold. Because both approaches have advantages and disadvantages, mixed approaches merit consideration. We turn to mixtures in Section 4.4.

4.2.2 Limited financial sophistication and capital income taxation

The implications of limited financial sophistication for capital income taxation are largely unexplored and likely complex. Even so, the literature points in a few interesting directions.

If consumers rigidly employ well-defined heuristics when making financial decisions, positive analysis becomes reasonably straightforward. Suppose, for example, that – consistent with common financial planning strategies – consumers aim to achieve fixed rates of earnings replacement after retirement. It would then follow that the interest elasticity of saving is negative and potentially substantial, and consequently that efforts to increase saving by reducing the rate of capital income taxation are counterproductive (Bernheim, 1994). Indeed, tax breaks provided through retirement savings accounts would simply constitute lump-sum subsidies. Alternatively, imagine that consumers employ fixed rules of thumb, such as saving 10% of earnings (Bernheim, 1994). In that case, the interest elasticity of saving would be zero. Of course, far from being fixed, heuristics and rules of thumb may respond to the economic environment in unknown ways, rendering policy analysis far more challenging.

In a few cases, research has identified specific biases arising from limited sophistication, such as the tendency to underestimate compounding, a phenomenon known as *exponential growth bias* (Wagenaar and Sagaria, 1975; Eisenstein and Hoch, 2007; Levy and Tasoff, 2016; Stango and Zinman, 2009; Almenberg and Gerdes, 2012). Models of this bias may have significant policy implications. Consider, for example, the possibility that people evaluate their intertemporal opportunities by computing simple interest, rather than compound interest (Levy and Tasoff, 2016). The resulting underestimation of returns could provide another justification for capital income subsidization, although it would appear to argue for subsidies that increase with the investment horizon, and consequently decline with age. Unfortunately, this simple model appears to have implausible implications, such as an infinite willingness to pay for any asset that makes a fixed periodic payment indefinitely.

4.3 Special savings accounts: commitment features

Next we examine an alternative strategy for addressing inefficiencies associated with imperfect self-control: create appropriate commitment opportunities, and possibly provide consumers with inducements to employ them. Discussions of commitment devices originate with Strotz (1955-1956). For a time-inconsistent consumer, the purpose of a commitment is to bring future choices in line with current objectives and intentions. In the example of Section 2.2.5, Norma might avoid eating pizza by making a social commitment to meet a friend for lunch at a restaurant that only serves salad.

Typically, policymakers imbed these opportunities into special savings accounts, such as IRAs and 401(k)s, and provide further inducements in the form of tax breaks. The nature of the associated commitments vary. Below, we draw an important distinction between provisions affecting the liquidity of invested funds and those that provide for delayed implementation of contribution decisions. IRAs and 401(k)s are both illiquid investments. With 401(k)s, employers implement changes in contributions with a significant lag (next pay period). IRAs do not share this feature.

Throughout this section, we assume the consumer is a quasi-hyperbolic discounter and adopt the same welfare standard as in Section 4.2. Alternative theories of self-control have similar implications for commitment opportunities. Dramatically different implications could follow from other welfare perspectives.

4.3.1 The case for illiquidity.

The main idea. We begin by illustrating how the existence of illiquid savings vehicles can help consumers overcome self-control problems. For this purpose, we consider a variant of the first model examined in Section 4.2: a quasi-hyperbolic consumer with Cobb-Douglas preferences lives for four periods ($t = 0, 1, 2, 3$) and must allocate her resources to consumption in periods 1 through 3. Instead of starting out with all of her income in period 1, she receives an income stream (y_1, y_2, y_3). She divides her savings between three accounts, one liquid, the other two illiquid. The rules of one illiquid account preclude withdrawals prior to period 3 and prevent her from using these funds for collateral to secure loans; the rules of the second are identical except that they preclude withdrawals prior to period 2. The liquid account pays an interest rate of zero, while the period-2 illiquid account pays $\varepsilon_2 > 0$, and the period-3 illiquid account pays $\varepsilon_3 > \varepsilon_2$. To avoid confounding the effects of illiquidity and subsidies, we focus on the limiting case in which $\varepsilon_3 \to 0$. The analytic purpose of these small subsidies is simply to break (perceived) ties rather than to offer meaningful incentives. At the outset, we will assume the consumer has access to perfect credit markets, so that liquid balances can be negative, up to the sum of future income.

Without an illiquid account, both sophisticates and naifs select the consumption profile c^U defined in Section 4.2 (with $\sigma = T = 0$). With the option of contributing to illiquid accounts, a sophisticate instead achieves the first-best by borrowing $y_2 + y_3$ in period 0 and investing $\frac{Y}{3}$ in both of the illiquid accounts. The analysis for naifs is essentially identical. A naif does not expect to misspend her resources, and therefore sees no need for illiquidity. However, in period 0, she prefers to implement her desired plan through the same strategy as the sophisticate because of the (tiny) subsidies. For the naif, commitment is incidental but nevertheless equally effective, provided the special accounts offer some small bonus. It is worth emphasizing that sophisticates will continue to use the illiquid account when the returns are taxed ($\varepsilon_2, \varepsilon_3 < 0$), but naifs will not.

Robustness. The strong conclusion of the previous paragraph – that illiquid accounts permit quasi-hyperbolic consumers to achieve the first-best allocation – hinges on several critical assumptions. First, we have assumed that the available investment instruments provide the consumer with flexible control over the duration of illiquidity through the mix of investments in the period-2 and period-3 illiquid accounts. In practice, special savings accounts offer little or no flexibility in this dimension. If the government only offers an illiquid "retirement account" targeting period 3, the consumer will be able to lock in her period 3 consumption as of period 0, but will not be able to prevent herself from overconsuming in period 1 at the expense of period 2.

Second, we have assumed that consumers have unlimited ability to borrow against future earnings at the market rate of return. The existence of credit constraints can significantly reduce the welfare benefits of offering illiquid savings accounts. To illustrate, imagine that borrowing is prohibitively expensive, so that all account balances must be non-negative. In that case, the consumer has no ability in period 0 to influence future consumption, and therefore can no longer achieve the first-best allocation. In the next period, her ideal is to achieve the period-1 full-committment allocation, defined as the solution to

$$\max\left[\ln(c_1) + \beta\ln(c_2) + \beta\ln(c_3)\right]$$

$$s.t.\ c_1 + c_2 + c_3 = Y$$

It is easily verified that the solution entails $c_1^{'} = \frac{1}{1+2\beta}Y$ and $c_2^{'} = c_3^{'} = \frac{\beta}{1+2\beta}Y$. Let's assume $c_1^{'} < y_1$ and $c_3^{'} > y_3$, so that the period-1 full-commitment solution remains feasible even with liquidity constraints. The consumer achieves that outcome if $c_2^{'} \geq y_2$ by consuming c_1 in period 1 and allocating $c_3^{'} - y_3$ to the period-3 illiquid account, but cannot achieve it if $c_2^{'} < y_2$.[68] Moreover, in cases where the consumer's period-1 saving in the period-3 illiquid account is too small to crowd out all her period-2 saving, offering illiquid accounts has no effect on her consumption.[69]

As an additional wrinkle, imagine that borrowing is possible but costly. For example, it might require the use of credit cards. A sophisticate might then become reluctant to invest too much in the illiquid account in period 0, for fear that he would thereby induce himself to borrow in period 1. In contrast, the naif would invest more heavily in the illiquid account in order to obtain the higher return, and suffer as a consequence. This example alerts us to the possibility that the creation of tax-favored commitment opportunities can actually harm unsophisticated consumers.

Third, we have assumed away all uncertainty concerning future income and cash needs, arising for example from major or minor emergencies that require ready access to liquid funds. Unconditional commitments entail costs because they require the consumer to sacrifice potentially useful flexibility.[70] Amador et al. (2006) show with reasonable generality that the consumer's optimal strategy nevertheless involves commitment to a minimum level of saving.

Fourth, we have assumed that external commitments are the only routes to self-control. As noted in the Appendix, an alternative view holds that people often achieve self-control through internal means, such as contingent self-reinforcement. Under that view, it is essential to evaluate the manner in which internal and external self-control strategies interact, and in particular to consider whether they reinforce or undermine each other. Bernheim et al. (2015c) explore these issues and draw out implications for the structure of savings plans. They demonstrate that optimal behavior involves a simple, intuitive, and behaviorally plausible pattern of self-reinforcement: failure to meet a self-set standard leads the individual to briefly "fall off the wagon," and then return to the preferred decision rule. A key insight from their analysis is that external strategies for exercising self-control, such as reducing liquidity, can undermine internal self-control by limiting the scope for self-reinforcement.

Provisions pertaining to withdrawals. So far, we have focused on fixed-term accounts that entirely proscribe early withdrawals. In practice, special savings accounts can offer a degree of liquidity by providing for limited withdrawals, possibly under specified conditions, or subject to penalties. The logic of such provisions is readily evident in settings with uncertainty, particularly

[68] These same conclusions hold regardless of whether the consumer is a sophisticate or a naif, assuming as above that the planner can favorably resolve the naif's indifference through the use of tiny subsidies.

[69] If the consumer saves in period 2, then the division of her period-1 saving between the liquid account and the period-3 illiquid account does not affect the resulting consumption profile on the margin. Consequently, the first-order conditions governing her first-period and second-period are the same as those that identify c^U.

[70] These costs are avoidable if consumers can make conditional (i.e., state-contingent) commitments. As a practical matter, most commitments are either unconditional or conditional on a limited range of events.

if consumers occasionally encounter unforeseen emergencies. An ideal approach would allow for hardship withdrawals under conditions meeting objective criteria, but in practice it is difficult to enumerate all meritorious hardships, and verification can be problematic. A less perfect but more practical solution is to penalize early withdrawals, setting the penalties and withdrawal limits by evaluating the marginal benefits of improved self-control and marginal costs of reduced flexibility.

The analysis of Bernheim et al. (2015c) makes a case for policies that permit unrestricted withdrawals once consumers' accumulated savings exceed preset thresholds. Their theory implies that effective internal self-control may be possible only when consumers have sufficient liquid resources. Those who have not yet accumulated much wealth may therefore be unable to save in the absence of external self-control devices such as illiquid savings accounts. However, once wealth rises above some critical threshold, continued illiquidity may prevent more effective internal self-control strategies from kicking in. From this perspective, illiquid accounts are most beneficial when their use is limited to "priming the pump."

4.3.2 The case for delayed implementation of decisions

We now turn our attention to commitment features that involve delayed implementation of decisions. One possibility is to impose a delay between the contribution decision and implementation (see, e.g., Laibson, 1996). This feature is extremely common in practice: when an employee changes her pension plan contribution rate, her employer typically implements the change in a subsequent pay period rather than immediately. Taking this idea a step further, employers could also allow households to specify savings trajectories, or to specify conditions for escalation of contributions to special accounts. The Save More Tomorrow plan devised by Thaler and Benartzi (2004) is an example of this approach. A final possibility is to allow for withdrawals with low or zero penalties contingent on advance notification (see, e.g., Laibson, 1997).

The main idea. To illustrate the potential benefits associated with delayed implementation of contribution decisions, we reexamine the model employed in Section 4.3.1, modified as follows: we replace the two illiquid accounts with a one-period savings instrument requiring a one-period-in-advance contribution election. In other words, the consumer can specify period-t contributions to the special account in period $t-1$, and can access the entire account balance in period $t+1$. Importantly, the consumer cannot reverse her period-t contribution election in period t. Nor is she permitted to accomplish this end indirectly by borrowing against the period $t+1$ account balance in period t. (The account is illiquid in that limited sense.) The liquid account pays a rate of return of zero, while the special account pays $\varepsilon > 0$ as a result of a tiny subsidy.

In this setting, a sophisticate achieves the first-best allocation. Through a standard argument involving backward induction, one can show that she commits herself to period-1 contributions of $\frac{2Y}{3}$ in period 0. The most she can consume in period 1 is then $\frac{Y}{3}$, which she achieves by borrowing $y_2 + y_3$ and spending all uncommitted resources. Given her present-focused preferences, that limit is binding: she spends $\frac{Y}{3}$ in period 1, and in addition commits herself to period-2 contributions of

$\frac{Y}{3}$. Upon reaching period 2, she again consumes as much as she can ($\frac{Y}{3}$), leaving $\frac{Y}{3}$ for period 3.

A naif expects to achieve the same allocation as the sophisticate, and is indifferent between doing so through regular or special savings accounts when both options pay the same rate of return ($\varepsilon = 0$). However, for any $\varepsilon > 0$, she strictly prefers special savings. Indeed, she sees the return differential as creating a pure arbitrage opportunity, and borrows as much as possible in order to finance greater contributions. As a result, despite seeing no value in commitments, the naif undertakes the same commitments as the sophisticate, and thereby achieves the first best.

The foregoing conclusions do not depend on the particular structure of special savings accounts, provided consumers have sufficient opportunities to make decisions that are implemented subject to delays. Suppose, for example, that the special account targets "retirement," in the sense that account balances become perfectly liquid in period 3. As before, consumers can commit to contributions one period in advance. Here we assume in addition that they can schedule penalty-free withdrawals one period in advance. In that case, a standard backward-induction argument reveals that the sophisticate commits to period-1 contributions of $\frac{2Y}{3}$ in period 0. Once period 1 arrives, she borrows $y_2 + y_3$ and consumes $\frac{Y}{3}$, and invests $\frac{2Y}{3}$ in the special savings account as before, but in addition schedules a withdrawal of $\frac{Y}{3}$ in period 2, leaving $\frac{Y}{3}$ for period 3.

Robustness We have seen above that the introduction of borrowing constraints reduces the welfare benefits of offering illiquid special savings accounts. In contrast, the case for delayed implementation of contribution decisions remains equally strong. To illustrate, we reintroduce the assumption that borrowing is prohibitively expensive, so that all account balances must be non-negative, while assuming that $y_1 \geq \frac{Y}{3} \geq y_3$, so that the first-best remains feasible. A standard backward-induction argument reveals that the sophisticate commits to period-1 contributions of $y_1 - \frac{Y}{3}$ in period 0, and commits to period-2 contributions of $\frac{Y}{3} - y_3$ in period 1, thereby achieving the first-best allocation. With a tiny subsidy, the naif does the same.

That said, uncertainty concerning future income and cash needs potentially reduces the benefits of delaying the implementation of contribution decisions. The issues are essentially the same as in the context of illiquid accounts. Likewise, the use of savings accounts with advance contribution election requirements may undermine internal self-control strategies by delaying self-punishment. In particular, when a consumer binges in period t, any period-t commitment she makes to her contribution for period $t + 1$ limits her ability to self-punish starting in period $t + 1$.

Commitments to consumption trajectories. As noted above, the Thaler and Benartzi (2004) Save More Tomorrow plan provides consumers with opportunities to commit in advance to allocating a portion of their future salary increases toward retirement saving. Despite the authors' informal claims, it is not clear that their proposal is well-founded in the formal theory of self-control. What matters for the theory is simply that decisions are made in advance, outside the window of present focus. In the context of our simple models, allowing the consumer to lock in period-1 and period-2 saving in period 0 offers no advantage over allowing her to lock in period-1 saving in period 0 and period-2 saving in period 1. The same principle holds in a more general setting with respect to commitments that are contingent on realizations of period-2 income.

To make a sound conceptual case for Thaler and Benartzi's approach, one requires a rather different theory of self-control. One possibility is that consumers discount hyperbolically, attaching the weight $\frac{1}{1+\alpha t}$ to outcomes t periods in the future, rather than quasi-hyperbolically (see Ainslie, 1992). Under that assumption, decisions pertaining to periods 2 and 3 are more future-oriented when made in period 0 than in period 1.

An entirely different case for Thaler and Benartzi's approach would proceed from the premise that consumers are imperfectly attentive: they may sometimes fail to elect higher contributions upon receiving a salary increase simply because they neglect the decision. Locking in a contingent plan for escalating contributions removes that possibility.

4.3.3 Implications of population heterogeneity

In contrast with taxes and subsidies, special savings accounts accommodate dimensions of population heterogeneity pertaining to the severity of self-control problems. To illustrate, suppose people differ with respect to β, which parameterizes the degree of present focus. In our basic models, a uniform system of special accounts with appropriate commitment opportunities permits every consumer to achieve her personal optimum, whereas a uniform system of taxes and subsidies does not.

That said, more complex models may implicate additional dimensions of population heterogeneity that are less amenable to uniform treatment within a system of special savings accounts. For example, in settings with uncertainty, consumers may value flexibility differently based on their exposures to short-term income and expenditure fluctuations, as well as their risk preferences. Because the costs and benefits of early withdrawal penalties are consumer-specific, optimizing these provisions for all consumers simultaneously is impossible.

An alternative is to provide consumers with opportunities to customize account provisions governing implementation delays, as well as the degree, duration, and/or conditions of illiquidity. For example, suppose we add uncertainty concerning income and/or expenditures to our simple model, thereby rendering commitments costly, but also allow consumers at the outset (period 0) to select the parameters governing limits on early withdrawals and associated penalties. According to the theory, each sophisticate will select the parameters that are optimal for her in light of her own circumstances and preferences. Bernheim et al. (2015c) emphasize that customizability may allow special savings accounts to complement internal self-control strategies more effectively.

While customizability offers potential advantages, it also raises concerns. An important question is whether consumers are sophisticated enough to make good decisions concerning the provisions of their accounts, let alone to optimize them. Despite some of our previous observations, inducing naifs to make optimal choices through tiny indifference-resolving subsidies is not generally possible. For instance, a naif will actively resist optimal early withdrawal penalties in settings with uncertainty. Additionally, optimizing account features can be mathematically complex, and consumers have little experience with those types of choices.

A second concern is that population heterogeneity usually goes hand-in-hand with private information, in the sense that each consumer knows more about her own circumstances than the government. When offering an option targeted at a particular type of consumer, the government has no way to prevent other types of consumers from selecting it. That limitation is potentially problematic when the option entails provisions with budgetary implications, such as penalties, fees, and subsidies. A menu of options that appears feasible (in the sense of budget balance) when each consumer is assigned to her intended option may become infeasible when consumers are free to pick any option on the menu. We discuss the implications of this observation in Section 4.4 below, where we consider mixed policies involving taxes, subsidies, and special accounts.

4.3.4 Evidence on the demand for commitments

For many years, evidence of a widespread demand for commitment proved elusive.[71] While anecdotes were plentiful (Laibson et al., 1998; Caskey, 1997; Beverly et al., 2003), there was little hard evidence concerning the prevalence of the cited practices, such as cutting up credit cards. A collection of relatively recent papers has begun to fill that gap.[72] Some of these focus specifically on financial choices; see, for example, Shipton (1992) on the use of lockboxes in Gambia, or Ashraf et al. (2006) on the demand for commitment savings products in the Philippines. Likewise, Aliber (2001), Gugerty (2007), Anderson and Baland (2002), and Ambec and Treich (2007) view ROSCA participation as a commitment device. Perhaps the cleanest evidence of a demand for commitment to saving comes from an experiment by Beshears et al. (2015), which documents a preference among many U.S. households for greater illiquidity when allocating funds among commitment accounts paying the same rate of return. Still, nagging doubts persist, partly because much of the evidence is equivocal, and partly because its scope is limited.[73] Skeptics continue to wonder why, if time inconsistency is so prevalent, the free market provides so few commitment devices, and unambiguous examples in the field are so difficult to find.[74] Indeed, some suggest that the fewness of the obvious exceptions proves the rule.

Why might time-inconsistent consumers exhibit limited demand for external commitment devices? One possibility is that they are stubbornly naive, in the sense that they fail to appreciate their own behavioral tendencies despite repeated experience. A second is that, in settings with uncertainty, commitments require consumers to sacrifice valuable flexibility (Laibson, 2015). This

[71] Most of the pertinent literature through 2010 echoes this evaluation. For example, Gine et al. (2010) write that "there is little field evidence on the demand for or effectiveness of such commitment devices. For recent surveys, see Bryan et al. (2010); DellaVigna (2009).

[72] Notable contributions on the use of commitment devices in non-financial contexts include Ariely and Wertenbroch (2002), Kaur et al. (2015) and Augenblick et al. (2015) on work effort, Houser et al. (2010) and Toussaert (2017) on temptation, Toussaert (2016) on weight loss, Gine et al. (2010) on smoking, and Bernheim et al. (2016) and Schilbach (2017) on alcohol consumption (which also includes a nice summary of previous work).

[73] For example, in Ariely and Wertenbroch's experiment, students may have been motivated by a misguided desire to signal diligence. Likewise, much of the evidence on the demand for commitment savings products in developing countries is potentially attributable to a desire for other-control (family and friends) rather than to self-control; see, e.g., Dupas and Robinson (2013).

[74] Many common financial products, such as mortgages and retirement accounts entail precommitments. However, those products offer other advantages, and it is not clear whether their inflexibility increases or reduces demand.

explanation assumes that consumers cannot make state-contingent commitments, which is reasonable if the difficulty of observing the relevant states (e.g., moods) renders them non-contractable. A third possibility is that private pensions, mortgages, and other long-term financial contracts happen to satisfy consumers' demand for commitment while also addressing other needs and objectives. A final explanation is that externally enforced commitments may undermine internal methods of self-regulation involving "contingent self-reinforcement." Foundations for the notion that people may self-impose contingent punishments and rewards to establish incentives for following desired plans of action are found in the literatures on self-regulation and behavior modification dating back to the 1960s.[75] Bernheim et al. (2015c) demonstrate that external constraints can undermine these internal mechanisms.

4.4 Special savings accounts with taxes and subsidies

Actual policies, such as the statutes that establish the frameworks for specialized retirement accounts, entail a mix of tax provisions and commitment features. Laibson et al. (1998), Angeletos et al. (2001), and Laibson et al. (2003) employ simulation methods to evaluate their effects. These papers study rich environments in which QHD consumers can contribute either to conventional liquid savings accounts or to illiquid tax-favored retirement accounts. Illiquidity is partial in the sense that withdrawals are permitted prior to age 60, but trigger penalties. The simulation models encompass other important factors such as income uncertainty, but abstract from internal self-control strategies. The authors demonstrate that reasonably parameterized QHD models can account for a number of otherwise puzzling behavioral patterns, such the observed comovements between income and consumption, including the sharp decline in consumption at retirement, and heavy reliance on costly revolving debt, such as credit cards. They also find that the welfare benefits of tax-favored retirement accounts may be substantial.

Ideally, we would like to determine the optimal mix of taxes, subsidies, and special savings account provisions in light of self-control problems, uncertainty with respect to income and expenditures, multiple dimensions of population heterogeneity, and asymmetric information between the consumer and account provider. Economists have only recently begun to make meaningful progress toward that ideal.

Galperti (2015) considers a setting in which consumers seek to provide for future consumption while retaining the flexibility to meet shorter-term needs, which are stochastic. Population heterogeneity takes a limited form: consumers either have limited or perfect self-control. The ideal contract for someone with limited self-control provides for a subsidized return on saving, an intermediate degree of commitment, and fixed fees that pay for the subsidies. Unfortunately, consumers with perfect self-control are also drawn to these contracts. They end up saving more on average, and therefore receive higher subsidies, which prevents the contract provider from breaking even. The

[75] According to Bandura and Kupers (1964), people "often set themselves relatively explicit criteria of achievement, failure to meet which is considered undeserving of self-reward and may elicit self-denial or even self-punitive responses..." See also Bandura (1971, 1976); Mischel (1973); Rehm (1977); Kazdin (2012); Ainslie (1975, 1991, 1992).

provider has to take this self-selection into account.

Galperti characterizes optimal contract provision for a monopolist and for a benevolent planner. In each case, the optimal menu specifies a contract for both types of consumers, and is designed so that those with and without self-control both prefer their intended option. The solution has some interesting and intuitive properties. First, the optimal contract for those without self-control may specify minimum and maximum levels of saving. The point of these provisions is to limit the contract's attractiveness to those with self-control. This finding suggests a possible rationalization for the observation that contributions to tax-favored savings accounts such as IRAs and 401(k)s are capped. Second, the optimal contract for those with self-control typically includes an unused detrimental alternative that those without self-control would find irresistibly tempting. The purpose of this provision is likewise to discourage imitation.

The special features of Galperti's model may limit its applicability. Most notably, there are no "outside" saving or borrowing options. Consumers must choose between one of the two contracts, and have no other means of moving resources across time. The absence of heterogeneity with respect to consumer sophistication is also likely important. Still, the crisp intuitions behind the key findings suggest the possibility of generalization.

The task of implementing optimal capital income tax analyses empirically for reasonably realistic settings with behavioral consumers would appear challenging. In addition to addressing various theoretical complexities, one would need to measure the joint distribution of present-focus, savings elasticities, and factors influencing the demand for flexibility. The necessary inputs for such an investigation are not found in existing empirical studies.

4.5 Special savings accounts: default options

Starting with Madrian and Shea (2001), a number of studies have found that changing the default contribution rate for a 401(k) pension plan has a powerful effect on employees contributions, particularly compared with conventional policy instruments such as capital income taxes; see also Choi et al. (2002, 2004, 2005, 2006); Beshears et al. (2008); Carroll et al. (2009).[76] Yet the selection of default options has received far less attention. Only a few studies, discussed below, have explicitly examined their use as policy instruments.

4.5.1 Theories of default effects

How should employers and policy makers exploit default effects, if at all? Several proposals have surfaced in the years since Madrian and Shea (2001) first documented the phenomenon.

One idea is to set 401(k) defaults so as to maximize contributions (Thaler and Sunstein, 2008). Support for this objective emanates from the belief that consumers save too little. While some unabashedly defend that judgment on paternalistic grounds, others insist that the inadequacy of saving is an objective consequence of self-control problems. The theoretical relevance of self-control

[76] Bronchetti et al. (2013) describe a related context in which no default effect is observed.

is questionable, however, in light of the fact that workers make 401(k) contribution elections well in advance of implementation, which generally occurs in a subsequent pay cycle, so that all consequences of these decisions lie outside the time window usually associated with present-focused tendencies.

A second idea is to set 401(k) defaults with the object of minimizing the frequency with which people opt out. Thaler and Sunstein (2003) advocate this approach, offering as informal justification a principle of ex post validation (meaning that those who stick with the default evidently consider it acceptable). However, they do not articulate an objective function that would rationalize this criterion.

A third idea is to structure 401(k)s so that all employees must make active decisions, with the object of ensuring that contribution rates reflect actual preferences (Carroll et al., 2009). A conceptual difficulty with this approach is that the result is necessarily contrary to the preferences of anyone who would rather avoid the costs of making a contribution election. It is also worth noting that an active-choice requirement is equivalent to *maximizing* the frequency with which people opt out of the default option. In that sense, the second and third proposals are diametrically opposed.

To sort out the welfare effects of default contributions rates, one must first understand the nature of default effects. Several theories merit consideration. First, defaults can influence the choices of rational consumers in settings where opt-out entails significant costs. However, according to DellaVigna (2009) and Bernheim et al. (2015a), 401(k) opt-out costs would have to be implausibly large to account for the magnitude of default effects. Second, to the extent opting out requires effort and workers are time-inconsistent, they may procrastinate with respect to making 401(k) elections. This theory also encounters difficulties. If consumers are sophisticated with respect to their time inconsistency, then for reasonable parameterizations of preferences, default effects would not be much larger than for the first theory. Naivete can rationalize much larger default effects under the assumption that little learning occurs: workers must cling to false beliefs about the likelihood of near-term action even though experience falsifies that belief pay period after pay period. Third, inertia may reflect inattention. While large default effects are equally problematic for theories of rational inattention, consumers may deploy their attention suboptimally. Finally, a default may provide a psychological "anchor" in a setting where workers are unclear about their own preferences.

Throughout most of this section, we will assume for the purpose of illustration that default effects arise from sophisticated time inconsistency (quasi-hyperbolic discounting), even though that is not the most plausible explanation. Here, however, we adopt a less restrictive perspective on welfare than in previous subsections, allowing instead for the possibility that decisions with (some) immediate consequences may have as much normative validity as decisions with (only) delayed consequences. We briefly discuss implications of other theories in the final subsection.

4.5.2 Optimal defaults with sophisticated time inconsistency

A simple model. A three-period model based on Bernheim et al. (2015a) suffices to illustrate the key insights concerning optimal defaults. The worker's task is to choose the level of some period-1 action, $x \in [x_{min}, x_{max}]$. She makes the decision either in period 0 (if commitments are allowed) or in period 1. Either way, her options are to take no action and accept a default, $x = D$, or expend period-1 effort to select an alternative. Active choice entails an immediate utility cost of γ. In period 2, she receives x along with income m, which together deliver utility of $v(x, x^*) + m$, where $x^* \in [x_{min}, x_{max}]$, her ideal point, varies across the population. For simplicity, we also assume each individual's preferences are single-peaked in x. The period-2 utility loss from receiving an option other than the ideal point is

$$\Delta(D, x^*) = v(x^*, x^*) - v(D, x^*)$$

With respect to intertemporal tradeoffs, the worker is a quasi-hyperbolic discounter, with $\delta = 1$ (for simplicity) and $\beta \in [0, 1]$.

Throughout, we assume that the same value of x^* governs contribution elections both in naturally occurring decision problems and within the welfare-relevant domain. As justification, we reiterate that the consequences of these decisions lie outside the time window usually associated with present-focused tendencies. For the purpose of this discussion, we abstract from the important possibility that consumers may misapprehend their ideal points due to a lack of financial sophistication.

Conditional on opting out, the worker will plainly choose x^*. Whether she opts out depends on the timing of her decision. Her optimal choice rule takes the following form: accept the default when $\Delta(D, x^*) < \gamma/\beta_c$, otherwise opt out.[77] When making the decision in period 0 (as a commitment), $\beta_c = 1$; when making it in period 1 (contemporaneously), $\beta_c = \beta$.

Evaluating outcomes. In Sections 4.2 through 4.4, we followed the common practice of treating β as a bias, which amounts to respecting choices made only in period 0. As noted in Section 6, the justification for this normative perspective is subject to debate. Here we explore the robustness of policy prescriptions by examining optimal defaults taking the welfare-relevant domain to be either period-0 choices, period-1 choices, or both.

To evaluate welfare, we compute equivalent variations (EVs) for changes in the default option, using $x = x^*$ (the first-best) as the baseline outcome. For those who do not opt out, the EV is $-\Delta(D, x^*)$. For those who do opt out, the EV is $-\frac{\gamma}{b_e}$, where b_e is the discount factor used for welfare evaluation and e is the frame of evaluation. We focus here on decisions made without commitment (that is, in period 1), and consider both possible evaluation frames: $e = 0$, which assesses outcomes based on period 0 choices (so $b_0 = 1$), and $e = 1$, which uses period 1 choices (so $b_1 = \beta$). Letting P denote the fraction of the population satisfying a stated condition, we can write the aggregate (average) EV from the perspective of evaluation frame $e = 1$ as:

$$EV_B = -\frac{\gamma}{\beta} P\left(\Delta(D, x^*) > \frac{\gamma}{\beta}\right) - P\left(\Delta(D, x^*) < \frac{\gamma}{\beta}\right) E\left(\Delta(D, x^*) \mid \Delta(D, x^*) < \frac{\gamma}{\beta}\right) \qquad (10)$$

[77] We adopt the convention of resolving indifference in favor of opting out for the case of equality.

For evaluation from $e = 0$, the analogous expression is

$$EV_A = EV_B + \gamma \left(\frac{1}{\beta} - 1 \right) P \left(\Delta(D, x^*) > \frac{\gamma}{\beta} \right) \tag{11}$$

Notice that $EV_A > EV_B$: the monetary equivalent of a failure to elect x^* is greater when evaluating outcomes according to the period-1 frame, because the worker attaches more importance to period-1 effort costs in period 1 than in period 0.

Optimal defaults. First we take the welfare-relevant domain to consist of period-1 choices ($e = 1$), so that the decision criterion and the welfare criterion agree, as in a setting with time consistency. From an inspection of equation 10, one can see that, as a general rule, EV_B tends to reach local maxima with respect to D within the most highly concentrated portions of the ideal-point distribution: when $P\left(\Delta(D, x^*) > \frac{\gamma}{\beta}\right)$ is smaller, fewer workers incur the maximal welfare loss, $\frac{\gamma}{\beta}$; in addition, the average loss among those who accept the default, $E\left(\Delta(D, x^*) \mid \Delta(D, x^*) < \frac{\gamma}{\beta}\right)$, tends to be smaller when the density of ideal points achieves a local maximum at D. Accordingly, the most natural candidates for optimal defaults include the central point of the ideal-point distribution, the smallest and largest allowable contributions (x_{min} and x_{max}), and any common kink-points in the function $v(\cdot, x^*)$ (arising, for example, from caps on matching contributions by employers).

Notably, the Thaler and Sunstein (2003) opt-out-minimization criterion, which prescribes maximization of $\mathrm{P}\left(\Delta(D, x^*) < \frac{\gamma}{\beta}\right)$, delivers similar policy recommendations. While the two criteria often agree in practice, they can also diverge significantly; see Bernheim et al. (2015a). However, the optimal policy converges to opt-out minimization as $\gamma \to 0$ (Bernheim and Mueller-Gastell, 2018).

As an illustration, consider the special case in which the loss function is quadratic ($\Delta(D, x^*) = \mu (D - x^*)^2$) and the distribution of x^* is single-peaked and symmetric around $\bar{x}$. Then it is easy to check that the first-order condition, $\frac{dEV_B}{dD} = 0$, is satisfied when the default coincides with the median bliss point ($D = \bar{x}$), in which case the opt-out frequency is minimized.

Next we take the welfare-relevant domain to consist of period 0 choices ($e = 0$). According to equation 11, the welfare criterion EV_A consists of two components. The first is simply EV_B, the criterion we applied above. The second is the opt-out frequency, $\mathrm{P}\left(\Delta(D, x^*) > \frac{\gamma}{\beta}\right)$, multiplied by a positive weight, $\gamma\left(\frac{1}{\beta} - 1\right)$. The presence of the second term shifts the welfare objective in the direction of opt-out *maximization*. Carroll et al. (2009) show that, for sufficiently low β, the solution involves an extreme default that compels active choice. In contrast, for higher values of β, the logic of maximizing EV_B takes over. As a result, in plausible special cases (e.g., with a quadratic loss function, as defined above), the optimal policy involves either the minimization or maximization of opt-out frequencies, depending on whether β is above or below a threshold (Carroll et al., 2009).

Bernheim and Mueller-Gastell (2018) argue that, with a richer and more realistic set of policy instruments, an employer should never seek to incentivize opt-out by setting undesirable defaults. According to their analysis, the optimal strategy is to correct the opt-out decision by imposing a fee on passive choosers while balancing the employer's budget through a general transfer, and then to set the default rate as if no bias exists.[78] For small opt-out costs (γ) and other natural special cases, it then follows that optimal defaults minimize the opt-out frequency (conditional on the optimal penalty) irrespective of decision bias.

Supposing once more that the employer only sets a default contribution rate, what if one remains agnostic about biases and treat all choices as welfare-relevant? In those settings, EV_A becomes an upper bound on the aggregate equivalent variation and EV_B becomes a lower bound. If we assume framing effects are large enough to account for the powerful influence of defaults on choices, then the region of indeterminacy between EV_A and EV_B is necessarily large. However, Bernheim et al. (2015a) use empirically calibrated models to show that the shapes of the EV_A and EV_B (versus D) functions are similar for 401(k) contribution rates below 20%. Within that range, the optimal default is generally insensitive to the decision frame.

4.5.3 Optimal defaults under other theories

As noted in Section 4.5.1, models of sophisticated time inconsistency have difficulty accounting for the observed magnitude of the default effect for 401(k) contribution. Theories involving naive time inconsistency, irrational inattention, and anchoring are potentially more plausible. Bernheim et al. (2015a) and Goldin and Reck (2017) explore their implications for optimal defaults. Because we view the Bernheim-Rangel framework as a generalization of the Behavioral Revealed Preference paradigm (see Section 2.2.3), our discussion will employ the vocabulary of the former.

For any particular theory, one must first make the potential decision frames explicit, and then take a stand on which frames are welfare-relevant. The current application raises no special issues concerning the definition of decision frames for theories involving sophisticated or naive time inconsistency. However, the cases of irrational inattention and anchoring are more complicated. For those theories, it is tempting to think of the default rate, D, as the frame, inasmuch as it may trigger attention or serve as a psychological anchor. However, that approach is conceptually problematic. By definition, decision frames are conditions that do not affect opportunities. Whenever opt-out entails non-negligible costs, changing D changes the opportunity set. Therefore, the default rate cannot be part of a properly defined decision frame.

One solution to this difficulty is to nest the problem of interest within a more general environment that separates the default framing from the practical consequences of establishing a default. In naturally occurring settings, one can describe those consequences by an effort-cost schedule that drops discontinuously at the default. More generally, however, one could contrive arbitrary effort-cost schedules, for instance by varying the processing requirements across the potential contribution rates, and possibly introducing burdens on passive choosers. Equipped with a choice mapping defined over this broader domain, one can easily identify properly defined frames and framing effects. Models of attention and anchoring permit one to infer this generalized choice mapping, which in turn enables applications of the Bernheim-Rangel apparatus.

[78] See also *Bernheim et al. (2015a)*, who consider dissipative penalties for passive choice.

Three of the theories we have mentioned – sophisticated time inconsistency, naive time inconsistency, and irrational inattention – have the property that the ideal outcome according to the consumer's perceptions, x^*, does not depend on the default framing. Bernheim et al. (2015a) show that these theories have similar implications for optimal defaults, for similar reasons, though the details differ. Anchoring belongs in a separate category, because it implies that x^* depends on D. This feature of the anchoring theory potentially induces a high degree of welfare ambiguity, and may preclude one from reaching useful conclusions absent a refinement of the welfare-relevant domain. One possible refinement is to evaluate welfare in a "neutral frame," corresponding to the default D that induces the same x^* as an active choice regime (one without a default). This refinement may be particularly appealing if, for example, anchoring effects reflect the incorrect belief that defaults embody authoritative advice. According to the empirical analysis in Bernheim et al. (2015a), this perspective leads to the conclusion that consumer surplus varies to only a small degree with the default. Because higher contributions entail costs to employers and the government via matching and tax breaks, the socially optimal default rate is then zero.

4.6 Financial education and choice simplification

As noted in Section 6, low levels of literacy raise concerns about the general quality of financial decision making. In this section, we discuss two types of policy responses: financial education, which aims to improve decisions by helping consumers acquire the basic knowledge and skills they need to understand the choices they face, and choice simplification requirements, which seek to render the consequences of financial choices more transparent.

4.6.1 The behavioral effects of financial education

The term "financial education" subsumes a wide range of diverse interventions. Most of these programs fall into two broad categories, according to whether they are employer-based or school-based. Employers provide the lion's share of adult financial education in the U.S.[79] They typically engage professional consultants whose offerings tend to be brief but highly polished.[80] Brevity is, in effect, a design constraint: thorough educational programs are not only costly but also time-consuming, which makes them unappealing to workers. To compensate for brevity, these programs generally focus on simple heuristics accompanied by highly motivating messages. The intent is to make the substantive material engaging, memorable, and actionable. In contrast, high school courses often span a full semester, permitting a more expansive and in-depth treatment of subject

[79] In a 2013 survey of 407 retirement plan sponsors covering more than 10 million workers by Aon Hewitt, 77% of providers offered on-site financial education seminars or meetings (Austin and Evens, 2013). In the 2015 FINRA National Financial Capability Study, 40.24% of respondents aged 20 - 65 who have received financial education did so through an employer.

[80] A meta-analysis by Fernandes, Lynch Jr and Netemeyer (2014) finds that the average financial education program involves only 9.7 hours of instruction. That time is usually divided among a long list of complex topics. For example, Skimmyhorn (2015) reports that a financial education program used by the U.S. military covers compound interest, the focus of our current study, along with a collection of several more complex topics – retirement concepts, the Thrift Savings Plan, military retirement programs, and investments – all within a single two-hour session.

matter, as well as more interactive pedagogy, including practice and discussion. However, teacher qualifications and experience vary considerably from school to school (Brown et al., 2014).

In light of this diversity, one would hardly expect all programs to affect behavior similarly. Even educational interventions that achieve similar improvements in tested comprehension may have dissimilar effects on behavior, depending on the particular manner in which each intervention motivates participants, and whether it helps them learn to internalize and operationalize conceptual knowledge rather than directional imperatives. From its inception, the literature has studied workplace and school-based programs separately (beginning with Bernheim, Garrett and Maki, 2001a, and Bernheim and Garrett, 2003), but has only recently begun to explore the heterogeneity of approaches within each category, and to examine how the effects of an intervention depend on its design and constituent components. Increasingly, the literature relies on controlled experiments rather than naturally occurring data. The experimental approach offers important advantages in settings where naturalistic interventions are highly composite and heterogeneous. Programmatic diversity may help to explain why different authors have reached different conclusions concerning the behavioral effects of financial education; see, for example, Duflo and Saez (2003), Bayer et al. (2009), Bayer, Bernheim and Scholz (2009), Goda, Manchester and Sojourner (2012), Cole and Shastry (2012), Cole, Sampson and Zia (2011), Skimmyhorn (2012), Servon and Kaestner (2008), Collins (2010), Lührmann, Serra-Garcia and Winter (2014), Mandell (2009), Bertrand and Morse (2011), Drexler, Fischer and Schoar (2014), Carlin, Jiang and Spiller (2014), Heinberg et al. (2014), Lusardi et al. (2014), and Brown et al. (2014), as well as the chapter on personal financial decision making in this volume, Beshears et al. (forthcoming). Recent surveys by Hastings, Madrian and Skimmyhorn (2013) and Lusardi and Mitchell (2014a) underscore the mixed nature of the available empirical evidence.

4.6.2 The welfare effects of financial education

The welfare effects of financial education are far from obvious. Discussions of this issue often proceed from preconceptions, such as the notion that people would be better off with high saving and balanced portfolios, or that a better understanding of financial concepts necessarily promotes better decisions. Yet it is also possible that particular interventions alter behavior through mechanisms that involve indoctrination, exhortation, deference to authority, social pressure, or psychological anchors. If so, their benefits are unclear.

These concerns are particularly acute for workplace interventions. As noted above, employer-sponsored programs typically compensate for brevity by offering simple heuristics and emphasizing motivational rhetoric. Compelling rhetoric may distract from substance and promote a one-size-fits-all response, which may be excessive for some and even directionally inappropriate for others.

Methods for evaluating the quality of financial decision making In principle, one can empirically evaluate the quality of decision making in the financial domain using any of the strategies discussed in Section 2.2.4. For example, Ambuehl et al. (2017) deploy the strategy of implementing reframed decision problems. They introduce a measure of *financial competence* based on discrepancies between choices in equivalent valuation tasks. Specifically, they compare a consumer's willingness to accept (WTA) for two equivalent claims on future income, where one is a simplified version of the other. The simple version states the future claim transparently. The complex version packages the claim as an income-generating asset, designed so that the consumer requires a knowledge of targeted financial principles to infer the claim, and hence to understand the equivalence between the simple and complex versions. Someone who both possesses and fully operationalizes that knowledge will consistently ascribe the same value to both claims regardless of their preferences and/or other decision biases. When consumers' WTAs for equivalent claims differ, the magnitude of the discrepancy provides an intuitively appealing measure of her competence to make good decisions in contexts involving the pertinent principles. Subject to the second-best considerations discussed in Section 2.2.6, it also has a precise welfare interpretation: it indicates the extent to which the consumer's incomplete operational command of the principles that govern the equivalence exposes her to losses.

To illustrate, say one is concerned that people poorly understand the concept of compound interest, and that this limitation causes them to make suboptimal investment decisions. To evaluate this possibility, one might assess the consumer's WTA for pairs of equivalent claims such as the following: the complex claim represents a \$10 investment that promises a return of 6% per day compounded daily for 15 days while the simple claim simply promises \$24 in 15 days. Ordinarily, a consumer will be willing to choose each asset over a fixed sum of money if and only if the sum does not exceed some threshold value, call it p^* for the first claim and q^* for the second. A quick calculation reveals that the two claims are equivalent, subject to rounding. Thus, swapping out one for the other in a decision problem changes framing while leaving opportunities intact. As a general matter, any education intervention that successfully provides subjects with an operational understanding of compound interest should bring p^* into closer alignment with q^*. Furthermore, $|p^* - q^*|$ bounds the magnitude of the welfare loss resulting from the consumer's poor comprehension of the complexly framed decision problem.

As discussed in Section 2.2.4, this method allows the analyst to measure decision-making quality in settings that implicate preferences without making strong assumptions about behavioral or cognitive processes. Also, as mentioned in Section 2.2.6, one can defend the resulting welfare measure against second-best critiques. The portability of the approach may be limited, however, because complex naturally occurring investment tasks do not necessarily lend themselves to transparent simplifications.

The literature on financial education has also explored other methods mentioned in Section 2.2.4. For example:

- Song (2015) deploys structural methods to evaluate the welfare effects of changes in retirement contributions resulting from an educational intervention targeting compound interest. His analysis hinges on the accuracy with which a particular life-cycle model, calibrated with data drawn from other choice domains, describes lifetime opportunities, unobserved future choices, and "true" preferences.

- A variety of studies, including Ernst et al. (2004), Calvet et al. (2007, 2009), Agarwal et al. (2009), Baltussen and Post (2011), and Choi et al. (2011) gauge the quality of financial decision making using dominance methods. Aufenanger et al. (2016) deploys this approach (and others) to evaluate the effects of financial education.
- Choi, Kariv, Mueller and Silverman (2014) assesses the quality of financial decision making by measuring the extent to which choices violate revealed preference axioms. The suitability of this method for evaluating financial education is unclear, because educational interventions do not target conformance with WARP directly, and non-conformance may result from a variety of considerations that are unrelated to the consumer's understanding of specific financial principles.

Welfare evaluations of financial education interventions Ambuehl et al. (2017) evaluate the welfare effects of an educational intervention on compound interest, one of the fundamental concepts in personal finance. It resembles typical employer-sponsored interventions with respect to its brevity, as well as its emphasis on heuristics and motivational messages. It also appears to be highly effective according to conventional outcome measures: treated subjects perform substantially better on an incentivized financial literacy test, they report applying their newly gained knowledge when performing the decision tasks we assign them, and their average WTAs for interest-bearing assets change in a direction that counteracts the previously documented tendency to underestimate compounding (exponential growth bias). Nevertheless, they find that the intervention does not, on average, improve the quality of decision making, because its effects are poorly correlated with initial biases.

A possible explanation for this finding is that subjects may interpret motivational rhetoric as substantive advice and, even when their tested knowledge improves, emerge with an insufficient *operational* understanding of financial concepts to make appropriate adjustments. To explore this hypothesis, the authors implement two additional variants of the intervention, one that retains its substantive elements but omits the motivational rhetoric, and another that retains the motivational rhetoric but omits almost all of the substance. They show that the effects on financial literacy and self-reported decision strategies are primarily attributable to the substantive elements of instruction, as one would hope. However, in sharp contrast, the effects on financial choices are primarily attributable to the non-substantive elements. In particular, the intervention's motivational rhetoric increases subjects' WTA for interest-bearing assets regardless of the extent to which any particular individual initially understates or overstates the effects of compounding. This indiscriminate response is beneficial in some cases and harmful in others; on average, there is no benefit. When stripped of motivational rhetoric, exclusively substantive instruction has some effect on behavior, and it does reduce reliance on simple interest calculations (the most common type of mistake), but it fails to promote reliance on correct compound interest calculations, instead increasing the prevalence of other mistakes. As a result, its impact on WTAs for interest-bearing assets is directionally haphazard and, on average, welfare-neutral.

Other studies have reached similarly discouraging conclusions concerning the welfare effects of financial education. For instance, using the structural approach, Song (2015) also finds that the effect of an educational intervention involving compound interest is indiscriminate: the impact on measured saving is not closely related to the gap between actual and optimal rates implied by a parameterized life-cycle consumption model, and the intervention induces some people to oversave. See also Aufenanger et al. (2016) and Bruhn et al. (2016).

4.6.3 Choice simplification

Choice simplification requirements aim to mitigate the consequences of low financial sophistication by rendering the consequences of financial choices more transparent. In the language of Ambuehl et al. (2017), such policies amount to replacing naturally occurring, complexly framed decision tasks with their simply framed counterparts on a widespread basis in the real world, rather than on a limited basis merely for the purpose of diagnosis and evaluation. As noted in Section 2.2.6, second-best considerations arising from the possible existence of other decision-making biases can undermine the general case for transparency. Consequently, formal justifications for choice simplification implicitly hinge on the perspective of idealized welfare analysis.

Field evidence on the effects of choice simplification is mixed. Beshears et al. (2013) show that simplified options for retirement plans that collapse a highly multidimensional problem into a simple binary choice can increase enrollment rates by 10 to 20 percentage points. It does not follow, however, that the increase reflects an improved understanding of consequences. Instead, it may simply involve an aversion to complexity. Indeed, in another context, Beshears et al. (2011) find no evidence that the providing information concerning mutual fund features through a simplified Summary Prospectus rather than a statutory prospectus meaningfully influences portfolio choices.

The main challenge facing advocates of choice simplification is the problem of determining which presentations of information actually render the consequences of complex, real-world choices more comprehensible to consumers. The welfare effects of ostensibly "simpler" presentations that are in fact contrived to nudge consumers in predetermined directions are unclear. Perhaps the most promising strategies for achieving neutral improvements in transparency involve the use of visualization tools that provide consumers with free reign to explore the consequences of available options (Lusardi et al., 2014).

4.7 Mandatory saving

The previous sections focus on policies that seek to induce "good" financial decision making by modifying consumers' incentives, information, and/or motivations. As an alternative, the government could simply take these choices out of consumers' hands and save on their behalf. This approach is widely used: developed economies generally mandate participation in public pension programs, which exist side-by-side with opportunities for private saving. An important branch of the literature explores the design of these programs.

When devising a universal system of mandatory saving, it is essential to bear in mind that the population is highly heterogeneous. Some dimensions of this heterogeneity are unobservable. People differ with respect to important characteristics that the government cannot directly measure, such as the degree of susceptibility to the cognitive biases that motivate the mandate. A simple saving requirement that employs a one-size-fits-all structure treats everyone identically, which is plainly not ideal. Other dimensions of heterogeneity are observable. Conditioning on measurable characteristics allows the government to achieve distributional objectives. The literature explores the ways in which the corrective and distributional aims of mandatory saving programs interact.

Cremer et al. (2008) study settings in which people earn different wages and exhibit differing degrees of myopia (defined as an assumed discrepancy between the discount rates governing decisions and normative judgments).[81] They choose labor supply and saving when young, and consume the returns to saving when old, possibly subject to a liquidity constraint. The government observes and taxes earnings to finance a public pension benefit, which is linear in earned income. At one extreme (a "Bismarckian system"), each individual receives the returns to the taxes they paid. At the other extreme (a "Beveridgean system"), everyone receives the same benefit. The planner's problem is to determine the size of the program (the tax rate) and the degree of redistribution (the slope of the linear function relating pension benefits to earnings). The authors investigate the manner in which the prevalence of myopic consumers affects the optimal policy parameters. Numerical simulations show that, in the absence of liquidity constraints, both the generosity and redistributiveness of the program increase as "myopics" become more numerous. However, only the first of those results survives when liquidity constraints are introduced. The degree to which these results depend on assumptions about functional forms is unclear.

A significant limitation of the Cremer et al. (2008) analysis is that it does not contemplate the relative merits of addressing the government's objectives through mandatory saving rather than through the various incentive strategies discussed in the preceding subsections. Moser and de Souza e Sivla (2015) fill this gap by examining a related model that likewise depicts heterogeneity in earnings as well as in the degree of present focus.[82] They demonstrate that the optimal policy offers low-income individuals a one-size-fits-all savings instrument resembling social security. In contrast, it offers high-income individuals a set of policies resembling specialized savings accounts that accommodate heterogeneous preferences. The system uses flexibility for high earners as a reward in order to generate the revenues required for redistribution. Moser and Silva conclude that the design of the existing U.S. system of retirement saving is inefficient.

Other papers explore additional dimensions of the design problem in settings with behavioral agents. For instance, Cremer et al. (2009) and Tenhunen and Tuomala (12) allow for nonlinear

[81] See also Findley and Caliendo (2009). The literature on the optimal level of social security benefits appears to originate with Feldstein (1985).

[82] See also Fehr and Kindermann (2010), who compared the merits of a standard social security program with a system of private savings accounts.

pension formulas; Imrohoroglu et al. (2003) study unfunded social security systems within an overlapping generations framework (see also Fehr et al., 2008); Pestieau and Possen (2008) add incentive problems arising from ex post altruism ("rational prodigality," also known as the Samaritan's dilemma); and Cremer et al. (2007) examine the political economy of program design. For surveys of this literature, see Findley and Caliendo (2008) and Cremer and Pestieau (2011).

4.8 Other policies

The preceding discussion of policies targeting saving is by no means complete. Here we briefly mention a few other classes of policy levers.

Some analysts argue that low levels of saving in the U.S. are at least partially attributable to policies that promote ready access to credit. Easy credit removes a consumer's ability to accumulate illiquid assets. Consequently, its effects are opposite those of providing commitment opportunities (see Section 4.3). Laibson (1997) analyzes the effects of access to credit for QHD consumers, limiting attention to Markov-perfect behavior. He demonstrates that an increased ability to borrow against otherwise illiquid assets reduces the steady-state capital-output ratio, and causes a substantial reduction in welfare. He points to the 1980s as a period of rapid expansion in U.S. consumer credit due to the spread of credit cards and ATM machines, and suggests that these developments may have undermined self-control. However, Bernheim et al. (2015b) demonstrate that easier access to credit can enhance a consumer's ability to self-regulate through personal strategies involving contingent reward and punishment. Karlan and Zinman (2010) present empirical evidence that calls Laibson's conclusions into question. They conducted a field experiment that expanded access to costly consumer credit in South Africa, and found that on average the intervention improved economic self-sufficiency, intra-household control, community status, and overall optimism.

Policies affecting the composition of income may also influence overall rates of saving. Shefrin and Thaler (1988) argue that the tendency for people to think of their assets and income streams as belonging to different "mental accounts," and to associate different accounts with different purposes, causes the marginal propensity to consume to differ according to the nature of the resources. As an example, imagine that people view dividends as spendable income and capital gains as long-term saving. A policy that induces corporations to reduce their dividend-payout rates will shift investors' earnings from the former category to the latter, thereby increasing saving under the Shefrin-Thaler hypothesis. Other policies that change the form and/or timing of cash receipts, such as bonuses and income tax withholding, may have similar effects. For example, Jones (2012) concludes that changes in withholding are likely non-neutral.[83]

To the extent consumers are periodically inattentive to financial decisions, policies that promote reminders may also improve their outcomes. Karlan et al. (2016a) provide evidence that reminders are indeed effective at improving follow-through on intentions to save. It is unclear, however, whether the mechanism involves attention or some form of social pressure (e.g., brow-beating).

Bernheim (1991) mentions the possibility that governments could also attempt to enhance the

[83] Jones (2012) attributes these non-neutralities to consumer inertia rather than mental accounting.

salience of saving decisions and the psychological appeal of future-oriented behavior through promotional campaigns. Unfortunately, evidence concerning the effectiveness of these policies is mostly limited to anecdotes, such as the experience of Japan after World War II. Even so, related evidence suggests that these types of promotional efforts may be effective. For example, a field experiment by Bertrand et al. (2010) shows that non-substantive promotional content, such as including a photo of an attractive woman, significantly increases the take-up rate for loan offers. See Sections 2.4 and 3.5 for discussions of how one might evaluate the welfare effects of these types of "nudges."

5 Policies Targeting Earnings

Although studies of optimal income taxation constitute one of the oldest and largest literatures in Public Economics, the field of BPE has only recently begun to explore these questions. We introduce a simple model of taxes on earnings in Section 5.1, which we use to study the implications of biases that intrinsically affect how people trade off labor costs against consumption (Section 5.2), or that involve misperceptions of the taxes (Section 5.3). The latter biases create important methodological difficulties for standard approaches to optimal income taxation—namely, the "mechanism design" approach—which we discuss in Section 5.4. The existence of perceptual and attentional biases can also overturn the classical Atkinson and Stiglitz (1976) results about the optimal use of commodity taxes in the presence of nonlinear income taxation, as we explain in Section 5.4.

We end by applying lessons about income taxation to questions concerning social insurance in Section 5.6, and by discussing other miscellaneous questions such as tax filing and tax compliance in Section 5.7.

5.1 A Stylized Model of Income Taxation with Behavioral Consumers

We formulate a behavioral extension of the Sheshinski (1972) model of social insurance, which simplifies the standard mechanism design problem (e.g., Mirrlees 1971; Saez 2001) by assuming that the tax-transfer schedule is linear. Farhi and Gabaix (2015) provide a general analysis of optimal nonlinear income taxation with behavioral agents, and generalize the basic lessons learned from an analysis of linear income taxation.

There is a continuum of individuals with differing skill levels θ, distributed according to a probability measure μ. Type θ must work x/θ hours to generate (before-tax) income x. The government imposes a linear tax rate τ on income, which it uses to fund a lump-sum grant R. Consumption is therefore $c = (1-\tau)x + R$. Welfare-relevant choices are governed by the utility function $V(c, x; \theta) = v(c - h(x/\theta))$. In the naturally occurring decision frame, type θ chooses a level of earnings $x_\theta(\tau)$ that may not maximize V.

The expression $\bar{x}(\tau) = \int x_\theta(\tau) d\mu$ represents average earnings, and $\varepsilon_{\bar{x},1-\tau}$ denotes the elasticity of average earnings with respect to the net-of-tax-rate. This response reflects moral hazard: consumers work less if they must pay high taxes and/or are provided with a generous social safety net. It is analogous to moral hazard arising from health insurance (e.g., people buy more medication than they should, or invest less in staying healthy) and unemployment insurance (e.g., people exert less effort to maintain or find employment), and thus the insights from this model apply to those settings as well.

As in Section 3 on commodity taxation, we define $\gamma_\theta := \frac{h'(x_\theta(\tau)/\theta)}{1-\tau} - 1$ as the price-metric measure bias. Because $x^*_\theta(\tau)$ is a function of the tax-rate τ, so is γ_θ, but we will typically omit the argument for brevity.

Because a taxpayer gets to keep the fraction $(1-\tau)$ of the marginal unit of income, she chooses earnings x_θ to satisfy $h'(x_\theta(\tau)/\theta) = 1-\tau$. Thus, positive γ_θ means labor supply is too high, and there are welfare gains from discouraging work. Conversely, negative γ_θ means that labor supply is too low, and there are welfare gains from encouraging work.

As before, γ_θ has a simple interpretation. Because $h'(x_\theta(\tau))/\theta = (1-\tau)(1+\gamma_\theta)$, γ_θ measures the proportionate increase in the income retention rate, $1-\tau$, that would induce a taxpayer who maximizes welfare, V, to choose the same level of labor supply that he chooses in the naturally-occurring (bias-inducing) frame.

The policymaker chooses the tax rate τ to maximize

$$W = \int v(c(x_\theta(\tau)) - h(x_\theta(\tau)/\theta) + \tau\bar{x})d\mu.$$

subject to government budget balance. Consider the welfare impact of increasing the tax rate τ by some small amount, $d\tau$. This variation has the following effects, where ε_θ denotes a type θ's elasticity of income with respect to the income retention rate $1-\tau$:

- A direct utility cost $-v'x_\theta(\tau)d\tau$ to each individual earning x_θ.
- A mechanical increase in tax revenue equal to $dM = \bar{x}(\tau)d\tau$, raising each individual's utility $\bar{x}(\tau)v'd\tau$.
- An indirect effect on revenue due to substitution toward leisure, given by $\tau dx_\theta = -\frac{\tau}{1-\tau}\varepsilon_{x_\theta,1-\tau}x_\theta(\tau)d\tau$ for each individual. Averaging across individuals, the total effect of substitution on tax revenue is $-\frac{\tau}{1-\tau}\varepsilon_{\bar{x},1-\tau}\bar{x}(\tau)$.
- An indirect cost (or benefit) to each individual, due to substitution toward leisure, given by $v' \cdot [(1-\tau) - h'/\theta]dx_\theta = \gamma_\theta x_\theta(\tau)\varepsilon_\theta v' d\tau$.

For the welfare formulas that follow, let $\bar{v'}$ denote the population average of marginal utilities, given the tax rate. Also define $\alpha(\theta) := \frac{v'x_\theta(\tau)\varepsilon_{x_\theta,1-\tau}}{\bar{v'}\bar{x}(\tau)\varepsilon_{\bar{x},1-\tau}}$, which measures how a taxpayer's marginal utility of consumption, as well as responsiveness to the tax rate, compare to the population averages.[84] Then

[84] By "responsiveness" we mean $\frac{dx_\theta}{d(1-\tau)}$. Note that $\frac{dx_\theta}{d(1-\tau)} \propto \varepsilon_{x_\theta,1-\tau}x_\theta$.

$$
\begin{aligned}
W'(\tau) &= \int \left[-v' x_\theta(\tau) + \gamma_\theta x^*_\theta(\tau) \varepsilon_\theta v' + v' \bar{x}(\tau) - v' \frac{\tau}{1-\tau} \varepsilon_{\bar{x},1-\tau} \bar{x}(\tau) \right] d\mu \\
&= \int \left[v' \cdot (\bar{x}(\tau) - x_\theta(\tau)) \right] d\mu - \frac{\tau}{1-\tau} \bar{v}' \bar{x}(\tau) \varepsilon_{\bar{x},1-\tau} + \bar{v}' \bar{x}(\tau) \varepsilon_{\bar{x},1-\tau} \int \gamma_\theta \frac{x^*_\theta(\tau) \varepsilon_\theta v'}{\bar{v}' \bar{x}(\tau) \varepsilon_{\bar{x},1-\tau}} d\mu \\
&= \underbrace{-Cov[v', x_\theta(\tau)]}_{\text{Redistribution/insurance}} \underbrace{- \frac{\tau}{1-\tau} \bar{v}' \bar{x}(\tau) \varepsilon_{\bar{x},1-\tau}}_{\text{Moral hazard}} + \underbrace{\bar{v}' \bar{x}(\tau) \varepsilon_{\bar{x},1-\tau} \left(E[\gamma_\theta] + Cov[\gamma_\theta, \alpha(\theta)] \right)}_{\text{Bias correction}} \qquad (12)
\end{aligned}
$$

Because $W'(\tau) = 0$ at the optimal tax rate, the previous expression implies that the optimal tax rate must satisfy

$$
\frac{\tau}{1-\tau} = \underbrace{\frac{1}{\bar{x}(\tau) \bar{v}' \varepsilon_{\bar{x},1-\tau}}}_{\text{Moral hazard}} \left[\underbrace{-Cov[v', x_\theta(\tau)]}_{\text{Redistribution/insurance}} + \underbrace{\bar{v}' \bar{x}(\tau) \varepsilon_{\bar{x},1-\tau} \left(E[\gamma_\theta] + Cov[\gamma_\theta, \alpha(\theta)] \right)}_{\text{Bias correction}} \right] \qquad (13)
$$

We can decompose formula (12) into three key terms that determine the optimal marginal tax rate. As indicated by the labels, one term represents the value of redistribution (or social insurance more generally): higher tax rates allow for more redistribution from those with high wages/wealth and therefore low marginal utility of income to those with low wages/wealth and therefore high marginal utility of income. Another term captures the effects of moral hazard: higher taxes and therefore higher levels of redistribution disincentivize taxpayers from working. These two terms capture the standard economic forces that shape the optimal level of income taxation and social insurance more generally.

The third term reflects behavioral considerations: it is the gain from counteracting taxpayers' biases. When people work too little, increasing the tax rate has the additional *cost* of reducing their labor supply even more. When they work too much, increasing the tax rate has the additional *benefit* of reducing their hours. As in the case of commodity taxes, what matters is not only the average bias, but also the extent to which those with large biases i) have high marginal utilities of income (leading the policy maker to care more about their mistakes) and ii) exhibit high elasticities with respect to the tax.

5.2 Intrinsic biases in the consumption-labor tradeoff

We first discuss consumption-based biases under which people improperly trade off consumption and labor—or, in richer environments, make improper tradeoffs between multiple dimensions of consumption—but correctly understand the tax system. For example, studies such as Kaur et al. (2015), DellaVigna and Paserman (2005) and Augenblick et al. (2015) suggest that time inconsistency may be present in labor supply. Incorrect beliefs about returns to labor may also play a role.

Building on this literature, Lockwood (2016) observes that present bias has important implications for optimal income taxation, since labor generates a more immediate cost than the delayed benefits accrued from consuming the earned income. Lockwood (2016) presents a model in which people maximize $U = \beta c - \psi(l)$, where β is the degree of present focus and $\psi(l)$ is the cost of labor. The policymaker believes that they should be maximizing $V = c - \psi(l)$; that is, the normative criterion corresponds to "long-run" utility.

Formulas (12) and (13) are easily adapted to such labor-supply biases. Under present-bias, choices satisfy $h'(x^*_\theta/\theta)/\theta = \beta_\theta(1-\tau)$. In the absence of present-bias, they would satisfy $h'(x^*_\theta/\theta)/\theta = (1-\tau)$. Thus for this particular bias, $\gamma_\theta = \beta_\theta - 1$. Because consumers under-supply labor, a tax increase is particularly costly, so the optimal tax rate is lower than in the standard model.

To see the implications of present bias most clearly, suppose that β and that elasticities are homogeneous. Then formula (13) becomes

$$\frac{\tau}{1-\tau} = \underbrace{\frac{1}{\bar{x}\bar{v}'\varepsilon_{\bar{x},1-\tau}}}_{\text{Moral hazard}} \left[\underbrace{-Cov[v', x_\theta]}_{\text{Redistribution/insurance}} + \underbrace{\bar{v}'\bar{x}\varepsilon_{\bar{x},1-\tau}(\beta - 1)}_{\text{Bias correction}} \right].$$

The simplified formula shows that the marginal tax rate is increasing in the present bias β. In fact, if the taxable earnings elasticity $\varepsilon_{\bar{x},1-\tau}$ is sufficiently high and the present bias β is sufficiently low, then the optimal tax rate τ may be negative.

Gerritsen (2015) provides more general formulas for an optimal nonlinear income tax rate that allow for other consumption-based biases.

5.3 Biases induced by tax misperceptions

While there is currently little direct evidence that quasi-hyperbolic discounting or limited self-control more broadly contribute to misoptimized earnings, a growing literature documents prevalent confusion, use of heuristics, and inattention in the context of income taxation.

Formally, for any given τ, suppose people behave as if the tax rate is actually $\hat{\tau}(\tau, x^*_\theta, \theta)$. As we discuss later, $\hat{\tau}$ may depend not only on τ, but also on factors such as the individual's average tax rate, which varies with her income x^*_θ. Each taxpayer chooses labor supply to satisfy the first-order condition $h'/\theta = 1-\hat{\tau}$. This condition implies that $\gamma_\theta = \frac{1-\hat{\tau}}{1-\tau} - 1 = \frac{\tau-\hat{\tau}}{1-\tau}$. Substituting γ_θ into formula (13) yields an expression for the optimal income tax rate as a function of the misperceptions.

To glean more intuition via a concrete example, suppose consumers underreact to tax rates by a factor $1-\sigma$, perhaps because the taxes are not always salient, as discussed in Section 3.4. Then $\hat{\tau} = \sigma\tau$, and equation (12) becomes

$$W'(\tau) = \underbrace{-Cov[v', x_\theta]}_{\text{Redistribution/insurance}} \underbrace{-\frac{\tau}{1-\tau}\bar{v}'\bar{x}\varepsilon_{\bar{x},1-\tau}}_{\text{Moral hazard}} + \frac{\sigma\tau}{1-\tau}\underbrace{\bar{v}'\bar{x}\varepsilon_{\bar{x},1-\tau}}_{\text{Bias correction}},$$

which implies that the optimal tax rate satisfies

$$\frac{\tau}{1-\tau} = \frac{1}{1-\sigma}\frac{-Cov[v', x_\theta]}{\bar{x}\bar{v}'\varepsilon_{\bar{x},1-\tau}}. \tag{14}$$

Equation (14) formalizes the intuition that moral hazard costs decrease when people underreact to the income tax, which allows the policymaker to set a higher tax rate and thereby achieve greater redistribution.

Next we survey the empirical literature on tax perception biases.[85]

Confusion

When surveyed about the key parameters characterizing their federal income tax burdens, such as their marginal tax rates, taxpayers regularly report values that deviate significantly from the truth (Fujii and Hawley, 1988; Blaufus et al., 2013; Gideon, 2014; Rees-Jones and Taubinsky, 2018a).

Analysis of observational data reveals that knowledge of the tax code varies widely: comparing across geographic neighborhoods, Chetty et al. (2013) find significant differences in bunching at the refund-maximizing kink point of the earned income tax credit (EITC) schedule. Moreover, those who move from low-bunching to high bunching neighborhoods increase their EITC refunds, apparently due to improved information.

Taxpayers also leave significant tax benefits "on the table" every tax year through, for example, failures to claim itemized deductions (Benzarti, 2016) or the EITC (Bhargava and Manoli, 2015). Attempts to "teach the tax code" are largely ineffective on average, but can work when paired with expert advice (as in, e.g., Chetty and Saez, 2013).

Feldman et al. (2016) show that taxpayers also confuse changes in lump-sum transfers with changes in marginal tax rates. They examine the effect of the Child Tax Credit (CTC), a transfer given to households that include a child younger than 17 during the calendar year. The age-17 cutoff introduces a discontinuity in the average tax credit received, as a household whose child "ages out" on December 31, 2010 could not claim the CTC for 2010, whereas a household whose child "ages out" on January 1, 2011 could. Using a regression discontinuity design, the authors find that the loss of the CTC is associated with a relative decline in reported wage income of roughly 0.5 percent. They also show that this effect is not driven by efforts to time earnings strategically. This effect is notable because the CTC is a lump sum, which means it does not affect incentives to work on the margin. The income effects generated by the loss of this lump-sum transfer would lead individuals to work more, not less.[86]

Adoption of heuristics

Liebman and Zeckhauser (2004) describe two heuristics for approximating a convex schedule such as the US income tax.

People who use the first heuristic, *ironing*, know their average tax rates and assume that taxes are proportional to income. The forecasted tax at income x is then given by $\tilde{T}(x|x^*, \omega) = A(x^*|\omega) \cdot x$, where x^* denotes the individual's own income, ω denotes all individual-specific characteristics that determine the applicable tax schedule, and $A(x^*|\omega)$ denotes the individual's average tax rate. This

[85] The discussion here borrows from Rees-Jones and Taubinsky (2018b).

[86] Although income effects are generally estimated to be small. See, e.g., Gruber and Saez (2002).

heuristic leads to reasonably accurate beliefs about the *levels* of taxes when considering small deviations from one's current income.

Feldman et al. (2013) argue that this first heuristic potentially accounts for confusion over marginal tax rates, which they document, and de Bartolome (1995) documents similar responsiveness to average tax rate shocks in the laboratory . In a survey experiment directly eliciting perceptions of tax schedules, Rees-Jones and Taubinsky (2018a) find that 43% of US tax filers adopt the ironing heuristic.

People who use the second heuristic, *spotlighting*, know their own *marginal* tax rates (as well as their total liabilities), and assume the tax schedule is linear. Using the spotlighting heuristic, the forecasted tax at income x is given by $\tilde{T}(x|x^*, \omega) = T(x^*|\omega) + MTR(x^*|\omega) \cdot (x - x^*)$, where x^* again denotes the individual's own income, $MTR(x^*|\omega)$ denotes the marginal tax rate at that income, and $T(x^*|\omega)$ denotes the true tax due at that income. Within one's own tax bracket, this heuristic leads to correct beliefs about the level and slope of the tax schedule. While this heuristic has received some theoretical attention, Rees-Jones and Taubinsky (2018a) find little experimental evidence that people adopt it. However, more empirical work on the spotlighting heuristic is needed.

Significantly, Rees-Jones and Taubinsky (2018a) find that the ironing heuristic explains most of the systematic misperceptions of the federal income tax that they document, including underestimation of marginal tax rates. For example, when they estimate the ironing propensity using only questions about incomes *outside of the respondents' own tax brackets*, the estimated model accurately predicts respondents' underestimation of *marginal tax rates within their own tax bracket*.

A nuance of formalizing the implications of biases such ironing and spotlighting involves the interdependence between perceptions and behavior. In the case of ironing, for example, perceived marginal tax rates depend on one's own average tax rate, which is a function of taxable income,—which in turn depends on how the individual perceives the federal income tax code. This circularity between perceptions and choices necessitates the application of a solution concept. The simplest possible solution concept, as formalized by Rees-Jones and Taubinsky (2018a), assumes that behavior and perceptions are in equilibrium: behavior is optimal given the perceptions that follow from the behavior. In dynamic settings, other possibilities arise, such as supposing that the perception in period t reflects the average tax rate (and thus behavior) in period $t - 1$.[87]

Salience Bias

While most of the evidence on tax salience involves commodity taxes, as summarized in Section 3.4, the core findings appear to apply to the income tax code as well. Miller and Mumford (2015) examine a salient and highly visible change to the Child and Dependent Care Credit (CDCC) introduced in 2003 that, when considered in isolation, increased the subsidization of child and dependent care administered through the income tax. This policy also interacted with provisions of the existing Child Tax Credit in a non-salient but offsetting manner, in many cases creating an overall reduction in subsidization. Miller and Mumford demonstrate that taxpayers respond as if they were aware of

[87] See Ito (2014) for evidence supporting this mechanism for the case of tiered electricity pricing.

the salient incentives and ignorant of the arguably non-salient interactions. The lack of bunching at kink points (Saez, 2010; Chetty et al., 2011) in the tax schedule could also reflect salience bias or the ironing and spotlighting heuristics discussed above, but there are other conventional explanations, such as adjustment costs (see Chetty et al., 2011).

5.4 Mechanism design approaches and implementation non-invariance

The growing evidence on perceptual biases violates a core assumption underlying standard optimal tax analysis: that behavior only depends on the choice set induced by the tax system. According to this assumption, behavior should not vary across the tax systems that could implement any given choice set. Rees-Jones and Taubinsky (2018b) call this assumption *implementation invariance*, and explain that it underlies the mechanism design approach to optimal taxation.

Various studies have used the mechanism design approach to characterize fully flexible tax systems that generate distortions due to taxpayers' private information (see, e.g., Mirrlees, 1971, for a static model and Golosov et al., 2006, for a review of applications to dynamic models). The classical optimal income tax problem, as formulated by Mirrlees (1971), allows the policymaker to select an arbitrarily nonlinear tax schedule, but assumes that taxpayers' skill levels, θ, are unobservable, so that the tax can only depend on earned income. Instead of optimizing over all possible tax schedules, it is often useful to restate this problem in terms of direct revelation mechanisms: each individual makes an announcement about his type (which does not have to be truthful), and receives the consumption and labor bundle specified for that announcement. The optimal direct mechanism maximizes welfare while satisfying i) the incentive-compatibility constraint that each type must wish to make a truthful announcement, and ii) the budget-balance constraint that total consumption must not exceed total before-tax earnings.

After finding the optimal direct mechanism, the second step is to solve an implementation problem: select a tax system that creates the same opportunities as the direct mechanism. Typically, implementation is non-unique in dynamic settings (Golosov et al., 2006).

Rees-Jones and Taubinsky (2018b) argue that the existence of perceptual and attentional biases implies that the implementation invariance assumption cannot hold in practice. Using the ironing and salience biases as examples, they formalize three implications. First, the presence of these biases precludes an application of the revelation principle, which is what normally allows the analyst to separate the task of identifying the optimal direct mechanism from problem of finding a tax system that implements the mechanism. As a result, the level of welfare attained under the optimal direct mechanism neither approximates nor bounds the welfare attainable with the optimal tax schedule. Second, some biases can preclude implementation of the optimal direct mechanism through taxes, and also preclude mimicking the optimal tax solution with a direct mechanism. Third, the presence of these biases can mitigate the role of information rents—a central concept in the mechanism design literature—and consequently yield results resembling those that follow from frameworks in which information asymmetries play no role, such as the Ramsey approach—a point we illustrate next in Section 5.5.

Rees-Jones and Taubinsky (2018b) argue that a more fruitful way forward is to optimize directly over the available tax instruments, which makes it feasible to account for non-standard responses to the specific tax instruments under consideration. This method is consistent with a modified version of the sufficient statistics approach, as exemplified by formulas (12) and (13). Farhi and Gabaix (2015) use this approach to characterize an optimal nonlinear-income tax, generalizing the sufficient statistics formulas of Saez (2001). Because they focus on a nonlinear income tax, their formulas provide a number of important nuances absent from (12) and (13); for example, that a change in the top marginal tax rate can affect everyone's perceptions of their tax rates, and thus change the behavior of low-income consumers, or that the optimal marginal tax rate may be negative for low-income consumers (but not high-income earners), in contrast to classical results. Such formulas generalize standard characterizations of the optimal tax rates, which involve conventional statistics such as elasticities, by adding a behavioral term that involves an empirically implementable price-metric measure of bias.

5.5 Consumption taxes versus income taxes

Allcott et al. (2018a) revisit a classic question in public economics: whether revenue generation and redistribution are best achieved through direct taxation—i.e., the income tax—or indirect taxation—e.g., commodity taxes or capital income taxes. Their starting point is the Atkinson and Stiglitz (1976) theorem, which demonstrates that for a broad class of utility functions, the optimal tax system uses income taxation to achieve all distributional objectives. That is, the use of differential commodity taxes to redistribute from rich to poor is suboptimal.

The logic of the Atkinson-Stiglitz theorem is that a tax on (say) some luxury good reduces the appeal of attaining high earnings—since one cannot purchase as much of that good—and thereby distorts labor supply in the same way as an income tax targeted at the high earners who consume that good. It is better to employ an income tax directly, which avoids distorting consumption choices.

Key to this reasoning is the assumption that all commodity taxes are fully salient when consumers make the decisions that determine income. As Allcott et al. (2018a) show, when this assumption is relaxed, consumer behavior depends not only on actual opportunity sets, but also on the particular combination of income and commodity taxes that generates those sets.

A key result of Allcott et al. (2018a) is that the canonical Ramsey-style formulas turn out to be relevant in the context of non-salient commodity taxes. Specifically, they show that the optimal commodity tax follows the Diamond (1975) "Many-person Ramsey tax rule,"with a scaling adjustment for the degree of inattention. That is, differential commodity taxes are useful when they are not fully salient, and their optimal magnitudes follow two intuitive principles that routinely surface in policy debates: commodity taxes should be lower when the price-elasticity of the taxed good is higher, and they should higher when the taxed good is more heavily consumed by the rich. With standard consumers, optimal commodity taxes have these properties only when an optimal income tax is unavailable.

This result contributes in an interesting way to the evolution of thinking concerning direct versus indirect taxation. The Ramsey framework once had a profound impact on Public Economics, but is now widely discounted because it ignores income taxation. The Allcott et al. (2018a) result shows that the rejection of the Ramsey framework may have been a premature consequence of rigidly adhering to the assumption of perfect rationality.

5.6 Social insurance

While we have thus far interpreted θ in our baseline model as a fixed characteristic (earnings ability), we can also interpret it as the realization of state of nature, as in a social insurance problem. We briefly discuss two important types of social insurance–unemployment insurance and health insurance–as they relate to the broader themes of this chapter. We refer the reader to the chapters in the upcoming second volume of this Handbook on behavioral issues in Labor Economics and Health Economics for further discussion. We also note in passing that social insurance problems sometimes introduce an additional wrinkle that is not present in optimal tax problems, in that private markets may also provide options for protection.

Unemployment insurance

In the case of unemployment insurance, moral hazard occurs because insurance diminishes the returns to searching for a new job and/or reduces incentives to keep a current job (Baily, 1978; Chetty, 2008). The literature has incorporated three different themes from behavioral economics. Spinnewijn (2015) studies incorrect beliefs about the returns to search (a "slope effect") and about the likelihood of finding a job (a "level effect"). The former primarily distorts search effort, while the latter distorts precautionary savings. Spinnewijn (2015) derives a modification of Baily-Chetty formula that allows for incorrect beliefs, and that is conceptually similar to (13).

Two other papers have emphasized the role of quasi-hyperbolic discounting and reference-dependence in job search, while not offering characterizations of optimal policies. DellaVigna and Paserman (2005) note that workers who are more impatient search less intensively but set lower reservation wages, and thus the overall effect of impatience on rates of exit from unemployment is generally unclear. However, the latter effect dominates for exponential agents, while the former dominates for quasi-hyperbolic agents. They provide evidence for the quasi-hyperbolic discounting model by showing that measures of impatience are negatively correlated with search effort and the unemployment exit rate, and are orthogonal to reservation wages.

DellaVigna et al. (2017) use Hungarian data to study how workers' hazard rates of exiting unemployment respond to changing benefit schedules. They show that the data support a reference-dependent model in which the reference point is a function of past consumption. They also argue that the data are most consistent with a model featuring high levels of impatience, which strongly suggests quasi-hyperbolic discounting.

Exploring the normative implications of these findings is a useful next step for future research. The implications of quasi-hyperbolic discounting accord with the principles discussed in Section

5.2, but the implications of reference dependence are not easily captured by the static frameworks discussed in this section.

Health insurance

In the case of health insurance, a classic consequence of moral hazard is the over-use of medical services, which insurance subsidizes. Various behavioral biases may lead patients to further overuse some medications such as painkillers but underuse others such as statins (Baicker et al., 2015). The formula for the optimal copay thus features all of the same tradeoffs introduced in the simple model studied in Section 5.1. See Baicker et al. (2015) for further details and implications.

An important issue not studied by Baicker et al. (2015) is that patients may misunderstand the price of utilization. As Brot-Goldberg et al. (2017) show, for example, people misunderstand the complicated dynamic incentives induced by deductibles and other provisions. With underestimation of utilization costs, the logic of formula (14), which we developed in the context of income taxation, would imply that the optimal amount of insurance is lower than with perfectly rational consumers. However, if people overestimate utilization costs because, for example, they react to spot prices rather than the effective prices in plans with deductibles, then the logic of formula (14) would imply that the optimal level of insurance is higher than with perfect rationality. These observations imply that plan features such as deductibles, which help to reduce moral hazard in classical models, may have additional effects associated with changing price perceptions, perhaps in a socially beneficial direction.

Another important topic concerns biases affecting choices of health insurance plans. A growing body of evidence suggests that people routinely make mistakes at the plan-choice stage (see, e.g., Abaluck and Gruber, 2011; Handel, 2013; Handel and Kolstad, 2015; Ericson, 2014; Bhargava et al., 2017). Handel et al. (2016) explore the implications of these “information frictions” for the efficiency of competitive insurance markets, and show that the mistakes can sometimes increase welfare by counteracting adverse selection. These results can have important implications for the design of subsides and other government interventions in health insurance markets.

5.7 Other issues

5.7.1 Correcting tax misperceptions

Common complaints that the U.S. tax code is so notoriously complex regularly lead to calls for simplification. A related question is whether “teaching the tax code,” as in Chetty and Saez (2013), might be desirable.

While the intuitive justification for helping people formulate more informed responses to the tax code may seem compelling, our discussion of nudges in Section 3.5 suggests that the issue is more nuanced. If the complexity of the tax code makes people underreact to the the disincentives that taxes create, then eliminating consumers’ mistakes might lead to lower labor supply and an undesirable reduction in tax revenue.

An additional consideration is that greater taxpayer competence could affect the progressivity of the tax burden. Rees-Jones and Taubinsky (2018a) analyze this possibility formally by simulating the effects of a hypothetical educational intervention that eliminates reliance on the ironing heuristic. Using their empirical estimates of the propensity to iron, they calculate the equivalent variation associated with eliminating misconceptions. Although the propensity to iron does not vary across the income distribution, it leads to greater underreaction among higher income taxpayers, because they face a higher discrepancy between the marginal and average tax rates. Thus, an intervention that eliminated ironing would be equivalent to a tax reform that reduced tax burden on the rich but not on the poor. In other words, an educational intervention that eliminated misconceptions about the tax schedule would have a highly regressive impact.

An analogous theme in recent studies on health insurance is that behavioral biases can improve market outcomes by combatting adverse selection (Handel, 2013; Handel and Kolstad, 2015; Handel et al., 2016; Spinnewijn, 2017) or moral hazard (Baicker et al., 2015).

5.7.2 Tax filing and tax compliance

In practice, taxpayers–especially the self-employed–have some control over the taxable income they report. They can reduce their liabilities either through tax evasion (deliberate misreporting) or tax avoidance (choices, such as charitable donations, that legally reduces their liabilities).

The classical compliance model of Allingham and Sandmo (1972) may fail to describe evasion and avoidance activities accurately for at least four behavioral reasons. First, taxpayers may hold incorrect beliefs about the likelihood of being audited (Chetty, 2009). Bergolo et al. (2017) provide suggestive evidence for this proposition by showing that IRS letters that provide information on audit statistics versus those that do not significantly affect firm-level tax reporting in Uruguay. Firms that hold correct beliefs about audit probabilities should disregard the information in these letters.

Second, social norms, feelings of duty, and the desire to avoid guilt or shame may motivate tax compliance above and beyond the threat of audits (Luttmer and Singhal, 2014). For example, Perez-Truglia and Troina (2016) show that increasing the salience of shame for tax delinquents significantly increases their compliance. Dwenger et al. (2016) find that taxpayers are intrinsically motivated to comply with a church tax in Germany.

Third, loss aversion may influence a taxpayer's motivation to pursue avoidance or evasion. If a positive "balance due" at the end of the year feels like a loss, while a negative "balance due" feels like a gain, people will be especially likely to engage in avoidance when their balance due is positive. Rees-Jones (forthcoming) estimates that taxpayers facing a payment on tax day reduce their tax liability by $34, relative to taxpayers owed a refund.

Fourth, because some forms of tax avoidance are costly, people may not take full advantage of opportunities to reduce their tax burdens, and behavioral biases such as procrastination may amplify this tendency. Benzarti (2016) studies taxpayers' propensities to reduce their tax burdens by itemizing deductions. A standard revealed preference analysis puts the total cost of filing at $200 billion (∼1.2% of GDP). However, Benzarti (2016) also provides evidence that much of the implied cost reflects procrastination, arguably from quasi-hyperbolic discounting.

5.7.3 Toward more general welfare criteria

The standard utilitarian criterion used for analyzing the optimal tax system throughout this section may be inconsistent with the nature of other-regarding preferences and attitudes toward redistribution among the general population. The literature on social preferences offers a variety of theories that could in principle inform the construction of more representative social objective functions.

Saez and Stantcheva (2016) provide a general theory of optimal taxation that is flexible enough to capture more nuanced preferences for redistribution via *generalized social marginal welfare weights*. A generalized weight captures the value that society places on increasing a particular individual's consumption by a unit, but is not necessarily tied to the individual's marginal utility of income, as are utilitarian weights. Instead, the generalized weights can depend on individual and aggregate characteristics, some of which result from the tax system itself. These weights allow for alternatives to utilitarianism such as libertarianism, equality of opportunity, and poverty alleviation. The weights can also capture nuanced preference such as a disdain for "freeloaders" who would work absent means-tested transfers.

6 Concluding Remarks

This chapter has reviewed basic conceptual frameworks for evaluating the welfare effects of public policies and for optimizing policy design, as well as empirical strategies for implementing these evaluations, when consumers do not behave in accordance with classical theories. Our discussion demonstrates the feasibility of extending the methods of public economics to allow for principled, quantitative policy evaluation under a wide variety of hypotheses about decision making.

As we have explained, choice-oriented methods of welfare analysis reduce each normative question to three basic positive questions:

- What is the scope of consumers' concerns?
- Which choices are welfare-relevant (i.e., free from characterization failure)?
- What is the choice mapping?

In applications, answers to the first two questions do not always receive as much careful consideration as warranted. We recommend attending to them as thoroughly as to the third when there is potential for controversy.

Choice-oriented formulas for welfare effects and optimal policies follow directly from the answers one provides to the three preceding questions. Conditional on those answers, neither the formulas nor the conclusions that flow from them are sensitive to assumptions about the underlying mechanisms. Thus, despite the aforementioned relationship between normative and positive analysis, welfare evaluation frequently does *not* require the type of finely nuanced understanding of mechanisms commonly sought in studies that pursue purely positive objectives. Normative analysis depends on these nuances only insofar as they affect answers to the three questions stated above.

While it is true that psychological mechanisms determine which choices are welfare-relevant, large classes of mechanisms have essentially the same implications concerning the scope of characterization failure, and hence there is no need to distinguish among them for this purpose. Often, one can refine the welfare-relevant domain based on general qualitative evidence – for example, according to whether the consumer properly understands some feature of a decision problem, rather than according to precisely how or why she misunderstands it.[88] Returning to one of our applications, one could formulate many cognitive models of underreaction to sales taxes when stores only post pre-tax prices. However, from a normative perspective, what matters is the existence of underreactions (which justifies removal of the associated choices from the welfare-relevant domain), rather than the particular mechanism that produces them.[89]

Finally, when conducting normative analysis, we do not require the type of broadly *generalizable* understanding of behavior commonly sought in studies that focus on positive questions. For normative purposes, it does not matter that two disparate behavioral phenomena may share the same underlying psychological cause, or that an understanding of cognitive mechanisms in one context may help us anticipate behavior in another. Rather, what matters is the incidence of characterization failure and the nature of the choice mapping within the context of interest. Of course, in some settings one cannot extrapolate the full choice mapping from limited data without adopting a specific cognitive model. However, better data would in principle render those structural assumptions superfluous. The assumptions of a specific cognitive model are thus best thought of as necessary compromises in the face of data limitations.

It follows from the preceding observations that the prevalent mode of analysis in positive behavioral economics, which emphasizes the broad (cross-domain) predictive and/or explanatory power of parsimonious models that depict specific psychological mechanisms one at a time, may be counterproductive in Behavioral Public Economics. Robust normative analysis requires a somewhat different mindset. Focusing on a particular mechanism, rather than a class of mechanisms that justify a particular welfare-relevant domain while rationalizing a given choice correspondence, can obfuscate the economic logic behind one's conclusions, as well as their robustness.

It goes without saying that there are numerous unresolved issues in Behavioral Public Economics. Below is a brief synopsis of a few open questions that strike us as particularly important.

Evaluating welfare. Identifying mistaken choices using objective evidence-based criteria is a critical component of choice-oriented behavioral welfare economics. Skeptics of this paradigm often point to what they see as weak and sometimes ad hoc justifications for particular normative perspectives. The process of systematizing principles and methods for identifying instances of characterization

[88] See Handel and Schwartzstein (2018) for a further discussion of when in-depth understanding of psychological mechanisms is needed for policy analysis, and when it is not.

[89] A parallel point arises in the literature on rational inattention; see the discussion of welfare in Caplin et al. (2018)..

failure is still in its infancy, and the feasibility of building a parallel empirical apparatus around the notion of optimization failure remains speculative. Fortunately, creative theoretical approaches to the problem of identification hold out the promise of significant progress; see, e.g., Benkert and Netzer (forthcoming) and Goldin and Reck (2015). Separately, the Non-comparability Problem (discussed in Section 2.2.2) potentially limits the applicability of choice-oriented welfare analysis to settings in which consumers either do not care about the conditions of choice, or only care about those conditions in well-defined circumstances. New methods that address these limitations would prove valuable.

Sin taxes. While it is commonly asserted that smokers overconsume cigarettes because of self-control problems, under-appreciation of nicotine's addictive properties, or incorrect beliefs about health risks, there is essentially no direct measurement of this tendency in the smoking domain, and no domain-specific estimates of the price-metric biases that one would need to implement an optimal tax formula. The same observation holds for most unhealthy foods and alcohol. There is also little evidence concerning the ideal level of incentives for physical exercise. This omission is unfortunate given the growing number of studies that examine various price and non-price levers for motivating exercise, and that proceed from the presumption that people do not exercise enough. The economics of optimal exercise incentives is further complicated by the fact that exercise facilities are often priced far below marginal costs in response to individuals' biases (DellaVigna and Malmendier, 2004; DellaVigna and Malmendier, 2006).

Policies affecting saving. We have seen that present focus can have strikingly different implications for the optimal treatment of capital income depending on whether it is "always on" or intermittently triggered by environmental cues. Yet as far as we know, existing empirical studies attempt to measure the average degree of present focus, rather than the extent to which it varies across decisions for a given subject, or the causes of that variation. We have also seen that certain policy approaches presuppose a demand for commitment. Yet there is almost no direct evidence concerning the existence or strength of this demand within the context of personal saving (Beshears et al., 2015, being an important exception). Finally, we have seen that consumers often fail to understand all the likely consequences of the various complex financial decisions that are integral to life-cycle planning, and that they frequently fail to act on pertinent knowledge even when they acquire it. Economists have not yet focused on the problem of identifying effective strategies for overcoming that failure.

Income taxes. While there is growing evidence that complicated income tax schedules confuse taxpayers, there is little understanding of how this confusion would evolve with possible income tax reforms and little quantitative measurement of the type that is necessary to implement optimal tax formulas. An improved understanding of the sources of confusion can aid with the former challenge. Moreover, while present-focus can in principle affect labor supply,[90] there is little understanding of what role, if any, it plays in determining individuals' incomes in developed economies, since the outcomes of many income-determining decisions—such as what job to take—are delayed. Other

[90] See, e.g., Kaur et al. (2015).

theories—such as the focusing model of Koszegi and Szeidl (2013)—might imply excessive work effort because the benefits (e.g., annual salary) may attract more attention than the costs (e.g., required hours of work each day). Another unaddressed question is whether people are fully attentive to income taxes when they make the choices that determine their incomes. Motivated by work on sales tax salience, one might conjecture that when, for example, choosing which job to take, people might simply compare before-tax salaries.

The many open questions and challenges remaining in Behavioral Public Economics are both conceptually fascinating and practically important. Beyond being a productive area for further research, we anticipate that this line of inquiry will generate lasting impact on public policy and social welfare.

Appendix: Behavioral themes pertaining to saving

Imperfect self-control

The notion that people exercise imperfect self-control resonates with experience and casual empiricism. While the idea is intuitive, formalizations involve conceptual subtleties. The literature provides two broad approaches, one based on the notion of time inconsistency, and another that posits the existence of *internal goods*.

Imperfect self-control with time-inconsistent preferences.

One leading school of thought associates imperfect self-control with time-inconsistent preferences defined over otherwise conventional goods. In the main text, we illustrated this idea through the example of a consumer, Norma, who chooses salad over pizza for lunch when deciding early in the morning, but reverses this decision at lunchtime.

Formalizations. The theory of time-inconsistent preferences originated with Strotz (1955-1956). Other early contributions clarified the appropriate notion of optimal planning within Strotz's framework (Pollak, 1968), resolved questions about existence (Peleg and Yaari, 1973; Goldman, 1980), and began to explore applications (Schelling, 1984). The framework gained considerable momentum in the 1990s based on the work of David Laibson (1997; 1998), who popularized a particular class of time-inconsistent preferences known as *quasi-hyperbolic discounting* (QHD, or, more colloquially, the $\beta\delta$ model), which he borrowed from a related experimental literature in psychology (Chung and Herrnstein, 1961).[91] The QHD model encapsulates a desire for immediate gratification, or *present focus*, within an elegant and simple framework that departs minimally from standard formulations of intertemporal preferences, and as a result has become one of the main workhorses of behavioral economics. That said, the literature has also explored other interesting preference formulations in the tradition of Strotz, including the possibilities that self-control problems arise only in particular

[91] The period-t objective function for a QHD consumer is $u_t + \beta \sum_{s=t+1}^{T} \delta^{s-t} u_s$, where $(u_t, ..., u_T)$ represents flow utility.

states of nature (Bernheim and Rangel, 2004; Dekel and Lipman, 2012), and that consumers apply different rates of discount to the experiences associated with different goods (Banerjee and Mullainathan, 2010). The Strotz framework may be particularly descriptive of household decision making, inasmuch as interpersonal aggregation naturally yields time inconsistency even when individual household members are time-consistent (Bernheim, 1999; Jackson and Yariv, 2014).

Choice reversals. One empirical hallmark of time inconsistency is the tendency to make systematically different choices among a fixed set of alternatives as the earliest consequences become more imminent. However, one can often contrive other explanations for these same patterns. To illustrate, suppose Norma prefers to eat salad when she is happy and pizza (a comfort food) when she is sad. When lunchtime arrives, she knows her mood and chooses accordingly. Assuming she is time-consistent and has the opportunity to form a mood-contingent plan, she selects the same options when making the decision a few hours in advance. Yet when we ask her to choose a single lunch option at 10am without stating contingencies, she performs an expected value calculation based on her anticipated mood. It is straightforward to construct examples in which those calculations systematically favor salad over pizza. In those cases, Norma chooses salad more frequently in advance than at lunchtime, even though she is time-consistent.

Methods of self-regulation. For sophisticated consumers – those who understand their own behavioral tendencies – the more telling hallmarks of time inconsistency involve strategies for exercising self-control. These fall into two broad categories, according to whether they involve externally enforced commitments or internal methods of self-regulation.

Analyses of externally enforced commitments originate with Strotz (1955-1956). We discussed this strategy at some length in the main text (Section 4.3).

Bernheim et al. (2015c) formalize notions of internal self-regulation through self-punishment and self-reward. They depict intertemporal choice as a dynamic game played by successive incarnations of a single decision maker with quasi-hyperbolic preferences, and interpret subgame-perfect, history-dependent equilibrium strategies as methods of exercising self-control through the credible deployment of contingent punishment and reward.[92] They explore the nature of optimal internal self-control, demonstrating that it has a simple and behaviorally plausible structure that is immune to self-renegotiation: in effect, if a consumer fails to meet her personal standard (“falls off the wagon”), she responds to her lapse with a temporary binge (“gets it out of her system”) before rededicating herself to her original objectives. Their main result demonstrates that, in the presence of credit constraints, low initial assets can limit self-control, trapping people in poverty, while people with high initial assets can accumulate indefinitely. They also show that external commitments can undermine internal self-regulation by limiting opportunities for self-reward and self-punishment. We mentioned these implications in Section 4.3.1.

[92] In contrast, other studies of quasi-hyperbolic discounting and time inconsistency focus almost exclusively on Markov-perfect equilibria, which involve no history dependence, and hence cannot capture the phenomenon of contingent self-reinforcement. Exceptions include Laibson (1994) and Benhabib and Bisin (2001).

Normative interpretations. The most common normative interpretation of the QHD model is that $\beta < 1$ represents a cognitive bias. We provided a critique of that perspective in Section 2.2.5.

An alternative normative interpretation of the QHD model holds that the consumer has a distinct "true" preference relation at each moment in time. Laibson et al. (1998) adopt this view and apply the Pareto criterion (as mentioned in Section 2.2.2). A conceptually problematic feature of their analysis is the assumption that the date-t "self" does not care about past consumption. In reality, most of us care about our memories of past consumption, but there is no way to elicit those preferences through choices, inasmuch as date t' consumption is fixed at all dates $t > t'$.

Bernheim and Rangel (2009) apply their framework to the QHD model and explore its implications under various definitions of the welfare-relevant domain. Among other results, they provide a precise characterization of normative ambiguity when all choices are deemed welfare-relevant.

Some additional normative issues arise in the context of naive time inconsistency. The choices of naive consumers depend on two aspects of the decision frame: timing (whether it is contemporaneous or forward-looking) and "transparency." Models of naive choice generally depict an "obscure" frame in which the decision maker must infer her future actions, but such models also allow one to deduce the choices she would make with "transparent" framing that renders the actual continuation paths, and hence ultimate consequences, readily apparent. Taking this interpretation literally, one would refine the welfare-relevant domain by excluding decisions with obscure framing, and retaining only those with transparent framing.

That said, caution may be warranted. Models are simply lenses through which we interpret and rationalize choice patterns. If we treat a model of naivete as an as-if representation that may happen to fit the choice data rather than as a literal depiction of cognitive processes, the argument for ignoring supposedly naive choices becomes less compelling. One may then wish to apply the Bernheim-Rangel framework in an agnostic manner, respecting all choice frames, irrespective of the model's labeling.

Imperfect self-control with internal goods.

A second important school of thought explains the notion of imperfect self-control by invoking unconventional "internal" goods, such as the psychological costs of exercising willpower or of experiencing temptation. Under this view, apparent choice reversals are in fact not reversals at all, but rather consequences of subtle changes in the available consumption bundles.

To illustrate, let's return to the example of Norma's lunch choices. Here we account for her behavior by positing the existence of a latent psychological good, call it "yearning," that depends on a comparison between the option she chooses and any available alternative that tempts her, and thereby encapsulates the internal costs of exercising willpower. When she expresses a preference for salad rather than pizza two hours before lunchtime, she has in mind a comparison between two bundles, one consisting of pizza with no yearning, the other consisting of salad with no yearning. When lunchtime arrives and she examines a menu listing both salad and pizza, the only available bundles consist of pizza with no yearning and salad with yearning (for pizza). Assuming yearning is sufficiently costly, she prefers the first to the second. A casual observer might make the mistake of inferring that she is time-inconsistent, choosing salad in advance and pizza in the moment. In fact, her preferences are entirely consistent, and appearances to the contrary simply reflect our inability to observe internal goods.

A potential limitation of this approach is that it cannot rationalize certain types of attitudes – for example, the possibility that, as of 10am, Norma wishes she could get herself to choose salad at lunchtime even if pizza is on the menu, regardless of how she expects to feel about the choice once lunchtime arrives. If one takes the view that such intertemporal "disagreements" are central to the psychology of self-control, then all theories of time-consistent choice are problematic, even when they incorporate internal goods.

Externally enforced commitments and internal self-regulation. Like time inconsistency, theories with internal goods naturally generate a demand for externally enforced commitments. Here the purpose of a commitment is to change the nature of the consumption bundles available in the future. In our example, Norma makes a social commitment to meet a friend at a restaurant that only serves salad, rather than one that serves both salad and pizza, in order to replace the options (salad, yearning) and (pizza, no yearning) with the single option (salad, no yearning).

With time inconsistency, Norma prefers to make a commitment only if it changes her behavior. In contrast, with internal goods, she may do so even when her behavior is unaffected – for example, when she prefers (salad, no yearning) to (salad, yearning), and prefers (salad, yearning) to (pizza, no yearning). In that case, her behavior indicates a preference for regulating her behavior through external rather than internal methods.

Formalizations. Some economists have attempted to formalize the preceding ideas by modeling internal goods explicitly, thereby providing explicit psychological microfoundations for the cognitive processes governing self-control. This approach originates with Thaler and Shefrin (1981), who formulated a "dual-self" representation of decision making in which behavior reflects two separate motivational systems. One system operates as a patient forward-looking "planner," the other as a myopic present-focused "doer." In the Thaler-Shefrin framework, the planner is in charge and controls the impulses of the doer by exercising willpower at a psychological cost. Behavior reflects the planner's time-consistent preferences over bundles that include both conventional goods and willpower expenditures. The doer's inclinations simply modify an otherwise standard optimization problem. Focusing only on the conventional goods, the consumer's choices appear to be menu-dependent, but that appearance is misleading because it ignores the internal goods. In this setting, the planner can reduce future willpower costs without encountering resistance from the present-focused doer by restricting future opportunities. Thus, the model generates a robust demand for commitment. Subsequent articulations and extensions of this approach include Shefrin and Thaler (1988) and Fudenberg and Levine (2006).

A possible criticism of the preceding approach is that its central assumptions concerning cognition may not be amenable to direct empirical investigation. An alternative strategy is to formulate the theory entirely in terms of observable choices by defining preferences over menus of conventional consumption bundles and the options selected from them, rather than over bundles of both conventional and internal goods. One can think of a preference ordering over menu/consumption pairs as a reduced form for preferences over the mental states the pairs induce.[93] In Norma's case, instead of saying she prefers (S, no yearning) to (P, no yearning) to (S, yearning), where S and P represent salad and pizza, respectively, we say she prefers $(S, \{S\})$ to $(P, \{S, P\})$ to $(S, \{S, P\})$. This is the approach taken by Gul and Pesendorfer (2001), who propose a collection of preference axioms that characterize the following class of utility functions:

$$U(X, x) = u(x) - \left[\max_{y \in X} v(y) - v\left(x\right) \right].$$

Here, X is the menu and x is the chosen option. One can interpret $u(x)$ as the utility derived from x, and $v(x)$ as a measure of the extent to which x tempts the consumer. In that case, $\max_{y \in X} v(y) - v\left(x\right)$ represents a temptation penalty, which the consumer incurs when she fails to choose the most tempting alternative. We mentioned the Gul-Pesendorfer model in Sections 2.2.2 and 4.2. See also Dekel, Lipman, and Rustichini (2001; 2009) for a related theory.

One can potentially criticize the latter approach by questioning the validity of evaluating the plausibility of preference axioms without theorizing explicitly about the cognitive processes governing self-control. Arguably, their axioms are reasonable if and only if they are consistent with a sensible model of process. A second potential criticism concerns the stability of the reduced-form preferences. To illustrate, compare two scenarios: in the first, Norma chooses between two restaurants, one that serves only salad, and a second that serves salad and pizza; in the second, her options are limited to a single restaurant, where she must choose upon arrival between two menus, one listing only salad, the other listing pizza and salad. Both scenarios provide Norma with the same (X, x) options. Yet it seems likely that, in the second scenario, temptation will adhere not only to the selection of an entree, but also to the choice of a menu. The Gul-Pesendorfer framework contains no element that could account for the hypothesized difference in behavior across these scenarios. In contrast, the doer-planner model can attribute behavioral discrepancies to the length of the doer's evaluation horizon.

Normative interpretations. Taking the Thaler-Shefrin model literally in the spirit of Behavioral Revealed Preference, welfare is arguably ambiguous because we can evaluate it from the perspective of either the planner or the doer. While one could attempt to argue directly that the doer's preferences are normatively invalid, it is hard to imagine a route to that conclusion involving empirical evidence and objective criteria. Alternatively, one can abandon literal interpretations, treat the doer-planner model as an as-if representation, and apply the Bernheim-Rangel framework. There are then two ways to proceed.

One approach is to assume that people actually care about the psychological costs associated

[93] This same perspective is implicit in standard consumer theory: one can think of a preference ordering over bundles of conventionally defined goods as a reduced form for preferences over the mental states the bundles induce.

with phenomena such as yearning and temptation. Because the model implies coherent choices (in the sense of WARP) over consumption bundles that subsume those internal goods, it delivers (in principle) an unambiguous welfare criterion, one that coincides with the hypothesized preferences of the as-if planner. Whether we can implement that criterion empirically is another matter: because the Thaler-Shefrin model implies that internal goods depend not just on what is chosen but also on the conditions of choice, it introduces the Non-comparability Problem, which can render welfare unrecoverable, at least without additional structure (see Section 2.2.2).

The other approach is to assume that consumers' concerns are limited to conventional goods, and to treat the internal goods as as-if representations. Because the model implies inconsistent choices over conventionally defined consumption bundles, one then arrives at an ambiguous welfare criterion, one that reflects the hypothesized conflict between the as-if doer and planner.

Normative ambiguity seems to disappear in the Gul-Pesendorfer framework, which posits a single coherent preference ordering over (X, x) pairs. We have seen, however, that preferences of this form also implicate the Non-comparability Problem, which means that welfare is unrecoverable without additional structure. Moreover, if we interpret this model as a reduced form for preferences over the mental states those pairs induce, potentially as the result of conflicts between opposing motivational systems, then its use simply obscures the normative ambiguity that exists in the Thaler-Shefrin framework without resolving it.

Limited financial competence

Another branch of the literature challenges the notion that the typical consumer makes deliberate financial decisions based on an accurate understanding of the relationship between choices and consequences. Reservations concerning this premise fall into the following categories.

Low financial literacy.

Many consumers appear to lack the knowledge and skills necessary for sound life-cycle planning. Early work in this area documented important deficiencies in both pertinent factual knowledge (Bernheim, 1988, 1995, 1998; Gustman and Steinmeier, 2004, 2005), for instance concerning pensions and Social Security, and comprehension of important financial principles, such as inflation, asset diversification, and compound interest (Bernheim, 1998). Subsequent research on financial literacy has corroborated these concerns; for reviews, see Lusardi, 2009; Lusardi and Mitchell, 2014b. In a few cases, research has identified specific biases, such as the tendency to underestimate compounding, a phenomenon known as *exponential growth bias* (Wagenaar and Sagaria, 1975; Eisenstein and Hoch, 2007; Levy and Tasoff, 2016; Stango and Zinman, 2009; Almenberg and Gerdes, 2012).

Financial literacy is strongly correlated with financial choices such as rates of saving (Bernheim, 1998; Lusardi and Mitchell, 2007, 2011). In principle, these correlations could reflect the causal effect of knowledge, reverse causation (e.g., those who save more have greater incentives to acquire financial information), or common causation (e.g., those with financial interests both save more and acquire more knowledge). Unfortunately, it is difficult to identify correlates of financial literacy that are independent of tastes, and that influence behavior only through knowledge. Financial education is a potential instrument, but it may affect the motivation to save through channels other than financial literacy. Consequently, while one can use instrumental variables to immunize the measured correlation against reverse causation, common causation is more problematic. Efforts to establish causation are therefore not entirely convincing.

Limited reliance on experts and use of planning tools.

Low financial literacy need not imply poor decision making. In principle, consumers can compensate for gaps in knowledge and analytic skills by relying on financial professionals, or by employing appropriate tools, such as planning software and financial calculators. However, in practice, relatively few consumers deploy these resources. In one survey, a majority of baby boomers reported relying primarily on parents, relatives, friends, or simply their own judgment, when making financial choices, while only 15% said they relied primarily on financial professionals (Bernheim, 1998; see also Lusardi, 2009; Lusardi and Mitchell, 2011). Observed correlations between financial literacy and behavior, mentioned above, are more troubling in light of these findings.

The superficiality of decision processes.

A large fraction of the population engages in no serious financial planning, and members of that same group tend to be low savers (Lusardi, 1999; Lusardi and Mitchell, 2007). Instead, households appear to fall back on simple heuristics and rules of thumb when making critical financial decisions, despite (or perhaps because of) their complexity. In one study (Bernheim, 1994), 62% of respondents said they formulated savings targets in terms of percentages of income, but nearly three-quarters of those reported targets that were even multiples of 5%. This pattern was equally prevalent among those who claimed to have formal financial plans. Moreover, stated targets were unrelated to critical economic variables such as earnings growth. Even professional financial advisors sometimes make rough-and-ready recommendations, such as maintaining an emergency fund equal to six months of household income, or saving 20% of gross income (Doyle and Johnson, 1991).

Problematic choices.

Another branch of the literature attempts to document limited financial competence by identifying mistakes in decision making. Early work in this area focused on behavioral patterns that either seem peculiar, such as the absence of a relationship between age and rates of saving among Japanese households (Hayashi, 1986), or that experts deem inadvisable, such as low rates of saving (Bernheim, 1993), low enrollment in pension plans that offer generous matches, naive diversification strategies, and the tendency for employees to invest in their employers' stock (Benartzi and Thaler, 1999, 2001, 2007). More recent work along these lines focuses on evidence of excessive inertia, suggestibility, and inattention (e.g., Madrian and Shea, 2001; Bernheim et al., 2015a; Karlan et al., 2016b).

In most of these cases, it is difficult to rule out all taste-based explanations for the observed phenomena (see, for example, Scholz et al., 2006, concerning wealth accumulation). This limitation has prompted the development of other methods for identifying financial mistakes; see Section 4.6.2 of the main text.

References

Abaluck, Jason and Jonathan Gruber, "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program," *American Economic Review*, 2011, *101* (4), 1180–1210.

Afriat, Sidney N., "Efficiency Estimation of Production Functions," *International Economic Review*, 1972, *13* (3), 568–98.

Agarwal, Sumit, John C. Driscoll, Xavier Gabaix, and David Laibson, "The Age of Reason: Financial Decisions over the Life Cycle and Implications for Regulation.," *Brookings Papers on Economic Activity*, 2009, *Fall*, 51–101.

Aguiar, Mark and Erik Hurst, "Deconstructing Life Cycle Expenditure," *Journal of Political Economy*, 2013, *121* (3), 437–492.

Ainslie, G., "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control," *Psychological Bulletin*, 1975, *56*, 383–396.

_ , "Derivation of 'Rational' Economic Behavior from Hyperbolic Discount Curves," *American Economic Review*, 1991, *81*, 134–140.

Ainslie, George W., *Picoeconomics*, Cambridge: Cambridge University Press, 1992.

Aliber, M., "Rotating Savings and Credit Associations and the Pursuit of Self-Discipline," *African Review of Money Finance and Banking*, 2001, pp. 51–72.

Allcott, Hunt, "Consumers' Perceptions and Misperceptions of Energy Costs," *American Economic Review*, 2011, *101* (3), 98–104.

_ , "Social norms and energy conservation," *Journal of Public Economics*, 2011, *95* (9), 1082 – 1095. Special Issue: The Role of Firms in Tax Systems.

_ , "The Welfare Effects of Misperceived Product Costs: Data and Calibrations from the Automobile Market," *American Economic Journal: Economic Policy*, 2013, *5* (3), 30–66.

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

_ **and Judd Kessler**, "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons," *American Economic Journal: Applied Economics*, forthcoming.

_ **and Nathan Wozny**, "Gasoline Prices, Fuel Economy, and the Energy Paradox," *Review of Economics and Statistics*, 2014, *96* (5), 779–795.

_ **and Todd Rogers**, "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," *American Economic Review*, October 2014, *104* (10), 3003–37.

_ **, Benjamin B. Lockwood, and Dmitry Taubinsky**, "Ramsey Strikes Back: Optimal Commodity Taxes and Redistribution in the Presence of Salience Effects," *American Economic Association Papers and Proceedings*, 2018.

__ , **Benjamin B Lockwood, and Dmitry Taubinsky**, "Regressive Sin Taxes, with an Application to the Optimal Soda Tax," *working paper*, 2018.

__ , **Christopher Knittel, and Dmitry Taubinsky**, "Tagging and Targetting of Energy Efficiency Subsidies," *American Economic Review Papers and Proceedings*, May 2015, *112*, 72–88.

__ , **Sendhil Mullainathan, and Dmitry Taubinsky**, "Energy policy with externalities and internalities," *Journal of Public Economics*, 2014.

Allingham, Michael G. and Agnar Sandmo, "Income Tax Evasion: A Theoretical Analysis," *Journal of Public Economics*, 1972, *1* (3-4), 323–338.

Almenberg, Johan and Christer Gerdes, "Exponential growth bias and financial literacy," *Applied Economics Letters*, 2012, *19* (17).

Amador, M., I. Werning, and G.-M. Angeletos, "Commitment vs. Flexibility," *Econometrica*, 2006, *74*, 365–396.

Ambec, S. and N. Treich, "Roscas as Financial Arrangements to Cope with Self-Control Problems," *Journal of Development Economics*, 2007, *82*, 120–137.

Ambuehl, Sandro, B. Douglas Bernheim, and Annamaria Lusardi, "A Method for Evaluating the Quality of Financial Decision Making, with an Application to Financial Education," *NBER working paper No. 20618*, 2017.

Anderson, S. and J.-M. Baland, "The Economics of Roscas and Intrahousehold Resource Allocation," *Quarterly Journal of Economics*, 2002, *117*, 963–995.

Ando, Albert and Franco Modigliani, "The 'Life Cycle' Hypothesis of Saving: Aggregate Implications and Tests," *American Economic Review*, 1963, *53* (1), 55–84.

Angeletos, George-Marios, David Laibson, Andrea Repetto, Jeremy Tobacman, and Stephen Weinberg, "The hyperbolic consumption model: Calibration, simulation, and empirical evaluation," *Journal of Economic Perspectives*, 2001, pp. 47–68.

Ariely, Daniel and Klaus Wertenbroch, ""procrastination, Deadlines, and Performance: Self-Control by Pre-Commitment," *Psychological Science*, 2002, *13* (3), 219–224.

Aristotle, *Aristotle's Nicomachean Ethics*, Translated by Robert C. Bartlett and Susan D. Collins. Chicago: University of Chicago Press, 2012, translation.

Arrow, Kenneth J., "A Note on Freedom and Flexibility," in Kaushik Basu, Presanta Pattanaik, , and Kotaro Suzumura, eds., *Choice, Welfare and Development: A Festschrift in Honour of Amartya K. Sen*, Oxford: Oxford University Press, 1995.

Ashraf, N., D. Karlan, and W. Yin, "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," *Quarterly Journal of Economics*, 2006, *121* (2), 635–672.

Atkinson, A. B. and A. Sandmo, "Welfare Implications of the Taxation of Savings," *Economic Journal*, 1980, *90* (359), 529–549.

Atkinson, Anthony B and J.E. Stiglitz, "Design of Tax Structure - Direct Versus Indirect Taxation," *Journal of Public Economics*, 1976, *6*, 55–75.

Attari, Shahzeen, Michael DeKay, Cliff Davidson, and Wandi Bruine de Bruin, "Public Perceptions of Energy Consumption and Savings," *Proceedings of the National Academy of Sciences*, 2010, *37*, 16054–16059.

Aufenanger, Tobias, Richter Friedemann, and Matthias Wrede, "Measuring Decision-Making Ability in the Evaluation of Financial Literacy Education Programs," *Unpublished Manuscript*, 2016.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger, "Working over time: Dynamic inconsistency in real effort tasks," *Quarterly Journal of Economics*, 2015, *130* (3), 1067–1115.

Austin, Rob and Winfield Evens, "2013 Trends & Experience in Defined Contribution Plans," *Aon Hewitt*, 2013.

Ayres, Ian, Sophie Raseman, and Alice Shih, "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage," *The Journal of Law, Economics, and Organization*, 2013, *29* (5), 992–1022.

Bagwell, Laurie Simon and B. Douglas Bernheim, "Veblen Effects in a Theory of Conspicuous Consumption," *American Economic Review*, 1996, *86* (3), 349–373.

Baicker, Katherine, Sendhil Mullainathan, and Josh Schwartzstein, "Behavioral Hazard in Health Insurance," *Quarterly Journal of Economics*, 2015, *130* (4), 1623–1667.

Baily, Martin N., "Some aspects of optimal unemployment insurance," *Journal of Public Economics*, 1978, *10* (3), 379–402.

Baltussen, Guido and Gerrit T. Post, "Irrational Diversification: An Examination of Individual Portfolio Choice," *Journal of Financial and Quantitative Analysis*, 2011, *5*, 1463 – 1491.

Bandura, A., *Social Learning Theory*, New York: General Learning Press, 1971.

_ , "Self-Reinforcement: Theoretical and Methodological Considerations," *Behaviorism*, 1976, *4* (2), 135–155.

_ **and C. J. Kupers**, "Transmission of Patterns of Self-Reinforcement Through Modeling," *Journal of Abnormal and Social Psychology*, 1964, *69* (1), 1–9.

Banerjee, A. and S. Mullainathan, "The Shape of Temptation: Implication for the Economic Lives of the Poor," *NBER Working Paper No. 15973*, 2010.

Banks, James, Richard Blundell, and Sarah Tanner, "Is There a Retirement Savings Puzzle?," *American Economic Review*, 1998, *88* (4), 769–788.

Bayer, Patrick J., B. Douglas Bernheim, and John Karl Scholz, "The Effects of Financial Education in the Workplace: Evidence from a Survey of Employers," *Economic Inquiry*, 2009, *47* (4), 605–624.

Benartzi, Shlomo and Richard H. Thaler, "Risk Aversion or Myopia? Choices in Repeated Gambles and Retirement Investments," *Management Science*, 1999, *45* (3), 364–381.

_ and _ , “Naive Diversification Strategies in Defined Contribution Savings Plans,” *American Economic Review*, 2001, *91*, 79–98.

_ and _ , “Heuristics and Biases in Retirement Savings Behavior,” *Journal of Economic Perspectives*, 2007, *21*, 81–104.

_ , John Beshears, Katherine L. Milkman, Cass R. Sunstein, Richard H. Thaler, Maya Shankar, Will Tucker-Ray, William J. Congdon, and Steven Galing, “Should Governments Invest More in Nudging?,” *Psychological Science*, 2018/03/05 2017, *28* (8), 1041–1055.

Benhabib, J. and A. Bisin, “Self-Control and Consumption-Saving Decisions: Cognitive Perspectives,” *Working Paper, Department of Economics, New York University*, 2001.

Benjamin, Daniel J., Alan Fontana, and Miles Kimball, “Reconsidering Risk Aversion,” *Mimeo*, 2016.

_ , Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones, “What Do You Think Would Make You Happier? What do You Think You Would Choose?,” *American Economic Review*, 2012, *102* (5), 2083–2110.

Benjamin, Daniel J, Ori Heffetz, Miles S Kimball, and Alex Rees-Jones, “Can marginal rates of substitution be inferred from happiness data? Evidence from residency choices,” *American Economic Review*, 2014, *104* (11), 3498–3528.

Benkert, Jean-Michel and Nick Netzer, “Informationa Requirements of Nudging,” *Journal of Political Economy*, forthcoming.

Benzarti, Youssef, “How Taxing Is Tax Filing? Leaving Money on the Table Because of Hassle Costs.,” *working paper*, 2016.

Bergolo, Marcelo, Rodrigo Ceni, Guillermo Cruces, Matias Giaccobasso, and Ricardo Perez-Truglia, “Tax Audits as Scarecrows: Evidence from a Large-Scale Field Experiment,” *working paper*, 2017.

Bernheim, B. Douglas, “On the Voluntary and Involuntary Provision of Public Goods,” *American Economic Review*, 1986, *76* (4), 789–793.

_ , “Social Security Benefits: An Empirical Study of Expectations and Realizations,” in E. Lazear and R. Ricardo-Campbell, eds., *Issues in Contemporary Retirement*, Stanford: Hoover Institution Press, 1988, pp. 312–345.

_ , *The Vanishing Nest Egg: Reflections on Saving in America*, Twenthieth Century Fund/Priority Press: New York, 1991.

_ , “Is the Baby Boom Generation Saving Adequately for Retirement? Summary Report,” New York: Merrill Lynch, Pierce, Fenner and Smith Inc. January 1993.

_ , “Person Saving, Information, and Economic Literacy: New Directions for Public Policy,” in “Tax Policy for Economic Growth in the 1990s,” American Council for Capital Formation: Washington, DC, 1994, pp. 53–78.

_ , “Do Households Appreciate Their Financial Vulnerabilities? An Analysis of Actions, Perceptions, and Public Policy,” in “Tax Policy and Economic Growth,” American Council for Capital Formation: Washington, DC, 1995, pp. 1–30.

_ , “Financial Illiteracy, Education, and Retirement Saving,” in Olivia S. Mitchell and Sylvester J. Scheiber, eds., *Living with Defined Contribution Pensions*, University of Pennsylvania Press, Pension Research Council, the Wharton School, University of Pennsylvania, 1998, pp. 38–68.

_ , “Comment on ’Family Barganing and Retirement Behavior’,” in Henry J. Aaron, ed., *Behavioral Dimensions of Retirement Decisions*, Washington, DC: Brookings Institution Press, 1999, pp. 273–281.

_ , “Behavioral Welfare Economics,” *Journal of the European Economic Association*, 2009, *7* (2-3), 267–319.

_ , “The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics,” *Journal of Benefit-Cost Analysis*, 2016, *7* (1), 12–68.

_ , *Behavioral Welfare Economics: From Foundations to Applications*, Oxford University Press, 2018.

Bernheim, B Douglas and Antonio Rangel, “Addiction and Cue-triggered Decision Processes,” *American Economic Review*, 2004, pp. 1558–1590.

Bernheim, B. Douglas and Antonio Rangel, “Behavioral Public Economics: Welfare and Policy Analysis with Fallible Decision-Makers,” in Peter Diamond and Hannu Vartianen, eds., *Behavioral Economics and its Applications*, Princeton University Press, 2007, pp. 7–77.

Bernheim, B Douglas and Antonio Rangel, “Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics,” *The Quarterly Journal of Economics*, 2009, *124* (1), 51–104.

Bernheim, B. Douglas and Daniel M. Garrett, “The effects of financial education in the workplace: Evidence from a survey of households,” *Journal of Public Economics*, 2003, *87.*

_ **and Jonas Mueller-Gastell**, “Default Options and the Case for Opt-Out Minimization,” *Working paper, Stanford University*, 2018.

_ **and Kyle Bagwell**, “Is Everything Neutral?,” *Journal of Political Economy*, 1988, *96* (2), 308–338.

_ **, Andrey Fradkin, and Igor Popov**, “The Welfare Economics of Default Options in 401 (k) Plans,” *American Economic Review*, 2015, *105* (9), 2798–2837.

_ **, Daniel Bjorkegren, Jeffrey Naecker, and Antonio Rangel**, “Non-Choice Evaluations Predict Behavioral REsponses to Changes in Economic Conditions,” *NBER Working Paper No. 19269*, 2015.

_ **, Daniel M. Garrett, and Dean M. Maki**, “Education and saving: The long-term effects of high school financial curriculum mandates,” *Journal of Public Economics*, 2001, *80*, 435–465.

Bernheim, B Douglas, Debraj Ray, and Sevin Yeltekin, “Poverty and Self-Control,” *Econometrica*, September 2015, *83* (5), 1877–1911.

Bernheim, B. Douglas, Jonathan Meer, and Neva K. Novarro, “Do Consumers Exploit Commitment Opportunities? Evidence from Natural Experiments Involving Liquor Consumption,” *American Economic Journal: Economic Policy*, 2016, *8* (4), 41–69.

__ , **Jonathan Skinner, and Steven Weinberg**, "What Accounts for the Variation in Retirement Wealth among U.S. Households?," *American Economic Review*, September 2001, *91* (4), 832–857.

Bertrand, Marianne and Adair Morse, "Information Disclosure, Cognitive Biases, and Payday Borrowing," *The Journal of Finance*, 2011, *LXVI* (6), 1865–93.

__ , **Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman**, "What's Avertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment," *Quarterly Journal of Economics*, 2010, *125* (1), 263–306.

Beshears, John, James J. Choi, Christopher Harris, David Laibson, Brigitte C. Madrian, and Jung Sakong, "Self Control and Commitment: Can Decreasing the Liquidity of a Savings Account Increase Deposits," *NBER Working Paper No. 21474*, 2015.

__ , __ , **David Laibson, and Brigitte C. Madrian**, "Early decisions: A regulatory framework," *Swedish Economic Policy Review*, 2005, *12*, 41–60.

__ , __ , __ , **and** __ , "The Importance of Default Options for Retirement Savings Outcomes: Evidence from teh United States," in S. J. Kay and T. Sinha, eds., *Lessons from Pension Reform in the Americas*, Oxford: Oxford University Press, 2008, pp. 59–87.

__ , __ , __ , **and** __ , "How does simplified disclosure affect individuals' mutual fund choices?," *Explorations in the Economics of Aging*, 2011, pp. 75–96.

__ , __ , __ , **and** __ , "Simplification and Saving," *Journal of Economic Behavior and Organization*, November 2013, *95*, 130–145.

__ , **James J Choi, David Laibson, and Brigitte C. Madrian**, "Behavioral Household Finance," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics*, Elsevier, forthcoming.

Beverly, Sondra G., Amanda Moore McBride, and Mark Schreiner, "A Framework of Asset-Accumulation Stages and Strategies," *Journal of Family and Economic Issues*, 2003, *24* (2), 143–156.

Bhargava, Saurabh and Day Manoli, "Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment," *American Economic Review*, 2015, *105* (11), 3489–3529.

__ , **George Loewenstein, and Justin Sydnor**, "Choose to Lose: Health Plan Choices from a Menu with Dominated Options," *Quarterly Journal of Economics*, 2017, *132* (3), 1319–1372.

Blaufus, Kay, Jonathan Bob, Jochen Hundsdoerfer, Christian Sielaff, Dirk Kiesewetter, and Joachim Weimann, "Perception of income tax rates: evidence from Germany," *European Journal of Law and Economics*, 2013, *40* (3), 457–478.

Bollinger, Bryan, Phillip Leslie, and Alan Sorensen, "Calorie Posting in Chain Restaurants," *American Economic Journal: Economic Policy*, 2011, *3* (1), 91–128.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience and Consumer Choice," *Journal of Political Economy*, 2013, *121*, 803–843.

Bronchetti, Erin Todd, Thomas S. Dee, David B. Huffman, and Ellen Magenheim, "When a Nudge Isn't Enough: Defaults and Saving among Low-Income Tax Filers," *National Tax Journal*, 2013, *66* (3), 609–34.

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

Brot-Goldberg, Zarek, Amitabh Chandra, Ben Handel, and Jonathan T. Kolstad, "What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics," *Quarterly Journal of Economics*, 2017, *132* (3), 1261–1318.

Brown, Alexandra, J. Michael Collins, Maximilian Schmeiser, and Carly Urban, "State Mandated Financial Education and the Credit Behavior of Young Adults," *Divisions of Research and Statistics and Monetary Affairs Federal Reserve Board, Washington D.C., Finance and Economics Discussion Series*, 2014, pp. 2017–68.

Bruhn, Miriam, Luciana de Souza Leao, Arianna Legovini, Rogelio Marchetti, and Bilal Zia, "The Impact of High School Financial Education: Evidence from a Large-Scale Evaluation in Brazil," *American Economic Journal: Applied Economics*, 2016, *8* (4), 256–295.

Bryan, Gharad, Dean Karlan, and Scott Nelson, "Commitment Devices," *Annual Review of Economics*, 2010, *2*, 671–698.

Busse, Meghan R., Christopher Knittel, and Florian Zettelmeyer, "Are Consumers Myopic? Evidence from New and Used Car Purchases," *American Economic Review*, 2013, *103* (1), 220–256.

_ , Devin G. Pope, Jaren C. Pope, and Jorge Silva-Risso, "The Psychological Effect of Weather on Car Purchases," *Quarterly Journal of Economics*, 2015, *130* (1), 371–414.

Butera, Luigi, Robert Metcalfe, and Dmitry Taubinsky, "The Welfare Effects of Social Recognition: Theory and Evidence from a Field Experiment with they YMCA," *working paper*, 2018.

Calvet, Laurent E., John Y. Campbell, and Paolo Sodini, "Down or Out: Assessing the Welfare Costs of Household Investment Mistakes," *Journal of Political Economy*, 2007, *115* (5), 707–47.

_ , _ , and _ , "Measuring the Financial Sophistication of Households," *American Economic Review: Papers & Proceedings*, 2009, *99* (2), 393–398.

Caplin, Andrew, Daniel Csaba, and John Leahy, "Rational Inattention and Psychometrics," *working paper*, 2018.

Carlin, Bruce I., Li Jiang, and Stephen A. Spiller, "Learning Millennial-Style," *Working Paper, Anderson School of Business, UCLA*, 2014.

Carrera, Mariana, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky, "Can Planning Prompts Change Repeated Behaviors? Evidence from a Randomized Field Experiment on Gym Attendance," *working paper*, 2018.

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

Carson, Richard and W. Michael Hanemann, "Contingent Valuation," in "Handbook of Environmental Economics, Volume 2," Elsevier, 2005, pp. 821–936.

Carson, Richard T., "Contingent Valuation: A Practical Alternative when Prices Aren't Available," *Journal of Economic Perspecitves*, 2012, *26* (4), 27–42.

Caskey, John P., *Beyond Cash-and-Carry: Financial Savings, Financial Services, and Low-Income Households in Two Communities*, Swarthmore, PA: Swarthmore College: Report written for the Consumer Federation of America and the Ford Foundation, 1997.

Chaloupka, Frank J, Kenneth E Warner, Daron Acemoğlu, Jonathan Gruber, Fritz Laux, Wendy Max, Joseph Newhouse, Thomas Schelling, and Jody Sindelar, "An evaluation of the FDA's analysis of the costs and benefits of the graphic warning label regulation," *Tobacco Control*, 2014.

Chamley, Christophe, "Optimal Taxation of Capital Income in General Equilibrium with Infinite Lives," *Econometrica*, 1986, *54* (3), 607–622.

Chetty, Raj, "Moral Hazard versus Liquidity and Optimal Unemployment Insurance," *Journal of Political Economy*, 2008, *116* (2), 173–234.

_ , "Is the Taxable Income Elasticity Sufficient to Calculate Deadweight Loss? The Implications of Evasion and Avoidance," *American Economic Journal: Economic Policy*, July 2009, *1* (2), 31–52.

_ , "Behavioral Economics and Public Policy: A Pragmatic Perspective," *American Economic Review Papers and Proceedings*, 2015, *105* (5), 1–33.

_ , Adam Looney, and Kory Kroft, "Salience and Taxation: Theory and Evidence," *NBER working paper No. 13330*, 2007.

_ , _ , and _ , "Salience and Taxation: Theory and Evidence," *American Economic Review*, 2009, *99* (4), 1145–1177.

_ and Emmanuel Saez, "Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients," *American Economic Journal: Applied Economics*, 2013, *5* (1), 1–31.

_ , John N Friedman, and Emmanuel Saez, "Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings," *American Economic Review*, 2013, *103* (7), 2683–2721.

_ , John N. Friedman, Tore Olsen, and Lugi Pistaferri, "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence From Danish Tax Records," *Quarterly Journal of Economics*, 2011, *126* (2), 749–904.

Choi, James J., David Laibson, and Brigitte C. Madrian, "$100 Bills on the Sidewalk: Suboptimal Investment in 401(k) Plans," *Review of Economics and Statistics*, 2011, *93* (3).

_ , _ , _ , and Andrew Metrick, "Defined Contribution Pensions: Plan Rules, Participant Choices, and the Path of Least Resistance," *Tax Policy and the Economy*, 2002, pp. 67–113.

_ , _ , _ , and _ , "For Better or Worse: Default Effects and 401(k) Savings Behavior," in David A. Wise, ed., *Perspectives on the Economics of Aging*, Chicago: University of Chicago Press, 2004, pp. 81–126.

_ , _ , _ , and _ , "Passive Decision and Potent Defaults," in David A. Wise, ed., *Analyses in the Economics of Aging*, Chicago: University of Chicago Press, 2005, pp. 59–78.

_ , _ , _ , and _ , "Saving for Retirement on the Path of Least Resistance," in Edward J. McCaffery and Joel Slemrod, eds., *Behavioral Public Finance: Toward a New Agenda*, New York: Russell Sage Foundation, 2006, pp. 304–351.

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

Chung, Shin-Ho and Richard J. Herrnstein, "Relative and Absolute Strengths of Resposne as a Function of Frequency of Reinforcement," *Journal of the Experimental Analysis of Animal Behavior*, 1961, *IV*, 267–72.

Cole, Shawn and Gauri Kartini Shastry, "Is high school the right time to teach self-control? The effect of education on financial behavior.," *Unpublished Manuscript, Harvard University.*, 2012.

_ , Thomas Sampson, and Bilal Zia, "Prices or Knowledge? What Drives Demand for Financial Services in Emerging Markets?," *The Journal of Finance*, 2011, *66* (6), 1933–1967.

Collins, J.M., "The impacts of mandatory financial education: Evidence from a randomized field study.," *Working Paper, Center for Financial Security, University of Wisconsin-Madison*, 2010.

Conn, Vicki S., Todd M. Ruppar, Maithe Enriquez, and Pam Cooper, "Medication adherence interventions that target subjects with adherence problems: Systematic review and meta-analysis," *Research in Social and Administrative Pharmacy*, 2016, *12* (2), 218–246.

Corneo, Giacomo and Olivier Jeanne, "Conspicuous Consumption, Snobbism and Conformism," *Journal of Public Economics*, 1997, *66* (1), 55–71.

Costa, Dora L. and Matthew E. Kahn, "Energy Conservation "Nudges" and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment," *Journal of the European Economic Association*, 2013, *11* (3).

Cremer, H., P. De Donder, D. Maldonado, and P. Pestieau, "Voting over type and generosity of a pension system when some individuals are myopic," *Journal of Public Economics*, 2007, *91*, 2041–2061.

_ , _ , _ , and _ , "Designing an optimal linear pension scheme with forced savings and wage heterogeneity," *Internaional Tax and Public Finance*, 2008, *15*, 547–562.

_ , _ , _ , and _ , "Non linear pension schemes with myopia," *Southern Economic Journal*, 2009, *76*, 86–99.

Cremer, Helmuth and Pierre Pestieau, "Myopia, redistribution and pensions," *European Economic Review*, 2011, *55* (165-175).

Davis, Lucas and Christopher Knittel, "Are Fuel Economy Standards Regressive?," *working paper*, 2016.

_ and Severin Borenstein, "The Distributional Effects of U.S. Clean Energy Tax Credits," *Tax Policy and the Economy*, 2016, *30* (1), 191–234.

de Bartolome, Charles A. M., "Which Tax Rate do People Use: Average or Marginal?," *Journal of Public Economics*, 1995, *56* (1), 79–96.

Dekel, E. and B. L. Lipman, "Costly Self-Control and Random Self-Indulgence," *Econometrica*, 2012, *80* (3), 1271–1302.

_ , _ , and A. Rustichini, "Representing Preferences with a Unique Subjective State Space," *Econometrica*, 2001, *69*, 891–934.

_ , _ , and _ , "Temptation Driven Preferences," *Review of Economic Studies*, 2009, *76* (3), 937–971.

DellaVigna, Stefano, "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, 2009, *47* (2), 315–372.

_ and M Daniele Paserman, "Job Search and Impatience," *Journal of Labor Economics*, 2005, *23* (3).

_ and Ulrike Malmendier, "Contract Design and Self-Control: Theory and Evidence*," *The Quarterly Journal of Economics*, 2004, *119* (2), 353–402.

_ and _ , "Paying Not to Go to the Gym," *American Economic Review*, June 2006, *96* (3), 694–719.

_ , Attila Lindner, Balazs Reizer, and Johannes F. Schmieder, "Reference-Dependent Job Search: Evidence from Hungary," *Quarterly Journal of Economics*, 2017, *132* (4), 1969–2018.

_ , John A. List, and Ulrike Malmendier, "Testing for Altruism and Social Pressure in Charitable Giving," *Quarterly Journal of Economics*, 2012, *127* (1), 1–56.

Diamond, Peter, "Optimal tax treatment of private contributions for public goods with and without warm glow preferences," *Journal of Public Economics*, 2006, *90*, 897–919.

Diamond, Peter A., "A many-person Ramsey tax rule," *Journal of Public Economics*, 1975, *4* (4), 335–342.

_ , "A Framework for Social Security Analysis," *Journal of Public Economics*, 1977, *8* (3), 275–298.

_ and Jerry A. Hausman, "Individual Retirement and Savings Behavior," *Journal of Public Economics*, 1984, *23*, 81–117.

Dolan, Paul, Richard Layard, and Robert Metcalfe, "Measuring Subjective Wellbeing for Public Policy: Recommendations and Measures," *Londong School of Economics and Political Science, Center for Economic Performance, Special Paper No. 23*, 2011.

Doyle, Robert J. and Eric T. Johnson, *Readings in Wealth Accumulation Planning*, Bryn Mawr, Pennsylvania: The American College, 1991.

Drexler, Alejandro, Greg Fischer, and Antoinette Schoar, "Keeping It Simple: Financial Literacy and Rules of Thumb," *American Economic Journal: Applied Economics*, 2014, *6* (2), 1–31.

Duflo, Esther and Emmanuel Saez, "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," *Quarterly Journal of Economics*, 2003, *118* (3).

Dupas, Pasaline and Jonathan Robinson, “Why Don’t the Poof Save More? Evidence from Health Savings Experiments,” *American Economic Review*, 2013, *103* (4), 1138–1171.

Dwenger, Nadja, Henrik Kleven, Imran Rasul, and Johannes Rincke, “Extrinsic and Intrinsic Motivations for Tax Compliance: Evidence From a Field Experiment in Germany,” *American Economic Journal: Economic Policy*, 2016, *8* (3), 203–332.

Dworkin, Gerald, “Paternalism,” in Richard A. Wsserstrom, ed., *Morality and the Law*, Wadsworth Publishing Company, 1971.

Easterlin, Richard A., “Does Economic Growth Improve the Human Lot? Some Empirical Evidence,” in P. A. David and W. R. Levin, eds., *Nations and Households in Economic Growth*, Stanford University Press, 1974, pp. 98–125.

Echenique, Frederico, Sangmok Lee, and Matthew Shum, “The Money Pump as a Measure of Revealed Preference Violations,” *Journal of Political Economy*, 2011, *119* (6), 1201–1223.

Eisenstein, Eric M. and Stephen J. Hoch, “Intuitive Compounding: Framing, Temporal Perspective, and Expertise,” *Unpublished Manuscript*, Dec 2007.

Enke, Benjamin and Florian Zimmermann, “Correlation Neglect in Belief Formation,” *Unpublished Manuscript*, 2015.

Ericson, Keith M., “Consumer Inertia and Firm Pricing in the Medicare Part D Prescription Drug Insurance Exchange,” *American Economic Journal: Economic Policy*, 2014, *6* (1), 38–64.

Ernst, Keith, John Farris, and Uriah King, “Quantifying the Economic Cost of Predatory Payday Lending,” Technical Report, Center for Responsible Lending 2004.

Erosa, Andres and Martin Gervais, “Optimal Taxation in Life-Cycle Economies,” *Journal of Economic Theory*, 2002, *105* (2), 338–369.

Farhi, Emmanuel and Xavier Gabaix, “Optimal Taxation with Behavioral Agents,” Working Paper 21524, National Bureau of Economic Research September 2015.

Fehr, H., C. Habermann, and F. Kindermann, “Social security with rational and hyperbolic consumers,” *Review of Economic Dynamics*, 2008, *11*, 884–903.

Fehr, Hans and Fabian Kindermann, “Pension Funding and Individual Accounts in Economies with Life-cyclers and Myopes,” *CESifo Economic Studies*, 2010, *56* (3), 404–443.

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

_ , Peter Katuscak, and Laura Kawano, “Taxpayer Confusion over Predictable Tax Liability Changes: Evidence from the Child Tax Credit,” *Finance and Economics Discussion Series Working Paper*, 2013.

_ , _ , and _ , “Taxpayer Confusion: Evidence from the Child Tax Credit,” *American Economic Review*, 2016, *106* (3).

Feldman, Naomi, Jacob Goldin, and Tatiana Homonoff, "Raisin the Stakes: Experimental Evidence on the Endogeneity of Taxpayer Mistakes," *working paper*, 2015.

Feldstein, Martin, "The opitmal level of social security benefits," *Quarterly Journal of Economics*, 1985, *100*, 303–321.

_ and Jeffrey B. Liebman, *Social security*, 1 ed., Vol. 4, Elsevier,

Fernandes, Daniel, John G Lynch Jr, and Richard G Netemeyer, "Financial literacy, financial education, and downstream financial behaviors," *Management Science*, 2014, *60* (8), 1861–1883.

Findley, T.S. and F. Caliendo, "The behavioral justification for public pensions: a survey," *Journal of Economics and Finance*, 2008, *32* (409-425).

_ and _ , "Short horizons, time inconsistency and optimal social security," *International Tax and Public Finance*, 2009, *16* (487-513).

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

Fluerbaey, Marc and Erik Schokkaert, "Behavioral Welfare Economics and Redistribution," *American Economic Journal: Microeconomics*, 2013, *5* (3), 180–205.

Frey, Bruno S. and Alois Stutzer, "Should National Happiness be Maximized," *CREMA Working Paper*, 2007.

_ , Matthias Benz, and Alois Stutzer, "Introducing Procedural Utility: Not Only What, but Also How Matters," *Journal of Institutional and Theoretical Economics*, 2004, *160* (3), 377–401.

_ , Simon Luechinger, and Alois Stutzer, "The Life Satisfaction Approach to Environmental Valuation," *Annual Review of Resource Economics*, 2010, *2*, 139–160.

Fudenberg, Drew and David Levine, "A Dual-Self Model of Impulse Control," *American Economic Review*, 2006, *96* (5), 1449–1476.

Fujii, Edwin T and Clifford Hawley, "On the Accuracy of Tax Perceptions," *The Review of Economics and Statistics*, 1988, *70* (2), 344–47.

Fujiwara, Daniel and Paul Dolan, "Happiness-Based Policy Analysis," in Matthew D. Adler and Marc Fleurbaey, eds., *The Oxford Handbook of Well-Being and Public Policy*, Oxford: Oxford University Press', 2016, pp. 286–317.

Gabaix, Xavier, "A Sparsity Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2014, *129*, 1661–1710.

_ , "Behavioral Inattention," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics*, Elsevier, forthcoming.

_ and David Laibson, "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets," *Quarterly Journal of Economics*, 2006, *121* (2).

Galperti, Simone, "Commitment, Flexibility, and Optial Screening of Time Inconsistency," *Econometrica*, 2015, *83* (4), 1425–1465.

Gerritsen, Aart, "Optimal taxation when people do not maximize well-being," *working paper*, 2015.

Gideon, Michael, "Survey Measurement of Income Tax Rates," 2014.

Gine, Xavier, Dean Karlan, and Jonathan Zinman, "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation," *Aerican Economic Journal: Applied Economics*, 2010, *2* (4), 213–25.

Glaeser, Edward L., "Paternalism and Psychology," *University of Chicago Law Review*, 2006, *73* (1), 133–156.

Goda, Gopi Shah, Colleen Flaherty Manchester, and Aaron Sojourner, "What Will My Account Really Be Worth? An Experiment on Exponential Growth Bias and Retirement Saving.," *NBER working paper*, 2012, *17927.*

Goldin, Jacob and Daniel Reck, "Preference Identification Under Inconsistent Choice," *Mimeo, University of Michigan*, 2015.

_ **and** _ , "Optimal Defaults and Normative Ambiguity," *working paper, Stanford University*, 2017.

_ **and Tatiana Homonoff**, "Smoke gets in your eyes: cigarette tax salience and regressivity," *American Economic Journal: Economic Policy*, 2013, *5* (1), 302–336.

Goldman, Steven M., "Consistent Plans," *Review of Economic Studies*, 1980, *47* (3), 533–537.

Golosov, Mikhail, Aleh Tsyvinski, and Ivan Werning, *New Dynamic Public Finance: A User's Guide*, MIT Press, 2006.

Graham, Carol, "Subjective Well-Being in Economics," in Matthew D. Adler and Marc Fleurbaey, eds., *The Oxford Handbook of Well-Being and Public Policy*, Oxford: Oxford University Press', 2016, pp. 424–450.

Griffin, James, *Well-Being*, Oxford: Clarendon Press, 1986.

Grossman, S. J. and O. D. Hart, "Disclosure Laws and Takeover Bids," *The Journal of Finance*, 1980, *35* (2), 323–334.

Grossman, Sanford J., "The Informational Role of Warranties and Private Disclosure about Product Quality," *The Journal of Law Economics*, 1981, *24* (3), 461–483.

Gruber, Jon and Emmanuel Saez, "The elasticity of taxable income: evidence and implications," *Journal of public Economics*, 2002, *84* (1), 1–32.

Gruber, Jonathan and Botond Kőszegi, "Is Addiction "Rational"? Theory and Evidence," *The Quarterly Journal of Economics*, 2001, *116* (4), 1261–1303.

_ **and** _ , "Tax incidence when individuals are time-inconsistent: the case of cigarette excise taxes," *Journal of Public Economics*, 2004, *88* (9), 1959–1987.

Gugerty, M., “You Can’t Save Alone: Commitment and Rotating Savings and Credit Associations in Kenya,” *Economic Development and Cultural Change*, 2007, *55*, 251–282.

Gul, Faruk and Wolfgang Pesendorfer, “Temptation and Self-Control,” *Econometrica*, 2001, *69* (6), 1403–35.

_ **and** _ , “The Case for Mindless Economics,” in Andrew Caplin and Andrew Schotter, eds., *The Foundations of Positive and Normative Economics: A Handbook*, Oxford: Oxford University Press‘, 2008, pp. 3–42.

Gustman, Alan L. and Thomas L. Steinmeier, “What People Don’t Know about Their Pensions and Social Security,” in Wiliam G. Gale, John B. Shoven, and Mark J. Warshawsky, eds., *Private Pensions and Public Policy*, Washington, DC: Brookings Institution Press, 2004, pp. 57–125.

_ **and** _ , “Imperfect Knowledge of Social Security and Pensions,” *Industrial REelations*, 2005, *44* (2), 373–397.

Hamermesh, Daniel S., “Consumption During Retirement: The Missing Link in the Life Cycle,” *Review of Economics and Statistics*, 1984, *66* (1), 1–7.

Handel, Ben and Joushua Schwartzstein, “Frictions or Mental Gaps: What’s Behind the Information We (Don’t) Use and When Do We Care?,” *Journal of Economic Perspectives*, 2018, *32* (1), 155–178.

Handel, Benjamin R, “Adverse selection and inertia in health insurance markets: When nudging hurts,” *The American Economic Review*, 2013, *103* (7), 2643–2682.

_ **and Jonathan T Kolstad**, “Health Insurance for “Humans”: Information Frictions, Plan Choice, and Consumer Welfare,” *American Economic Review*, 2015, *105* (8), 2449–2500.

Handel, Benjamin R., Jonathan T. Kolstad, and Johannes Spinnewijn, “Information Frictions and Adverse Selection: Policy Interventions in Health Insurance Markets.,” *working paper*, 2016.

Harrison, Glenn W. and E. Elisabet Rutstrom, “Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods,” in Charles R. Plott and Vernon L. Smith, eds., *Handbook of Experimental Economic Result*, Vol. 1, Elsevier, 2008, pp. 752–267.

Harsanyi, John, “Rule Utilitarianism and Decision Theory,” in H. Gottinger and W. Leinfellner, eds., *Decision Theory and Social Ethics*, Dordrecht: Reidel, 1978.

Hastings, Justine S., Brigitte C. Madrian, and William L. Skimmyhorn, “Financial Literacy, Financial Education, and Economic Outcomes,” *Annual Review of Economics*, 2013, *5*, 347–373.

Hausman, Daniel M., *Preference, Value, Choice, and Welfare*, Cambridge: Cambridge University Press, 2012.

Hausman, Jerry A., “Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables,” *Bell Journal of Economics*, 1979, *10* (1), 33–54.

Hayashi, Fumio, “Why is Japan’s Saving Rate So Apparently High?,” in Stanley Fisher, ed., *NBER Macroeconomics Annual*, Cambridge, Mass.: MIT Press, 1986.

Heinberg, Aileen, Angela A. Hung, Arie Kapteyn, Annamaria Lusardi, Anya Savikhin Samek, and Joanne K. Yoong, "Five steps to planning success. Experimental Evidence from U.S. Households," *Oxford Review of Economic Policy*, 2014, *30* (4), 697–724.

Helliwell, John F. and Christopher P. Barrington-Leigh, "Measuring and Understanding Subjective Well-Being," *Canadian Journal of Economics*, 2010, *43* (3), 729–753.

_ , Richard Layard, and Jeffrey Sachs, *World Happiness Report 2013*, New York: The Earth Institute, Columbia University, 2014.

Heutel, Garth, "Optimal Policy Instruments for Externality-Producing Durable Goods under Present Bias," *Journal of Environmental Economics and Management*, 2015, *72*, 54–70.

Hochman, Harold M. and James D. Rodgers, "The Optimal Treatment of Charitable Contributions," *National Tax Journal*, 1977, *30* (1), 1–18.

Houser, Daniel, Daniel Schunk, Joachim Winter, and Erte Xiao, "Temptation and Commitment in the Laboratory," *Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 488*, 2010.

Imrohoroglu, A., S. Imrohoroglu, and D. H. Joines, "Time-inconsistent preferences and social security," *Quarterly Journal of Economics*, 2003, *118*, 745–783.

Ireland, Norman, "On Limiting the Market for Status Signals," *Journal of Public Economics*, 1994, *53* (1), 91–110.

Ito, Koichiro, "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing," *American Economic Review*, February 2014, *104* (2), 537–63.

Jackson, Matthew and Leaat Yariv, "Present Bias and Collective Dynamic Choice in the Lab," *American Economic Review*, 2014, *104* (12), 4184–4204.

Jones, Damon, "Inertia and Overwithholding: Explaining the Prevalence of Income Tax Refunds," *American Economic Journal: Economic Policy*, 2012, *4* (1), 158–85.

Judd, Kenneth L., "Redistribution Taxation in a Simple Perfect Foresight Model," *Journal of Public Economics*, 1985, *28* (1), 59–83.

Kagan, Shelly, *Normative Ethics*, Boulder, Colorado: Westview Press, 1998.

Kahneman, Daniel, Alan B. Krueger, David Schkade, Norbert Schwarz, and Arthur Stone, "Toward National Well-Being Accounts," *American Economic Review*, 2004, *94* (2), 429–434.

Kaplow, Louis, "A Note on Subsidized Giving," *Journal of Public Economics*, 1995, *58* (3), 469–477.

Karlan, Dean and Jonathan Zinman, "Expanding Credit Access: Using Randomized Decisions to Estimate the Impacts," *Review of Financial Studies*, 2010, *23*, 433–464.

_ , Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman, "Getting to the Top of Mind: How Reminders Increase Saving," *Management Science*, 2016, *62* (12), 3393–3411.

_ , _ , _ , and _ , "Getting to the Top of Mind: How Reminders Increase Saving," *Management Science*, 2016, *62* (12), 3393–3411.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan, "Self-control at work," *Journal of Political Economy*, 2015, *123* (6), 1227–1277.

Kazdin, A. E., *Behavior Modification in Applied Settings*, Long Grove, IL: Waveland Press, 2012.

Koszegi, Botond and Adam Szeidl, "A Model of Focusing in Economic Choice," *Quarterly Journal of Economics*, 2013, *128*, 53–104.

_ and Matthew Rabin, "Choices, Situations, and Happiness," *Journal of Public Economics*, 2008, *92*, 1821–1832.

_ and Matthew Rabinew Rabin, "Revealed Mistakes and Revealed Preferences," in Andrew Caplin and Andrew Schotter, eds., *The Foundations of Positive and Normative Economics: A Handbook*, Oxford: Oxford University Press, 2008, pp. 193–209.

Krusell, P., B. Kuruscu, and A. A. Smith, "Temptation and Taxation," *Econometrica*, 2010, *78*, 2063–2084.

Kunda, Ziva, "The Case for Motivated Reasoning," *Psychological Bulletin*, 1990, *108* (3), 480–98.

Kurz, Mordecai, "On the structure and diversity of rational beliefs," *Economic Theory*, Nov 1994, *4* (6), 877–900.

Lacy, Heather P., Angela Fagerlin, George Loewenstein, Dylan M. Smith, Jason Riis, and Peter A. Ubel, "Are They Really That Happy? Explaining Scale Recalibration in Estimates of Well-Being," *Health Psychology*, 2008, *27* (6), 669–675.

Laibson, David, "Self-Control and Saving," *Working Paper, Department of Economics, Harvard University*, 1994.

_ , "Hyperbolic Discount Functions, Undersaving, and Savings Policy," *NBER Working Paper No. 5635*, 1996.

_ , "Golden eggs and hyperbolic discounting," *The Quarterly Journal of Economics*, 1997, *112* (2), 443–478.

_ , "Life-Cycle Consumption and Hyperbolic Discount Functions," *European Economic Review*, 1998, *42*, 861–871.

_ , "Why Don't Present-Baised Agents Make Commitments?," *American Economic Review*, 2015, *105* (5), 267–272.

_ and Keith Marzilli-Ericson, "Intertemporal Choice," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics*, Elsevier, forthcoming.

_ , Andrea Repetto, and Jeremy Tobacman, "Self-Control and Saving for Retirement," *Brookings Papers on Economic Activity*, 1998, pp. 91–196.

_ , _ , and _ , "A Debt Puzzle," in P. Aghion, R. Frydman, J. Stiglitz, and M. Woodford, eds., *Knowledge, Information and Expectations in Modern Macroeconomics*, Princeton University Press, 2003.

Larsen, Randy L. and Barbara L. Fredrickson, "Measurment Issues in Emotion REsearch," in Daniel Kahneman, Ed Diener, and Norbert Schwarz, eds., *Well-Being: The Foundations of Hedonic Psychology*, Russell Sage Foundation, 1999.

Levy, Matthew R. and Joshua Tasoff, "Exponential Growth Bias and Life Cycle Consumption," *Journal of the European Economics Association*, 2016, *14* (3), 545–583.

Lichtenstein, Sarah and Paul Slovic, *The Construction of Preference*, Cambridge: Cambridge University Press, 2006.

Liebman, Jeffrey B. and Richard Zeckhauser, "Schmeduling," *Working Paper*, 2004.

Lipsey, R. G. and Kelvin Lancaster, "The General Theory of Second Best," *Review of Economic Studies*, 1956-57, *24* (1), 11–32.

List, John A., Robert P Berrens, Alok K. Bohara, and Joe Kerkvliet, "Examining the Role of Social Isolation on Stated Preferences," *American Economic Review*, 2004, *94* (3), 741–752.

Lockwood, Benjamin B., "Optimal taxation with present bias," *working paper*, 2016.

_ and Dmitry Taubinsky, "Regressive Sin Taxes," *NBER working paper No. 23085*, 2017.

Loewenstein, George, "Out of Control: Visceral Influences on Behavior," *Organizational Behavior and Human Decision Processes*, 1996, *65* (3), 272–92.

_ and Ted O'Donoghue, ""We can do this the easy way or the hard way": Negative emotions, self-regulation, and the law," *University of Chicago Law Review*, 2006, *73* (1), 183–206.

Long, Michael W, Deirdre K Tobias, Angie L Cradock, Holly Batchelder, and Steven L Gortmaker, "Systematic Review and Meta-analysis of the Impact of Restaurant Menu Calorie Labeling," *American Journal of Public Health*, 05 2015, *105* (5), e11–e24.

Lührmann, Melanie, Marta Serra-Garcia, and Joachim Winter, "The Impact of Financial Education on Adolescents' Intertemporal Choices," *Unpublished Manuscript, University of Munich*, 2014.

Lusardi, Annamaria, "Information, Expectations, and Savings for Retirement," in Henry J. Aaron, ed., *Behavioral Dimensions of Retirement Economics*, Washington, DC: Brookings Institution Press and Russell Sage Foundation, 1999, pp. 81–115.

_ , "U.S. Household Savings Behavior: The Role of Financial Literacy, Information and Financial Education Programs," in C. Foote, L. Goette, and S. Meier, eds., *Policymaking Insights from Behavioral Economics*, Federal Reserve Bank of Boston, 2009, pp. 109–149.

_ and Olivia Mitchell, "Financial Literacy and Planning: Implications for Retirement Well-being," in Annamaria Lusardi and Olivia Mitchell, eds., *Financial Literacy. Implications for Retirement Security and the Financial Marketplace*, Oxford University Press, 2011, pp. 17–39.

_ and _ , "The Economic Importance of Financial Literacy: Theory and Evidence," *Journal of Economic Literature*, 2014, *52* (1), 5–44.

_ and Olivia S. Mitchell, "Baby Boomer Retirement Security: The Roles of Planning, Financial Literacy, and Housing Wealth," *Journal of Monetary Economics*, 2007, *51* (1), 205–224.

_ and _ , "The Economic Importance of Financial Literacy: Theory and Evidence," *Journal of Economic Literature*, 2014, *52* (1), 5–44.

_ , Anya Savikhin Samek, Arie Kapteyn, Lewis Glinert, Angela Hung, and Aileen Heinberg, "Visual Tools and Narratives: New Ways to Improve Financial Literacy," *NBER Working Paper*, 2014, *20229.*

Luttmer, Erzo F.P. and Monica Singhal, "Tax Morale," *Journal of Economic Perspectives*, 2014, *28* (4), 149–168.

Madrian, Brigitte C and Dennis F Shea, "The power of suggestion: Inertia in 401 (k) participation and savings behavior," *The Quarterly Journal of Economics*, 2001, *116* (4), 1149–1187.

Mandell, Lewis, "The Financial Literacy of Young American Adults: Results of the 2008 National Jump$tart Coalition Survey of High School Seniors and College Students," *Jump$tart Coalition*, 2009, *Washington, D.C.*

Mariger, Randall P., "A Life-Cycle Consumption Model with Liquidity Constraints: Theory and Empirical Results," *Econometrica*, 1987, *55* (3), 533–557.

Meier, Armando, Lukas Schmid, and Alois Stutzer, "Rain, Emotions and Voting for the Status Quo," *IAZ Discussion Paper No. 10350*, 2016.

Milgrom, Paul R., "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 1981, *12* (2), 380–391.

Milkman, Katherine L., John Beshears, James J. Choi, David Laibson, and Brigitte C. Madrian, "Using implementation intentions prompts to enhance influenza vaccination rates," *Proceedings of the National Academy of Sciences*, 2011, *108* (26), 10415–10420.

Mill, John Stuart, *Utilitarianism*, Renaissance Classics, 2012, reprinted.

Miller, Benjamin and Kevin Mumford, "The Salience of Complex Tax Changes: Evidence From the Child and Dependent Care Credit Expansion," *National Tax Journal*, 2015, *68* (3), 477–510.

Mirrlees, James A, "An exploration in the theory of optimum income taxation," *The Review of Economic Studies*, 1971, pp. 175–208.

Mischel, W., "Toward a Cognitive Social Learning Reconceptualization of Personality," *Psychological Review*, 1973, *80* (4), 252–283.

Moser, Christian and Pedro Olea de Souza e Sivla, "Paternalism vs Redistribution: Designing Retirement Savings Policies with Behavioral Agents," *Working paper, Princeton University*, 2015.

Mullainathan, Sendhil, Joshua Schwartzstein, and William J Congdon, "A Reduced-Form Approach to Behavioral Public Finance," *Annual Review of Economics*, 2012, *4*, 1–30.

New, Bill, "Paternalism and Public Policy," *Economics Philosophy*, 1999, *15*, 63–83.

Ng, Yew-Kwang, "A Case for Happiness, Cardinalism, and Interpersonal Comparability," *The Economic Journal*, 1997, *107* (445), 1848–1858.

Nordhaus, William, "Measuring REal Income with Leisure and Household Production," in Alan B. Krueger, ed., *Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being*, University of Chicago Press, 2009, pp. 125–144.

Nozick, Robert, *Anarchy, State, and Utopia*, Basic Books, 1974.

O'Donoghue, Ted and Matthew Rabin, "Doing It Now or Later," *American Economic Review*, 1999, *89* (1), 103–124.

_ **and** _ , "Optimal sin taxes," *Journal of Public Economics*, 2006, *90* (10), 1825–1849.

OECD, *Behavioural Insights and Public Policy: Lessons from Around the World* OECD Publishing, Paris: OECD Publishing, 2017.

Olafsson, Arna and Michaela Pagel, "The Retirement-Consumption Puzzle: New Evidence from Personal Finances," *Working Paper, Columbia Business School*, 2018.

Parfit, Derek, *Reasons and Persons*, Oxford: Oxford University Press, 1984.

Peleg, Bezalel and Mehahem E. Yaari, "On the Existence of a Consistent Course of Action When Tastes are Changing," *Review of Economic Studies*, 1973, *40* (3), 391–401.

Perez-Truglia, Ricardo and Ugo Troina, "Shaming Tax Delinquients: Evidence from a Field Experiment in the United States," *working paper*, 2016.

Pestieau, P. and U. Possen, "Prodigality and myopia: Two rationales for social security," *Manchester School*, 2008, *76*, 629–652.

Piket, Thomas and Emmanuel Saez, "A Theory of Optimal Inheritance Taxation," *Econometrica*, 2013, *81* (5), 1851–1886.

Pollak, R. A., "Consistent Planning," *Review of Economic Studies*, 1968, *35* (2), 201–208.

Read, Daniel and Barbara van Leuwen, "Predicting Hunger: the Effects of Appetite and Delay on Choice," *Organizational Behavior and Human Decision Processes*, 1998, *76* (2), 189–205.

Rees-Jones, Alex, "Quantifying Loss-Averse Tax Manipulation.," *The Review of Economic Studies*, forthcoming.

_ **and Dmitry Taubinsky**, "Measuring Schmeduling," *Working Paper*, 2018.

_ **and** _ , "Taxing Humans: Pitfalls of the Mechanism Design Approach and Potential Resolutions," *Tax Policy and the Economy*, 2018, *1*.

Rehm, L. P., "A Self-Control Model of Depression," *Behavior Therapy*, 1977, *8*, 787–804.

Roemer, John E., *Equality of Opportunity*, Harvard University Press, 1998.

Saez, Emmanuel, "Using elasticities to derive optimal income tax rates," *The Review of Economic Studies*, 2001, *68* (1), 205–229.

_ , "The desirability of commodity taxation under non-linear income taxation and heterogeneous tastes," *Journal of Public Economics*, 2002, *83* (2), 217–230.

_ , "The optimal treatment of tax expenditures," *Journal of Public Economics*, 2004, *88*, 2657–2684.

_ , "Do taxpayers bunch at kink points?," *American Economic Journal: Economic Policy*, 2010, pp. 180–212.

_ **and Stefanie Stantcheva**, "Generalized social marginal welfare weights for optimal tax theory," *American Economic Review*, 2016, *106* (1), 24–45. American Economic Review.

Schelling, Thomas C., "Self-Command in Practice, in Policy, and in a Theory of Rational Choice," *American Economic Review*, 1984, *74* (2), 1–11.

Schilbach, Frank, "Alcohol and Self-Control: A Field Experiment in India," *Mimeo, MIT*, 2017.

Scholz, John Karl, Ananth Seshadri, and Surachai Khitatrakun, "Are Americans Saving 'Optimally' for Retirement?," *Journal of Political Economy*, 2006, *114* (4), 607–643.

Sen, Amartya K., "Plural Utility," *Proceedings of the Aristotelian Society, New Series*, 1980-1981, *81*, 193–215.

_ , *Commodities and Capabilities*, North Holland, 1985.

_ , *Inequality Reexamined*, Harvard University Press, 1992.

_ , "Internal Consistency of Choice," *Econometrica*, 1993, *61* (3), 495–521.

Servon, L.J. and R. Kaestner, "Consumer financial literacy and the impact of online banking on the financial behavior of lower-income bank customers," *Journal of Consumer Affairs*, 2008, *42*, 271–305.

Shefrin, Hersh and Richard H. Thaler, "The Behavioral Life-Cycle Hypothesis," *Economic Inquiry*, 1988, *26*, 609–643.

Sheshinski, Eytan, "The Optimal Linear Income-tax," *The Review of Economic Studies*, 1972, *39* (3), 297–302.

Shipton, P., "The Rope and the Box: Group SAvings in the Gambia," *Report, Department of Anthropology, Boston University*, 1992.

Shogren, Jason, "Experimental Methods and Valuation," in "Handbook of Environmental Economics, Volume 2," Elsevier, 2005, pp. 969–1027.

Skimmyhorn, William L., "Essays in behavioral household finance." PhD dissertation, Harvard Kennedy School, Cambridge, MA 2012.

_ , "Assessing Financial Edcuation: Promising Evidence From Boot Camp.," *USMA Working Paper*, 2015.

Smith, Alec, B. Douglas Bernheim, Colin Camerer, and Antonio Rangel, "Neural Activity Reveals Preferences Without Choices," *American Economic Journal: Microeconomics*, 2014, *6* (2), 1–36.

Soll, Jack B., Katherine L. Milkman, and John W. Payne, "A User's Guide to Debiasing," in Gideon Keren and George Wu, eds., *Wiley-Blackwell Handbook of Judgment and Decision Making*, Wiley-Blackwell Publishing, forthcoming.

Song, Changcheng, "Financial Illiteracy and Pension Contributions: A Field Experiment on Compound Interest in China," *Unpublished Manuscript*, March 2015.

Spinnewijn, Johannes, "Unemployed but optimistic: Optimal insurance design with biased beliefs," *Journal of the European Economic Association*, 2015, *13* (1), 130–167.

_ , "Heterogeneity, Demand for Insurance, and Adverse Selection," *American Economic Journal: Economic Policy*, 2017, *9* (1).

Stango, Victor and Jonathan Zinman, "Exponential Growth Bias and Household Finance," *Journal of Finance*, 2009, *64* (6).

Stevenson, Betsey and Justin Wolfers, "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox," *Brookings Papers on Economic Activity*, 2008, *2008* (1), 1–87.

Straub, Ludwig and Iván Werning, "Positive Long Run Capital Taxation: Chamley-Judd Revisited," 2015.

Strotz, R. H., "Myopia and Inconsistency in Dynamic Utility Maximization," *The Review of Economic Studies*, 1955-1956, *23* (3), 165–180.

Stutzer, Alois and Bruno Frey, "Stress that Doesn't Pay: The Commuting Paradox," *Scandanavian Journal of Economics*, 2008, *110* (2), 339–366.

Sugden, Robert, "The Opportunity Criterion: Consumer Sovereignty Without the Assumption of Coherent Preferences," *American Economic Review*, 2004, *94* (4), 1014–1033.

Sunstein, Cass R., "Nudging: A Very Short Guide," *Journal of Consumer Policy*, 2014, *37* (4), 583–588.

_ **and Richard H. Thaler**, "Libertarian Paternalism Is Not an Oxymoron," *The University of Chicago Law Review*, 2003, *70* (4), 1159–1202.

Taubinsky, Dmitry and Alex Rees-Jones, "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment," *Review of Economic Studies*, forthcoming.

Tenhunen, S. and M. Tuomala, "On optimal lifetime redistribution policy," *Journal of Public Economic Theory*, 12, (171-198).

Thaler, Richard H. and Cass R. Sunstein, "Libertarian Paternalism," *American Economic Review*, 2003, *93* (2), 175–79.

_ **and** _ , *Nudge: Improving Decisions About Health, Wealth, and Happiness*, New Haven: Yale University Press, 2008.

_ **and Hersh Shefrin**, "An Economic Theory of Self-Control," *Journal of Political Economy*, 1981, *89*, 392–406.

_ **and Shlomo Benartzi**, "Save More TomorrowTM: Using Behavioral Economics to Increase Employee Saving," *Journal of Political Economy*, February 2004, *112* (S1), S164–S187.

Toussaert, Severine, "Connecting Commitment of Self-Control Problems: Evidence from a Weight Loss Challenge," *Working paper, London School of Economics*, 2016.

_ , "Eliciting Temptation and Self-Control through Menu Choices: A Lab Experiment," *Mimeo, London School of Economics*, 2017.

Tsvetanov, Tsvetan and Kathleen Segerson, "Re-evaluating the Role of Energy Efficiency Standards: A Behavioral Economics Approach," *Journal of Environmental Economics and Management*, 2013, *66* (2), 347–363.

Wagenaar, William M. and Sabato D. Sagaria, "Misperception of Exponential Growth," *Perception and Psychology*, 1975, *18* (6), 416–422.

Weitzman, Martin L., "Prices vs. Quantities," *The Review of Economic Studies*, 1974, *41* (4), 477–491.