J. Benefit Cost Anal. 2016; 7(1):12–68
doi:10.1017/bca.2016.5

B. Douglas Bernheim

The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics[1]

Abstract: This paper discusses the ways in which behavioral economics challenges the premises of conventional welfare economics. It proposes revised premises that survive those challenges and sets forth a welfare framework derived from that foundation. It argues that the proposed framework is practical, in the sense that it lends itself to applications, as well as unifying, in the sense that it subsumes other approaches and illuminates the relationships between them.

1 Introduction

Over the past twenty years, theories and models of decision making from behavioral economics have come to play increasingly important roles in economic analyses of policy-relevant questions. Though far from universal, the deployment of some behavioral concepts, such as time inconsistency, has become reasonably routine and no longer elicits head scratching or eye rolling from mainstream audiences.

The undeniable success and growing influence of behavioral economics presents us with an important challenge. Normative questions are central to economics. For well over half a century, the dominant approach to those questions has been rooted in the paradigm of revealed preference, which instructs us to infer

1 My views on this topic have been shaped by many years of intense conversation and debate with Antonio Rangel, to whom I owe an obvious intellectual debt. This paper is an abbreviated version of my Clarendon Lectures, which will be published once I finish the manuscript (soon, I hope). Many thanks to Vince Crawford for encouraging me to collect my thoughts on this topic for the purpose of giving those lectures. Luca Braghieri, Shengwu Li, Robert Metcalfe, and two anonymous referees read and commented on a preliminary version of this paper; I am profoundly grateful for their insightful and often challenging observations, which have improved the paper immeasurably. I am also grateful to seminar participants Oxford University and the University of Wyoming, as well as to many individuals whose comments on my earlier papers have influenced the ideas set forth in this one. Finally, I acknowledge financial support from the National Science Foundation through grants SES-0752854 and SES-1156263.
B. Douglas Bernheim: Stanford University, Department of Economics, Stanford Institute for Economic Policy Research, and National Bureau of Economic Research,
e-mail: bernheim@stanford.edu

objectives and welfare ("the good and the bad") from choices. But behavioral economics teaches us that choices are not always consistent ("the ugly"). 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?

The literature contains a hodgepodge of approaches to normative questions in behavioral economics. These include (1) the use of *ad hoc*, context-specific criteria; (2) attempts to extend the revealed preference paradigm to settings where choices express preferences imperfectly; (3) proposals for classifying particular types of choices as "mistakes"; (4) efforts to devise methods entirely rooted in the principle of respect for choice; (5) the use of responses to questions about happiness and other aspects of subjective well-being; and (6) arguments for exercising paternalistic judgments in settings where behavior responds to "nudges." Reading the literature, one can take the impression that behavioral welfare economics has become a bit of free for all.

My own views on this topic have evolved considerably over the past fifteen years. As I no longer find myself in complete agreement with all the positions I have taken previously, this paper is *not* simply a reiteration of my prior efforts (such as Bernheim & Rangel, 2009, and Bernheim, 2009*a*).[2] Nevertheless, my main premise is that, with some different grounding and reinterpretation, the Bernheim–Rangel apparatus can serve as the foundation for a practical and unified approach to behavioral welfare economics that encompasses all of the alternatives mentioned above and illuminates the relationships among them.

The paper covers a great deal of territory, and as a result I have had to skip some important issues and discuss others less thoroughly than I would like. I am in the process of writing a more complete version based on my Clarendon Lectures, which I hope will be available sometime soon, and I refer interested readers there (eventually) for a more complete treatment.

Before diving into details, I will provide a brief outline of the overall argument. Meaningful measurement requires a clear conceptual understanding of what one is trying to measure. Accordingly, I begin in Section 2 with a foundational question: what is economic welfare? I articulate the premises underlying standard welfare economics and consider the ways in which behavioral economics challenges their validity. Most obviously, if people do not reliably exercise good judgment, then perhaps they are not the best arbiters of their own well-being. I reject that broad inference, arguing instead that economic research calls our judgments into question only when it shows they are based on incorrect understandings of the relationships between options and outcomes. Ultimately, I arrive at revised premises that provide

2 For precursors, see Bernheim and Rangel (2007*a*,*b*).

the foundation for a framework based on selective deference to choice. Significantly, I argue that those premises should not presuppose the existence of what behavioral economists often call "true preferences" or aggregate "experienced utility" as those concepts are commonly understood. Rather, they should allow for the possibility (indeed, I would say probability) that we aggregate the many diverse aspects of our subjective experiences only when called upon to do so for particular purposes, such as making a choice or answering a question about our well-being. I also explain why sensible premises anchor the conceptual foundations of a welfare framework in choice rather than in alternatives such as self-reported well-being. Even so, I emphasize that self-reported well-being and other nonchoice data can serve useful purposes in a choice-oriented framework because they help us understand what people would likely choose in settings where correctly informed choices are not observed.

When contemplating the design of a choice-oriented welfare framework, economists are often initially attracted to an approach I call *behavioral revealed preference*, which I discuss in Section 3. The object of this paradigm is to formulate a complete model of choice that separately specifies the consumer's underlying judgments and the cognitive processes that map those judgments to choices. Its intellectual appeal is undeniable, and there are certainly contexts in which its application is compelling. Consequently, a unified normative framework should subsume it. At the same time, an application of that framework demands a thorough structural understanding of choice processes. Unfortunately, our understanding of many (perhaps most) choice phenomena remains partial and imperfect. The uncomfortable aspect of this approach is that it requires us to take strong stands concerning the nature of preferences and decision processes, even when – if we are honest with ourselves – we must acknowledge that we have too little basis for doing so.

How then can we proceed? In Section 4, I outline a two-step approach to evaluating economic welfare. In the first step, we identify all choices that merit deference; in the second, we construct a welfare criterion based on the properties of those choices. As I explain, one can reinterpret behavioral revealed preference as a special case of this approach. It entails serious challenges because it places demanding restrictions on the inputs for the second step: we cannot "recover preferences" unless choices are consistent. This requirement effectively forces us, in the first step, to take uncomfortably strong stands on the nature of judgments and the apparatus that maps them into choices. Consequently, to avoid the need for those stands, we must dispense with the consistency requirement. A key feature of what I will call the *Unified Framework* is that the second step flexibly accommodates inconsistencies among the choices that merit deference. This alternative approach to the second task fundamentally alters the nature of the first, because we are not *compelled* to settle on welfare-relevant domains within which all choices are internally consistent.

Subsequent sections elaborate on the first and second steps of the Unified Framework. With respect to the first step, one recurrent theme in the literature is the notion that fallible consumers can make *mistakes*. There seems to be some agreement that behavioral welfare economics should allow for this possibility and not instruct policy makers to mimic errors. Unfortunately, common definitions of mistakes can lead to circularity: we identify mistakes by looking for choices that conflict with "true preferences," and we infer "true preferences" from choices that are not mistaken. In Section 5, I offer a definition of mistakes that is consonant with the foundational principles set forth in Section 2, and discuss various ways to identify faulty choices. Sometimes this objective requires us to know something about the cognitive apparatus driving choice, but even so we can often make do with more limited information than in the behavioral revealed preference paradigm. After illustrating these points through a recent empirical application involving financial education, I explain how this framework subsumes other studies, as well as concepts such as "biased beliefs" and "libertarian paternalism."

Of course, the viability of the enterprise hinges on our ability to accommodate inconsistencies among choices when constructing the welfare criterion in the second step of the Unified Framework. In Section 6, I set forth some simple properties that the welfare criterion ought to possess, and then observe that they uniquely characterize the *unambiguous choice relation*: we say that one alternative is unambiguously superior to another if and only if the second is never chosen when the first is available. This criterion allows us 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. Thus, for example, if someone is willing to accept various amounts between $3 and $5 for an object in decision frames that merit deference, we do not try to resolve that conflict. Instead, we reach a partial conclusion and live with the ambiguity: owning the object improves the consumer's well-being at least as much as $3, and no more than $5. As I explain, the criterion lends itself to applied analysis, and yields generalizations of standard concepts, such as equivalent variation, compensating variation, and Pareto efficiency. I also summarize a recent empirical application involving default options in 401(k) plans.

2 What is economic welfare?

A common view among economists is that normative evaluations inherently boil down to subjective judgment. As a result, one often hears the assertion that normative questions have no right or wrong answers, or that we cannot resolve them by

evaluating evidence.[3] Regrettably, that perspective excuses a lack of rigor and fosters an "anything goes" approach to behavioral welfare economics that undermines its usefulness.

How can we avoid turning normative inquiry into a free for all? One approach is to agree on a set of fixed and generally applicable philosophical premises. Ideally, those premises will direct our attention to positive issues that we can address objectively, and from which we can derive normative conclusions.

Standard welfare economics embraces this approach. To make headway, it invokes general premises that associate welfare with choices. I would articulate them as follows:[4]

Premise 1: Each of us is the best judge of our own well-being.

Premise 2: Our judgments are governed by coherent, stable preferences.

Premise 3: Our preferences guide our choices: when we choose, we seek to benefit ourselves.

When mainstream economists evaluate the costs and benefits of actual or hypothetical policy interventions, these premises direct their attention to positive questions about the choices people would make if they had the opportunity. Those *are* questions they can address objectively.

To illustrate, suppose we have determined that a policy intervention will result in Norma eating a salad rather than a sandwich. Is she better or worse off? Some might argue that the answer is entirely a matter of opinion, and that one cannot address it objectively. The three premises of standard welfare economics are powerful because they allow us to replace this vexing normative question with a positive one: would Norma *choose* a salad over a sandwich? If she would, then we conclude she is better off with the salad.

One can certainly raise objections to these premises (see, for example, Parfit, 1984, Kagan, 1998 and Hausman, 2012), and indeed philosophers have hotly debated the definitions of welfare and well-being for millennia. From a practical perspective, the possibility that the conventional notion of economic welfare may fall short of a philosophical ideal should not trouble us excessively, as long as that notion captures important aspects of well-being and lends itself to useful implementation. As Kagan (1998) observes, "... from a practical standpoint, at least, our

3 For example, a leading introductory textbook by Paul Samuelson and William Nordhaus explains that "[t]here are no right or wrong answers to [normative] questions because they involve ethics and values rather than facts," while in contrast "[p]ositive economics deals with questions... [that] can all be resolved by reference to analysis and empirical evidence" (Samuelson & Nordhaus, 2001, pp. 7–8).

4 There are several philosophical alternatives to this preference-satisfaction account of well-being; see Parfit (1984) or Kagan (1998). See also Hausman (2012) for a slightly different characterization of the standard economic perspective.

inability to resolve the theoretical dispute may not be debilitating." One can always note qualifications and additional considerations when there is reason to think they are important.

I am concerned here with challenges to these premises that emerge not from philosophy, but rather from behavioral economics, which forces us to reexamine our conventional preconceptions about judgment and choice. In the next three subsections, I consider each premise in turn.

A. Deference to individual judgment

In standard welfare economics, why do we proceed from the premise that each of us is the best judge of our own well-being? As I see it, the argument has two components. The first consists of justifications for individualism and self-determination in the tradition of classical liberalism: my views about my life are paramount because it is, after all, *my* life. The second entails the central Cartesian principle that subjective experience is inherently private and not directly observable.[5] This state of affairs renders each of us uniquely qualified to assess our own well-being. We know how we feel; others can only make educated guesses. These considerations create a strong presumption in favor of deference to our judgments.

According to one argument, behavioral economics overturns this presumption, and thereby challenges the validity of the first premise, by showing that people do not reliably exercise good judgment. As I explain below, that argument is faulty because it conflates what I will call direct and indirect judgments. A direct judgment pertains to outcomes we care about for their own sake – our "ultimate objectives" – whereas an indirect judgment involves alternatives that lead to those outcomes. Behavioral economics impugns various indirect judgments, but not direct ones.

A simple example serves to illustrate the issue. Suppose Norma must choose between two closed boxes, a red one containing apples, and a yellow one containing pears. For the moment, to keep the example as transparent as possible, assume her ultimate goal is to obtain a particular type of fruit rather than to achieve particular mental states such as satiation and satisfaction. On an earlier occasion she looked inside both boxes, but her memory is faulty and she now thinks the yellow box contains bananas. She chooses the yellow box because she likes bananas more than apples. However, if she peered inside the boxes once again, she would choose the red one because she likes apples better than pears.

5 According to Thornton (2004), the principle that "[t]he experiences of a given person are necessarily private to that person" is "of unmistakable Cartesian origin, and... widely accepted by philosophers and nonphilosophers alike."

In this example, a choice between the red and yellow boxes involves an indirect judgment. Norma does not care about the boxes; each one is simply a means to an end. In contrast, under our assumptions, her choices among different types of fruit involve direct judgments. Clearly, when Norma's poor memory causes her to choose the yellow box over the red one, she is exercising poor indirect judgment. However, there is nothing wrong with her direct judgment. Indeed, we consider her indirect judgment faulty precisely between her direct and indirect judgments diverge.

Now let us add a wrinkle: assume Norma's ultimate goal is to achieve certain mental states ("internal goods"). From that perspective, all consumption items ("external goods") are means to ends, and choices among them always involve indirect judgments.[6] Moreover, just as Norma may misjudge the contents of a box, she may also misapprehend the relationships between consumption goods and mental states. However, assuming she is sufficiently familiar with apples, pears, and bananas to understand the consequences of eating each, her indirect judgments among open boxes will be correctly informed, and hence will faithfully reflect her direct judgments.

Behavioral economics and psychology provide us with ample reason to question certain types of indirect judgments. No doubt some would claim there is also evidence that people exercise poor direct judgment – for example, that they like certain goods or experiences "too much" and others "not enough." Let us be clear, however: existing economic research demonstrates nothing of the sort.[7] Specific claims concerning instances of poor judgment usually turn out upon close examination to involve indirect evaluations. The occasional objection to a direct judgment entails nothing more than a difference of opinion between the analyst and the consumer as to what constitutes a good or fulfilling life. If Norma thinks she is better off with apples than with pears knowing full well the consequences of consuming each, an analyst who weighs the various dimensions of experience differently can certainly disagree, but there is no objective foundation for overturning the presumption articulated at the start of this section and declaring the analyst's perspective superior.

6 To be clear, I take no stand on the question of whether consumers ultimately judge outcomes based on internal goods, external goods, or both. The framework described herein encompasses all of these possibilities.

7 One can of course take the position that certain direct judgments are morally flawed. Economists occasionally adopt this perspective; for example, see Harsanyi (1978), who argues against respecting judgments motivated by sadism, resentment, and the like, or Sen (1980-1981). The conventional economic framework seeks to assess well-being without factoring in these types of moral considerations, concerning which economists have no special expertise. I follow that tradition.

A determined critic might argue that (1) observed judgments are generally indirect, and (2) defects in human reasoning are so pervasive that they taint *all* indirect judgments. In that case, the first premise is useless: as analysts, we may aspire to respect each person's direct judgments, but we lack the information required to do so. Some form of paternalism then becomes unavoidable. For the most part, however, behavioral economists do not credit the view that human decision making is universally defective, and indeed our research generally points to concerns that arise within limited domains.

Thus, despite initial appearances, behavioral economics does not overturn the first premise of standard welfare economics. It does, however, potentially qualify that premise. Each of us may be the best judge of our own well-being, but all our indirect judgments are not created equal.

B. Preferences versus contextual aggregation

Choice anomalies are the bread and butter of behavioral economics. Some economists try to rationalize them while preserving the assumption of stable, coherent preferences, either by creatively redefining the objects of choice, or by positing imperfect decision-making processes that project preferences into choices with distortions. According to this view, "true preferences" actually exist inside our heads, and we access them (perhaps imperfectly) each time we are called upon to make a choice. When formulating models, economists usually posit the existence of a utility function, U, that embodies these preferences. Sometimes behavioral economists also assume that, upon receiving a particular alternative, call it x, we actually experience $U(x)$ as a subjective sensation, which some would call "experienced utility."[8]

An alternative perspective on choice anomalies assumes that each of us acts on the basis of *multiple* objective functions, which we harmonize inconsistently. This view likewise holds that true preferences actually exist inside our heads, but it allows for the possibility that different versions of those preferences may either coexist or successively replace each other. I tend to equate this view with the hypothesis that we all suffer from a mild form of multiple personality disorder, and indeed behavioral economists often describe these theories as envisioning "multiple selves."

The concepts of preferences and utility are so ingrained among economists that we naturally gravitate to one of the two preceding alternatives. However, there is

8 Unfortunately, this phrase does not have a precise definition, and different people appear to use it differently. For example, to some, it references a stream of instantaneous hedonic sensations, and may omit various considerations affecting choice.

a third possibility, one derived from psychologists' notion of "constructed preference," that for me has the greater ring of truth. According to this view, I aggregate the many diverse aspects of my experience only when called upon to do so for a given purpose, such as making a choice or answering a question about my well-being.[9] For instance, at a given point in time, I may be troubled by a financial setback, happy about some recent professional success, worried about a conflict with a family member, pleased by the taste of a good wine, and irritated by the itch of a mosquito bite, but nevertheless experience no comprehensive sensation of well-being.[10] To answer a question about my overall welfare, or to choose between alternatives without deploying a previously constructed rule of thumb,[11] I must weigh the positives against the negatives and construct an answer *de novo*. I cannot simply access an aggregate sensation that is already part of my subjective experience, or consult an overall preference ordering that already resides inside my head, because neither of these aggregates exist until I am called upon to deliberate and aggregate. From this perspective, the concepts of "true preferences" and "experienced utility" are fictions; they may play useful roles in "as-if" representations of behavior, but we should not take them literally.

This third perspective potentially attributes particular choice anomalies to the vagaries of aggregation. In particular, when I deliberate and aggregate, the weights I attach to the various dimensions of my subjective experience may be sensitive to context. For example, circumstances may render one aspect of experience more psychologically salient than another. In that case, even my direct judgments may be context-dependent. To be sure, each time I aggregate, I may well deploy similar principles and criteria, and come to conclusions that are at least roughly consistent, in which case an analyst may be tempted to infer that I actually have well-defined preferences that imperfectly govern my choices. However, under this third view, my consistency simply means that I use similar aggregation procedures in different contexts, not that my aggregation procedures are all derived from my "true preferences."

9 Lichtenstein and Slovic (2006) write as follows: "One of the main themes that has emerged from behavioral decision research during the past three decades is the view that people's preferences are often constructed in the process of elicitation." See also the discussion in Hausman (2012).

10 The notion that life consists of highly disaggregated subjective experiences has a long philosophical tradition; see, for example, Aristotle (2011, translation), Mill (2012, reprinted), and more recently Sen (1980-1981), who advocates a vector view of utility.

11 Consumers may adopt rules of thumb to streamline decision making. Imagine, for example, that careful deliberation would lead Norma to choose apples over pears in almost all states of nature. To avoid the costs of deliberation, she may adopt a simple habitual rule such as "always choose apples over pears." In that case, she may appear to make decisions by accessing her true preferences, but she actually does so by accessing a previously constructed rule of thumb, and thereby deploying a cognitive shortcut.

In light of the foregoing, I find the second premise of conventional welfare economics untenable. Fortunately, it is also unnecessary. The first (modified) premise instructs us to defer to consumer's direct judgments and correctly informed indirect judgments.[12] The preceding discussion alerts us to the fact that those judgments may not be monolithic. In deferring to the individual, we may simply have to live with the possibility that some evaluations will be inconclusive.[13]

C. Why choice?

As explained in the previous section, there are sound reasons for skepticism concerning a literal interpretation of the hypothesis that "preferences guide our choices." Even so, a behavioral economist can comfortably endorse the principle that "when we choose, we seek to benefit ourselves." Accordingly, while the third premise of conventional welfare economics does not emerge unscathed from this discussion, we can still treat consumers' choices as shedding useful light on their judgments.

To be sure, as emphasized in Section 2.A, a judgment is not worth respecting if it is based on a misunderstanding. Choices are often susceptible to misunderstandings because they generally involve indirect judgments, for instance about physical goods we value for the mental states they deliver, rather than direct ones. It follows that there may be valid reasons for deferring to the judgments behind some but not all of our choices.

The question remains, why should we draw the line at choices? Why not accord equal status to other types of judgments, such as evaluations of happiness and life satisfaction? Alternatively, why not try to build an even better welfare framework around self-reported well-being (henceforth abbreviated SRWB)?[14] 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, one could argue that other types of constructed judgments should be equally admissible for the purpose of evaluating welfare.

12 In a similar spirit, Griffin (1986) proposes a theory of welfare based on "ideal preferences," which putatively reflect the desires we would have if we were fully informed, clear-headed, unbiased, free from prejudice, and the like.

13 I see no foundation for Brandt's (1979) assertion that conflicts among an individual's judgments render notions of welfare based on desire satisfaction "unintelligible." Indeed, in subsequent sections, I explain how one can coherently accommodate those conflicts.

14 The phrase "subjective well-being" (abbreviated SWB) is more commonly used in the literature. I prefer the phrase "self-reported well-being" (SRWB) because it reminds us that subjective well-being is actually unobserved.

Traditionally, economists have been dismissive of SRWB. As a group, we derive this perspective from the revealed preference tradition, which holds that choice is observable and measurable, while well-being is inherently unobservable and not directly measurable.[15] Proponents of SRWB view this position as overly simplistic and easily refuted.[16] Certainly, *self-reported* well-being is no less observable or measurable than choice.[17] Moreover, the purported premises for a welfare framework based on SRWB – that people tend to know how they feel and are generally willing to say – strike its devotees as no less reasonable than those justifying a choice-based approach. Indeed, some would argue that SRWB elicits evaluations of actual subjective experience more directly than choice. Many economists have become increasingly sympathetic to this view and, as a result, SRWB has made significant inroads into economic research.

While choice and SRWB may both entail judgments that the individual constructs without referencing "true preferences," this state of affairs does not necessarily place them on an equal footing. When conducting normative analysis, deference to a constructed judgment is warranted only if the purposes of the analysis and the judgment are conformable.

Naturally, different people may have different purposes in mind when they make normative evaluations. That said, in my experience, economists typically 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*. Significantly, when people make choices for themselves, they aggregate over the many dimensions of their experience *for precisely the same purpose* – that is, to make a selection that promotes their well-being. Thus, the purposes of constructing judgments for normative analysis on the one hand, and for making choices on the other, are conformable. When advising policy makers on the selection of an alternative that affects Norma, we defer to Norma's choices because they reveal the alternatives that, in her judgment, would provide her with the greatest overall benefit if selected.

Critically, other types of constructed judgments aggregate experience for entirely different purposes. Granted, if we interpreted the purpose of normative economic analysis differently, we might construe it as sharing those objectives –

15 See, for example, Gul and Pesendorfer (2008), whose position on this issue is uncompromising.

16 Indeed, few economists adhere rigidly to this position. For example, when asked to defend the assumption that choice reflects well-being, they may point to stated intentions as corroboration. Likewise, Gul and Pensendorfer's (2008) notion that one can use nonchoice information to "motivate" a particular model is difficult to distinguish from the premise that such information allows one to differentiate empirically between that model and potential alternatives.

17 Regrettably, many SRWB practitioners have obfuscated this point by using the phrase "subjective well-being," which suggests to skeptics that the object is to measure the unmeasurable.

for example, we might say that its purpose is to aid the selection of alternatives that induce people to report the highest level well-being in response to survey questions. However, on close examination such positions prove difficult to defend. To make this point more concrete, in the next subsection I evaluate the conceptual case for building a welfare framework around SRWB instead of choice.[18]

D. A closer look at self-reported well-being

Possible normative foundations

The most direct conceptual route to an SRWB-oriented welfare framework, and the one I sense most proponents of the approach implicitly have in mind, posits the existence of a utility function, U, that not only embodies "true preferences," but also describes the aggregate subjective well-being the consumer would actually experience upon receiving each alternative. Under this view, we can try to apprehend the "underlying truth," U, by examining either choices or reports of well-being. Because each alternative can in principle reveal U, neither can claim the conceptual high ground. Instead, one is entitled to argue for the practical superiority of either approach based on the plausibility of the assumptions that tie the observations – either choices or reports of well-being – to U.

I reject that perspective. As I explained previously, there is no compelling reason to think that people aggregate their inclinations or experiences except when called upon to do so for specific purposes. Under my view, choice and SRWB both involve the construction of judgments, not the apprehension of underlying truths, because "true preferences" and aggregate "experienced utility" do not actually exist. Instead, each judgment expresses its own truth concerning aggregation. To determine its relevance for normative analysis, one must ask whether the principles of its construction match the analyst's objectives.

While the purpose of choice is to make a selection, the purpose of SRWB is 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.

18 The literature contains a number of excellent commentaries concerning the limitations of SRWB. Examples include Frey and Stutzer (2007), Nordhaus (2009), Dolan et al. (2011), and Dolan and Metcalfe (2012). Much of this literature focuses on the measurement of the flow of well-being at a point in time (for instance, the National Well-Being Accounts of Kahneman et al., 2004). As a result, it tends to emphasize somewhat different issues, such as interpersonal comparability and what Nordhaus calls the "zero problem." Still, others have previously raised several of the issues I discuss here, albeit sometimes in different guises.

Accordingly, the respondent's underlying motivation for selecting one set of words rather than another is murky at best.

That said, let us begin with the most favorable assumption: for whatever reason, when people answer such questions, they feel obliged to respond truthfully. Even then, their narrow purposes in constructing expressed judgments depend on how they interpret the question. As analysts, we can try to promote a particular interpretation by carefully crafting the question's wording, but what interpretation should we intend? Having dispensed with the notion that people can access underlying pre-existing truths about their aggregate well-being, two possibilities remain.

- First, we may intend for respondents to mimic the aggregation principles implicit in their choices, for example by imagining what they would choose.[19]
- Second, we may have in mind some normative ideal other than correctly informed choice, which we are trying to invoke by using particular words and phrases.

The first possibility endows SRWB with normative significance only through its correlation with choice. Thus, if there is a case for a welfare framework in which SRWB plays a primary conceptual role rather than a derivative one, it lies in the second possibility. To pursue it, one would have to both impugn the principle of deference to correctly informed choice and articulate a concrete alternative. Certainly, choice-based normative standards are not above criticism; for example, one can argue against deference to sadism. However, to justify some alternative normative ideal, one would have to wade into thousands of years of philosophical controversy and emerge with a specific proposal.

Conceptual problems with elicitation

Determined advocates for SRWB might be willing to take on this challenge, or alternatively they might simply insist that aggregate well-being is an actual sensation that people can access and report. In either case, we must still ask ourselves whether questions about SRWB are likely to invoke the intended concepts. There are two conceptual problems.

First, to arrive at an appropriately worded survey question, we have to take a stand on the thorny issue of what constitutes well-being. Different philosophical traditions point in different directions. Should we ask people to evaluate the balance

19 For example, when people wonder whether others are better off than they are, they often asks themselves the question, "would I switch places with him (or her)?" Such thinking turns questions about self-reported well-being into hypothetical choices.

between pleasure and pain, in the spirit of quantitative hedonism? To focus more broadly on their mental states? To factor in moral considerations apart from their direct implications for those states? Or to consider the correspondence between the way the world is, and the way they would like it to be? In contrast, a choice-based framework allows us to finesse this issue by evaluating each individual's welfare according to their own conception of what constitutes the good and the bad.

The second conceptual problem involves linguistics. The phrases that economists, psychologists, and philosophers use to describe normative ideals, such as "experienced utility," are terms of art. Necessarily, the SRWB method attempts to elicit them through questions that employ related natural language. But people construe common words and phrases according to their own experiential associations, rather than the rigorous principles the analyst intends. Thus, the entire enterprise hinges on the vagaries of meaning attributed to particular natural words and phrases, none of which specifically conjure the concept of interest.

To make this abstract point more concrete, consider an example wherein choice and self-reported well-being conflict.[20] While attending a party, Norma says she would be happier drinking wine than soda, but she nevertheless chooses soda. She explains this apparent conflict by noting that she is better off drinking soda because she has to drive home. The purpose of this example is to make the point that an answer to a question about happiness depends on how the respondent construes the word "happy." Here, Norma's construction is narrow, so she leaves out dimensions of experience that are important both for her well-being and for choice.

This problem is endemic, because linguistic constructions are governed by the vicissitudes of an individual's experience, rather than by systematic normative principles. Norma may have learned the meaning of the word "happy" early in her linguistic development by hearing her parents describe her that way whenever she smiled. For Norma, welfare analysis based on self-reported happiness would then favor the types of subjective experiences she had when she was learning to speak and her parents saw her smiling. Fostering child-like joy is certainly not a horrible objective, but we should not be too troubled in cases where it conflicts with choice.

Some economists and psychologists view this issue as a practical problem rather than a conceptual one, and believe they can address it by asking broad questions with the object of encompassing everything that might contribute to overall well-being. For example, many studies ask about "life satisfaction" rather than happiness. It is important to realize, however, that the conceptual problem has two components. The first is scope: the SRWB question must elicit responses concerning the full range of subjective experience. The second is weighting: the SRWB question must induce the respondent to aggregate the various dimensions of subjective

20 Instances where they agree provide no basis for preferring one to the other.

experience using the same weights that are implicit in the intended normative ideal. Asking broad questions potentially addresses the issue of scope, but not the issue of weighting. Weights will still depend upon the types of subjective experiences the respondent happens to associate with whatever word or phrase is used. For example, if Norma's parents repeatedly told her she "must feel very satisfied" each time she completed some difficult task, the aggregator she implicitly invokes when thinking about satisfaction may place disproportionate weight on those types of feelings. Thus, Norma may choose wine over soda because wine will promote her self-assessed well-being more effectively, while nevertheless acknowledging that she would get greater satisfaction from choosing soda because abstinence would require her to exercise will power successfully.

Admittedly, proponents of an SRWB-oriented framework could devise other examples in which the choice, rather than the assessment of well-being, appears problematic. For example, when confronted with her apparent inconsistency, Norma might say, "I wish I could get myself to stop drinking so much!" Note, however, that this statement conjures up another decision problem along with an associated choice that *is* consistent with her assessment of well-being. Specifically, assuming it is sincere, it implies that Norma would choose soda over wine if she could make that commitment in advance. Accordingly, such examples only serve to make the point, acknowledged by all behavioral economists, that choices are not always internally consistent. Because self-reported well-being is also subject to internal inconsistency, this observation is not a valid reason to prefer one approach to the other. Regardless of which approach we choose, we will require methods for dealing with such inconsistencies.

So far, I have focused on examples involving ambiguity concerning the principles governing aggregation of subjective experiences at a single point in time. Aggregation over time is even more conceptually problematic for an SRWB-oriented framework. To illustrate, let us imagine for the moment that people actually experience sensations of aggregate well-being, which we would like to elicit. When formulating intertemporal preferences, economists typically assume that utility at time t, written U_t, depends on the stream of "instantaneous utilities" enjoyed from time t forward, $(u_t, u_{t+1}, u_{t+2}, \ldots) \equiv w_t$. Accordingly, we write $U_t = A(w_t)$, where A is the intertemporal aggregator employed at time t. For example, with geometric discounting, $A(w_t) = \sum_{s=0}^{\infty} \delta^t u_t$.

The object of some SRWB procedures is to assess u_t at each point in time, for example by periodically "pinging" people on their mobile phones with a question such as, "how are you feeling at this moment?" (See, for example, the discussion of Experience Sampling Methods in Kahneman, Krueger, Schkade, Schwarz & Stone, 2004). Critically, this approach performs no intertemporal aggregation. In fact, it

tells us nothing at all about the aggregator A. Even in the best-case scenario, we only recover the stream of instantaneous utilities, which does not by itself permit us to conduct welfare analysis.

One can try to address this problem within the SRWB framework by asking people to report how they feel not only about events that are happening concurrently, but also about their future prospects. Here we encounter a thorny point of interpretation: what precisely do we mean by "feelings about the future"? An important line of research in behavioral economics acknowledges that anticipatory emotions, such as anxiety, hope, and fear, contribute to our hedonic states (see, for example, Caplin & Leahy, 2001, Bernheim & Thomadsen, 2005, Koszegi, 2006). We model that phenomenon by allowing instantaneous utility, u_t, to depend not only on concurrent consumption, c_t, but also on subsequent experiences: $u_t = h(c_t, u_{t+1}, u_{t+2}, \ldots)$.[21] Thus, u_t captures one type of "feelings about the future." Critically, U_t encompasses a second type of "feelings about the future": how we weigh current versus future instantaneous utility when evaluating our overall well-being.

Now we come to the critical question: once we acknowledge the fact that people have anticipatory emotions, how do we formulate a question that elicits overall utility, U_t, and not merely instantaneous utility, u_t? What language would allow respondents to understand that we do not simply mean the first type of feelings about the future, and that we also want them to report feelings of the second type? If we wanted to elicit u_t instead, how would we change the question? The model does not offer useful guidance. Furthermore, the problem here goes well beyond the limitations of natural language: I am not even sure how I would phrase a question to another economist to elicit U_t rather than u_t, other than by invoking choice.[22]

My reflexive instinct to pose questions concerning U_t in terms of choice is far from accidental. Feelings about the future that are captured by U_t but not by u_t presumably involve intellectual judgments rather than hedonic sensations; otherwise, u_t would already subsume them. Indeed, in this setting, the term "experienced utility" evokes the temporally disaggregated stream of instantaneous utilities, $(u_t, u_{t+1}, u_{t+2}, \ldots)$, not overall utility, U_t, which is never hedonically experienced. As far as I can tell, even if the individual experiences u_t as a continuous,

21 In writing this specification, I have implicitly assumed perfect foresight for convenience.

22 A similar problem arises with respect to moral judgments. When making choices, the moral implications of our alternatives may matter to us both because we expect them to affect our mental states, and because we care directly about acting morally. What language would allow respondents to understand that we are asking about both concepts, and not just the first?

atemporally aggregated sensation, U_t does not exist unless we go through the intellectual exercise of making a choice or constructing an answer to a question based on linguistic associations.[23]

Motives

All of the preceding was predicated on the optimistic assumption that people feel obliged to answer questions about well-being truthfully. Do they? Economists sometimes criticize SRWB on the grounds that there are no consequences for giving incorrect answers, but that is not entirely accurate. As a general matter, respondents give one answer rather than another because they perceive differential consequences. To understand the principles governing the construction of the judgments embedded in SRWB, one would have to determine the nature of those perceptions.

A respondent who anticipates feeling guilty about lying (a consequence) plainly has an incentive to tell the truth. That said, answers may have other incidental consequences that provide people with incentives to misreport their true feelings. For example, I may be tempted to provide responses that speak well of my character.[24] My innate honesty tempers that tendency, but only to a degree. Indeed, an aversion to lying is of no help whatsoever if I talk myself into believing an answer that helps me sustain a self-serving personal narrative. Incidental consequences can also create incentives for respondents to give questions only cursory consideration. After all, deliberation is costly, and people may be particularly averse to contemplating negative sensations. There is no reason to think that honesty counterbalances those considerations, inasmuch as one can report superficial judgments truthfully.

Admittedly, economists routinely rely on other types of survey data that are subject to similar biases. For instance, we use survey responses to questions about income and charitable contributions even though we know people may exaggerate both to look good. That said, we certainly recognize the potential severity of the

23 To be fair, one can imagine preference formulations for which these problems do not arise. For example, let us assume we can separate instantaneous utility into two components, one reflecting the hedonic value of current activities, u_t^c, the other capturing feelings about the future, u_t^f, so that $u_t = u_t^p + u_t^f$. If we assume in addition that people do not have anticipatory feelings about future anticipatory feelings (for example, they do not hope to be hopeful or fear being afraid), we can write $u_t^f = \sum_{s=1}^{\infty} \delta^s u_{t+s}^p$. In that case, $u_t = \sum_{s=0}^{\infty} \delta^s u_{t+s}^p$, in which case one does not need to distinguish between hedonic utility, u_t, and overall utility, U_t. However, if our object is to formulate a general framework, the possibility of avoiding the conceptual problem in a special case offers relatively little comfort.

24 Another possibility is that I may have an incentive to exaggerate my preferences if I think the resulting SRWB analysis will be politically impactful; see Frey and Stutzer (2007).

problem, and rarely use surveys when more reliable alternatives, such as administrative records, are available. Moreover, when there are no alternatives, we typically insist on validating survey data by comparing it to reliable benchmarks. In contrast, there is no way to validate SRWB. We simply cannot tell whether honesty meaningfully tempers other incidental incentives in settings where the truth is subjective and inherently unverifiable.

Preliminary findings based on some pilot surveys I have fielded as part of an ongoing research project with Jim Andreoni underscore the potential severity of these concerns. One curious finding is that, when reporting happiness on a scale of 1 to 7 (where 7 is extremely happy and 1 is extremely unhappy), very few people select an answer below 4. And yet, when asked about happiness at some previous point in time, such answers are common. Apparently, people are willing to admit that they remember being sad in the past, but not that they are currently sad. A second finding involves a comparison between two groups of subjects randomly selected from the same population. One is asked to report the level of happiness they would expect to feel after some adverse event. The second is asked to say how they think they would respond to a question about their level of happiness after the same adverse event. Curiously, the level of happiness the second group says they would report is significantly higher than the level the first group says they would feel. Presumably, this discrepancy reflects an awareness that reports of well-being are systematically skewed. For example, people may recognize that they are hesitant to admit sadness.

Scaling and recoverability

Thus far, my critique of SRWB has focused on the mismatch between the purposes behind these constructed judgments and the objectives of normative economic analysis. Setting that issue aside, the pertinent literature also explores various practical objections to the use of SRWB as a welfare measure. One of these strikes me as particularly problematic: we always measure SRWB on a unitless scale. As a result, respondents have to decide what the numbers mean, and the nature and context of the question may affect that decision. For example, people may adopt different uses of the scale according to whether the question asks about current, past, anticipated, or hypothesized experiences. Even focusing narrowly on questions about current well-being, the most natural meanings of the numerical responses are context-specific. For example, the respondent might treat 4 as "typical" because it is in the middle of the 1-to-7 range. If an event occurs that alters what is typical, the manner in which he normalizes the scale would then change. Consider, for example, the case of colostomies (Loewenstein & Ubel, 2008). Understandably, most people say

they think they would be extremely unhappy if they had to have a colostomy. Yet people who have had colostomies report feeling just about as happy as the rest of us. Could this reflect a general tendency to underestimate adaptability? Probably not: colostomy patients are often willing to pay large sums to reverse the procedure. It is more likely that their measured happiness reflects "the new normal," and possibly a reluctance to admit sadness. Likewise, celebrated results in the literature, such as the absence of a strong relationship between SRWB and income (the Easterlin paradox),[25] may be attributable to confounding changes in normalizations.

More generally, the problem here is that we have no good way to distinguish between changes in underlying well-being and changes in the way people interpret a unitless scale. As a formal matter, absent additional assumptions, these two effects are not separately recoverable, in the sense that one cannot identify their separate effects even with ideal data (see Bernheim, 2009*a*, for an extended discussion of this point). While the SRWB literature acknowledges the possibility that changes in the interpretation of the well-being scale may confound comparisons, such commentaries generally fail to address the question of recoverability; see, for example, the discussion of scaling in Dolan, Layard and Metcalfe (2011). Some studies claim to measure rescaling separately from effects on happiness, but they rely on supposedly intuitive assertions rather than rigorous accounts of identification, and close examination reveals that their conclusions hinge on unstated and arbitrary assumptions.[26] Absent a formal treatment of recoverability and identification, it is difficult to know what to make of the conventional SRWB agenda.

Is SRWB nevertheless useful?

Some economists are unimpressed by the preceding considerations. They insist that SRWB is worth studying for practical reasons: common sense tells us that answers to questions about well-being are meaningful, and research confirms that they are correlated with other welfare measures derived from choices and biometrics. According to this argument, the concept may not be ideal, but it is practical and useful. To be clear, I largely agree with this perspective. By no means do I intend to imply that SRWB responses are meaningless or useless. Rather, my point is that the conceptual foundations of our welfare framework should invoke choice rather than SRWB. Moreover, once we establish those foundations, we can think more

25 See Easterlin (1974). More recent research suggests that there is a relationship between happiness and income, but that it is too weak for Easterlin to have detected in his small sample; see Stevenson and Wolfers (2008).

26 For example, Lacy et al. (2008) implicitly assume that people use the same scale when rating their own current experiences and others' hypothetical experiences.

rigorously about the *legitimate* uses of SRWB data. Though there are such uses, they may not include taking survey responses at face value as reliable measures of well-being. I would draw an analogy here to the literature on hypothetical choices. No one disputes the easily demonstrated fact that hypothetical choices are highly correlated with real choices. And yet, it is widely acknowledged that hypothetical choices are subject to systematic biases that are not easily corrected, likely because these tasks are only incidentally consequential (as are SRWB questions).[27] In some settings, these biases render responses highly unreliable when taken at face value (see Bernheim, Bjorkegren, Naecker & Rangel, 2015, for an example). While we cannot directly evaluate the accuracy of SRWB responses in the same way as hypothetical choice, there is no reason to think that the issues would be any different.

E. Strategies for integration

A fan of SRWB might react to the preceding observations by pointing out that choice is no picnic either. I could not agree more: practical problems crop up in both contexts. One problem with a choice-oriented approach is that we cannot always observe choices within the pertinent domain. Consider the problem of evaluating the welfare impact of policies that reduce the likelihood of oil spills. The typical individual makes no choices that directly and measurably impact that probability. Another concern mentioned above and discussed at length in Section 5 below is that some choices are faulty, for example because people misunderstand or ignore the connections between options and outcomes.

Sometimes we can "fill in" the missing choices and/or "correct" the faulty ones by extrapolating from observed decisions or by conducting experiments, but in other cases the decision of interest is too far removed from prior experience and too costly to implement. That is when measures of SRWB, as well as other types of subjective evaluations and even biometrics (including facial expression, skin conductance, blood pressure, brain activity, and the like), may be particularly helpful. Because they are highly correlated with behavior, we can use them to predict the choices people would make if they had the opportunity in settings where no actual choices are observed. In addition, because we can measure reactions to actual outcomes, we can also in principle predict the selections people would have made had

27 See, for example, Murphy, Allen, Stevens and Weatherhead (2005). Scholars working in this area have tried to design various protocols to "fix" the hypothetical questions. One strategy is to emphasize the importance of the research and entreat respondents to be serious and honest (Cummings & Taylor, 1999). Another is to ask them to take "solemn oaths" (Jacquemet, Joule, Luchini & Shogren, 2013). One could deploy the same techniques in the SRWB setting. However, their effectiveness in the hypothetical choice setting is controversial (see, for example, Bernheim et al., 2015).

they understood the consequences of their actions correctly in settings where they were actually confused.

Consider the following examples.

- By estimating functions relating SRWB to environmental conditions, income, and other covariates, we may be able to predict the average willingness to pay for environmental goods, even though no individual makes choices that meaningfully influence the levels of those goods (see Frey, Luechinger & Stutzer, 2009).
- If we have reason to believe that people systematically misunderstand how they will feel about lengthy commutes and consequently make faulty decisions about where to live, we may be able to infer their fully informed choices from the corresponding levels of SRWB (see Stutzer & Frey, 2008).
- Even if we are interested in completely novel choice settings, so that there is no opportunity to observe hedonic reactions to the outcomes of interest, we can use assessments of anticipated well-being for particular outcomes, along with other prospective subjective evaluations and biometric reactions, to predict the choices people would make in those settings (see Bernheim, Fradkin & Popov, 2015).

Most existing research on SRWB fits easily into a unified choice-oriented welfare framework, but it requires a different interpretation. Instead of taking SRWB at face value as a generally reliable measure of overall well-being, we construe it 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).

To illustrate the advantages of this alternative interpretation, suppose we want to know whether a particular policy improves welfare. Imagine that some jurisdictions have implemented the policy while others have not, and that we can survey residents of both. To perform SRWB-oriented welfare analysis, we must ask the "right" question about aggregate well-being. If different questions (for instance, about "happiness" versus "life satisfaction") yield different answers, we have to select one, despite having no objective basis for doing so.

In contrast, within a choice-oriented welfare framework, we view the answers to those same questions as predicting which policy regime people would choose, assuming they correctly anticipate the consequences. One option is to proceed exactly as in the SRWB-oriented approach: try to settle on the "right" question, and assume people would choose the alternative that delivers the greatest SRWB. This strategy involves the same analytic steps as the SRWB-oriented approach; we simply reinterpret the findings. While it may on occasion yield reasonably accurate predictions, it is also simplistic and in many contexts demonstrably biased. A better

option is to treat the problem as one of optimal statistical prediction, and deploy the various tools that economists and statisticians have developed for this purpose. With this approach, there is no need to resolve which of two SRWB measures is "correct," because one can use them as co-predictors of choice. Indeed, to obtain even greater accuracy, one can expand the set of co-predictors to include other types of subjective evaluations, and possibly even biometric reactions.[28] Experimental evidence indicates that this strategy can substantially improve upon the practice of taking measures of SRWB at face value, in the sense that it dramatically reduces both average prediction error and bias; see Bernheim et al. (2015).

D. Summary

The preceding discussion leaves us with two revised premises that survive the various challenges from behavioral economics and provide us with a foundation for the choice-oriented welfare framework outlined in Sections 4 through 6. They are:

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.

Notice that, in stating these premises, I refer to "judgments" rather than "preferences." My object is to avoid confusion among multiple distinct interpretations of "preference," including: (i) the notion that every choice tautologically expresses a preference; (ii) the notion that every choice derived from a direct or correctly informed indirect judgment tautologically expresses a preference, and (iii) the (in my view problematic) assumption that we choose by consulting our "true preferences," the supposed existence of which does not depend on the act of choice or the construction of judgment.[29] Henceforth, whenever I refer to "preference" rather than "true preference," I intend the *second* interpretation.

To formulate a welfare framework based on our revised premises, we must grapple with two main issues. First, how do we distinguish between choices that reflect correctly and incorrectly informed judgments? Second, how do we accommodate inconsistencies among the judgments that merit deference? In the next section, I discuss an approach that has gained some traction among behavioral economists, but that often requires a more complete understanding of decision

28 See Smith, Bernheim, Camerer and Rangel (2014) for an application involving biometric reactions.
29 Additional interpretations include (iv) the notion that every judgment (whether choice or nonchoice) tautologically expresses a preference, and (v) the notion that every direct or correctly informed indirect judgment tautologically expresses a preference.

processes than we currently possess, or perhaps are ever likely to achieve. However, careful consideration of the challenges presented by that approach leads us to a tractable alternative, which I describe in subsequent sections.

3 Behavioral Revealed Preference

When contemplating the design of a choice-oriented welfare framework, an economist's first instinct is usually to adapt the familiar revealed preference paradigm. This thought process often leads to some variant of a general framework that I will call *behavioral revealed preference* (BRP). The framework boils down to an assumption and a principle. Let us start with the assumption:

> ***The BRP assumption***: People have well-behaved unitary or conglomerate preferences, which play an important role in the process that generates choices.

To be clear, this approach does not require us to assume that "true preferences" (defined as in Section 2) actually exist. Instead, we can interpret a consumer's "preferences" as an analytic representation of all the direct judgments he or she constructs when making choices.

The BRP assumption differs from the central premise of standard revealed preference in two ways. First, it allows for the possibility that people try to respect multiple (conglomerate) preference orderings, rather than a single (unitary) ordering. Second, while it envisions a process through which preferences influence choices, it does not imply that people always choose the option they most prefer.

Next we turn to the principle:

> ***The BRP principle***: If enough is known about the process generating choices, then one can invert it conditional on its unknown parameters, and recover both those parameters and preferences from choices.

In essence, the object is to formulate a complete model of choice that separately specifies the consumer's direct judgments and the elements of cognition that map those judgments into choice. A good example appears in Koszegi and Rabin (2008). They model biased beliefs (the gambler's fallacy) in a setting where a decision maker bets on repeated flips of an objectively fair coin, and show that one can in principle recover both beliefs and risk preferences from choices.

A. Building a BRP model

Building a BRP model requires one to make strong assumptions. Broadly, they fall into the following three categories.

- First, one must take a stand on the aspects of experience that contribute to well-being. Formally, by specifying the domain of preferences, which is implicit in the dimensions of the consumption set, $\boldsymbol{X}$, we implicitly provide an answer to the question, *what do people care about*?
- Second, one must take a stand on whether the decision maker has unitary or conglomerate preferences over the consumption set. Are her direct and correctly informed indirect judgments always mutually consistent? Alternatively, does she arrive at different conclusions under different conditions? Does she hold conflicting views simultaneously?
- Third, one must take a stand on the nature of the apparatus that maps our direct judgments into decisions. Are those judgments expressed directly into choices, or are they distorted by limited cognition, biases, and other psychological phenomena? Does the answer depend on context, and if so, how? If we assume the decision maker has conglomerate preferences, does the context determine which judgments are "in charge," or is there a process through which conflicts between objectives play out?

Classes of theories

When we encounter puzzling choice patterns (*anomalies*), each of these categories offers potential routes to theoretical explanations. To rationalize choices, we can assume that people have nonstandard concerns, as in Gul and Pesendorfer's (2001) formulation of temptation preferences; we can imagine that people have conglomerate preferences which they harmonize inefficiently, as in Laibson, Repetto, and Tobacman's (1998) formulation of time-inconsistent choice; or we can assume that the apparatus of decision making scrambles preferences, as in Rubinstein and Salant's (2006) model of choices from lists.

Of the three BRP strategies mentioned in the last paragraph, the third is the most common. The second is rarely pursued because it raises some additional difficulties, including the question of how one balances the competing objectives of "multiple selves." For example, should we treat each preference ordering as a distinct individual and apply the Pareto criterion? That solution leaves one wondering whether we have taken the multi-self metaphor too seriously and stretched it too far. Moreover, in the context of intertemporal decision making, it raises new difficulties,

because the preferences of each "self" are only recoverable on limited domains: a choice at time t cannot affect consumption at time $t - k$ for any $k > 0$. Some have addressed that issue by assuming the time-t self is indifferent with respect to all past consumption, thereby extending that self's preference ordering to the full choice domain (see Laibson et al., 1998). Unfortunately, that assumption is entirely arbitrary and likely false, inasmuch as we value our past experiences.

Decision frames

One important difference between BRP and the conventional welfare paradigm is that, in the former, we allow for the possibility that decisions depend on conditions that have no direct bearing on well-being. The common term for any such condition is a *decision frame*.[30] 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 he is asked to decide at 1 pm after he has just eaten, or at 4 pm when he is hungry (Read & van Leuwen, 1998). Here, the natural assumption is that the preference domain encompasses various food items, which means we interpret the time at which Norman makes his choice as the decision frame, f. To explain his behavior, we must then assume that the frame either influences the construction of direct judgments (in which case we posit conglomerate preferences), or distorts the expression of those judgments into choices. Notice, however, that this fact pattern admits another interpretation: Norman's well-being depends not only on the food he eats, but also on what he orders and when he orders it.[31] Under that assumption, $\boldsymbol{X}$ consists of bundles specifying both orders and meals, and there are no decision frames. This interpretation shifts the explanation for Norman's behavior from the second and third categories to the first.

How do we determine what people care about, and thereby draw a line between consumption bundles and frames? Sometimes we may rely on information about the mechanism through which a given condition affects choice. For example, any condition that demonstrably leads to confusion about the decision problem is properly considered an aspect of the decision frame rather than a characteristic of consumption bundles. Another strategy is simply to ask people what they care about, or to introspect. Some reliance on self-reports is probably unavoidable. Even so, we do not encounter the same conceptual problems as with SRWB, because our object

30 Bernheim and Rangel (2009) use the term *ancillary condition*.

31 Sen (1993) makes a more general 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)."

here is merely to learn *whether* people care about various aspects of experience, rather than to measure *how much* they care.

It is also important to draw the line between consumption bundles and frames in a way that permits useful welfare analysis. To illustrate, let us return to the problem of Norman's lunch. Suppose a "planner" is charged with ordering his lunch on his behalf. The spirit of choice-oriented welfare analysis is to ask what Norman would choose for himself. If we assume he cares only about which meal he eats, we are in business: the planner can in principle mimic his choice. But if instead we allow for the general possibility that Norman also cares about what he orders and when he orders it, his choices will only reveal his preferences over more complex bundles. They will shed no light on the question of whether he is better or worse off with a sandwich or a salad when he places no order.

This simple illustration points to a guiding principle: the least problematic route to a usable choice-oriented welfare framework is to assume the consumer does not care about conditions pertaining specifically to the experience of choosing (henceforth, *conditions of choice*) except insofar as they affect what is chosen, and to treat them as part of the decision frame. Absent this assumption, choices cannot directly reveal preferences among the alternatives available to the planner. We cannot simply ask what the consumer would choose if he faced "the same" decision as the planner, because in that case the alternatives (correctly defined) would be different.

B. Challenges

The idealized "recipe" for BRP analysis follows standard scientific practice. We formulate theories, then evaluate and refine them by confronting them with new data and testing their implications. For successful theories, we recover unknown parameters including preferences, and use the parameterized models to conduct both positive and normative analysis.

The intellectual appeal of this approach is undeniable, and there are certainly contexts in which its application is compelling. Consequently, a unified normative framework should incorporate it. At the same time, an application of the BRP framework demands a thorough structural understanding of choice processes. Unfortunately, our understanding of many (perhaps most) choice phenomena remains partial and imperfect. The uncomfortable aspect of this approach is that it *requires* us to take strong stands concerning the nature of preferences and decision processes, even when – if we are honest with ourselves – we must acknowledge that we have very little basis for doing so.

Advocates of the BRP approach respond that its successful application will become less challenging over time because the state of our knowledge will steadily improve. While I hope that rosy prognosis turns out to be accurate, I also recognize that it might not. The associated challenges may prove surmountable, but they are nonetheless daunting.

The first challenge is that behavioral models often have multiple normative interpretations. Consequently, even if we can satisfy ourselves that we have arrived at the right positive model, welfare analysis may remain problematic. For example, suppose we become convinced that people make their choices according to the following model:

$$\max_{x \in X} s(x) \text{ s.t. } u(x) \geqslant t(X)$$

where x is a consumption bundle, X is an opportunity set, and s, u, and t are functions. Norman (who by now has completed his Ph.D. in economics) offers the following interpretation: s measures the salience of an alternative, u measures the utility it delivers, and t establishes a threshold. In other words, people "satisfice" by selecting the most salient alternative from among those that clear some utility hurdle. Norma (similarly credentialed) disagrees. She believes Norman has mislabeled the model, and offers an alternative interpretation: u measures the salience of an alternative, and s measures the utility it delivers. In other words, people select their preferred alternative from among those that are sufficiently salient. Their daughter Felicity (who is unschooled but precociously clever) insists that they are both mistaken. She asserts that people have conglomerate preferences (both u and s) which they struggle to reconcile. Even though Norman, Norma, and Felicity agree about the positive model, they disagree about welfare.

To be fair, there may be ways to resolve this disagreement. For example, Norma and Norman may agree that SRWB is at least a rough proxy for the quality of subjective experience. If they find that it is highly correlated with $u(x)$ and uncorrelated with $s(x)$, Norman would declare victory, and Norma might acquiesce. However, they are more likely to discover that SRWB is positively correlated with both, because more attractive alternatives tend to be more salient. In that case, they may spend the rest of their careers writing a succession of journal articles that debate the merits of alternative SRWB measures, without resolving much of anything.

In some cases, many economists appear to think that the correct normative interpretation of a positive behavioral model is obvious. Consider, for instance, the familiar formulation of quasi-hyperbolic discounting (also known as $\beta\delta$ preferences), popularized by Laibson (1997). Discussions of this model often employ heavily value-laden language, including phrases such as "present bias" and "self-control problems." Consistent with the judgments implied by this language, they frequently assume people have unitary preferences and equate welfare with

δ-discounted utility ("the long-run criterion"). And yet, as I have explained elsewhere (Bernheim, 2009*a*), this model admits a large number of disparate normative interpretations. For example, one could take the position that true happiness is achieved by living in the moment, and that we suffer from a tendency to overintellectualize when making decisions about the future.

Advocates of the long-run criterion have been known on occasion to scoff at this objection. They ask incredulously whether I reject the medical and psychiatric consensus that substance addiction is a problem. To be absolutely clear, I accept that consensus, because in that context there is a reasoned, evidence-based foundation for the normative conclusion (see Bernheim & Rangel, 2004, and Section 5 below). That said, substance addiction is a distinctive neurobiological phenomenon. Consequently, the foundation is narrow, and does not justify the same normative judgment in all contexts where time inconsistency is observed.

Nor can one dismiss my reservation concerning the general application of the long-run criterion as an abstract philosophical issue. Many cultures emphasize the importance of living in the moment. Moreover, according to popular wisdom, no one wishes on their deathbed that they had put in more hours at the office. On the contrary, people tend to regret having worked too hard and spent too little time with their families. Thus, when economists advocate the long-run criterion as a general normative principle, one has to wonder whether this is simply a case of successful workaholics believing that everyone else ought to be more like them.

The second challenge is that, as a general rule, many distinct nonstandard positive models can account for the same choice mapping. Practitioners of the BRP approach face a difficult dilemma. On the one hand, to rationalize nonstandard behavior, they have to broaden the class of potential explanations. On the other hand, to identify preferences, they have to limit that class. Once one steps away from the standard framework, it is difficult to know where to place those limits, or how to justify them. Thus, the problem becomes one of having too many degrees of freedom.

The practical implications of this observation are readily evident in the pertinent literature. Over the past decade or so, we have seen a growing proliferation of theories purporting to explain classic anomalies such as behavior in the ultimatum game or the endowment effect. In principle, the proliferation of theories could be scientifically healthy, but only if there is also a winnowing. Unfortunately, precious little winnowing occurs. In behavioral economics, theories are hard to kill.

In some cases, different theories about decision processes have observationally equivalent implications for choice patterns. Consequently, there is no hope of distinguishing one from another unless we find ways to incorporate rigorous analysis

of nonchoice data into formal tests. So far, that agenda has met with only limited success (see Bernheim, 2009*b*).

Of course, in many cases, competing theoretical explanations for the same anomaly have distinctive implications for other choices, and consequently we could in principle distinguish between them. Yet even then, theories rarely die, due to the third challenge: human behavior is extraordinarily complex.

One of the main lessons I take from the empirical literature in behavioral economics concerns the prevalence and complexity of context-dependent choice patterns. There is growing evidence that the details of decision problems not only matter, but do so in ways that are difficult to systematize outside of limited domains. Even choice patterns that behavioral economists consider "well established" appear to be context-dependent.[32] As a result, our theories often perform rather poorly when we test their predictions in contexts that do not closely resemble those in which they are calibrated.[33]

Poor predictive performance sounds like a good justification for some winnowing. And yet the theories in question usually survive, for two reasons. First, given the acknowledged complexity of human behavior, all models are regarded as parsimonious approximations. The theorist aims to capture an important and systematic aspect of the decision process, but never pretends to describe that process comprehensively. Consequently, one can always construe a "failure" as an indication that something else is also going on in a given setting, rather than as proof that the model is fundamentally wrong-headed. Because we implicitly or explicitly allow for the possibility that special contextual details can bring other mechanisms into play, proponents of different theories can find themselves arguing endlessly and unproductively over which experimental setting most effectively illuminates the "fundamental" mechanisms.[34] Second, with limitless degrees of freedom, models are easily tweaked. Instead of vanishing in favor of their competitors, they morph.[35] The impetus to winnow is thereby once again defeated.

32 See, for example, Plott and Zeiler (2005) on the endowment effect, Andreoni and Sprenger (2012) on time inconsistency, and Harbaugh, Krause and Vesterlund (2010) on the "four-fold pattern of risk taking."

33 See, for example, Kagel and Wolfe (2001), who find that leading theories of fairness, which were originally formulated to explain results in two-person bargaining problems, fail to predict behavior in the three-person ultimatum games. The literature contains many other examples.

34 See, for example, the exchange between Isoni, Loomes and Sugden (2011) and Plott and Zeiler (2011) concerning the endowment effect, or that between Engelmann and Strobel (2006), Fehr, Naef and Schmidt (2006), and Bolton and Ockenfels (2006) concerning models of fairness.

35 Cumulative Prospect Theory (Tversky & Kahneman, 1992) is perhaps the best known example: it was developed in response to the recognition that Prospect Theory (Kahneman & Tversky, 1979) violates dominance, and is therefore easily falsified.

To make matters even more challenging, some theories are formulated in ways that make them virtually impossible to falsify. To take an example, the theory of reference-dependent preferences can rationalize virtually any choice pattern. Many of its vaunted successes largely consist of showing that, by making fortuitous assumptions about reference points, one can "account" for otherwise anomalous patterns. Consider the oft-repeated claim that reference dependence successfully predicts daily income targeting by taxicab drivers (Camerer et al., 1997). This "prediction" hinges on two additional assumptions: first, the reference point applies to income rather than to leisure (or both); second, drivers evaluate each day's earnings separately relative to the referent ("narrow framing"). With respect to the first assumption, one can show that, if the reference point applies to leisure instead of income, the within-day labor supply elasticity will typically be positive rather than negative. This is of course convenient for fans of reference dependence, given that the findings of Camerer et al. (1997) are now disputed (see, for example, Farber, 2008).

Will we eventually discover ways to overcome these challenges? I certainly hope so. For my own part, I continue to conduct research concerning structural models of decision making, and I would expect a unified normative framework to accommodate progress along these lines. However, we are not yet at the point where we can always comfortably hang our hats on particular structural models with clear normative interpretations, and we may not arrive at that point any time soon. Therefore, a unified normative framework should also offer us an alternative.

4 Toward a unified framework

Our object is to find a way to proceed in the general spirit of the choice-oriented BRP paradigm, but using a framework that does not force us to take so many potentially uncomfortable stands. That said, it is plainly impossible to do away with all assumptions. As soon as we use the phrase "choice data" or write down the symbol for the consumption set (X), we are implicitly assuming that we know how to describe the objects of choice. Accordingly, we cannot avoid taking a stand on the aspects of experience that contribute to well-being. The Unified Framework therefore preserves this feature of the BRP approach, and we arrive at our understanding of the consumption set in precisely the same way. However, having bitten that bullet, we dispense with the need to take additional stands concerning the nature

of preferences and decision mechanisms, while at the same time permitting such stands when there is an adequate foundation.[36]

A. A useful reinterpretation of BRP analysis

What is "the general spirit of the choice-oriented BRP paradigm"? Some might say that it involves the process of understanding preferences and choices through structural models. I take a somewhat different view. When we perform welfare analysis within the BRP framework, we implicitly accomplish two tasks:

- First, we identify collections of decision problems for which choices express preferences (that is, direct or correctly informed indirect judgments) without distortion.
- Second, we construct a welfare criterion by conducting standard revealed preference analysis on those restricted domains.

In most cases, it is easy to see how one unpacks BRP analyses into these two components. To illustrate, take the case of quasi-hyperbolic discounting (the $\beta\delta$ model). As I have noted, a common normative interpretation of this model is that δ-discounting governs true preferences, and that β reflects a bias. In that case, decisions made at the earliest possible date with full commitment express preferences without distortion. One can therefore recover that ordering directly from those choices.[37]

In some cases, this point is less obvious. For certain models, it may appear that choices *always* distort preferences. However, I would argue that this appearance reflects a failure to envision the entire choice domain.

As an example, consider decisions involving ordered lists of options, such as candidates enumerated on a ballot. Various studies document a tendency to pick the alternatives listed first (Miller & Krosnick, 1998)). Rubinstein and Salant (2006)

36 An interesting recent paper by Goldin and Reck (2015) pursues an intermediate approach. They take stands not only on the aspects of experience that contribute to well-being, but also on the existence of unitary preferences, thereby ruling out explanations for behavioral anomalies involving context-dependent direct judgment. However, they largely dispense with the need to take a stand concerning the nature of the process mapping preferences to choices. Their main result shows that it is possible to recover preferences provided those processes satisfy weak properties. This approach also fits within the Unified Framework, but I question whether there is a legitimate foundation for the assumption of unitary preferences.

37 An alternative interpretation of the same model holds that the preferences of every time-dated "self" merit deference. That interpretation lends itself to similar unpacking. Decisions made at any given date with full commitment express an undistorted preference ordering, which one can recover from those choices. Once all such orderings are recovered, one can construct a welfare criterion based on the Pareto relation; see, for example, Laibson et al. (1998).

formulate a theoretical BRP model of this phenomenon in which choice problems consist of ordered lists. Here, the ordering defines the decision frame, and *all* orderings potentially distort the expression of "true preferences." At first this may appear contrary to what I have written above, but the explanation is simple: as formulated, the theory only pertains to the limited choice domain within which the distortion occurs.

The nub of this theory is that cognitive limitations lead people to simplify certain choices by applying a heuristic. It is relatively easy to envision other decision frames engineered so that they do not trigger this simplification. I can imagine a number of testable possibilities. The bias might not apply to sufficiently short lists, such as pairs of alternatives, or to settings where the decision maker is not permitted either to rush or to skip over descriptions of particular options. Alternatively, one could simply avoid presenting the menu in list form. Imagine, for example, making a choice between two options that appear on a screen, one to the northeast of the other. Presumably one can adjust the positions so that there is demonstrably no bias toward one or the other. Each of these possibilities points to a set of decision problems that could reveal preferences without distortion.

As a purely logical matter, one can of course imagine environments in which choice mechanisms *always* distort preferences. However, we cannot accept such formulations without implicitly licensing all manner of mischief. Consider the following illustration. Upon learning that Norman never orders salad, an economist theorizes that he actually prefers salad, but is consistently governed by habit. If there is truly no setting that would overturn the purported habit, are we really prepared to accept the proposition that Norman is better off, according to his own judgment, with a salad? I am not. Absent *any* setting that is free from an alleged distortion, we ought to question whether the associated conception of preference is merely a contrivance.

B. Overall structure of the Unified Framework

The reinterpretation of the BRP framework discussed in the previous section is useful because it motivates the two core tasks of the Unified Framework:

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

The BRP approach entails serious challenges because it places demanding restrictions on the inputs for the second step: we cannot "recover preferences"

unless choices are consistent. This requirement is what forces us, in the first step, to take the strong stands on the nature of direct judgments and the apparatus that maps them into choices, as discussed in Section 3A. Consequently, to avoid the need for those stands, we must dispense with the consistency requirement.

A key feature of the Unified Framework is that the second step flexibly accommodates inconsistencies among the choices that merit deference. As we will see, we derive the criterion used in that step from the *unambiguous choice relation*: we say that one alternative is unambiguously superior to another if and only if the second is never chosen when the first is available. Intuitively, this criterion instructs us to respect choice whenever it provides clear normative guidance, and to live with whatever ambiguity remains. Thus, it allows us 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. Significantly, when all choices are mutually consistent, the criterion specializes to the standard notion of revealed preference.

This alternative approach to the second task fundamentally alters the nature of the first task. In Step 1 of the Unified Framework, we can in principle identify choices that reflect direct or correctly informed indirect judgments by entertaining the same evidence, arguments, and modeling strategies as in the BRP framework. However, we are not *compelled* to settle on welfare-relevant domains (or, in the case of compound preferences, broad subdomains) within which all choices are internally consistent. Consequently, unlike the BRP approach, the Unified Framework does not *force* us to go out on a limb and take strong stands concerning the nature of preferences and decision processes when we lack an adequate foundation for doing so, or to ignore the important possibility that the construction of direct judgments may be highly context-dependent. Instead, we can set objective and appropriate criteria for evaluating whether a choice merits deference. If, after applying those criteria, we fail to arrive at an internally consistent set of choices, we are not compelled to "try harder"; instead, the Unified Framework lets us proceed. In contrast to the BRP approach, it also allows us to perform welfare analysis provisionally under different views of which choices do and do not merit deference, and thereby provide a better understanding of the assumptions upon which particular normative conclusions depend.

I turn next to the details of the Unified Framework. First, I discuss the types of data that will play a role (Section 4.C). Then I explain how one might arrive at the welfare-relevant domain (Section 5). Finally, I describe the derivation and application of the welfare criterion (Section 6).

C. Data inputs

The main data input for a choice-oriented normative framework is the choice mapping $C(X, f)$, which tells us the alternative(s) selected from the opportunity set X when frame f prevails. Economists typically learn about the choice mapping by observing certain choices and statistically interpolating or extrapolating others, sometimes using structural models. Often these methods suffice, but sometimes they do not. In any given setting, it is entirely possible that one will take a stand on the domain of preferences, only to discover that observed decisions shed inadequate light on choices within that domain.[38]

To understand the problem, consider an example. When Norman has lunch at a restaurant, his subjective enjoyment naturally depends on what he eats. Suppose it also depends on the entire menu, including the items he does not order. That might make sense if he is on a diet and becomes irritated when he has to turn down tempting but fattening alternatives. In that case, a properly defined consumption bundle is a pair, (X, x), where X is the menu and x is the selected item (as in Gul & Pesendorfer, 2001). If Norman lives in an area with a wide variety of restaurants, we can in principle observe his choices among these bundles. If instead he lives in a small town with a single restaurant that changes its menu from day to day, we can observe choices among bundles involving the same menu – for instance, ({pasta, salad}, pasta) versus ({pasta, salad}, salad) – but not among bundles involving different menus, such as ({pasta, salad}, pasta) versus ({pizza, salad}, pizza). To see why that is a problem, suppose a "planner" has to select Norman's lunch for him. When the planner picks x, Norman "chooses" from the degenerate menu $\{x\}$, which means he ends up with the bundle $(\{x\}, x)$. Unfortunately, Norman never has the opportunity to choose between ({pasta}, pasta) and ({salad}, salad), so the planner cannot take guidance from his choices.

What can economic analysts do in such situations? One possibility is to adopt restrictive structural assumptions that allow us to extrapolate from the choices we do observe. If our object is to avoid taking strong stands for which we have little foundation, that may well be an unattractive alternative. A second possibility is to fill in the missing choices through experiments. Sometimes that is an excellent strategy, but it can also be logistically complicated and/or prohibitively expensive. Thus, when allowing for the possibility that people have nonstandard concerns, we will likely encounter settings in which we need to learn about the choice mapping at least in part through nonstandard methods. As noted at the end of Section 2.C, a third (nonstandard) possibility is to draw inferences from nonchoice data, including SRWB, stated preferences, and biometrics. For example, even if Norman lives in

38 To be clear, this same problem arises in the BRP framework.

a small town, we can elicit SRWB when the menu is limited to pasta, and when the menu is limited to salad, and on that basis perhaps make a reasonable inference about his choice between ({pasta}, pasta) and ({salad}, salad). It is worth reiterating that those inferences do not need to be simplistic, as they are when one takes SRWB responses at face value. Instead, we can treat the problem as one of optimal statistical prediction, using a variety of subjective responses as predictors, as in Bernheim et al. (2015).

The third possibility mentioned in the previous paragraph defines one important role for nonchoice data, including SRWB, within a unified normative framework. I discuss other roles below. Accordingly, the potential inputs for this framework are not limited to choice mappings. As a general matter, we will allow for the use of all other information about decisions and the processes that generate them, including anything one might employ, even qualitatively, in the course of BRP analysis to arrive at the correct model of choice. Any well-reasoned inference one might make in the BRP framework bearing on the scope of the welfare-relevant domain will also be permitted in this one.

5 Welfare-relevant choices

What justifies a declaration that a given choice does not merit deference? Obviously, we cannot blithely ignore someone's decisions, even odd ones, simply because we would choose something else. After all, the justification for the entire undertaking is that people are the best arbiters of their own well-being, at least with respect to their direct judgments and correctly informed indirect judgments. What matters is their own likes and dislikes, not ours. Clearly, we need a more objective way to proceed.

A. Principles

The Unified Framework allows the analyst to offer any evidence-based justification for limiting the welfare-relevant domain, provided it is made explicit so others can evaluate it for themselves. If, for example, the foundation for a BRP model is compelling, then the welfare-relevant domain implied by the model will be equally compelling within the Unified Framework once one spells out the reasoning that supports it.

That said, the structure of the BRP framework does not inherently focus attention on the identification of the welfare-relevant domain. As explained in Section 4, the issue is often implicit – sometimes even buried – rather than explicit, and

little thought has been given to systematizing the principles and criteria used for this purpose. It is therefore natural to wonder whether the restricted domains implied by some BRP models would withstand open scrutiny.

Defining mistakes

One recurrent theme in the literature is the notion that fallible consumers can make *mistakes*.[39] There seems to be some agreement that behavioral welfare economics should allow for this possibility, and should not instruct policy makers to mimic errors.[40] But what exactly is a mistake? Most studies implicitly define a mistaken choice as one that is contrary to true preferences. Unfortunately, this definition can quickly lead to circularity: we identify mistakes by looking for choices that conflict with true preferences, and we infer true preferences from choices that are not mistaken. Thus, we face a challenge: how do we identify mistakes without presupposing a knowledge of preferences?

Consideration of an example involving an "obvious" mistake helps us make some headway. I find the following illustration from my work with Antonio Rangel (Bernheim & Rangel, 2004) particularly useful:

> "American visitors to the United Kingdom suffer numerous injuries and fatalities because they often look only to the left before stepping into streets even though they know traffic approaches from the right. One cannot reasonably attribute this to the pleasure of looking to the left or to masochistic preferences. The pedestrian's objectives – to cross the street safely – are clear and the decision is plainly a mistake."

The "optimal policy" in this setting seems equally obvious: if I see someone looking to the left while stepping in front of a bus, I will grab him and pull him back. I am willing to wager that his response will be to thank me earnestly, and not to accuse me indignantly of suppressing the expression of his preferences through his choices.

Now comes the challenging part: how can we *objectively* justify classifying the pedestrian's action as a mistake, without assuming we know his objectives? We

39 Another theme, albeit less common, involves the notion of *meta-choices*. The proposal is simple: if frames A and B lead to different outcomes in otherwise equivalent decision problems, resolve this conflict by giving the individual a choice between the two problems. The logic behind this proposal is, however, far from clear. The meta-choice is just another choice involving the same alternatives but with different framing. Agreement across two of three frames does not mean that the third frame is necessarily problematic.

40 I acknowledge that some may disagree with this premise. Indeed, the notion of a mistake is anathema to the standard revealed preference paradigm, which treats choices as the only observable manifestation of preferences.

cannot simply disregard this choice as a guide for welfare because we consider it "obviously stupid," as that designation (though tempting) is neither objective nor generalizable. Other potential criteria are equally problematic. For example, relying on expressions of regret arbitrarily favors the ex post perspective over the ex ante outlook.

Setting aside paternalistic judgments, we tend to classify a decision as a mistake when it has two distinctive features. First, 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. In the terminology of Section 2, it involves an incorrectly informed indirect judgment. Elsewhere, I have called this *characterization failure* (Bernheim, 2009*a*). By itself, a failure of this type raises the *possibility* that a mistake may have occurred, but does not guarantee it, because one can make the right decision for the wrong reason. That brings us to the second distinctive feature: there is some other option in the opportunity set that the decision maker would select over the mistakenly chosen one in settings where characterization failure does *not* occur.

To illustrate, let us return to an example from Section 2: Norma must choose between two closed boxes, a red one containing apples, and a yellow one containing pears. She recalls this information incorrectly and chooses the yellow box because she thinks it contains bananas, which she likes better than apples. Although she suffers from characterization failure, her choice is not necessarily mistaken. After all, she may also prefer pears to apples. However, if she would choose the red box over the yellow one after looking inside to refresh her memory, then her original choice was plainly made in error. A planner acting on her behalf should ignore that choice and pick the red box, not the yellow one.

Identifying mistakes

An important feature of the definition given above is that it makes no reference to divergences between choices and preferences, and thereby avoids circularity. Instead, it references the decision maker's understanding of the available options and the outcomes they imply. While economists do not traditionally use that type of data, it is certainly available, and we can collect more of it.

One possibility is to evaluate whether people properly understand concepts central to the proper characterization of certain varieties of choice problems. In Section 5.B, I discuss an empirical application involving intertemporal choice. When, for example, an appreciation of the intertemporal budget constraint requires

an understanding of compound interest, those who lack that understanding will tend to mischaracterize their alternatives, for example by making simple interest calculations.

A second possibility is to 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. This strategy is implicit in the BRP approach, but is rarely made explicit because the empirical foundations for most BRP models are incomplete.

As an example, consider Rubinstein and Salant's (2006) model of choices from lists. Their normative analysis presupposes that the model correctly depicts the process through which "true preferences" influence decisions. However, they are conspicuously silent concerning the nature of the evidence that might validate that depiction; they simply make an assumption and leave factual verification to the empiricists. I can imagine various types of verification. For example, eye-tracking studies may show that people are less likely to attend to items that appear lower on lists, and incentivized questionnaires may reveal poor recollection of those items. Such evidence would implicitly support the conclusion that cognitive shortcuts lead people to truncate their opportunity sets, a form of characterization failure.[41]

Bernheim and Rangel's (2004) analysis of substance addiction explicitly adopts this second strategy. We point to research showing that a specific neurobiological mechanism (the mesolimbic dopamine system, or MDS) measures correlations between environmental cues and subsequent rewards. To establish that choices made in the presence of substance-related cues involve characterization failure, we note that addictive substances cause the MDS to malfunction in a way that exaggerates those correlations.

A third possibility is to evaluate whether people understand (or have understood) particular decision problems by posing factual questions with objectively verifiable answers. For example, before Norma makes her choice, we could ask her to tell us what she thinks the boxes contain.[42] If she says the yellow box contains bananas, we will know there is a problem. Certain types of ex post statements are also potentially useful. Consider the following two possibilities:

> ***Scenario A:*** Norma opens the yellow box, sees the pears, and says, "I really wish I had picked the red box. I would have been happier with apples."

41 Oddly, Rubinstein and Salant (2006) claim that the Bernheim–Rangel approach delivers the wrong normative criterion for their model. But that is only because they misapply the approach by skipping what I have called Step 1. Implicitly, they assume that the available evidence justifies their model and the normative interpretation they give it, but then apply our framework ignoring that evidence.

42 We can even incentivize her answer to ensure truthful revelation.

Scenario B: Norma opens the yellow box, sees the pears, and says, "I forgot there were pears in this box – I thought there were bananas! If I had remembered, I would have picked the red box."

In scenario A, we cannot tell whether Norma has experienced characterization failure. Alternatively, she may feel regret because her ex ante and ex post judgments differ. In contrast, her statement in scenario B identifies her mistake: it helps us tie her choice of the yellow box to her memory lapse and her mistaken understanding of its contents.

With these principles in mind, how might we address the case of the American pedestrian in London? Statements such as "I thought it was safe to cross," or "I just wasn't thinking about consequences," indicate an operational misunderstanding of the relationship between actions and outcomes. Evidence that people routinely rely on habituated, semi-automatic responses, along with the observation that our subject looked only to the left before stepping into the street, corroborate this judgment. Combining these symptoms of characterization failure with confirmation that he would have made a different choice had he noticed the traffic (such as statements to that effect and observations of his actions in similar settings when he looks both ways), we can comfortably classify his choice as a mistake.

Contrasting the frameworks

While the Unified and BRP frameworks both potentially involve exclusions from the welfare-relevant domain, one important difference merits emphasis: unlike the BRP approach, the Unified Framework does not require the analyst to have a complete understanding of characterization failure. Take Norma's case: it is enough to know that she forgot the yellow box contains pears, and that she would have picked the red box had she remembered. Because the welfare-relevant domain excludes the mistaken choice, the analyst does not need to know that Norma expected to find bananas in the yellow box. Similarly, in the context of choices from lists, if we find that order effects are present in decisions involving three or more alternatives but not in binary choices, we can restrict the welfare-relevant domain to the latter without reaching a complete understanding of the cognitive mechanisms generating those effects in the former. As a result, the Unified Framework can be much simpler and less demanding to apply than the BRP approach. The next subsection provides a practical illustration.

B. An empirical application

Putting the concepts discussed in the preceding subsection into practice can be reasonably straightforward. To illustrate, I will describe a recent empirical application involving financial education due to Ambuehl, Bernheim and Lusardi (2015).

Low levels of financial literacy have raised concerns about the quality of financial decision making. Financial education seeks to improve decisions by helping consumers understand the principles governing the connections between choices and consequences. Traditionally, evaluations of financial education focus on measured literacy, self-reported decision strategies, and directional effects on behavior. Normative claims are generally based on strong preconceptions ("literacy must help") or paternalism ("people ought to save more"), rather than rigorous welfare analysis. In fact, the effects of financial education on the quality of decision making are far from obvious, given that it may influence behavior through mechanisms involving indoctrination, deference to authority, social pressure, and the like.

Plainly, one cannot evaluate the welfare effects of financial education within the standard revealed preference paradigm, because all choices tautologically serve the objectives they reveal. A more nuanced view holds that characterization failures occur whenever the relationships between choices and outcomes hinge on principles the individual does not understand. To evaluate the welfare loss from characterization failure, we need to construct a normative criterion based on choices in settings where such failures do not occur. It would of course be circular to assume that educated choices are free of these failures. How then can we proceed?

One way forward is to identify decisions that do *not* hinge on the principles one suspects the individual misunderstands. Within the BRP framework, one can recover preferences from those choices, and use them to evaluate the quality of more complex and potentially problematic decisions. For example, Song (2015) parameterizes a life-cycle model based on risk and time preferences elicited from subjects through simple choice experiments, and then uses it to evaluate retirement saving. This strategy involves some heroic assumptions: one must believe the chosen model of risk and time preference is the right one in radically different types of tasks. The strength of these assumptions points to a more general concern: the BRP approach to evaluating financial education requires a better understanding of decisions involving risk and time than we actually possess.

In Ambuehl et al. (2015), we adopt a different approach derived from the Unified Framework that allows us to measure the effects of financial education on the quality of decision making without adopting a particular structural model of choice. The easiest way to understand our approach is through a simple illustration.

Say we are 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, we have to specify a decision context. Accordingly, suppose Norman has an opportunity to buy one of two financial assets. Asset A represents a $10 investment that promises a return of 6% per day, compounded daily for 15 days. Asset B simply promises $24 in 15 days. To make matters simple, assume Norman is liquidity constrained (so that his decision does not depend on market interest rates), and that he only cares about the time path of his consumption. Ordinarily he will be willing to purchase each asset if and only if its price does not exceed some threshold value, call it p^* for the first asset and q^* for the second. A quick calculation reveals that the two assets are equivalent, subject to rounding. Thus, swapping out one for the other in a decision problem changes framing while leaving opportunities intact. If we find that $q^* \neq p^*$, we would conclude that Norman's decisions are frame-dependent.

A discrepancy between p^* and q^* raises the possibility that Norman errs when making decisions involving one or both of the assets. If we are correct in assuming that he evaluates assets by trying to figure out their implications for future cash flows, then the description of asset B is transparent whereas the description of asset A is not. That observation tells us that characterization failure is more likely in decisions involving asset A, but it does not by itself imply that such failures occur. After all, the alternative frames could simply trigger different processes for constructing direct judgments, for example by rendering salient different aspects of anticipated experience. To exclude choices from the welfare-relevant domain, we need to provide evidence that frame dependence goes hand in hand with a failure to appreciate the relationship between choices and outcomes in a particular frame. In this instance, we can administer a test to evaluate Norman's command of the principles governing compound interest. Assuming he fails it, we would then have a foundation for inferring that he mischaracterizes his opportunity sets in decision problems involving asset A, and hence for excluding those choices from the welfare-relevant domain, but not for excluding decision problems involving asset B.

Having reduced the welfare-relevant domain to decision problems involving asset B, and assuming those choices are internally consistent, welfare calculations are straightforward. The greatest welfare loss Norman can sustain when deciding whether to purchase asset A is $|q^* - p^*|$. For example, if $q^* > p^*$, he may fail to purchase the asset at price $p^* + \varepsilon$, even though it is actually worth q^* to him. In an environment where the price of asset A is drawn from a uniform distribution, Norman's expected welfare loss is proportional to $(q^* - p^*)^2$, and this formula remains valid to a second-order approximation for other distributions.

One might object that I have outlined a framework for evaluating welfare losses in a somewhat artificial class of decision problems rather than, for example, actual decisions involving retirement saving. The interest here is not, however, in the magnitude of the welfare loss, but rather in a comparative static: how does financial education effect that magnitude? Unless it improves the quality of decision making in simple contexts to which the pertinent principles are easily applied, there is little hope that it will do so in more complex settings, except by accident.

In Ambuehl et al. (2015), we implement these ideas by presenting consumers with pairs of equivalent valuation problems involving "simply framed" and "complexly framed" assets, and comparing performance across groups receiving different educational interventions, including a control. We find that the main intervention substantially improves subjects' knowledge and conceptual understanding of compound interest. Subjects report using the newly gained knowledge in their decisions without displacing other potentially reliable methods. Directionally, effects on valuations of complexly framed investments counteract a widely suspected bias (underestimation of compounding). Yet despite these indications of apparent success, the intervention does *not* increase aggregate welfare. While improvements in knowledge result from the substantive elements of instruction, it turns out that behavior is responsive only to rhetorical elements aimed at motivation. Consequently, while the behavioral response is directionally appropriate, it is also indiscriminate, and ultimate unconstructive.

C. Reinterpreting the literature

The perspective on welfare outlined in this paper is unifying because it encompasses, rationalizes, and usefully refines a wide variety of ideas and analyses that appear in the literature. In this section, I offer some examples pertaining to the identification of welfare-relevant choices.

Consistency within the welfare-relevant domain

Many recent applications of behavioral welfare economics either implicitly or explicitly define welfare-relevant domains within which choices are mutually consistent. Often these analyses invoke the notion of "true preferences," which choices purportedly express without distortion on the restricted domain. The absence of welfare-relevant inconsistencies obviously simplifies the construction of the normative criterion, rendering the analytics largely conventional. Naturally, the legitimacy

of the enterprise hinges on the implicit or explicit justification for discounting other choices. Studies differ in their attentiveness to this issue, and few are adequately disciplined by the systematic application of overarching principles.

Perhaps the best known example of this approach is Chetty, Looney and Kroft's (2009) analysis of tax salience.[43] The essence of their main finding is that people buy less when stores post tax-inclusive prices than when they calculate tax at the register. Methodologically, there is a close connection to the analysis of financial education described above. 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. In effect, the authors conduct welfare analysis based on that premise.

From the perspective of the Unified Framework, this paper pursues a conceptually legitimate approach, but fails to meet the appropriate burden of proof for refining the welfare-relevant domain. In the spirit of the BRP paradigm, it embraces a structural model of bounded rationality with *well-behaved unitary preferences*, but does not explain the justification for that restriction. With unitary preferences, when two choices conflict, at least one of them *must* be in error. If one accepts that statement, then the authors' assumption – that the mistakes occur when stores do not transparently post tax-inclusive prices – is entirely palatable. However, the notion that a choice conflict implies at least one mistake is problematic under the plausible view that choices may reflect contextually constructed judgments. The paper does not rule out the possibility that the alternative frames influence those judgments by rendering different aspects of anticipated experience salient. For example, people may be more inclined to focus on their opportunity costs when forced to think explicitly about giving their money to the government rather than to the store. Thus, the difficulty with this analysis is that, ultimately, characterization failure is essentially asserted and not proven.

One way to justify the paper's restriction on the welfare-relevant domain would be to show that people are not aware of unposted taxes. But in fact, the authors

43 Other recent examples include Fahri and Gabaix (2015), Handel, Kolstad and Spinnewijn (2015), Alcott and Kessler (2015), and Taubinsky and Rees-Jones (2015). One can interpret all of them as applications of the Unified Framework in which, once one refines the welfare-relevant domain by excluding choice problems that purportedly trigger characterization failure, all remaining choices are mutually consistent. An important benefit of viewing them through the lens of the Unified Framework is that it forces us to make the associated assumptions about welfare relevance, as well as their justifications and weaknesses, explicit. The discussion of Chetty et al. (2009) illustrates this point.

demonstrate precisely the opposite using a survey administered to shoppers exiting the store. Consequently, one cannot dismiss the concern expressed in the previous paragraph as a mere conceptual quibble. An alternative justification would invoke the hypothesis that shoppers call the relevant taxes to mind when making their purchases only if posted prices are tax-inclusive A complete analysis of tax salience and welfare would need to offer support either for this hypothesis or some appropriate alternative.[44]

Biased beliefs

A large and growing literature focuses on mistakes involving "biased beliefs." As an illustration, consider Koszegi and Rabin's (2008) analysis of the gambler's fallacy, which I mentioned above. Imagine that we flip a fair coin repeatedly, in each instance assessing Norman's willingness to pay (WTP) for a bet on heads. We find that his WTP declines after a string of heads, ostensibly because he believes that "tails is due." Likewise, it rises after a string of tails. Assuming Norman does not care about the state of nature, we can use these data to recover his preferences and his mistaken beliefs.

While this example provides a compelling illustration of the BRP framework, the particular approach it employs has limited applicability. It only works when preferences are not state-dependent and we have sufficient information about objective probabilities. It does not apply at all in settings where probability assessments are entirely subjective, which covers much of the territory of interest to applied economists.

Let us take another look at the Koszegi–Rabin coin-flipping problem through the lens of the Unified Framework, and determine whether we can characterize Norman's choices as mistakes according to the criteria discussed above. Simple factual questions can reveal gaps in his grasp of conditional probability, and thereby create the presumption that this decision environment induces characterization failure. To determine whether that failure leads to a mistake, we ask whether there are other settings in which choices among the same alternatives would differ. Consider an otherwise identical experiment in which we repeatedly rename the two sides of the coin. In the first round, we call the outcomes "heads" and "tails," but use "cats" (for tails) and "dogs" (for heads) in the second round, "Wilma" (for heads) and "Ferd" (for tails) in the third, "pinot noir" (for tails) and "cabernet" (for heads) in the forth, and so on. Critically, after the first round, we do not tell Norman which label refers to which side of the coin. From an objective point of view, this problem is identical

44 See Taubinsky and Rees-Jones (2015), who make useful progress on this front.

to the original one: each consists of a series of lotteries involving payoffs in two equally probable payoff-irrelevant states of nature. Here, however, Norman's WTP will likely bear no relation to past outcomes. Moreover, because this setting stops Norman from thinking about conditional probability, the source of the characterization failure is removed. Thus, the relabeled problem belongs in the welfare-relevant domain, while the original one does not.

This may seem like a roundabout way to arrive at precisely the same welfare criterion as the BRP approach. The benefit of the unified perspective is that it more easily generalizes to settings in which preferences may be state-dependent and/or objective probabilities are unknown.

To illustrate, imagine Norma is enrolled in a physics class and has just taken her first test. Suppose we determine that she is willing to spend $60 for a security that pays $100 if she passes and $0 otherwise. How can we tell if she is under- or over-confident? We cannot measure her objective probability of passing, which depends on the features of this particular test. Even if we could, her desire for cash may be state-dependent; for example, she may want to celebrate success or console herself in failure.

Consider Norma's decision through the lens of the Unified Framework. Simple questions can reveal whether she has properly processed pertinent factual information to which she has access. For example, we might ask her to state the percentage of science tests she has passed over the prior three years. If she finds adverse outcomes more salient and memorable than favorable ones, she may say 70% when the answer is in fact 90%. We can then infer that her memory of pertinent events is faulty, which creates a presumption that her valuation of the security involves characterization failure. Now consider an otherwise identical valuation problem in which we remind her that she has passed 90% of her previous science tests. If she is then willing to spend $80 for the security instead of $60, we can infer that her faulty memory led her to feel under-confident in the original setting. Suppose that, despite further probing, we are unable to identify any other pertinent factual information that she has improperly processed. In that case, the task that yields the $80 valuation belongs in the welfare-relevant domain, while the one that yields the $60 valuation does not. Using this information, we can conduct welfare analysis: assuming she passes on the offered security at a price of $70, she foregoes $10 in potential subjective value; hence the welfare cost of her under-confidence is $10.

Spinnewijn's (2015) empirical analysis of unemployment insurance, in which excessive optimism concerning reemployment prospects plays a central role, proceeds in this spirit, and is consonant with the Unified Framework.

Libertarian paternalism and nudges

Another important strand of the literature advocates nudges, defined as noncoercive changes in "choice architectures" that minimally impact opportunities, but nevertheless incline people toward "good" decisions (Thaler & Sunstein, 2003). Such policies are 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.

As an illustration, consider the issue of saving for retirement. Suppose Norman has little or no knowledge of personal finance, and is ill-equipped to determine how current investments translate into future standards of living. Left to his own devices ("decision frame A"), he would save nothing. The government is concerned about people like him and is considering a pro-saving initiative consisting of advertisements depicting happy retirees on cruise ships ("decision frame B"). These ads are substantively uninformative but highly motivational, and they would induce Norman to adopt a "good" heuristic, saving 10% of his income for retirement. A policy that implements decision frame B as the choice architecture is a pure nudge, in the sense that it has no impact on the opportunity set.

The Unified Framework allows us to evaluate nudges without invoking paternalistic judgments. Take Norman's problem. Given our assumptions, frames A and B both induce characterization failure. A libertarian paternalist officiates between them by imposing his or her own judgment. Within the Unified Framework, we might instead consider a third decision frame, C, in which Norman carefully works through his saving decision with an expert advisor who provides objective, easily understood information about the relationships between options and consequences, but scrupulously avoids recommending or discouraging any particular solution.[45] Imagine that Norman demonstrates an operational understanding of this information after the counseling session, and chooses to save 8% of his income. Moreover, if his only options are 0% and 10%, he selects the latter. Given this collection of hypothesized facts, there is ample justification for including frame C in the welfare-relevant domain, while excluding frames A and B. The ideal policy is then to deploy frame C as the choice architecture. If that proves prohibitively costly or otherwise impractical to accomplish on a large scale, the next best policy is a nudge: replace frame A with frame B. Critically, we reach the conclusion that the nudge is welfare-improving without relying on paternalistic judgments by the analyst or the policy maker.

45 As a practical matter, in many contexts it may be necessary to extrapolate choices in frames such as C from other information, including simpler related choices.

6 The welfare criterion

Once we complete Step 1 of the Unified Framework, we may find that all welfare-relevant choices are consistent with each other. In that case, we can construct a normative criterion using the familiar principles of revealed preference. Indeed, we can then also interpret the analysis as a BRP exercise. Here I am concerned with the more challenging possibility that choices remain less than fully consistent even within the welfare-relevant domain. How then do we arrive at a coherent normative criterion, and how do we use it?

A. Settling on a normative criterion

Let us start with a more basic question: What, exactly, is a normative criterion? As I use the term, it is a rule that tells us whether one outcome is better than another. Mathematically, that means it is a binary relation. If W is the welfare relation, and if x and y are outcomes, the statement "xWy" means that x is a better outcome than y.[46]

In principle, we could arrive at a welfare relation by enumerating and comparing a multitude of alternatives. Instead, let us first narrow down the possibilities by establishing the minimal requirements for a sensible criterion.

The first requirement is that W should be coherent. Operationally, I take coherence to mean that we can identify at least one best element within any opportunity set.[47] In standard consumer theory, we ensure coherence by assuming completeness and transitivity of the preference ordering. However, weaker conditions will also suffice. In the spirit of setting minimal requirements, I will insist instead on the weakest possible property that delivers coherence:

Property #1 (coherence): W is acyclic.[48]

Acyclicity simply rules out cycles – for example, it tells us that, if an apple is better than a pear and a pear is better than a banana, a banana cannot be better than an apple.[49]

46 Obviously, x cannot be better than itself, so W is necessarily irreflexive.
47 Formally, $x \in X$ is a best element according to W if there is no $y \in X$ such that yWx.
48 W is acyclic iff $x_1 W x_2 \ldots W x_N$ implies $\neg x_N W x_1$. Sen (1970) showed that acyclic relations have maximal elements on finite sets, and Bergstrom (1975) extended this result to compact sets. Obviously, cyclicity implies the existence of a finite set (specifically, the collection of the elements in the cycle) for which there is no maximal element. Consequently, acyclicity is the weakest possible coherence criterion.
49 Transitivity would also tell us that an apple is better than a banana.

The welfare relation also ought to depend on choices in a reasonable way. Certainly, if Norma consistently chooses an apple over a pear, and never chooses a pear when an apple is available, the only reasonable conclusion one can draw in a choice-oriented framework is that she is better off with an apple than a pear. Thus, we also require:

Property #2 (respect for unambiguous choice): If, within the welfare-relevant domain, y is never chosen when x is available, then xWy.[50]

Finally, there ought to be a degree of consistency between Step 2, which generates the welfare relation, and Step 1, which yields the welfare-relevant domain. Specifically, we require:

Property #3 (consistency with the welfare-relevant domain): If x is chosen in some setting (X, f) within the welfare-relevant domain, then x is not improvable within X according to W.

To declare x improvable within X would mean that choosing x in the choice problem (X, f) is a mistake. But a central purpose of Step 1 was to weed out all identifiable mistakes, and no data or inferential methods admitted in Step 2 were 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.

Fortunately, it is possible to satisfy all three of these requirements simultaneously. Consider, for example, the *unambiguous choice relation*, denoted P^*. Formally, xP^*y if and only if the welfare-relevant domain contains no decision problem in which x is available but y is chosen. Obviously, P^* satisfies Property #2 by construction. The other two properties are easily checked.[51]

What other possibilities besides P^* might we consider in our search for a welfare relation? The following theorem tells us that there are no others.[52] As long as we want our welfare criterion to satisfy Properties 1–3, P^* is the only game in town.

Theorem. A binary relation W satisfies Properties 1–3 if and only if $W = P^*$.

50 Throughout this section, I assume that for any set X (including $\{x, y\}$) the welfare-relevant domain contains at least one decision problem of the form (X, f) for some fame f. Thus, if y is never selected when x is available, there is a welfare-relevant decision problem in which x is chosen when y is available.

51 If $x_1P^*x_2 \ldots P^*x_N$, then x_1 must be chosen from any decision problem of the form $(\{x_1, x_2, \ldots, x_N\}, f)$, which implies $\neg x_N P^* x_1$, so P^* is acyclic. Furthermore, if x is chosen in some setting (X, f) within the welfare-relevant domain, then for all $y \in X \backslash x$, it is not the case that x is never chosen within the welfare-relevant domain when y is available; therefore $\neg yP^*x$, which means x is unimprovable within X (Property 3).

52 This result follows from Theorem 2 in Bernheim and Rangel (2009).

This theorem makes our lives fairly simple. It means, for example, that although we have not explicitly ruled out the possibility of using nonchoice information about well-being, choice processes, and the like in Step 2, there is in fact no room for it. Once we arrive at Step 2 of this choice-oriented framework, choice necessarily dictates the normative criterion.

Having reached this conclusion, it is reassuring to note that the Unified Framework nests the BRP approach, as we intended. To see why this is so, note that P^* coincides with the usual revealed preference relation in the special case where choice is fully consistent within the welfare-relevant domain. In fact, the definition of P^* generalizes that relation to nonstandard settings in an intuitive and transparent way.

What happens when there are inconsistencies within the welfare-relevant domain? Mathematically, the answer is that P^* becomes incomplete. Speaking practically, that means we treat certain comparisons as normatively ambiguous. Suppose, for example, that Norma never chooses a pear, banana, or orange when an apple is available, and never chooses an orange when an apple, pear, or banana is available. However, depending on the decision frame, she sometimes chooses a pear and sometimes chooses a banana when both are available. In that case, we can draw a number of potentially useful conclusions about her welfare. For example, if she initially has an orange and we replace it with any of the other three alternatives, we can say that she is better off. However, if she initially has a pear and we replace it with a banana, the effect on her well-being is ambiguous.

In settings where choice inconsistencies are pervasive, P^* may not be very discerning. Whether the resulting ambiguity undermines our ability to draw useful welfare conclusions depends on the context. As noted in Section 5.C, I have discovered that it is sometimes possible to reach sharp conclusions about important policy questions even when P^* entails a great deal of normative ambiguity. That said, a lack of discernment will certainly prove problematic in some instances. When that occurs, it is important to acknowledge that our inability to make precise normative statements reflects the limits of our knowledge. Admitting this ambiguity is intellectually honest. If we wish to sharpen our conclusions, the Unified Framework appropriately directs us back to Step 1, and focuses our attention on the empirical issues we need to resolve.

B. Applying the criterion

So far, this discussion of normative criteria has been more than a little abstract. How would the typical applied economist, who probably has not thought much

about binary relations since his or her first year of graduate school, implement these ideas?

Fortunately, implementation is neither complex nor mysterious. To illustrate, 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.

Now consider a more elaborate example that has the look and feel of familiar applied economics. Suppose consumers care about a bundle of goods, $\boldsymbol{z}$. Our object is to evaluate the welfare effects of policies that change that bundle. As in classical economics, we measure welfare in units of a numeraire good, y, and use $\boldsymbol{x}$ to denote the rest of the bundle (so that $\boldsymbol{z} = (\boldsymbol{x}, y)$).

Imagine first that, after studying data on consumers' choices, we conclude that they behave as if they maximize a standard utility function of the form $y + U(\boldsymbol{x})$. To measure the welfare effects of a policy that ends up switching a consumer from an initial bundle $(\boldsymbol{x_0}, y_0)$ to an alternative $(\boldsymbol{x_1}, y_1)$, we can calculate the amount of the numeraire we would need to remove from the alternative bundle, call it c, to make the consumer indifferent between $(\boldsymbol{x_0}, y_0)$ and $(\boldsymbol{x_1}, y_1 - c)$:

$$c = (y_1 - y_0) + [U(\boldsymbol{x_1}) - U(\boldsymbol{x_0})]. \tag{1}$$

Now imagine instead that we conclude consumers behave as if they maximize a function of the form $y + U(\boldsymbol{x}, f)$, where f, the decision frame, belongs to some set F.[53] While we are comfortable assuming that consumers do not actually care about the frame, our understanding of the mechanism through which it affects their choices is poor. Consequently, we treat all of the frames as equally relevant for normative analysis, and avoid interpreting the as-if objective function literally as "utility."

The unambiguous choice relation, P^*, provides us with two ways to measure the welfare effect of replacing $(\boldsymbol{x_0}, y_0)$ with $(\boldsymbol{x_1}, y_1)$. First, we can calculate the smallest (infimum) amount of the numeraire good we could remove from the new bundle, call it c_A, such that the consumer would unambiguously choose the initial bundle:

$$c_A = \inf\{c \mid (x_0, y_0) P^* (x_1, y_1 - c)\}.$$

Second, we can calculate the largest (supremum) amount of the numeraire good we could remove from the new bundle (call it c_B) such that the consumer would

53 For simplicity, I assume throughout this discussion that one can pair any decision frame f with any consumption bundle $\boldsymbol{z}$. That is not always the case.

unambiguously choose that bundle:

$$c_B = \sup\{c \mid (x_1, y_1 - c) P^*(x_0, y_0)\}.$$

Critically, because we have specified an as-if objective function, we can easily translate these conditions into analytic expressions for c_A and c_B:

$$c_A = \max_{f \in F}[(y_1 - y_0) + [U(\boldsymbol{x_1}, f) - U(\boldsymbol{x_0}, f)]] \tag{2}$$

and

$$c_B = \min_{f \in F}[(y_1 - y_0) + [U(\boldsymbol{x_1}, f) - U(\boldsymbol{x_0}, f)]]. \tag{3}$$

Clearly, $c_A \geqslant c_B$.

Once we solve for c_A and c_B, we can make a variety of discerning welfare statements. For example, we can say that the policy is definitely more valuable than c_B units of the numeraire and definitely less valuable than c_A units. If it turns out that the implementation costs (in units of the numeraire) are less than c_B, it is definitely beneficial, and if they are greater than c_A, it is definitely harmful. However, if those costs lie between c_A and c_B, we must acknowledge ambiguity as to whether the policy is helpful or harmful.

What if we gather additional information about choice processes that justifies restricting the welfare-relevant domain to some smaller set $F^* \subset F$? The only change is that we replace F with F^* in equations (1) and (2). As result, c_A falls while c_B rises (weakly in both cases), shrinking the region of ambiguity.

The striking similarity between equations (2) and (3) on the one hand and equation (1) on the other highlights the fact that applied welfare analysis within the Unified Framework closely resembles its familiar counterpart within the standard framework. Indeed, the standard tools of applied welfare analysis, including equivalent and compensating variation, consumer surplus, aggregate consumer surplus, the Pareto criteria, Pareto efficiency, and methods of making interpersonal comparisons, all have close counterparts within the Unified Framework. A complete discussion of these issues would consume many pages; I refer the reader to Bernheim and Rangel (2009), Bernheim et al. (2015), and Fluerbaey and Schokkaert (2013).

C. An empirical application

Abstract discussions of economic welfare are all well and good, but a useful framework also lends itself to empirical applications. Here I summarize a recent study that exemplifies Step 2 of the Unified Framework.

Starting with Madrian and Shea (2001), a number of studies have found that changing the default contribution rate for a 401(k) pension plan, for example from 0% to 3%, has a powerful effect on employees' contributions, particularly compared with conventional policy instruments such as capital income taxes. Indeed, increases in these default rates are often cited as highly successful examples of nudges.

Formal normative analysis of 401(k) default options is tricky. Due to its magnitude, the default effect is generally regarded as a nonstandard behavioral phenomenon (Della Vigna, 2009). Possible explanations include a tendency to procrastinate 401(k) elections due to time inconsistency (either sophisticated or naive), inattention, and psychological anchoring on the default as a target.

The BRP approach requires the analyst to adopt one of these models, parameterize it, and take a stand on which decisions reveal "true preferences." For example Carroll et al. (2009) analyze the welfare effects of 401(k) default options using a model of sophisticated time inconsistency. They also assume that true preferences are revealed only by full-commitment choices with no immediate consequences. For the reasons discussed in Section 3, even if that model is correct, this "long-run" welfare standard may not be compelling. Unfortunately, any alternative standard may seem equally arbitrary, and the BRP approach does not allow us to acknowledge that ambiguity. Similar concerns arise with respect to other potential explanations for the default effect. For example, interventions that are intended to focus attention on 401(k) elections may simply browbeat workers into making choices they would prefer to avoid.

Bernheim et al. (2015) evaluate the welfare effects of 401(k) default options empirically using the Unified Framework. Observed behavior illuminates the choice mapping in a "naturally occurring" decision frame, and each potential theory of default effects extends the mapping to additional decision frames that have not yet been observed – for example, in the case of sophisticated time inconsistency, day-in-advance commitments to making 401(k) elections. Using these choice mappings as inputs, the paper analyzes the welfare effects of setting particular defaults, and evaluates optimal defaults, under various assumptions about which decisions frames are welfare-relevant. When the welfare-relevant domain encompasses conflicting choices, the analysis identifies the range of normative ambiguity.

A complete summary and explanation of the various findings in Bernheim et al. (2015) is beyond the scope of this paper. Instead I will focus on a single finding that illustrates an important point. According to the empirical analysis, in the naturally occurring frame workers act as if the average cost of making 401(k) elections is extremely high. Most of the behavioral explanations for default effects envision alternative frames in which the as-if opt-out cost would be very small. One would

therefore tend to think that welfare conclusions would be highly dependent on the choice of the welfare-relevant domain, in which case the analysis would only be discerning and useful if one took a stand on "true preferences," as in the BRP approach. As it turns out, that is not the case. Whether we evaluate welfare from the perspective of decisions made in the naturally occurring frame or an alternative frame, we subtract opt-out costs only for those people who actually opt out, and not for those who stay with the default. Those who do opt out in the naturally occurring frame have much lower as-if opt-out costs, on average, than those who do not. Whether we count those incurred as-if opt-out costs in full (to respect decisions made in the naturally occurring frame) or discount them heavily (to respect decisions made in an alternative frame) therefore makes surprisingly little difference. Welfare calculations and optimal defaults are not entirely independent of the evaluation frame, but the region of ambiguity turns out to be surprisingly small. Thus, to reach useful conclusions, one simply does not need to take the strong stands required by the BRP approach.

7 Conclusions

As the theories, models, and ideas from behavioral economics propagate through the rest of the field, it is important to avoid an "anything goes" approach to welfare. In this paper, I have offered a unified perspective on normative inquiry that integrates a range of approaches, while nevertheless adhering to integrated principles for thinking about welfare systematically, thereby imposing much needed structure and discipline. The framework does not supply an end-to-end "turn-the-crank" procedure; on the contrary, like empirical analysis generally, it is compatible with a range of assumptions concerning the processes studied, and consequently one can apply it in different ways, generating different answers. Even so, the structure forces us to be more explicit about our assumptions, as well as our justifications for them, and thereby facilitates more meaningful discourse about welfare with those who would proceed from different premises.

References

Alcott, Hunt & Kessler, Judd B. (2015). The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons. NBER Working Paper No. 21671.

Ambuehl, Sandro, Bernheim, B. Douglas & Lusardi, Annamaria (2015). The Effect of Financial Education on the Quality of Decision Making. NBER Working Paper No. 20618.

Andreoni, J. & Sprenger, C. (2012). Estimating Time Preferences with Convex Budgets. *American Economic Review*, *102*(7), 3333–3356.

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

Bergstrom, Theodore (1975). Maximal Elements of Acyclic Relations on Compact Sets. *Journal of Economic Theory*, *10*(3), 403–404.

Bernheim, B. Doulgas (2009*a*). Behavioral Welfare Economics. *Journal of the European Economic Association*, *7*(2–3), 267–319.

Bernheim, B. Douglas (2009*b*). Neuroeconomics: A Sober (but Hopeful) Appraisal. *AEJ: Microeconomics*, *1*(2), 1–41.

Bernheim, B. Douglas, Bjorkegren, Daniel, Naecker, Jeffrey & Rangel, Antonio (2015). Non-Choice Evaluations Predict Behavioral Responses to Changes in Economic Conditions. NBER Working Paper No. 19269.

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

Bernheim, B. Douglas & Rangel, Antonio (2004). Addiction and Cue-Triggered Decision Processes. *American Economic Review*, *94*(5), 1558–1590.

Bernheim, B. Douglas & Rangel, Antonio (2007*a*). Behavioral Public Economics: Welfare and Policy Analysis with Fallible Decision-Makers. In Peter Diamond & Hannu Vartianen (Eds.), *Behavioral Economics and its Applications* (pp. 7–77). Princeton, NJ: Princeton University Press.

Bernheim, B. Douglas & Rangel, Antonio (2007*b*). Toward Choice-Theoretic Foundations for Behavioral Welfare Economics. *American Economic Review Papers and Proceedings*, *97*(2), 464–470.

Bernheim, B. Douglas & Rangel, Antonio (2008). Choice-Theoretic Foundations for Behavioral Welfare Economics. In Andrew Caplin & Andrew Schotter (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (pp. 155–192). Oxford: Oxford University Press.

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

Bernheim, B. Douglas & Thomadsen, Raphael (2005). Memory and Anticipation. *The Economic Journal*, *115*, 271–304.

Bolton, Gary E. & Ockenfels, Axel (2006). Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment. *American Economic Review*, *96*(5), 1906–1911.

Brandt, Richard (1979). *A Theory of the Good and the Right*. Oxford: Clarendon Press.

Camerer, Colin, Babcock, Linda, Loewenstein, George & Thaler, Richard (1997). Labor Supply of New York City Cabdrivers: One Day at a Time. *Quarterly Journal of Economics*, *112*(2), 407–441.

Caplin, Andrew & Leahy, John (2001). Psychological Expected Utility Theory and Anticipatory Feelings. *Quarterly Journal of Economics*, *116*(1), 55–79.

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

Chalmers, David (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York and Oxford: Oxford University Press.

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

Crosby, Donald A. (2013). *The Philosophy of William James: Radical Empricism and Radical Materialism*. Lanham, Maryland: Rowman & Littlefield Publishers, Inc.

Cummings, R. G. & Taylor, L. O. (1999). Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method. *American Economic Review*, *89*(3), 649–665.

Della Vigna, Stefano (2009). Psychology and Economics: Evidence from the Field. *Journal of Economic Literature*, *47*(2), 315–372.

Dolan, Paul, Layard, Richard & Metcalfe, Robert (2011). Measuring Subjective Wellbeing for Public Policy: Recommendations on Measures. *London School of Economics and Political Science, Center for Economic Performance*. Special Paper No. 23, March.

Dolan, Paul & Metcalfe, Robert (2012). Measuring Subjective Wellbeing: Recommendations on Measures for Use by National Governments. *Journal of Social Policy*, *41*(2), 409–427.

Easterlin, Richard A. (1974). Does Economic Growth Improve the Human Lot? Some Empirical Evidence. In P. A. David & W. R. Levin (Eds.), *Nations and Household in Economic Growth* (pp. 98–125). Stanford University Press.

Engelmann, Dirk & Strobel, Martin (2006). Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Reply. *American Economic Review*, *96*(5), 1918–1923.

Fahri, Emmanuel & Gabaix, Xavier (2015). Optimal Taxation with Behavioral Agents. *Mimeo*, Harvard University.

Farber, Henry S. (2008). Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers. *American Economic Review*, *98*(3), 1069–1082.

Fehr, Ernst, Naef, Michael & Schmidt, Klaus M. (2006). Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment. *American Economic Review*, *96*(5), 1912–1917.

Fluerbaey, Marc & Schokkaert, Erik (2013). Behavioral Welfare Economics and Redistribution. *American Economic Journal: Microeconomics*, *5*(3), 180–205.

Frey, Bruno S., Luechinger, Simon & Stutzer, Alois (2009). The Life Satisfaction Approach to Environmental Valuation, CESifo Working Paper No. 2836, October.

Frey, Bruno S. & Stutzer, Alois (2007). Should National Happiness be Maximized? CREMA Working Paper, March.

Griffin, James (1986). *Well-Being*. Oxford: Clarendon Press.

Goldin, Jacob & Reck, Daniel (2015). Preference Identification Under Inconsistent Choice. *Mimeo*, University of Michigan.

Gul, Faruk & Pesendorfer, Wolfgang (2001). Temptation and Self-Control. *Econometrica*, *69*(6), 1403–1435.

Gul, Faruk & Pesendorfer, Wolfgang (2008). The Case for Mindless Economics. In Andrew Caplin & Andrew Schotter (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (pp. 3–42). Oxford: Oxford University Press.

Handel, Benjamin R., Kolstad, Jonathan T. & Spinnewijn, Johannes (2015). Information Frictions and Adverse Selection: Policy Interventions in health Insurance Markets. *Mimeo*, U.C. Berkeley.

Harbaugh, William T., Krause, Kate & Vesterlund, Lise (2010). The Fourfold Pattern of Risk Attitudes in Choice and Pricing Tasks. *Economic Journal*, *120*(545), 595–611.

Harsanyi, John (1978). Rule Utilitarianism and Decision Theory. In H. Gottinger & W. Leinfellner (Eds.), *Decision Theory and Social Ethics*. Dordrecht: Reidel.

Hausman, Daniel M. (2012). *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.

Isoni, Andrea, Loomes, Graham & Sugden, Robert (2011). The Willingness to Pay-Willingness to Accept Gap, the 'Endowment Effect,' Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: Comment. *American Economic Review*, *101*, 991–1011.

Jacquemet, Nicolas, Joule, Robert-Vincent, Luchini, Stephane & Shogren, Jason F. (2013). Preference Elicitation Under Oath. *Journal of Environmental Economics and Management*, *65*(1), 110–132.

Kahneman, Daniel, Krueger, Alan B., Schkade, David, Schwarz, Norbert & Stone, Arthur (2004). Toward National Well-Being Accounts. *American Economic Review*, *94*(2), 429–434.

Kahneman, Daniel & Tversky, Amos (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, *47*(2), 263–292.

Kagan, Shelly (1998). *Normative Ethics*. Boulder, Colorado: Westview Press.

Kagel, John H. & Wolfe, Katherine Willey (2001). Tests of Fairness Models Based on Equity Considerations in a Three-Person Ultimatum Game. *Experimental Economics*, *4*(3), 203–219.

Koszegi, Botond (2006). Emotional Agency. *Quarterly Journal of Economics*, *121*(1), 121–155.

Koszegi, Botond & Rabin, Matthew (2008). Revealed Mistakes and Revealed Preferences. In Andrew Caplin & Andrew Schotter (Eds.), *The Foundations of Positive and Normative Economics: A Handbook* (pp. 193–209). Oxford: Oxford University Press.

Lacy, Heather P., Fagerlin, Angela, Loewsenstein, George, Smith, Dylan M., Riis, Jason & Ubel, Peter A. (2008). Are They Really That Happy? Exploring Sclae Recalibration in Estimates of Well-Being. *Health Psychology*, *27*(6), 669–675.

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

Laibson, David, Repetto, Andrea & Tobacman, Jeremy (1998). Self-Control and Saving for Retirment. *Brookings Papers on Economic Activity*,(1), 91–196.

Loewenstein, George & Ubel, Peter A. (2008). Hedonic Adaptation and the Role of Decision and Experience Utility in Public Policy. *Journal of Public Economics*, *92*(8–9), 1795–1810.

Lichtenstein, Sarah & Slovic, Paul (2006). The Construction of Preference. In Sarah Lichtenstein & Paul Slovic (Eds.), *The Construction of Preference* (p. i). Cambridge: Cambridge University Press.

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

Mill, John Stuart (2012). *Utilitarianism*. Renaissance Classics.

Miller, J. M. & Krosnick, J. A. (1998). The Impact of Candidate Name Order on Election Outcomes. *Public Opinion Quarterly*, *62*(3), 291–330.

Murphy, James J., Allen, P. Geoffrey, Stevens, Thomas H. & Weatherhead, Darryl (2005). A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics*, *30*, 313–325.

Nordhaus, William (2009). 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, Chapter 5* (pp. 125–144). Chicago: University of Chicago Press.

Parfit, Derek (1984). *Reasons and Persons*. Oxford: Oxford University Press.
Plott, Charles R. & Zeiler, Kathryn (2005). The Willingness to Pay-Willingness to Accept Gap, The 'Endowment Effect,' Subject Misconceptions, and Experimental Procedures for Eliciting Valuations. *American Economic Review*, *95*(3), 530–545.
Plott, Charles R. & Zeiler, Kathryn (2011). The Willingness to Pay-Willingness to Accept Gap, the 'Endowment Effect,' Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: Reply. *American Economic Review*, *101*, 1012–1028.
Rangel, Antonio, Camerer, Colin & Montague, P. Read (2008). A Framework for Studying the Neurobiology of Value-Based Decision Making. *Nature Reviews Neuroscience*, *9*, 545–556.
Read, Daniel & van Leuwen, Barbara (1998). Predicting Hunger: the Effects of Appetite and Delay on Choice. *Organizational Behavior and Human Decision Processes*, *76*(2), 189–205.
Rubinstein, Ariel & Salant, Yuval (2006). A Model of Choice from Lists. *Theoretical Economics*, *1*, 3–17.
Samuelson, Paul & Nordhaus, William (2001). *Macroeconomics*. (17th ed.). New York, NY: McGraw-Hill.
Sen, Amartya (1970). *Collective Choice and Social Welfare*. San Francisco: Holden-Day.
Sen, Amartya (1980–1981). Plural Utility. *Proceedings of the Aristotelian Society, New Series*, *81*, 193–215.
Sen, Amartya (1993). Internal Consistency of Choice. *Econometrica*, *61*(3), 495–521.
Smith, Alec, Bernheim, B. Douglas, Camerer, Colin & Rangel, Antonio (2014). Neural Activity Reveals Preferences Without Choices. *American Economic Journal: Microeconomics*, *6*(2), 1–36.
Song, C. (2015). Financial Illiteracy and Pension Contributions: A Field Experiment on Compound Interest in China. *Mimeo*.
Spinnewijn, Johannes (2015). Unemployed by Optimistic: Optimal Insurance Design with Biased Beliefs. *Journal of the European Economic Association*, *13*(1), 130–167.
Stevenson, Betsey & Wolfers, Justin (2008). Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox. *Brookings Papers on Economic Activity*,(2), 1–87.
Stutzer, Alois & Frey, Bruno (2008). Stress that Doesn't Pay: The Commuting Paradox. *Scandinavian Journal of Economics*, *110*(2), 339–366.
Taubinsky, Dmitry & Rees-Jones, Alex (2015). Attention variation and welfare: theory and evidence from a tax salience experiment. *Mimeo*, U.C. Berkeley.
Thaler, Richard H. & Sunstein, Cass R. (2003). Libertarian Paternalism. *American Economic Review, Papers and Proceedings*, *93*(3), 174–179.
Thornton, Stephen P. (2004). Solipsism and the Problem of Other Minds. *The Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/solipsis/, accessed 8/25/2015.
Tversky, Amos & Kahneman, Daniel (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.