CHAPTER 8

REVEALED MISTAKES AND REVEALED PREFERENCES

BOTOND KÖSZEGI AND
MATTHEW RABIN

People often depart from the narrowest sense of rationality traditionally assumed in economics: they take actions that they would not take could they fully assess the distribution of all relevant consequences of those actions. We cannot rule out that such mistakes are common in domains of activity that economists care about, and if they are, this would obviously have many important implications. To design policy with regard to harmful and addictive substances, for instance, a planner should ideally know whether 18-year-olds who launch into a lifelong addiction are doing so as a well-calibrated accomplishment of maximizing their expected lifetime well-being. Hence, it would be useful to study mistakes until we either learn that they are not very common, or understand them and can incorporate them into our models and policy prescriptions.

Yet incorporating mistakes into economic models raises the methodological concern that research loses touch with behavioral data of economic interest. On the positive side, the fear is that mistakes are difficult or impossible to observe in economic data, so they become a free parameter to fit any situation. On the normative side, the fear is that if economists allow for mistakes, they lose the ability to employ the revealed preference approach to extract welfare measures from behavior, and

have to (paternalistically) impose a welfare measure themselves. Some researchers consider these concerns prohibitive enough to justify ignoring mistakes altogether.

In this chapter, we propose general methods to deal with these concerns and illustrate our methods through a number of examples. Our view is simple: despite mistakes, human behavior is not random. The mistakes people make are systematic, and often they can be easily identified in behavior. And most important, there is a strong association between what people do and what they intend to accomplish. So if we understand how people intend to accomplish their goals, including understanding their mistakes, we can identify their goals from behavior. Preferences are revealed in behavior, even if they are not implemented by behavior.

We begin with recalling that all economic theories, including standard theories, must make some assumptions to gain traction. Consider, for instance, the central question of welfare analysis—what policies a planner should choose on behalf of individuals to maximize their well-being. The classical revealed preference approach proposes a simple answer: whatever a person would choose for herself. This approach is usually justified by assuming that people choose correctly. Yet even taking such rationality as given, "action-unobservable" assumptions are always necessary to make sure the proposed choice reflects the welfare question we are interested in.[1] As an extreme example, assuming only rationality, behavioral evidence cannot reject the hypothesis that imposing painful death on a person is the best policy. Even if she always chooses to avoid painful death when given a choice, her very favorite alternative could be painful death being imposed *without* a choice.

Recognizing that all theories make action-unobservable assumptions to connect preferences to behavior, we outline our proposal to follow this tradition: make reasonable assumptions that render mistakes and preferences jointly observable from behavior. In many situations, very minimal assumptions on preferences ensure that explicit or implicit "bets" on an event reveal a person's beliefs about the likelihood of the event. These beliefs may be found to be objectively incorrect. Much like we write theories to explain other types of behavior, the goal is then to look for a theory of behavior incorporating the systematic mistakes. This theory can and should tie preferences closely to behavior.

We next illustrate how our approach can be used to study mistakes about exogenous events such as a statistical process, the stock market, or a natural phenomenon. Suppose that when observing a series of coin flips, a person bets on (i.e., accepts a lower payoff contingent on) the event that the next flip will be different from recent ones. Under the painfully reasonable assumption that her preference for money is independent of coin flips, this means that she is making a mistake. We could try to maintain rationality and explain his behavior as a preference to bet on changes in coin flips, but the much more plausible theory is that he suffers from the gambler's fallacy: she believes that if the same side of the coin has come up a number of times, the other one is "due." Not only is this theory a natural

and parsimonious explanation of the facts, but it helps make sense of behavior in other contexts, including using the behavior to identify the person's goals. That is, it facilitates application of the revealed-preference method. Suppose, for example, that after observing six flips of heads in a row, the person chooses to get blueberries if tails comes up next and huckleberries if heads comes up next—rather than vice versa—and he has the opposite preference after observing six flips of tails in a row. With a theory based on the gambler's fallacy, we can conclude that he likes blueberries more than huckleberries. With a theory based on a preference to bet money on changes in coin flips, we would be able to make no meaningful inference from this behavior.

Following this, we consider beliefs about one's own future behavior. For a given future choice set X from which an individual will choose, we can ask her to bet on her choice—to identify an option in X and be paid a small amount if and only if she chooses that option later. As long as the monetary payment does not change what she will actually choose, under minimal assumptions her bet on an option reveals a belief that she will choose that option. Consider, for instance, a person's future choice between an appetizer and a full entree. She may be hungry or satiated in the future, and independently of this state she may be hungry or satiated today. For simplicity, posit that these states are observable or can be induced. Suppose that if she is hungry today, she bets that she will choose the full course, and if she is satiated today, she bets she will choose the appetizer. Yet independently of the current state, she ends up choosing the entree if she is hungry in the future and the appetizer if she is satiated in the future. Under the excruciatingly reasonable assumption that she prefers to receive money with a chosen option rather than an unchosen option, this means that she has incorrect beliefs about future behavior whenever the future state is different from the current state. A parsimonious and portable theory that explains the mistake is that the person suffers from projection bias: she underestimates how changes in her state will change her preferences. And as with the gambler's fallacy, we show below how such a theory can be used to identify the person's preferences, even though she herself may not act in accord with those preferences.

Our approach to revealed preference is inconsistent with some—in our opinion, extreme—views of the role of normative analysis in economics. In one view, the role of normative analysis is to analyze what institutions might emerge if people are to agree to those institutions. Welfare analysis is then a part of positive analysis, addressing what institutions are stable. Theorizing about preferences hidden by mistakes may very well be useless in this analysis, because such preferences will never be expressed in the process of institutional design. In contrast, we feel (and believe most economists feel) that some social planners might be motivated at least in part to design institutions that are sensitive to people's values, even if people do not always choose according to those values. Researchers helping such planners can use our methods to identify preferences that are hidden by mistakes, and hence

help in making sure the preferences *are* expressed in the process of institutional design.[2]

Within this view, our approach is a direct answer to the concern that normative analysis based on goals other than what people would choose for themselves necessarily involves the researcher (or policy maker) imposing her own values.[3] Since we propose to derive a person's preferences from her own behavior, any normative conclusions we draw respect the person's values—even if they do not coincide with her choices. Indeed, in most of our examples, our approach does not a priori rule out any option in the person's choice set, and could (depending on the person's behavior) lead a planner to prefer that option.

The Impossibility of Skinnerian Welfare Economics

Because our approach relies on some action-unobservable assumptions and this may give it the appearance of inferiority relative to the revealed preference method, we begin with some simple examples to illustrate that even classical theories make crucial action-unobservable assumptions beyond rationality. We do not claim that the assumptions made in standard theories are unreasonable (we believe they are often *very* reasonable), only that they are there, and there is no logical or methodological reason why they could not in some cases be fruitfully replaced by other reasonable assumptions.

A textbook illustration of welfare conclusions derived from revealed preference is based on a choice between two simple consumption goods. If we observe, say, a person choosing blueberries over huckleberries, we can conclude that she likes blueberries more than huckleberries. This leads to the positive prediction that the person will choose blueberries over huckleberries in a similar situation, and to the normative implication that if someone else must make the choice on the person's behalf, the welfare-maximizing policy is to choose blueberries. While the positive prediction "merely" leaves it completely unspecified what constitutes a similar situation, the welfare conclusion also relies on an important implicit assumption: it requires that having to make the choice herself (rather than having someone else make it) does not affect the person's ranking between the two options. Much psychology indicates, and models such as Gul and Pesendorfer [2001] predict, that often this is an unreasonable assumption.

An extreme example clearly illustrates this point. Even supposing rational utility maximization, without further assumptions behavioral evidence cannot reject the hypothesis that a person is happier having painful death rather than a cake

imposed on her. Her preferences could be such that she likes painful death if it is imposed on her without choice, but not if she has any choice, so observing that she never chooses painful death is not evidence that she does not like it. To see this, suppose that a person's utility is defined over the choice set she faces and the choice from it. In particular, for any decision problem she is facing—including arbitrarily complicated, dynamic decision problems—her preferences are over the set of final outcomes available and the outcome ultimately chosen. If the person's utility function satisfies $u(\text{death}|\{\text{death}\}) > u(\text{cake}|\{\text{cake}\})$, and $u(\text{cake}|\{\text{cake, death}\}) > u(\text{death}|\{\text{cake, death}\})$, she will choose cake whenever given the opportunity, but the welfare-maximizing policy is to impose death on her. More generally, the relationship between $u(\text{cake}|\{\text{cake,death}\})$ and $u(\text{death}|\{\text{cake, death}\})$ places no restriction on the relationship between $u(\text{cake}|\{\text{cake}\})$ and $u(\text{death}|\{\text{death}\})$, so observed behavior tells us nothing about the welfare-maximizing option.

It may appear that asking the person to make a decision over choice sets rather than final outcomes gets around this problem, because she can reveal whether she likes the singleton choice set {death} or the singleton choice set {cake}. But if the person's preferences are over available final outcomes as we assumed—a perfectly consistent set of preferences—such a choice is meaningless, because her choice set over ultimate outcomes includes both options.

Revealed Mistakes and Revealed, but Unimplemented Preferences

The starting position of this chapter is that some deviations from patterns of behavior that would be expected based on standard models are due to mistakes in implementing preferences rather than due to unexplored types of preferences. Yet we do not want to abandon the idea that behavior reveals many economically important characteristics of a person, including her preferences as well as her mistakes. In this section, we outline our unsurprising proposal for how to proceed: to make assumptions so that preferences and mistakes become jointly action-observable. Because we do not impose rationality, we need alternative assumptions that are not necessary in the standard revealed preference setting. Our assumptions are often action-unobservable—but as we discussed in the preceding section, so are crucial assumptions underlying standard analysis.

Although this is not strictly necessary for our approach, we will interpret beliefs as a way of summarizing what a person thinks about the likelihood of events, and utility as a way of summarizing her experience with outcomes. This means that beliefs and utility are not solely abstractions that represent behavior, but rather

abstractions that capture real mental states and experiences. Our interpretation seems natural, and simplifies many of our discussions and statements.

Our framework is designed for situations where mistakes about *measurable* variables—both exogenous events and one's own behavior—are possible. The first step is to make minimal assumptions about preferences so that a set of observable choices—essentially, a set of explicit or implicit bets on an event—reveals beliefs about the relevant variable. This often involves little more than assuming in some form that the person strictly prefers more money to less. The second step is to compare the beliefs elicited from behavior to objective probabilities. If the two are different, there are exactly two ways to explain the behavior. One can either abandon even the minimal assumptions on preferences that were used to infer beliefs, or one can accept the logical conclusion that the person has revealed a mistaken belief. Because the former option often leads the researcher into wacky theories, in many situations the latter option will be the more fruitful way to proceed. The third step, therefore, is to write a generally applicable theory of preferences and mistakes that explains the behavior in this as well as the largest possible number of other settings, and that ties welfare to behavior. An important component of such a theory will often be a set of assumptions about circumstances where mistakes do *not* happen, because these circumstances provide the most direct way to elicit preferences in the face of mistakes.

The usefulness of this general approach depends on its workability in specific settings. In the following sections, we demonstrate how the framework can help deliver essentially smoking-gun evidence for mistakes in a few types of settings, and how this understanding can help motivate a theory of preferences and mistakes.

Mistaken Beliefs About Exogenous Objective Events

In this section we provide a simple way to elicit what can naturally be interpreted as mistaken beliefs about objective events, and give examples of how a theory of mistakes can be integrated into a coherent theory of preferences and behavior. Our arguments follow the logic of the framework for integrating mistakes into economic models that we have proposed above: (i) make minimal assumptions about preferences so that mistaken beliefs can be detected for any decision maker with preferences in this class; (ii) compare beliefs to objective probabilities; and (iii) if the two are different, write a theory of preferences and mistakes that explains the behavior. A lot of the ideas here for eliciting beliefs have been around in the literature on eliciting subjective beliefs in expected utility theory. Our contribution is only in observing that many of the same ideas work for a much broader

class of preferences, and that they can be used to spot mistakes regarding *objective* uncertainty.

As a simple illustration, suppose a person may have observed some flips of a fair coin, and we now offer her the following option to "bet" on the next flip of the coin. If she bets on tails (T) and T comes up, she wins \$1 with probability 0.45; and if she bets on heads (H) and H comes up, she wins \$1 with probability 0.55. To avoid complications regarding compound lottery reduction, all uncertainty is resolved at the same time as the coin flip. If the decision maker chooses to bet on T despite the lower probability payoff, that strongly suggests she believes T is more likely. More precisely, if her preferences are independent of the coin flip as well as the random process generating the above probabilities, and she prefers to win \$1 with a higher rather than lower overall probability, then betting on T reveals a belief that T is more likely.

In general, we can elicit a person's precise beliefs regarding the probability of the event E using a variation of the Becker-DeGroot-Marschak procedure [Becker, DeGroot, and Marschak, 1964]. We inform the person that a "relative price" $r \in [0, 1]$ for E will be drawn randomly, and ask her to indicate a maximum price, q, for which she is willing to bet on E. If $r < q$, she wins \$1 with probability $(1 - r)$ if event E occurs, and nothing if $\neg E$ occurs. If $r \geq q$, she wins \$1 with probability r if event $\neg E$ occurs, and nothing if E occurs. All uncertainty is resolved at the same time. Intuitively, for each r the person is choosing whether to place a $(1-r)$ bet on E or an r bet on $\neg E$. The cutoff value of r for which she switches her bet, q, is an indication of her beliefs about E. Specifically, if her preference for money is independent of E and the random process generating the probabilities r and $1 - r$, and she prefers to win money with higher probability, q is exactly the subjective probability she places on E.

Example: Gambler's Fallacy

To continue with our example, let E be the event that T comes up on the next flip. Suppose we find the following pattern in the person's betting behavior:

Fact 1: If she has observed no flips, she chooses $q = 0.5$.
Fact 2: If she has observed the sequence of flips HHHHHH, she chooses $q = 0.55$.
Fact 3: If she has observed the sequence of flips TTTTTT, she chooses $q = 0.45$.

Confronted with this empirical pattern, we have two options: we either have to conclude that the person incorrectly believes that the likelihood of T depends on previous flips, or we have to abandon even the minimal assumptions on preferences that allow us to interpret bets as reflections of beliefs. This is a case where the former seems to be a much more fruitful way to proceed. In order to explain all these facts in terms of preferences, an economist would have to assume that the person likes

betting on T after HHHHHH, that she likes betting on H after TTTTTT, and that she is indifferent if she has observed no previous flips. This would be a wacky set of preferences indeed. When a preference-based methodology leads down on such a silly path, and when an intuitive mistake-based explanation is available, we are dumbfounded why economists should still restrict themselves to investigating and theorizing about only the preference-based alternative.

Indeed, to explain the same pattern of behavior in terms of mistakes, we can maintain the assumption that the decision maker's preferences are independent of the next flip, but also suppose that she believes in the "gambler's fallacy": she thinks that if H has come up a number of times, T is "due." Not only does this theory intuitively and parsimoniously explain the empirical facts, and is a theory that carries easily across contexts, but it also improves one's ability to make conclusions about preferences from behavioral data—that is, it increases the power of revealed preference.[4] For example, suppose we observe that after HHHHHH, the person strictly prefers to get blueberries if T comes up and huckleberries if H comes up, rather than vice versa, that he has the opposite preference after TTTTTT, and that she is indifferent if she has observed no flips. Having made the assumption that the decision maker believes in the gambler's fallacy and her preferences are independent of the coin flip, we can conclude that she likes blueberries more than huckleberries. Under the assumption that preferences are state dependent, we would be able to make no meaningful inference from this behavior.

As an economically more important domain of choice than blueberries and huckleberries, suppose that we observe a series of choices by both Fiona and Giles about whether they would rather have (x_h, x_t) if the next coin flip is heads or tails, or (y_h, y_t). Independently of the coin flips he has seen, Giles chooses (x_h, x_t) over (y_h, y_t) whenever $.5\sqrt{x_h} + .5\sqrt{x_t} > .5\sqrt{y_h} + .5\sqrt{y_t}$. Fiona's choices are more complicated:

1. If she has observed no flips, she chooses (x_h, x_t) over (y_h, y_t) whenever $.5\sqrt{x_h} + .5\sqrt{x_t} > .5\sqrt{y_h} + .5\sqrt{y_t}$.
2. If she has observed HHHHHH, she chooses (x_h, x_t) over (y_h, y_t) whenever $.45\sqrt{x_h} + .55\sqrt{x_t} > .45\sqrt{y_h} + .55\sqrt{y_t}$.
3. If she has observed TTTTTT, she chooses (x_h, x_t) over (y_h, y_t) whenever $.55\sqrt{x_h} + .45\sqrt{x_t} > .55\sqrt{y_h} + .45\sqrt{y_t}$.

How do we interpret this pattern of choices? The answer seems (to us) obvious: Giles does not succumb to the gambler's fallacy, and Fiona does. Through her pattern of choices, Fiona has made implicit bets that T is more likely than H to follow the sequence HHHHHH and H is more likely than T to follow the sequence TTTTTT. We are inclined to interpret this as a mistaken view about the way the world works, rather than a preference for betting (and losing money) on changes in coin flips.

But our point is not simply that we think it is useful to identify Fiona as making a mistake. It is also that we think we can use this understanding to identify Fiona's preferences. Indeed, we can identify her preferences as firmly as Giles's preferences: both are expected-utility maximizers with log utility (at least over binary choices). We admit to not really understanding why we would at all be inclined to ban the study of Fiona from economics departments, but we are especially chagrined at that prospect in light of the fact that we can use the same powerful tools of economics to study Fiona as Giles in this case. Fiona has well-ordered and coherent preferences. She is making an error in statistical reasoning. The two can be jointly identified. It is useful to do so.

If one insists on the bad-psychology assumption that people of economic interest do not succumb to the gambler's fallacy, one is apt to misidentify and certainly underestimate the coherence of Fiona's preferences. To the anti-psychological-insight eye, Fiona's preferences may look a bit random, or stochastic; for instance, sometimes Fiona prefers less money to more [e.g., $(x_h, x_t) = (9, 11)$ to $(y_h, y_t) = (11, 10)$]. Abandoning our natural capacity to identify errors in this case means diminishing our capacity to apply the tools of revealed preference.

Is This Economically Important?

Our example of betting on coin flips is admittedly of little or no immediate economic relevance. It is intended not as an economically important setting, but rather as a clean platform to bring out and discuss some of our ideas and objections to them. Guessing that many researchers have tried to talk friends or relatives out of the gambler's fallacy, we hope that most will agree with the points we have made in this context. Yet many will disagree that economists should worry about mistakes in what they study. The extent of the disagreement may be that some believe mistakes—although they can be studied with economic tools—are empirically unimportant to actually engage. In that case, we are partly happy, and although we worry somewhat about a possible doctrinaire stand that maintains the lack of mistakes as a null hypothesis, we look forward to research and arguments to see who is right. But partly we find it painfully obvious that in some situations mistakes are so plausible that they should be considered carefully by any economist studying the question. We give an example that is closely related to betting on coin flips, but is economically more important.

Consider the empirical regularity that Odean [1999] and Barber and Odean [2001] documented and interpreted as overtrading: small investors pay substantial transaction costs—and thereby substantially decrease their returns—to keep moving their money between investments. That is, if they made similar trades but did so less frequently, their return would be much higher. In order to explain this behavior, we must either assume that investors are making a mistake, or assume that their

investment behavior is motivated by something other than financial gain. As a plausible preference-based explanation, we may conjecture that they enjoy the process of trading and are willing to give up a large part of their financial return to be able to do it. In order to test this conjecture, we could (now hypothetically, since the above authors' data sets do not have this information) use our belief-elicitation method to see what investors seem to believe about the return of stocks in their portfolios. Suppose we find that investors tend to bet that the stocks they purchase will outperform other stocks. While one could maintain that investors have a preference to bet on these stocks (but will like to bet on other stocks soon), this finding would strongly suggest that they are misassessing the profitability of these stocks and are buying them for that reason. This possibility is sufficiently plausible empirically, and its implication that investors retire tens of thousands of dollars poorer sufficiently important economically, to warrant the time of economists to investigate. Perhaps it is very difficult to find data analogous to our betting experiment, but if so, economists should find the best data and methods to test between the different plausible theories.

In addition, we completely disagree with researchers who, rightly claiming that mistakes easily observed in the lab are often difficult or impossible to observe in economic data, propose to ignore them from the analysis altogether on the grounds that they constitute a free parameter to fit any situation. It is logical nonsense to respond to the difficulty of action-observing mistakes by making the very strong (as a corollary, action-unobservable) assumption that they do not exist, especially if they have been demonstrated robustly in other settings. We, of course, fully agree that economists should aim to write generally applicable models with few free parameters, but do not see why a model that incorporates a theory of mistakes cannot have that property.

Mistaken Beliefs About Future Behavior

If a person cannot predict what she will do in the future, she may make wrong decisions today. Hence, an important class of mistakes is about one's own future behavior. For instance, a teenager might falsely believe that she will end a possible period of experimentation with cigarettes by quitting, and so start smoking too easily. In this section we provide a (partial) way to elicit beliefs about future behavior, and again provide examples of how an understanding of mistakes can be incorporated into a theory of mistakes and preferences.

To elicit a person's beliefs about what she will choose at a given future date from the finite choice set X—and to see whether these beliefs are correct—we ask her

to wager on her future choice. More precisely, we offer her a choice between the different choice sets $X_x = \{x + \$\epsilon\} \cup (X \backslash x)$ generated by all elements $x \in X$. That is, the decision maker can choose which option to attach a small monetary payment to, in essence placing a bet on one of the choices. If the following two conditions are satisfied, the person will select the decision set X_x if and only if she believes she would choose x from X:

(i) $\$\epsilon$ does not change the most preferred option in X.
(ii) It is better to choose $x + \$\epsilon$ from X_x than to choose x from any of the choice sets X_y for $y \in X, y \neq x$.

Condition (i), which is action-testable, ensures that the person cannot use the betting situation to provide herself incentives. If the bet was large enough to change her behavior, and she did not like how she thought she would behave, she might choose to wager on an option not because she thought she would choose it but to give herself a reason to choose it. Condition (i) is satisfied for a sufficiently small ϵ whenever the person does not believe she will be indifferent between any two options in X. Even if indifferences are possible, as long as condition (ii) holds, letting $\epsilon \to 0$ in the limit reveals an option the person believes she would not refuse to choose from X.

Condition (ii) says that the person prefers to attach money to a chosen option rather than an unchosen option. This condition is satisfied for all types of preferences with which we are familiar, including possibly time-inconsistent consequentialist preferences, and preferences that may depend on the choice set, such as temptation disutility.[5]

Example I: Naivete about Self-Control

Our approach can help identify an important class of mistaken beliefs, those about one's own future self-control. Suppose a person always commits a particular type of revealed mistake: she bets that she will choose exercise from the choice set {exercise, television}, but then she always chooses television. Furthermore, when asked ex ante whether to have the singleton choice set {exercise} or the singleton choice set {television} in the future—that is, when asked which option to commit to—she prefers {exercise}. A parsimonious explanation for this set of observations can be based either on models of hyperbolic discounting such as Laibson [1997] and O'Donoghue and Rabin [1999], or on models of temptation disutility by Gul and Pesendorfer [2001]: the person would like herself to exercise in the future, but may not have enough self-control to actually do so. In addition, her mistake is in overestimating her self-control, either because she underestimates her future short-term impatience or because she underestimates the strength of temptation disutility.

An important issue with our framework arises when the decision maker's eventual choice is not perfectly predictable, for instance, because random events affecting her valuation for different options are realized before she makes her choice. In this case, a bet on an option does not necessarily reflect a belief that that option is most likely. To continue with our example, suppose that 75% of the time the person chooses television from the choice set {exercise, television}, yet she bets on exercise. While she loses money with this bet, it is not necessarily a mistake. The bet can serve as an incentive to exercise, and under a time-inconsistent taste for immediate gratification, such an incentive is valuable. And under temptation disutility, the bet can decrease the temptation to watch television, increasing utility in states when exercise is chosen. Unfortunately, we have not been able to figure out a way to elicit beliefs in these random settings.

Example II: Projection Bias

As another example of mistaken beliefs about future behavior, consider a person's choice between an appetizer and a full entree at some given future date. She may be hungry or satiated in the future, and independently of this state she may be hungry or satiated today. Posit for now that these states are known to the observer of the person's behavior (or can be induced); below we turn to situations where the state is unknown to the observer. Suppose that if the person is hungry today, she bets that she will choose the full course, and if she is satiated today, she bets she will choose the appetizer. Yet independently of the current state, she ends up choosing the entree if she is hungry in the future and the appetizer if she is satiated in the future. Hence, she has revealed a mistaken belief about future behavior in situations where the future state is different from the current state.

A parsimonious and intuitive theory that explains the mistakes combines state-dependent utility with a partial inability to predict that utility. Denote the decision maker's hunger state by s, and consumption by c. Her utility can then be written as $u(c, s)$. Presumably, she has higher marginal utility for food when she is hungry, and that is why she selects the entree when hungry and the appetizer when satiated. But in addition to this, she suffers from projection bias as summarized and modeled in Loewenstein, O'Donoghue, and Rabin [2003]: she underestimates how changes in her hunger state will change her preferences. In its most extreme form, projection bias means that if the person is currently in state s, she believes her preferences in the future will be given by $u(c, s)$, even when she knows her future state will be different from her current one.

Unlike with the systematic misprediction of self-control above, in this example it is very hard to interpret the bets purely in terms of self-imposed incentives, even when the person's choice is not perfectly predictable. There does not seem to be a form of self-control problem such that the behavior the person would like to

commit herself to depends on the current state but not on the future state. Hence, some misprediction must be going on.

In addition to being easy to spot when states are known, projection bias is sometimes apparent even when states are unknown. Suppose that based on observing the behavior of a large number of people, we establish the following empirical patterns:

1. When people make choices from the large finite choice set X, we observe an empirical distribution of choices $f(x)$, where each $f(x)$ is very small.
2. When people can choose ex ante whether to face the same choice set X or the choice set $X_x \equiv (x + \$\epsilon) \cup \{y - \$\epsilon\}_{y \in X}$ for some $x \in X$ of their choosing, they all select an $X(x)$ for some $x \in X$, and choice set X_x is chosen with probability $f(x)$. But independent of these ex ante choices, for each X_x the population chooses $x + \$\epsilon \in X_x$ with probability $f(x)$ and $y - \$\epsilon \in X_x$ with probability $f(y)$.

Selecting the set X_x ex ante is only beneficial if the person chooses $x + \$\epsilon$ ex post. Hence, this choice reveals a person's confident belief that she will prefer x. But people typically make a different choice ex post, so their belief is revealed to be incorrect. Projection bias provides an explanation: because people think their current preferences are indicative of their future preferences, they think they know what they will prefer in the future. But because their future state and preferences are random, their beliefs have little predictive power.

Eliciting State-Contingent Utility

When a person has mistaken beliefs about a relevant random variable or fails to correctly predict her own behavior, her behavior generally does not correspond to the strategy that maximizes her expected well-being. Hence, the standard revealed preference methodology for assessing welfare must be modified. In this section we propose a simple and intuitive methodology to measure welfare for a particular class of preferences—time-consistent state-contingent utility—when a person may make systematic mistakes in assessing that utility. Our example is motivated by projection bias but works for other kinds of biases, as well.

Our method for eliciting a person's state-contingent preferences $u(c, s)$ relies partly on finding circumstances in which mistakes are unlikely. More precisely, one of our key assumptions is that for any state s, there is a state s' such that in state s', the decision maker correctly understands her preferences in contingency s. In the projection-bias example, $s = s'$ seems like a reasonable assumption: when hungry, a person accurately perceives the value of eating on an empty stomach; and when satiated, the person understands what it is like to eat on a full stomach. In other situations—for example, with impulsive consumption—it may be easier to accurately perceive one's preferences in a state when not in that state. Nevertheless, for notational simplicity we shall assume that when in state s, the person understands

her utility in state s. This implies that it is easy to recover each of the cardinal preferences $u(\cdot, s)$ up to an affine transformation using standard revealed-preference techniques.

This, however, will not be sufficient if we also want to ask questions about trade-offs between states. For example, suppose we want to know whether to give a person an entree when she is not very hungry or an appetizer when she is hungry. That is, for some c, c', c'', c''' and s, s', we want to know the ranking of $u(c', s) - u(c, s)$ and $u(c''', s') - u(c'', s')$. If the person suffers from projection bias, we cannot rely on her own choice in the trade-off to make this judgment. When she is hungry, she does not appreciate that food will feel less good once she is less hungry, and hence she may incorrectly choose the entree. And when she is not so hungry, she does not appreciate how much better food will feel once she is hungry, and she may again incorrectly choose the entree.

To gain additional leverage, we assume that there is a "numeraire" good with state-independent value. In the projection-bias example, this could be retirement savings or another form of generic consumption far removed from the current state. We can then value the willingness to pay to move from c to c' in state s, and the willingness to pay to move from c'' to c''' in state s', in terms of the numeraire, and get a comparison of true utilities. Comparisons such as this will be sufficient to make all trade-offs whenever states are determined exogenously.

Our methodology for eliciting true preferences in the face of mistakes requires some assumptions that are not necessary in the standard revealed preference framework. Crucially, however, the reason we need more assumptions is that we have dropped the key assumption of the standard method, that choices always correspond to welfare.

Caveats, Discussion, and Conclusion

While we believe our framework helps start incorporating theories of mistakes systematically into economic analysis—and bring those theories in line with standard economics methodology—several important issues remain unresolved. A major problem is that there may be situations where behavior clearly reveals a mistake, but the source of that mistake is unclear. That is, there may be multiple natural theories of preferences and mistakes that can explain a person's revealed mistakes. Yet the welfare implications of an observed mistake could depend fundamentally on its source.[6] The following example is based on our discussions of self-control problems and projection bias above. Suppose a pregnant woman predicts that she will give birth without anesthetics, and that if she could choose now, she would also make that choice. Yet when the time comes, she actually decides to give birth with anesthetics. One theory that explains this revealed mistake is

projection bias: when not experiencing the pains of labor, it is difficult to appreciate just how bad it is. Another explanation is self-control problems: the mother would like herself to give birth naturally, but when confronted with the immediate pain of doing so, she cannot carry this through. The two theories give diametrically opposed welfare implications: projection bias says that the woman's later preference to give birth with anesthetics maximizes welfare, whereas hyperbolic discounting and temptation disutility say that the optimal policy is to commit to her early choice.

While this problem should be taken very seriously in any particular instance, we feel it is not a fundamentally new problem or one economists have no tools to deal with. Similarly to how economists attempt to distinguish theories in the standard framework, one can look for predictions that distinguish the two theories of mistakes in behavior.

Furthermore, our approach for eliciting mistakes in beliefs works only in certain circumstances. The method above for eliciting beliefs about exogenous events, for instance, assumes that utility for money is independent of the realization of uncertainty in question. As recognized by researchers such as Kadane and Winkler [1988] and Karni [1999], situations where a person already has "stakes" in the random process do not satisfy this assumption. For example, if an investor has money in the stock market, she will be poorer if the stock market does badly, and presumably have more value for money in that case. Hence, her bets regarding stock market returns reflect not just her beliefs about the stock market, but also her need for money in different contingencies.

NOTES

We thank Andrew Caplin for many fascinating conversations and helpful suggestions.

1. To clarify what we mean and to highlight that we believe nonbehavioral evidence is also important, throughout this chapter we use the term "action-observable" for assumptions or predictions that can be linked directly to choice. By dint of calling such assumptions simply "observable," standard economics suggests either that nothing else is observable, or (more likely) that such observations are unimportant for economics. Oral and written statements about beliefs, serotonin and dopamine levels, smiles, and brain activity are observable, often much more directly than choice. These observations are and should be used in economics. But they are not the focus of this chapter.

2. We are not advocating that economists be directly involved in all policy design. But we do advocate that they be involved in developing the conceptual underpinnings of policy design. And normative analysis that allows for the possibility of mistakes is an important part of that conceptual underpinning.

3. As we emphasize, our approach relies on some action-unobservable assumptions, and these assumptions may implicitly involve value judgments. But since the standard

approach also relies on action-unobservable assumptions, in this sense it is equally problematic.

4. For formalizations of the gambler's fallacy and examples of the applicability of such a theory, see Rabin [2002] and Rabin and Vayanos [2005].

5. Identifying beliefs about choice from an infinite set is only slightly more complex. Suppose X is a compact set in a metric space, and utility is continuous with respect to this metric. The only complication relative to a discrete choice set is that there may be choices that are arbitrarily close in preference to the favorite option in X, and a small payment may induce the person to choose one of these near-optima instead. Once again, however, letting $\epsilon \to 0$ in the limit identifies an option in X that the person believes she would be willing to choose.

6. This issue is all the more important in light of some recent work on paternalism, e.g., Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin [2003], O'Donoghue, and Rabin [2003], and Sunstein and Thaler [2003], whose major theme is to design policies that aid people avoiding a particular kind of mistake while doing little plausible harm to any previously known type of rational person. While it is often easy to design such policies, it is far less clear that a proposed policy will not do more harm than good to other types of irrational agents.

REFERENCES

Barber, Brad M., and Terry Odean. 2001. Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment. *Quarterly Journal of Economics* 116(1): 261–292.

Becker, Gordon M., Morris H. DeGroot, and J. Marschak. 1964. Measuring Utility by a Single-Response Sequential Method. *Behavioral Science* 9: 226–32.

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin. 2003. Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism." *University of Pennsylvania Law Review* 151(3): 1211–1254.

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

Kadane, Joseph B., and Robert L. Winkler. 1988. Separating Probability Elicitation from Utilities. *Journal of the American Statistical Association* 83(402): 357–363.

Karni, Edi. 1999. Elicitation of Subjective Probabilities When Preferences Are State-Dependent. *International Economic Review* 40(2): 479–486.

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

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin. 2003. Projection Bias in Predicting Future Utility. *Quarterly Journal of Economics* 118(4): 81–123.

Odean, Terry. 1999. Do Investors Trade Too Much? *American Economic Review* 89: 1279–1298.

O'Donoghue, Ted, and Matthew Rabin. 1999. Doing It Now or Later. *American Economic Review* 89(1): 103–124.

———. 2003. Studying Optimal Paternalism, Illustrated by a Model of Sin Taxes. Mimeo, UC Berkeley.

Rabin, Matthew. 2002. Inference by Believers in the Law of Small Numbers. *Quarterly Journal of Economics* 117(3): 775–816.
Rabin, Matthew, and Dimitri Vayanos. 2005. The Gambler's and Hot-Hand Fallacies in a Dynamic Inference Model. Working Paper, London School of Economics.
Sunstein, Cass, and Richard Thaler. 2003. Libertarian Paternalism. *American Economic Review* 93(2): 175–179.